Dual Scaling of Dominance Data and its Relationship to Correspondence Analysis

MICHAEL GREENACRE & ANNA TORRES[1]

Facultat de Cipncies Econpmiques i Empresarials

UNIVERSITAT POMPEU FABRA

[1] Working Paper number 430 (December 1999), Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas, 23{27, E-08005 Barcelona, Spain. E-mail: michael@upf.es and anna.torres@econ.upf.es

Abstract

Dual scaling of a subjects-by-objects table of dominance data (preferences, paired comparisons and successive categories data) has been contrasted with correspondence analysis, as if the two techniques were somehow different. In this note we show that dual scaling of dominance data is equivalent to the correspondence analysis of a table which is doubled with respect to subjects. Both methods are in turn equivalent to a principal components analysis of the undoubled dominance table which is centred with respect to subject means.

Journal of Economic Literature Classification: C19, C88

Keywords: Correspondence analysis, dominance data, dual scaling, paired comparisons, preferences, principal component analysis, ratings.

1 Introduction

Dual scaling (Nishisato 1980, 1994) is a multivariate method for assigning scale values to the rows and columns of a table of data, with certain optimal properties. Correspondence analysis (Benz¶cri 1973, Greenacre 1984, Lebart et al 1984) is a method for assigning optimal spatial positions to the rows and columns of a data table. It is well-known that dual scaling (DS) and correspondence analysis (CA) are mathematically equivalent techniques in the cases of a two-way frequency table and a multiple indicator matrix (or response pattern matrix) { see Greenacre (1984) or Nishisato (1994), for example. Nishisato and Gaul (1988) talk of many methods (such as DS and CA) having a \common starting point" and \moving into a phase of unique advancement". He says that \the phenomenon of branching out will continue, and we may see the day when the name dual scaling will no longer be used synonymously with correspondence analysis". The same text appears in the introduction of Nishisato's 1994 book.

One particular example often cited as indicative of a di®erence between DS and CA is the case of so-called \dominance data", which includes pairwise comparisons, rank orders (or preferences) and successive categories (or ratings) data { see Nishisato 1978. Again in Nishisato (1994, p.185) it is stated that DS has the advantage over CA in being able to handle data with negative elements, in contrast to CA which is known to be applicable to nonnegative input data only. Quoting Nishisato and Gaul (1988, p. 152): \As is well known, correspondence analysis is not applicable to a data table which contains negative numbers. One of the distinct aspects of dual scaling, however, lies in the extension to a data table which contains negative numbers." In this note we show that for all three types of dominance data the two methods are actually equivalent, the only di®erence being that CA analyzes the uncentered data (with no negative elements) whereas the dual scaling formulation analyzes the centered data (with some negative elements). Furthermore, both of these equivalent analyses are in turn recoverable from an unstandardized principal components analysis of the preference matrix which has been centred with respect to subject (usuall row) means.

2 Preference data and dual scaling

We start with the simplest and most structured form of dominance data, that of preference data. We consider the same data set as Nishis

	6	1	5	2	4	3	-5	5	-3	3	-1	1
4	3	5	2	4	1	6	1	-3	3	-1	5	-5
5	3	4	2	6	1	5	1	-1	3	-5	5	-3
6	5	3	1	4	6	2	-3	1	5	-1	-5	3
7	1	2	4	5	3	6	5	3	-1	-3	1	-5
8	4	3	2	6	5	1	-1	1	3	-5	-3	5
9	2	1	4	5	3	6	3	5	-1	-3	1	-5
10	6	1	4	3	5	2	-5	5	-1	1	-3	3

Party Plans: A = potluck in the group room during the day; B = pub/restaurant crawl after work; C = reasonably priced lunch in an area restaurant; D = evening banquet at a hotel; E = potluck at someone's home after work; F = ritzy lunch at a good restaurant. Most preferred is 1, least preferred is 6.

