Analysis of Matched Matrices

Michael Greenacre

Facultat de Ciències Econòmiques i Empresarials
Universitat Pompeu Fabra
Ramon Trias Fargas, 25–27, E-08005 Barcelona, Spain

E-mail: michael@upf.es

Keywords: correspondence analysis, International Social Survey Program (ISSP), matched matrices, principal component analysis, singular-value decomposition.

Journal of Economic Literature Classification: C19, C88.

Abstract

We consider the joint visualization of two matrices which have common rows and columns, for example multivariate data observed at two time points or split according to a dichotomous variable. Methods of interest include principal components analysis for interval-scaled data, or correspondence analysis for frequency data or ratio-scaled variables on commensurate scales. A simple result in matrix algebra shows that by setting up the matrices in a particular block format, matrix sum and difference components can be visualized. The case when we have more than two matrices is also discussed and the methodology is applied to data from the International Social Survey Program.

1 Introduction

Principal components analysis (PCA), correspondence analysis (CA) and canonical variate analysis (CVA) are techniques based on the singular-value decomposition (SVD). In each case the geometric interpretation of the SVD permits the data matrix to be visualized in a low-dimensional Euclidean space (Greenacre and Underhill, 1982). This display, or \map", often contains points representing both the rows and the columns of the matrix, in what is known as a \biplot" of the matrix (Gabriel 1971). In a biplot the columns, say, which are often variables, are depicted by direction vectors which can be calibrated according to the original scales of the variables. The row points may then be projected onto these \biplot axes" in order to estimate the values in the original data matrix (Gabriel and Odoroff 1990 ; Greenacre 1992; Gower & Hand 1996). The success of the recovery of the data by their projections is measured by the percentage of explained variance in the map.

Often we would like to interpret and compare two or more matrices of the same size, which have rows and columns referring to the same entities (here we discuss the case of two matrices, we shall consider the general case later on). For example, we might have two cases-by-variables matrices where the first matrix contains observations at time point 1 and the second matrix contains observations on the same cases at time point 2. Or the two matrices may contain averages, or frequency counts, according to two separate subsets of the sample, for example males and females. We refer to such data matrices in general as matched matrices, in the sense of being repetitions over time or being split into subsamples.

One approach to the visual interpretation of matched matrices is to concatenate them row-wise or column-wise and then apply the usual analysis, be it PCA or CA, whichever is the more appropriate. In the first example mentioned above, a PCA, say, of the two matrices, one stacked on top of the other, would result in two points for each case. Each pair can be connected in the display to show that the data are paired observations and to allow the interpretation of each case's change over time. In the second example we would stack the male and female tables and a CA, say, would lead to a set of male points and a set of female points. In a similar way, male{female differences would be interpreted by comparing pairs of points for the same row object: for example if the rows were education groups, then we would be interpreting differences between males and females for each category of education.

Notice that although we have described the above joint analyses of the two matrices as a study of differences, in neither of the analyses is this an explicit objective of the visualization. The display optimally displays the individual points, which may or may not lead to an accurate display of the differences, which are the vectors joining the pairs of points. Clearly, if the quality of display (i.e., percentage of explained variance) is very high, then the points are accurately displayed and their difference vectors as well. However, when the quality of display is not high, it can often turn out that differences are poorly represented. In fact, it is our experience that in this case the map concentrates more on the sum of the matrices than their differences. We should be careful, therefore, in our interpretation of concatenated matrices by also calculating measures of quality of the display of differences, just as we measure qualities of display of individual points. If differences are of specific interest, then it would be appropriate to perform a separate analysis of the matrix of differences. For both PCA and CA analyzing the matrix difference implies performing an uncentred analysis, an option which is generally not available in PCA and CA software packages. Fortunately, we will show that an analysis of the sum and the difference between two matrices can be achieved in a single application of PCA or CA, with the usual centring, where the two matrices are set up in a certain block format. This requires no extra calculations or special software, only a versatile text editor to prepare the data. We will also show how the block matrix idea generalizes to more than two matrices, depending on the nature of the repetitions.

