Behaviormetrika
Vol.30, No.1, 2003, 87-98

A NOTE ON BETWEEN-SET DISTANCES IN DUAL SCALING AND CORRESPONDENCE ANALYSIS

Shizuhiko Nishisato* and Jose G. Clavel**

In quantifying a two-way table of data, one derives weights (spacings) for rows and columns so as to maximize the correlation of data weighted by row weights and those by column weights. It is well known that one can calculate the distance between any two rows (columns), called the within-set distance, but the distance between a row and a column, the between-set distance, cannot be calculated because the row weights span the space different from that of the column weights. The present paper shows (1) that one can calculate the between-set distances, and (2) that the data as a whole require an additional dimension to accommodate the discrepancy of the row space and the column space. Since the information contained in the between-set distances is an integral part of the data structure, it is not a matter of whether the study advocates the use of the between-set distance information, but rather it is a problem that dual scaling and correspondence analysis must deal with through development of an analytic method to tap into the entire information in the data. Until then, the current formulation of these methods which ignore the information contained in the between-set distances provide at best simplified approximations to the decomposition of data structure. This further development is not easy because these methods are by the very nature of categorical data based on the chi-square metric, making an entire inter-point (within-set and between-set) distance matrix not readily amenable to any currently available analytic method.

1. Introduction

It is well known in quantification of categorical data that the space for row variables and the space for columns variables are generally different (e.g., Lebart, Morineau & Tabard, 1977; Nishisato, 1980, 1994, 1996; Greenacre, 1984, 1989, 1993; Carroll, Green & Schaffer, 1986, 1987, 1989; Gifi, 1990; Greenacre & Blasius, 1994, 1997). Therefore, the general conclusion is that although within-set (i.e., within-row or within-column) distances can be calculated from the corresponding coordinates, the so-called between-set distances (i.e., between rows and columns) cannot, hence requiring some cautions in interpreting them when rows and columns are plotted in the same graph. To quote, "a great deal of caution is needed in interpreting the distance between a variable point and an individual point, because these two points do not belong to the same space" (Lebart, Morineau & Warwick, 1984, p.19).

It is also well known (see the aforementioned references) that we can always project row (column) variables onto the space of column (row) variables, leading

Key Words and Phrases: Quantification of rows and columns, within-set and between-set distances, Chi-square distance, projection, symmetric scaling, exhaustive analysis.

* University of Toronto, Canada. Any correspondence should be directed to S. Nishisato, CTL, OISE/UT, 252 Bloor Street West, Toronto, Ontario, Canada M5S 1V6.
** University of Murcia, Spain.

to the so-called non-symmetric scaling. This provides a true joint graph of row variables and column variables. From a practical point of view, however, this exact joint graph of rows and columns is often not satisfactory because a set of the projected variables has generally a smaller norm than the original set, thus tending the projected variables to cluster closer to the origin of the graph. The discrepancy between normed weights and projected weights widens as the corresponding singular value decreases. In other words, the exact plot of normed and projected weights remains reasonable when the first two singular values are both close to one, which often occurs when one plots the first two solutions (components), but if one wants to use this exact plot for solution 3 against solution 5, for instance, it is likely that the gap in the variance of projected and that of normed weights may no longer allow us to make reasonable comparisons of distances.

The most popular compromise has been to use symmetric scaling for a joint plot of projected variables, with the caution that one should be aware that it is an overlay of two discrepant spaces onto one, hence not an exact description, but a practical approximation. Within this context, some statistics to describe "badness of a joint graph" have been proposed (Nishisato, 1988), which is a passive way to quantify the amount of "cautions" needed in interpreting a joint graph. The symmetric scaling would make sense only when the two singular values are close to one, for otherwise the large discrepancy between the row space and the column space will make the comparisons of between-set points utterly meaningless.

An attempt to place rows and columns in the same space was presented under the name of CGS scaling (Carroll, Green, & Schaffer, 1986, 1987), which, however, fell short of a sound theoretical basis (Nishisato, 1980; Greenacre, 1989). Another attempt to accommodate information about both rows and columns was presented (Nishisato, 1990, 1997), of which practical merits are well demonstrated in one study (So, 1997), but its applicability to a large data set is yet to be assessed. The current paper is only a note to point out that it is possible to calculate between-set distances. Because of its simplicity, this computation has perhaps never been considered worthy of publication, but simple or not it has an important implication for dual scaling and correspondence analysis that the mathematical aspect of these methods has left out an important future problem, that is, the exact joint analysis, or, exhaustive analysis, of a two-way categorical table, using both within-set and between-set distance information.