We refer readers to Nishisato (1980, 1978, 1994) for an explanation and justification of the methodology. Here we merely summarize the result that to perform a dual scaling of the preference matrix, the first step is to calculate what is called a “dominance table”, and then effectively perform a singular-value decomposition of the table. Specifically, let the preference data, in Nishisato's notation, be denoted by the matrix $K(N \times n)$, with general element K_{ij}, where N is the number of subjects (respondents) and n the number of objects to rank order (thus in our example, $N = 10$ and $n = 6$). The dominance table, with general element e_{ij}, is defined as:

$$e_{ij} = n + 1 - 2K_{ij}$$

The optimal score vectors for subjects and objects solve the following pair of linear relationships:

$$x = \frac{c}{\rho}Ey \qquad y = \frac{c}{\rho}E^T x$$

where

$$c = \sqrt{\frac{1}{Nn(n-1)^2}}$$

These equations have several solutions, and we can subscript x_k, y_k and ρ_k by k to denote the k-th solution, or k-th dimension. The normalization proposed for these solutions is given by Nishisato (1978, 1980, p. 130) as $x_k{}^T x_k = y_k{}^T y_k = n$. There is a slight error, however, in the case of the subject score vectors, because as we shall see later, the normalization for x_k should be $x_k^T x_k = N$, which is indeed the normalization of the solution reported by Nishisato (1998, Table 1). These are what Greenacre (1984, pp. 93–95) calls the "standard coordinates" (see also Greenacre 1993, pp. 59–61, for a less technical, more substantive explanation). For graphical representation of the solution, Nishisato (1998) proposes the "asymmetric map" with the subject scores in standard coordinates and the object points in principal coordinates, that is the joint display of the x_k's and $\rho_k y_k$'s.

3 Correspondence analysis and doubling

Whereas CA is primarily a method for categorical data in the form of two-way and multi-way tables, it has been extended to many other types of data types (Benzécri 1973, Greenacre 1984, 1996). In the case of rating scales, variables are usually "doubled" to create a pair of variables, which can be called the positive and negative poles of the rating scale (see also Greenacre 1993, chapter 19). The convention is to translate the 1 to 5 rating scale, for example, to a 0 to 4 scale first, which is the positive pole of the doubled pair, and then subtract this value from 4 to obtain the negative pole. An alternative name for the variable and its doubled counterpart is "variable" and "anti-variable". In the case of a preference table, where the rows add up to a constant, there is not such an obvious reason for doubling the table and Greenacre (1984, pp. 183–184) has shown the difference between an undoubled and doubled correspondence analysis. There is another possibility for doubling a preference table, however, and this is to double the rows of the table, creating two rows for each subject which can be referred to as the "subject" and the "anti-subject". Again, we adopt the convention that ranks from 1 to 6, in our present example, are simply redefined to vary from 0 to 5. Then we subtract this ranking from 5 to obtain the doubled counterpart. For example, subject 1 has ranked party plan C with a 5, that is it is his or her second last preferred alternative. Subject 1 will be denoted by two rows of the doubled table, in which the third element corresponding to C will be $5 - 1 = 4$ in the first row and $5 - 4 = 1$ in the second row. These values can be thought of as counts, since in the former row the value 4 is the number of alternatives preferred to C and in the latter row the value 1 is the number of alternatives which C is preferred to. Thus, from a $N \times n$ table of 1-to-n preferences K_{ij}, let us define a $2N \times n$ table of doubled preferences where the first N rows are

$$L_{ij} = K_{ij} - 1$$

and the N additional rows are defined by:

$$M_{ij} = n ¡ 1 ¡ L_{ij}$$

(Table 2). We have assigned labels to the rows where a negative sign is attached to the subject numbers in the top half of the matrix since these are the counts against the respective object, whereas the lower half has a positive sign because these are the counts in favour of the respective object and in this sense indicating a positive association between the subject and the object.

Table 2
Preference Data in Doubled Format

	A	B	C	D	E	F
1-	5	0	4	3	2	1
2-	1	5	2	4	3	0
3-	5	0	4	1	3	2
4-	2	4	1	3	0	5
5-	2	3	1	5	0	4
6-	4	2	0	3	5	1
7-	0	1	3	4	2	5
8-	3	2	1	5	4	0
9-	1	0	3	4	2	5
10-	5	0	3	2	4	1
1+	0	5	1	2	3	4
2+	4	0	3	1	2	5
3+	0	5	1	4	2	3
4+	3	1	4	2	5	0
5+	3	2	4	0	5	1
6+	1	3	5	2	0	4
7+	5	4	2	1	3	0
8+	2	3	4	0	1	5
9+	4	5	2	1	3	0
10						