In Section 2 we summarize the basic mathematical result which serves as the basis for our methodology, and show how this result can be exploited in a principal component analysis and a correspondence analysis. In Section 3 we discuss an application to data concerning attitudes to whether women should stay at home or work after they get married, comparing responses from men and women across several countries. In Section 4 we discuss how the method extends to more than two matched matrices.

2 A simple result with many applications

2.1 SVD of a block matrix

A general result which we use throughout is the following:

> If A and B are two $n \times m$ matrices, then the SVD of the sum $A + B$ and the difference $A - B$ can be recovered in the SVD of the block matrix:
>
> $$\begin{bmatrix} A & B \\ B & A \end{bmatrix}$$

It is easily verified that if the SVDs of $A + B$ and $A - B$ are respectively:

$$A + B = UD_\alpha V^T \qquad A - B = XD_\beta Y^T \tag{1}$$

then the SVD of the $2n \times 2m$ block matrix is (up to an ordering of the singular values and corresponding singular vectors):

$$\begin{bmatrix} A & B \\ B & A \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} U & X \\ U & -X \end{bmatrix} \begin{bmatrix} D_\alpha & 0 \\ 0 & D_\beta \end{bmatrix} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} V & Y \\ V & -Y \end{bmatrix} \right)^T \tag{2}$$

Notice the following:

1. The left and right singular vectors are all orthogonal to one another thanks to the orthogonality of the vectors in the SVDs (1) and the change in sign of the matrices X and Y.

2. The presence of the factor $1/\sqrt{2}$ multiplying the left and the right singular vectors ensures the correct normalization of the solution: for example, since $U^TU = I$, we have for the corresponding columns of the left singular vectors in (1):

   $$\left(\frac{1}{\sqrt{2}} \begin{bmatrix} U \\ U \end{bmatrix} \right)^T \frac{1}{\sqrt{2}} \begin{bmatrix} U \\ U \end{bmatrix} = \frac{1}{2}U^TU + \frac{1}{2}U^TU = I$$

   Hence the left and right singular vectors in (2) are orthonormal.

3. The SVDs of the sum $A + B$ and of the difference $A - B$ do not appear separated as indicated in (1), but interleaved according to the magnitude of the corresponding singular values. In the SVD of the block matrix it is easy to distinguish the solution

   vectors corresponding to the sum and the difference: left and right singular vectors corresponding to the sum have two identical copies of a vector stacked on top of each other, whereas singular vectors corresponding to the difference have a vector stacked on top of the negative of the vector.

The total sum-of-squares in the block matrix is thus decomposed into two components, one component due to the matrix sum and one due to the matrix difference:

$$2\sum_i \sum_j a_{ij}^2 + 2\sum_i \sum_j b_{ij}^2 = \sum_i \sum_j (a_{ij} + b_{ij})^2 + \sum_i \sum_j (a_{ij} - b_{ij})^2 \tag{3}$$

2.2 Principal component analysis

If the columns of A and B are interval-scaled variables, then PCA would be the method of choice for visualizing the matrices. A and B are optionally standardized prior to analysis, depending on the variables' measurement scales and their inherent variances. The average row of A is $\bar{a}^T = (1/n)1^T A$ and the average row of B is $\bar{b}^T = (1/n)1^T B$. Centring in the joint analysis of A and B is with respect to the average of A and B: $\bar{c} = (1/2)(\bar{a} + \bar{b})$. Performing the PCA of the block matrix involves centring with respect to the row $\left[\bar{c}^T \;\; \bar{c}^T\right]$:

$$\begin{bmatrix} A - 1\bar{c}^T & B - 1\bar{c}^T \\ B - 1\bar{c}^T & A - 1\bar{c}^T \end{bmatrix} \tag{4}$$

The sum-of-squares decomposition corresponding to (3) is thus:

$$2\sum_i\sum_j (a_{ij} - \bar{c}_j)^2 + 2\sum_i\sum_j (b_{ij} - \bar{c}_j)^2 = \sum_i\sum_j (a_{ij} - \bar{c}_j + b_{ij} - \bar{c}_j)^2 + \sum_i\sum_j (a_{ij} - b_{ij})^2 \tag{5}$$

so that the matrix sum component is in centred form and the component corresponding to the matrix difference is in uncentred form.