2. Procedure

Let $\mathbf{F}$ be an $n \times m$ contingency table, $\mathbf{y}_k$ be an $n \times 1$ vector of scaled (i.e., normed to a constant) weights for rows, and $\mathbf{x}_k$ an $m \times 1$ scaled vector for columns, both associated with Solution (component) k. The task for the scaling is to determine $\mathbf{y}_k$ and $\mathbf{x}_k$ in such a way that the responses weighted by $\mathbf{y}_k$ and the responses weighted by $\mathbf{x}_k$ attains a maximal product-moment correlation ρ. This is only one of many ways to formulate dual scaling and correspondence analysis (see different objective functions for optimization in Nishisato, 1980; Greenacre, 1984; Gifi, 1990). It is well known, however, that the optimal weight vectors for Solution k satisfy the following:

$$\mathbf{y}_k = \frac{1}{\rho_k}\mathbf{D}_r^{-1}\mathbf{F}\mathbf{x}_k, \quad \mathbf{x}_k = \frac{1}{\rho_k}\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{y}_k, \tag{1}$$

where $\mathbf{D}_r$ and $\mathbf{D}_c$ are respectively diagonal matrices of row marginals and column marginals of $\mathbf{F}$, and ρ_k is a singular value of $\mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2}$. The weight vectors are so scaled that responses weighted by respective vectors are summed to zero. The above set of equations is called *dual relations* (Nishisato, 1980).

2.1. Normed versus projected weights

In the above formulas, $\mathbf{y}_k$ and $\mathbf{x}_k$ are called normed weights (or, standard coordinates) because the weighted sums of squares, $\mathbf{y}_k'\mathbf{D}_r\mathbf{y}_k$ and $\mathbf{x}_k'\mathbf{D}_c\mathbf{x}_k$, are set equal to a pre-determined constant. Normed weights multiplied by the singular value, that is, $\rho_k\mathbf{y}_k$ and $\rho_k\mathbf{x}_k$, are referred to as projected weights (or, principal coordinates). The distinction between standard coordinates and principal coordinates can be stated in three respects: (1) normed weights have a constant norm, irrespective of solution k, that is, the singular value, and (2) projected row weights are orthogonal projections of the normed row vector onto the column space and projected column weights are projections of the column weights onto the row space, and (3) the original data cannot be reconstructed from normed weights, but projected weights can reproduce the data. Two vectors $\mathbf{y}_k$ and $\rho_k\mathbf{x}_k$ span the same space, and so do $\rho_k\mathbf{y}_k$ and $\mathbf{x}_k$. Then, we know from the above that the row axis and the column axis of a solution k (e.g., $k = 1, 2$) are separated by an angle of $cos^{-1}\rho_k$ (see Nishisato, 1996). This means that the true joint configuration of row variables and column variables, represented by one solution, spans not over one-dimensional space, but two-dimensional space, the second dimension being due to the discrepancy between the row space and the column space; similarly, the true joint configuration of row variables and column variables, depicted by two solutions, spans three-dimensional space. In other words, the discrepancy between the row space and the column space necessitates one additional dimension for a complete description of the solution set. This aspect is fully explained in Section 2.3.

The third point mentioned above is a very important distinction between the two kinds of weights, that is, the data matrix can be reconstructed from projected weights, and not from normed weights. As is well known in statistics, the information in the data can be summarized by the distribution of eigenvalues (or singular values), and because normed weights do not contain information about singular values, they are not informative enough to reproduce the data.

2.2. Chi-square distance

In the literature on quantification theory, it is well known (see, for example, Lebart, Morineau & Warwick, 1984) that the within-set distance, for instance, is given by the so-called chi-square distance. The squared distance between two rows $\mathrm{y_i}$ and $\mathrm{y_{i'}}$ is given by:

$$d_{ii'}^2 = \sum_{k=1}^{K} \rho_k^2 \left(\frac{y_{ik}}{\sqrt{p_{i.}}} - \frac{y_{i'k}}{\sqrt{p_{i.'}}} \right)^2 , \tag{2}$$

where $p_{i.}$ and $p_{i.'}$ are proportions of the two marginals. If the Euclidean distance between point A and point B is 10 meters, and if point A contains 9 observations and point B one observation, then the midpoint in the chi-square metric is closer to point A than to point B. Or, more precisely, the chi-square distance between point A and the midpoint times 9 (the frequency of point A) is equal to the chi-square distance between point B and the midpoint times 1 (the frequency of point B). It is easy to see that a point with a greater mass has a greater pull than a point with a smaller mass. When all the points have an equal mass, then the chi-square distance becomes equivalent to the Euclidean distance.