–

1). The correspondence

matrix P is thus:

$$P = \begin{bmatrix} L \\ M \end{bmatrix} / Nn(n-1)$$

the row and column masses are uniform in each case:

$$r = \frac{1}{2N}1 \qquad c = \frac{1}{n}1$$

and the diagonal matrices of the row and column masses are:

$$D_r = \frac{1}{2N}I \qquad D_c = \frac{1}{n}I$$

CA is then the singular-value decomposition (SVD) of:

$$\begin{aligned} S &= D_r^{-1/2}(P - rc^T)D_c^{-1/2} \\ &= (2N)^{1/2}\left(\begin{bmatrix} L \\ M \end{bmatrix} / Nn(n-1) - \frac{1}{2Nn}11^T \right) n^{1/2} \\ &= \frac{(2Nn)^{1/2}}{2Nn(n-1)} \left(2\begin{bmatrix} L \\ M \end{bmatrix} - (n-1)11^T \right) \\ &= \frac{1}{2^{1/2}(Nn)^{1/2}(n-1)} \begin{bmatrix} -E \\ E \end{bmatrix} \\ &= 2^{-1/2}c \begin{bmatrix} -E \\ E \end{bmatrix} \end{aligned} \tag{1}$$

where E is the dominance matrix defined above and

$$c = \frac{1}{(Nn)^{1/2}(n-1)} = \sqrt{\frac{1}{Nn(n-1)^2}}$$

is the same constant in Nishisato's formulation. Notice that the dominance matrix appears in the lower, or doubled, part of the augmented matrix, while its negative appears in the top half.

The SVD of S takes the form:

$$S = \begin{bmatrix} -U \\ U \end{bmatrix} D_\alpha V^T$$

where $U^TU + U^TU = V^TV = I$.

The row standard coordinates are:

$$\Phi = D_r^{-1/2}U = (2N)^{1/2}U$$

for the last N rows and $-\Phi$ for the first N rows, and the column principal coordinates are:

$$G = D_c^{-1/2}VD_\alpha = n^{1/2}VD_\alpha$$

We now show that this solution is identical to that of the dual scaling problem. From (1) we can write:

$$\begin{bmatrix} -E \\ E \end{bmatrix} = \begin{bmatrix} -2^{1/2}U \\ 2^{1/2}U \end{bmatrix} \left(\frac{1}{c} D_\alpha \right) V^T$$

which implies that the SVD of E is:

$$E = (2^{1/2}U) \left(\frac{1}{c} D_\alpha \right) V^T$$

since $2U^TU = I$. This also shows that the singular values α are the same as the ρ's in Nishisatos dual scaling definition. To satisfy the normalization of Nishisato's definition, the left and right vectors have to be multiplied by $N^{1/2}$ and $n^{1/2}$ respectively to obtain the standard coordinates and the right vectors have to be scaled by the singular values in $D_\alpha = D_\rho$ to become principal coordinates. This gives:

$$X = N^{1/2}(2^{1/2}U) \qquad Y = n^{1/2}VD_\alpha$$

which are identical to the CA coordinates.

The results of the CA are given in Table 3, and are identical to those given by Nishisato (1998, Table 1).

4 Equivalence to principal components analysis

Not only is DS equivalent to CA, but in this case they are both equivalent (up to scaling factors only) to an unstandardized principal components analysis (PCA) of the original preference table, centered with respect to row means, that is centering the column vectors. This is trivially seen by calculating the centered preference table, with elements $K_{ij} - \frac{1}{2}(n+1)$, that is in our example subtracting the constant row mean of $3\frac{1}{2}$ from every element of the table. The result is a table equal to $-\frac{1}{2}E$. Since PCA is just the SVD of the centered matrix, the results must be identical apart from a change of sign and a scaling factor, since the singular values of the centred matrix will not yield directly the values of ρ needed for the computation of principal coordinates.