Notice that the factor $1/\sqrt{2}$ in (2) only appears in the intermediate calculations of the PCA and disappears in the graphical display when one imposes the usual scaling on the principal coordinates to have variance equal to the corresponding squared singular value. Since there are 2n rows in the block matrix, each row has an identical weight of $1/(2n)$. The principal coordinate matrix F is thus:

$$F = \sqrt{2n}\,\frac{1}{\sqrt{2}} \begin{bmatrix} U & X \\ U & -X \end{bmatrix} \begin{bmatrix} D_\alpha & 0 \\ 0 & D_\beta \end{bmatrix}$$

which gives identical principal coordinates to those obtained in the analyses of the matrix sum and matrix difference, where there are n points, each with weight $1/n$, leading to principal coordinates $\sqrt{n}\,U D_\alpha$ and $\sqrt{n}\,X D_\beta$ respectively. In all cases the variance of the points (2n points in the case of the block matrix, n points for the sum and n points for the difference) is equal to the squared singular value.

2.3 Correspondence analysis

Apart from the fact that points have different weights in CA and there is an inherent standardization in the form of the chi-square metric, everything goes through in an analogous fashion. Principal coordinates of row points and column points of the block matrix recover exactly the row and column points of the centred matrix sum and the uncentred matrix difference. However, it is the substantive issue in CA which is more relevant here. If A and B are two cross-tabulations based on different subsamples (e.g., males and females), then $A + B$ is the accumulation of the two subsamples and $A - B$ is the difference, cell by cell, of the frequencies. If the subsamples differ in overall frequency, or in marginal frequencies, then these differences will be displayed in the analysis of the matrix difference. For example, if there are many more males than females in the data set, then the matrix difference will mostly reflect this difference in sample sizes, not the differences between the male and female responses. It would be preferable here to reweight the samples or to analyse the data at the profile level, that is expressed in row or column percentages. Studies where the subsamples are of equal size do not present such a difficulty, especially when the data have been collected according to some fixed design which gives the matrix difference a substantive interpretation. For example, suppose that the rows of the tables are education groups and that the columns are the responses to a question in an opinion survey. If equal numbers of males and females in each education group are included in the survey, then we have no difference in the row margins and the matrix difference reflects gender differences in responses and not differences in education. Of course, if we were interested in educational differences as well, we would only impose the restriction that the overall sample sizes be equal.

3 Application: PCA of male–female attitudes across countries

To illustrate the approach, consider the data in Table 1 extracted from the International Social Survey Programme (ISSP) database. The data are percentage reponses to four questions related to woman staying at home or working after marriage and at different points of time in their married lives. The response percentages have been calculated separately for males and females.

The usual way to display these data is to analyze the 16×4 table, for example using PCA[1]. The two-dimensional map of the data is given in Figure 1 (all figures are in the Appendix). The row points are country–gender points, for example Germany–female and Germany–male are indicated by Df....m. Each pair of points has been connected and the differences generally coincide with the horizontal axis, with male

[1] Since these are percentages, we might want to apply some transformation to the data before performing PCA, for example a logarithmic or arcsine transformation (e.g., Aitchison 1986). Here we apply PCA directly to the data, without any standardization, to simplify the illustration of our approach.

points on the left and female points on the right. The four columns are depicted by vectors showing the directions of the biplot axes, with points 1 and 4 more in a vertical direction and 2 and 3 contributing strongly to both axes (the contributions could be confirmed more formally by looking at the column contributions to the principal axes).