2.3 Discrepancy between row space and columns space

Suppose we extract two solutions from the data with two singular values ρ_1 and ρ_2. Then, the two planes corresponding to the row space and the column space can be depicted as shown in Figure 1, where the two planes are not in the same space but in different spaces with angles related to the respective singular values (e.g., Nishisato, 1994). Now, once we know the angle of the discrepancy between the row space and the column space, one can use the well-known cosine law to calculate the exact distance between row i and column j.

Consider two points A and B in a two-dimensional space, with the distance between the origin (O) and point A being a , the distance between O and B being b, and the angle AOB being θ. Then, the cosine law states that the distance between A and B is given by

$$d_{\mathrm{AB}} = \sqrt{a^2 + b^2 - 2ab\cos\theta}$$

Let us apply the cosine law to the calculation of the between-set distance, which is the chi-square metric as indicated above. Noting that a^2 and b^2 in the above expression correspond to $\rho_k^2 x_{jk}^2/p_j$, and $\rho_k^2 x_{ik}^2/p_i$, respectively, and $\cos\theta$ to ρ_k, it is straightforward to obtain the following expression of the between-set chi-square distance between row i and column j in K-dimensional space,

$$d_{ij} = \sqrt{\sum_{k=1}^{K} \rho_k^2 \left[\frac{x_{jk}^2}{p_{.j}} + \frac{y_{ik}^2}{p_{i.}} - 2\rho_k \frac{x_{jk} y_{ik}}{\sqrt{p_{.j} p_{i.}}} \right]} \tag{3}$$

where $p_{\mathrm{i.}}$ and $p_{\mathrm{.j}}$ are marginal proportions of row i and column j, respectively.

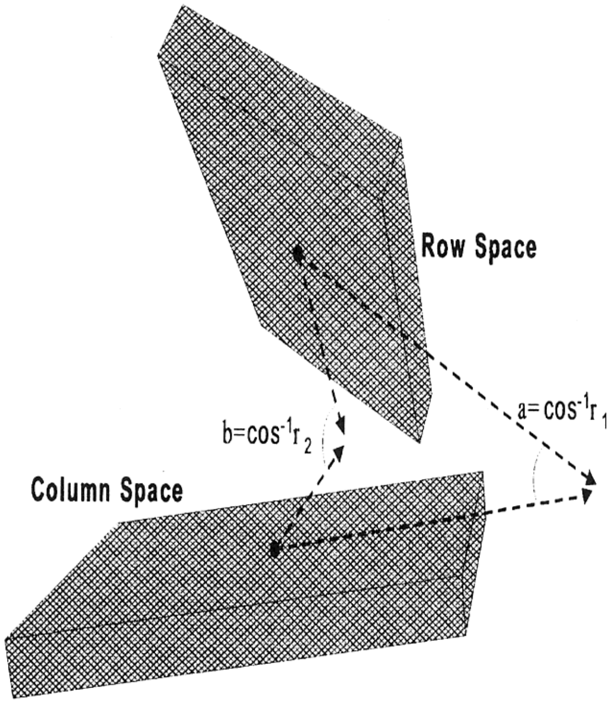

Figure 1: Two-dimensional DS solutions with row and column spaces

3. Numerical Illustration

The numerical example was adopted from Stebbins (1950): Eleven varieties of barley were planted at 10 agricultural experimental stations in the United States; an equal number of seeds of each variety were planted; at the harvesting time, all seeds were mixed, 500 seeds were randomly sampled, and were then sorted into the varieties; this process was repeated a number of times. Table 1 is a part of the data, and shows the final counting of six varieties of barley at six locations. This data set was analyzed by Nishisato (1994), with the following results. Five correlation ratios associated with five solutions are respectively. 0.735 (38%), 0.632 (33%), 0.480 (25%), 0.058 (3%), and 0.019 (1%), showing three dominant solutions, accounting for 96% of the total information in the data. Thus, one would typically be satisfied with joint graphs of solutions [I,II], solutions [I,III], and solutions [II,III], or just two of the three graphs.