Table 3

Results of Correspondence Analysis of Table 2

```
INERTIAS AND PERCENTAGES OF INERTIA
-----------------------------------

1 0.208971  44.78%  ************************************************
2 0.125402  26.87%  ******************************
3 0.067526  14.47%  ****************
4 0.037876   8.12%  *********
5 0.026892   5.76%  ******
  --------
  0.466667

ROW RESULTS (last 10 rows only)
-----------

+---+-----+------------+-------------+-------------+
|  I| NAME| QLT MAS INR|  k=1 COR CTR|  k=2 COR CTR|
+---+-----+------------+-------------+-------------+
| 11| A   | 698  50  50|  486 507  57|  299 191  36|
| 12| B   | 826  50  50|  -75  12   1| -616 814 151|
| 13| C   | 977  50  50|  535 614  69|  412 363  68|
| 14| D   | 738  50  50| -586 737  82|   16   1   0|
| 15| E   | 557  50  50| -509 555  62|  -27   2   0|
| 16| F   | 586  50  50|  407 356  40| -328 230  43|
| 17| G   | 672  50  50| -452 439  49|  330 234  43|
| 18| H   | 508  50  50|  289 179  20| -392 329  61|
| 19| I   | 650  50  50| -305 199  22|  458 450  84|
| 20| J   | 955  50  50|  641 880  98|  187  75  14|
+---+-----+------------+-------------+-------------+

COLUMN RESULTS
--------------

+---+-----+------------+-------------+-------------+
|  J| NAME| QLT MAS INR|  k=1 COR CTR|  k=2 COR CTR|
+---+-----+------------+-------------+-------------+
|  1| 1   | 778 167 186| -627 755 313| -110  23  16|
|  2| 2   | 810 167 209|  384 252 117|  571 558 433|
|  3| 3   | 540 167 106| -149  75  18| -371 465 183|
|  4| 4   | 243 167 129|  242 163  47|  170  80  38|
|  5| 5   | 694 167 140| -483 595 186|  197  99  52|
|  6| 6   | 940 167 231|  632 617 319| -457 323 278|
+---+-----+------------+-------------+-------------+
```

Results are taken verbatim from program SimCA (Greenacre 1986). The principal coordinates (multiplied by 1000) are given in the columns headed k=1 and k=2 respectively. Thus the results in Table 1 of Nishisato (1998) are identical for the objects (columns) which are in principal coordinates. To obtain the standard coordinates for the subjects, the principal coordinates should be divided by the square root of the corresponding principal inertias, i.e. $\sqrt{0.208971}$ and $\sqrt{0.125402}$ respectively. For example, the first coordinate of 1.06 reported by Nishisato for subject 1, dimension

1, can be obtained by calculating $0.486/\sqrt{0.208971}$. Notice that the coordinates of the first 10 rows are the same as the last 10, but with a change of sign.

5 Pairwise comparisons and ratings

The equivalence between dual scaling and correspondence analysis for these two other types of dominance data are easily deduced in a similar fashion.

In the case of paired comparisons, the data are set up directly in the format of the upper or lower ha

matrix will still sum to a constant $\frac{1}{2}n(n-1)$ because amongst the $n(n-1)$ ordered pairs of objects there is an equal number of preferences and dispreferences.

The only difference between the doubled matrix in this case and the doubled matrix in the case of the preference data, is that the data in each row may contain some identical values, whereas in the case of preferences the data in each row are a strict permutation of the integers $0, 1, 2, \ldots, n-1$.

As far as dual scaling of ratings is concerned, Nishisato (1980) proposed a way of treating ratings when the rating scale is identical for several items, e.g. each item has the ordinal scale bad, fair, good, excellent, i.e. an $m = 4$ point scale. The method is also discussed by Nishisato and Sheu (1984). The idea boils down to a rank-ordering of the $m-1$ cutpoints between the successive categories and the n items themselves, i.e. it is identical analytically to a preference table with $n+m-1$ objects. The novelty is in the coding of the data, which are then subject to the usual analysis of rank orders, which we have already shown to be equivalent to a doubled CA.

6 References

Benz¶cri, J.-P. & coll. (1973), L'Analyse des Donn¶es. Tome 2: l'Analyse des Correspondances, Dunod, Paris.

Greenacre, M.J. (1984), Theory and Applications of Correspondence Analysis, Academic Press, London.

Greenacre, M.J. (1986), \SimCA: A Program to Perform Simple Correspondence Analysis", American Statistician, 51, 230{231.

Greenacre, M.J. (1993), Correspondence Analysis in Practice, Academic Press, London.

Nishisato, S. (1978), \Optimal scaling of paired comparison and rank order data: an alternative to Guttman's formulation", Psychometrika, 43, 263{271.

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

Nishisato, S. (1994), Elements of Dual Scaling: An Introduction to Practical Data Analysis, Lawrence Erlbaum, New Jersey.

Nishisato, S. (1998), \An interpretable graph for rank order data", in Blasius, J. & Greenacre, M.J. (eds), Visualization of Categorical Data, Academic Press, San Diego, pp. 185{196.

Nishisato, S. & Sheu, W.-J. (1984), \A note on dual scaling of successive categories data, Psychometrika, 49, 493{500.

Nishisato, S. & Gaul, W. (1988), \Marketing data analysis by dual scaling", International Journal of Research in Marketing, 5, 151{170.