Table 1

Male and female views of working wives in eight countries (ISSP, 1989)

Country		Should wife stay at home...? (response percentage)			
		...before first child	...after first child	...when first child is at school	...when all children are at school
	D	6.3	78.3	51.4	14.6
	GB	3	74.7	15.3	4
M	USA	7.6	61.1	16.2	7.1
A	A	5.1	75.4	45.7	12.2
L	H	18.9	58.4	22.1	8.7
E	NL	3	60	17.3	3.6
	I	11.1	49.6	23.6	21.7
	IRE	7	56.4	33.5	9.2
	D	6.1	73.9	47.7	14.5
F	GB	2.4	66.6	10	1.9
E	USA	4	50	10.3	3.8
M	A	2.9	69.4	40.5	7.3
A	H	7.2	46.5	14.9	3.4
L	NL	1.5	52.2	10	2.3
E	I	3.8	38.3	12	10
	IRE	5.8	54.6	20.7	5.9

Each value is the percentage of respondents who are in favour of the wife staying at home in the following four periods: (1) before the first child is born; (2) after the birth of the first child; (3) after the first child has gone to school; and (4) after all children are at school. For example, 6.3% of German males in the sample do not want their wives to work after they are married (before the first child), while 6.1% of German females do not want to work just after they are married.

The interpretation would thus be as follows. Irrespective of gender of the respondents, Germans and Austrians more frequently want women to stay at home, while the British distinguish themselves by their high frequency of responses in favour of women staying at home in the early years of the first child. As far as male{female differences are concerned, men always want their wives to stay at home more than the wives themselves do. In some countries these differences are greater than others, for example the Italian and Hungarian differences look larger than the Austrian and German differences. It is not easy in this map to diagnose any strong tendency for the difference to be concentrated in any particular period. t does look like the Irish have a male–female difference which is along the vector (3) and not along vector (2). Looking at Table 1 we can confirm that there is a very small difference between males and females about whether a wife should work after the first child is born (males: 56.4%, females: 54.6%), but they disagree somewhat when the first child goes to school (males: 33.5%, females: 20.7%).

To clarify the quality of display of the differences we calculated the sums of squares corresponding to equation (3), where the a_{ij}'s and b_{ij}'s are centred with respect to the overall mean:

$$11878 = 10364 + 1514$$

Of the total sum of squares on the left, the joint map explains a total of 11090, that is an explained variance of 11090/11878, or 93.4%, which seems an excellent result at a first glance. However, by calculating the sum of and difference of each pair of vectors in the map and computing their sums of squares separately we see that the explained amount of 11090 has a component of 9934 explaining the sum and 1153 explaining the difference. This gives the following results for the sum and difference respectively:

$$\text{sum} : 100 \times 9934/10364 = 95.9\% \qquad \text{difference} : 100 \times 1153/1514 = 76.2\%$$

The component of the total sum of squares due to the matrix sum is much larger than that due to the difference, and the solution is clearly dominated by this fact. The overall quality of 93.4% is a combination of a 95.9% explanation of the larger sum component and a 76.2% explanation of the smaller difference component.

If we are specially interested in the differences, which is usually the case in a study like this one, then a specific analysis of the differences should be carried out. There are two ways of executing the analysis. The first way is to calculate the matrix difference and then perform an uncentred PCA of this difference. Clearly we do not want to remove the average difference from each column of the difference matrix, but we want to analyze and thus visualize the male–female differences in a space where the origin refers to the value zero, not the average of the differences. The second way is to apply the convenient result of Section 2, setting up the male and female matrices twice each in the 2 £ 2 block format and then perform a standard PCA of this 16 £ 8 block matrix. There is some superfluous computation involved in this latter approach, but the great advantage is that no special data preparation or additional programming is required.