There are three purposes for the numerical illustrations: (1) to show examples of exact within-set distances and exact between-set distances; (2) using two combinations of two dual scaling (DS) solutions [I,II] and [II,III], to show that a joint graph of each combination requires three dimensions, rather than two, and; (3)

Table 1: Numbers of seeds of a variety of barleys at different locations

	1.Arlington	2.Ithaca	3.St.Paul	4.Mocasin	5.Moro	6.Davi
1. Coast	446	57	83	87	6	362
2. Hanchen	4	34	305	19	4	34
3. White	4	0	4	241	489	65
4. Manchuria	1	343	2	21	0	0
5. Gatami	13	9	15	58	0	1
6. Meloy	4	0	0	4	0	27

to see how the contribution of the additional dimension increases as the singular value decreases. The third point is related to how good the traditional symmetric scaling for a joint graph is.

3.1 Results

As reported in Nishisato (1994, p.129-132), standard coordinates of three scaling solutions are as listed in Table 2.

Table 2: Standard coordinates for rows and columns of Table 1 in first three solutions

ROWS (yi)	Sol. 1	Sol. 2	Sol. 3	COL (xj)	Sol. 1	Sol. 2	Sol. 3
Eigenvalue	0.7352	0.6325	0.4802	Eigenvalue	0.7352	0.6325	0.4802
1. Coast	0.128	0.857	-0.879	1. Arlington	0.132	1.021	-1.174
2. Hanchen	0.577	0.927	2.144	2. Ithaca	1.750	-1.435	-0.197
3. White	-1.278	-0.820	0.201	3. St. Paul	0.520	1.063	2.064
4. Manchuria	1.865	-1.706	-0.250	4. Mocasin	-0.700	-0.419	0.060
5. Gatami	-0.187	-0.094	0.250	5. Moro	-1.454	-0.988	0.294
6. Meloy	-0.120	0.864	-1.038	6. Davi	-0.049	0.802	-0.768

Let us first calculate within-set (i.e., within-row and within-column) distances in two dimensional space, defined by DS solutions I and II. The correlation ratios associated with these solutions are $\rho_1^2 = 0.735$ and $\rho_2^2 = 0.632$, respectively, hence the corresponding singular values are the square roots of them, namely, 0.857 and 0.795. This means that the angle between the axis of the rows and the axis of the columns in solution I is $cos^{-1}(0.857) = 34.7$ degrees. Similarly, the angle between the row and the column axes in solution II is 41.5 degrees. One should note that even relatively high values of singular values as in the current example reveal substantially different angles from the ideal value of zero, the angle when the singular value is 1.

Table 3: Both within-set and between-set distances based on dual scaling solutions I and II

1. Coa	0.000	1.388	3.193	6.384	1.827	5.094	1.291	5.207	1.782	2.490	4.161	0.958
2. Han	1.388	0.000	4.566	6.421	3.171	4.702	1.646	5.274	1.497	3.788	5.428	1.810
3. Whi	3.193	4.566	0.000	6.868	1.419	7.372	3.773	5.923	4.458	1.297	1.950	3.223
4. Man	6.384	6.421	6.868	0.000	6.185	11.123	6.812	3.182	6.574	6.500	7.471	6.692
5. Gat	1.827	3.171	1.419	6.185	0.000	6.483	2.552	5.154	3.178	1.066	2.712	1.999
6. Mel	5.094	4.702	7.372	11.123	6.483	0.000	4.809	9.649	4.958	6.826	7.937	5.035
1. Arl	1.291	1.646	3.773	6.812	2.552	4.809	0.000	5.927	0.909	3.333	4.974	0.593
2. Ith	5.207	5.274	5.923	3.182	5.154	9.649	5.927	0.000	5.650	5.617	6.730	5.797
3. StP	1.782	1.497	4.458	6.574	3.178	4.958	0.909	5.650	0.000	4.038	5.732	1.426
4. Moc	2.490	3.788	1.297	6.500	1.066	6.826	3.333	5.617	4.038	0.000	1.725	2.744
5. Mor	4.161	5.428	1.950	7.471	[illegible].712	7.937	4.974	6.730	5.732	1.725	0.000	4.381
6. Dav	0.958	1.810	3.223	6.692	1.999	5.035	0.593	5.797	1.426	2.744	4.381	0.000

Table 4: Both within-set and between-set distances based on dual scaling solutions II and III

1. Coa	0.000	4.947	2.625	4.842	2.436	7.327	1.944	3.861	4.682	2.131	3.117	1.298
2. Han	4.947	0.000	4.796	7.128	3.769	11.065	5.577	6.129	3.267	4.645	5.053	4.987
3. Whi	2.625	4.796	0.000	2.608	1.046	9.848	3.705	2.092	4.784	0.766	1.199	2.954
4. Man	4.842	7.128	2.608	0.000	3.591	11.428	5.662	2.276	6.914	3.131	2.654	5.091
5. Gat	2.436	3.769	1.046	3.591	0.000	9.757	3.539	2.797	4.019	1.032	1.688	2.731
6. Mel	7.327	11.065	9.848	11.428	9.757	0.000	6.992	10.496	10.370	9.344	10.158	7.462
1. Arl	1.944	5.577	3.705	5.662	3.539	6.992	0.000	3.596	5.722	2.649	3.837	0.781
2. Ith	3.861	6.129	2.092	2.276	2.797	10.496	3.596	0.000	4.697	4.889	6.227	3.672
3. StP	4.682	3.267	4.784	6.914	4.019	10.370	5.722	4.697	0.000	4.368	4.969	5.097
4. Moc	2.131	4.645	0.766	3.131	1.032	9.344	2.649	4.889	4.368	0.000	1.356	1.894
5. Mor	3.117	5.053	1.199	2.654	1.688	10.158	3.837	6.227	4.969	1.356	0.000	3.142
6. Dav	1.298	4.987	2.954	5.091	2.731	7.462	0.781	3.672	5.097	1.894	3.142	0.000

Because of these angles between rows and columns, it is obvious that the joint plot of rows and columns for solution I requires two dimensional space. Likewise, we need three dimensional space to accommodate the joint graph of rows and columns for two DS solutions. With the additional dimension in each of the above example, it becomes possible to calculate the exact distance between row i and column j. For instance, suppose that we want to calculate the distance between the third row (y_3 = *White* with $p_3 = 0.293$, the latter being the proportion of the marginal of row 3 to the total frequency) and the fourth column (x_4 = *Mocasin* with $p_4 = 0.157$) in dimension I. Then, using the formula given earlier, the squared distance between them, d_{34}^2 , can be calculated by:

$$d_{34}^2 = 0.735 \times \left[\frac{-0.700^2}{0.157} + \frac{-1.278^2}{0.293} - 2\sqrt{0.735}\frac{(-0.700)(-1.278)}{\sqrt{0.157 \times 0.293}}\right] \tag{4}$$

$$+0.632 \times \left[\frac{-0.419^2}{0.157} + \frac{-0.820^2}{0.293} - 2\sqrt{0.632}\frac{(-0.419)(-0.820)}{\sqrt{0.157 \times 0.293}}\right] = 1.681$$

and hence the distance d_{34} is the positive square root of 1.681, which is 1.297.

The same computations can be carried out with respect to the space defined by DS solution II and solution III, noting that the correlation ratio for DS solution III is 0.480.

Within-set distances can be calculated by the usual way of the chi-square distance as mentioned earlier. For example, for solutions I and II the square of the chi-squared distance between column 5 (x_5 = *Moro*) and column six (x_6 = *Davis*) in two solutions, d_{56}^2 , can be calculated as:

$$d_{56}^2 = 0.735 \times \left(\frac{-1.454}{\sqrt{0.182}} - \frac{0.049}{\sqrt{0.178}}\right)^2 + 0.6325 \times \left(\frac{-0.988}{\sqrt{0.182}} - \frac{0.802}{\sqrt{0.178}}\right)^2 = 20.352 \tag{5}$$

The distance is the square root of this, that is, 4.511. Tables 3 and 4 show both within-set and between-set distances, based on solutions I and II, and solutions II and III, respectively.

4. Discussion

It was mentioned earlier that the discrepancy between the row space and the column space necessitates an additional dimension to accommodate all the data points. Thus, in the current examples, the above matrices of distances, based on two solutions, can be represented not in two-dimensional space, but in three-dimensional space. To examine this aspect, those matrices of distances were subjected to nonmetric multidimensional scaling (MDS), using ALSCAL in SPSS. When we have two categorical variables (barleys and locations), the entire data can be arranged into the form of an incidence matrix of observations (seeds in the current example) by barleys, $\mathbf{F}_1$, and that of observations by locations, $\mathbf{F}_2$. Then, the augmented data $(\mathbf{F}_1, \mathbf{F}_2)$ can be subjected to scaling to obtain a multidimensional configuration of barleys and locations. In this case, the distance matrix of the two sets of categories yield the matrix consisting of within-set and between-set distances. From this distance matrix, we can calculate the matrix of inner products, which can then be subjected to the nonmetric MDS. As is known, the initial configuration of nonmetric MDS is the least squares approximation to the principal axis configuration, as is the case with dual scaling (e.g., Nishisato, 1994). Table 5 contains the results associated with Table 3, that is, coordinates for rows and columns in three MDS dimensions and the contributions of the respective dimensions.

Table 5: Coordinates for rows and columns of three MDS dimensions, based on DS Solutions I and II

Name	Dimension1	Dimension2	Dimension3
1 COA	-0.371	0.164	-0.072
2 HAN	-0.488	0.860	-0.147
3 WHI	0.172	-1.386	0.140
4 MAN	2.864	0.915	0.443
5 GAT	0.078	-0.650	0.059
6 MEL	-2.921	0.826	-0.532
7 ARL	-0.656	0.405	0.470
8 ITH	2.118	0.886	-0.916
9 STP	-0.548	0.861	0.481
10 MOC	0.102	-1.060	-0.075
11 MOR	0.198	-1.949	-0.234
12 DAV	-0.549	0.129	0.383
Contribution	63%	31%	6%

The final Kruskal's S-stress of 0.02673 was reached in only three iterations, and RSQ of 0.998 was obtained. Since the distance matrix was generated by two dual scaling solutions and the two singular values are relatively large, we expect that the contribution of the third MDS solution would be small, which was demonstrated by the contribution being only 6% (see Table 5). Three MDS solutions were also obtained from the distance matrix in Table 4, which is based on DS solutions II and III (Table 6). Because of the smaller singular values, we see a larger contribution of the third MDS solution, which is now 12%. From these two sets of results, it is clear that the contribution of the additional dimension, due to the space discrepancy, will monotonically increase as the singular values become smaller, even to the extent that the additional dimension might contribute to the total decomposition more than some of the "genuine" dimensions. Should this happen, however, we will have to face yet another problem of identifying which solution is due to the space discrepancy.

Table 6: Coordinates for rows and columns of three MDS dimensions, based on DS II and III Solutions

Name	Dimension1	Dimension2	Dimension3
1 COA	-0.506	0.082	-0.056
2 HAN	0.857	-1.882	-0.246
3 WHI	0.702	0.446	-0.065
4 MAN	1.311	1.471	-0.033
5 GAT	0.658	-0.014	-0.095
6 MEL	-4.053	0.082	0.064
7 ARL	-0.837	0.312	0.168
8 ITH	0.748	0.624	1.371
9 STP	0.622	-1.644	0.744
10 MOC	0.366	0.156	-0.636
11 MOR	0.656	0.218	-1.139
12 DAV	-0.525	0.15	-0.078
Contribution	63%	25%	12%

The current study has shown (1) that not only within-set distances, but also between-set distances can be calculated, (2) that the discrepancy between the two spaces requires an additional space to accommodate all the data points, and (3) that the contribution of this additional dimension increases as singular values decrease.

The last point (3) presents an interesting topic for further investigation. For, it means that when the discrepancy between the row space and the column space is large, interpretations based on symmetric scaling of the DS solutions may not be appropriate. How large the discrepancy between the two spaces (e.g., 45 degrees, or 70 degrees or what not) is large enough to abandon the popular symmetric scaling interpretation is a difficult question to answer, but it seems that one should at least indicate some empirical guidelines on this point for symmetric scaling. In other words, when can we say about the symmetric joint graph that "graphing is believing" (Nishisato, 1997)?

We should also note the following point. In the current study, DS solutions were first obtained using the chi-square distances, while MDS solutions were obtained by treating them to be Euclidean distances, that is, as if each variable contained the same number of responses. Thus, in comparing the DS configurations and the MDS configuration with one extra dimension, the above difference in distance measures should be kept in mind. MDS was needed to make a point that one extra dimension is created by the discrepancy between the row and the column spaces.

This last point is perhaps the most noteworthy message of the present paper: Currently we do not have a method to analyze simultaneously both within-set and between-set chi-square distances. It seems crucial for data analysis to recognize that the current formulations of dual scaling and correspondence analysis ignore the information contained in between-set distances. Faced with the fact that the row space and the column space of a single data set are discrepant and the fact that the chi-square metric is used for analysis, we have not yet developed a method that can tap into both within-set and between-set distance information in the data.

The current symmetric scaling is logically problematic (i.e., overlaying two different spaces) and practically so, too, especially when singular values are relatively small. The present non-symmetric scaling of one set projected onto the other set is not satisfactory either because the two corresponding norms can be vastly different, and both sets are not equally treated.

It is strongly urged that further development of dual scaling and correspondence analysis must be made so as to analyze both within-set and between-set distance information. We can only then say that dual scaling and correspondence analysis provide exhaustive analysis of information in categorical data. The same extension is needed for principal component analysis of continuous data. The latter may be easier than the categorical case because of the use of the Euclidean distance, instead of the chi-square metric.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to the first author, and a fellowship to the second author from Fundacion Seneca, Consejeria de Cultura y Educacion de la Region de Murcia during his Visiting Professorship at the Ontario Institute for Studies in Education (OISE) of the University of Toronto from July to December 1997. The revision of the paper was done while the first author was a Visiting Professor at Kwansei Gakuin University, Nishinomiya, Japan. The authors wish to thank anonymous reviewers for their useful comments and Mr. Ichiro Shishido for pointing out some errors.

References

Carroll, J. D., Green, P. E., & Schaffer, C. M. (1986). Interpoint distances comparisons in correspondence analysis. *Journal of Marketing Research*, **23**, 271-280.

Carroll, J. D., Green, P. E., & Schaffer, C. M. (1987). Comparing interpoint distances in correspondence analysis: A clarification. *Journal of Marketing Research*, **24**, 445-450.

Carroll, J. D., Green, P. E., & Schaffer, C. M. (1989). Reply to Greenacre's commentary on the Carroll-Green-Schaffer scaling of two-way correspondence analysis solutions. *Journal of Marketing Research*, **26**, 366-368.

Gifi, A. (1990). *Nonlinear multivariate analysis.* New York: Wiley.

Greeacre, M. J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Greenacre, M. J. (1989). The Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal. *Journal of Marketing Research*, **26**, 358-365.

Greenacre, M. J. & Blasius, J. (1994). *Correspondence analysis in the social sciences.* London: Academic Press.

Greenacre, M. J. & Blasius, J. (1997). *Visual display of categorical data.* London: Academic Press.

Lebart, L., Morineau, A., & Tabard, N. (1977). *Techniques de la Description Statistique: Methodes et Logiciels pour l'Analyse des Grands Tableaux* [Techniques of statistical description: Methods and softwares for analysis of large tables of data]. Paris: Dunod.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis.* New York: Wiley.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications.* Toronto: University of Toronto Press.

Nishisato, S. (1988). Assessing quality of joint graphical display in correspondence analysis and dual scaling. In E. Diday, Y. Escoufier, L. Lebart, J. Page, Y. Schektman, & R. Tommasone (Eds.), *Data analysis and informatics, V.* (pp. 409-416). Amsterdam: North Holland.

Nishisato, S. (1990). Dual scaling of designed experiments. In M. Schader & W. Gaul (Eds.), *Knowledge, data and computer-assisted decisions* (pp. 115-125). Berlin: Springer-Verlag.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Nishisato, S. (1996). Gleaning in the field of dual scaling. *Psychometrika*, **61**, 559-599.

Nishisato, S. (1997). Graphing is believing. In Greenacre, M.J. & Blasius, J. (Eds.), *Visual display of categorical data* (Chapter 15, pp. 185-196). London: Academic Press.

So, S. (1997). Writing to make meaning or to learn the language? A descriptive study of multi-ethnic adults learning Japanese-as-a-foreign language. *Unpublished doctoral dissertation*, University of Toronto.

Stebbins, C. L. (1950). *Variations and evolution.* New York: Columbia University Press.

Received May 2001. Final version accepted April 2002.