We thus illustrate the second approach and show the complete results of the SVD of the block matrix in Table 2.

Table 2
SVD of 16 £ 8 block matrix of male and female data

DIMENSION	1	2	3	4	5	6	7	8
SINGULAR VALUES								
	89.35	44.55	36.39	15.39	12.63	10.30	7.60	5.05
LEFT SINGULAR VECTORS								
FIRST BLOCK	0.47775	0.07196	-0.09808	-0.04665	-0.19901	0.05281	0.24409	-0.05367
	-0.07161	-0.50671	-0.17897	-0.11884	-0.28202	0.01612	0.32953	0.14185
	-0.19302	-0.08330	-0.25178	-0.00014	-0.01181	-0.14144	0.28653	0.06735
	0.35262	-0.01539	-0.18301	-0.00683	0.20322	0.02646	-0.02169	0.24145
	-0.15466	0.12845	-0.34560	0.55586	-0.20904	-0.36809	-0.17683	-0.43645
	-0.18635	-0.15312	-0.19809	-0.10391	0.39216	0.09579	0.31663	-0.10909
	-0.21515	0.41173	-0.40427	-0.38957	-0.20777	0.04213	-0.32753	0.36381
	-0.00958	0.14638	-0.19836	0.11007	0.31428	0.57428	-0.10225	-0.28182
SECOND BLOCK	0.47775	0.07196	0.09808	-0.04665	-0.19901	-0.05281	-0.24409	0.05367
	-0.07161	-0.50671	0.17897	-0.11884	-0.28202	-0.01612	-0.32953	-0.14185
	-0.19302	-0.08330	0.25178	-0.00014	-0.01181	0.14144	-0.28653	-0.06735
	0.35262	-0.01539	0.18301	-0.00683	0.20322	-0.02646	0.02169	-0.24145
	-0.15466	0.12845	0.34560	0.55586	-0.20904	0.36809	0.17683	0.43645
	-0.18635	-0.15312	0.19809	-0.10391	0.39216	-0.09579	-0.31663	0.10909
	-0.21515	0.41173	0.40427	-0.38957	-0.20777	-0.04213	0.32753	-0.36381
	-0.00958	0.14638	0.19836	0.11007	0.31428	-0.57428	0.10225	0.28182
RIGHT SINGULAR VECTORS								
FIRST BLOCK	-0.02333	0.17371	-0.24637	0.54362	-0.41684	-0.34932	-0.37862	-0.41704
	0.39287	-0.52517	-0.44946	-0.01754	-0.26372	-0.30802	0.43078	0.13244
	0.58004	0.31295	-0.40974	0.10429	0.23397	0.53145	0.02071	-0.22190
	0.09308	0.30999	-0.26343	-0.43964	-0.44940	0.02561	-0.41310	0.50921
SECOND BLOCK	-0.02333	0.17371	0.24637	0.54362	-0.41684	0.34932	0.37862	0.41704
	0.39287	-0.52517	0.44946	-0.01754	-0.26372	0.30802	-0.43078	-0.13244
	0.58004	0.31295	0.40974	0.10429	0.23397	-0.53145	-0.02071	0.22190
	0.09308	0.30999	0.26343	-0.43964	-0.4494	-0.02561	0.41310	-0.50921

Notice the block form of the singular vectors, as given by (2). Looking at the relative signs only of the left (or right) singular vectors, and marking the sign of the first block throughout as positive, the vectors are arranged in the following pattern:

$$\begin{bmatrix} + & + & + & + & + & + & + & + \\ + & + & \text{¡} & + & + & \text{¡} & \text{¡} & \text{¡} \end{bmatrix}$$

Hence dimensions 1, 2, 4 and 5 correspond to the matrix sum component and dimensions 3, 6, 7, and 8 correspond to the matrix difference. This is what we meant previously when we said the components would be interleaved in the solution of the block matrix.

The total sum of squares due to the sum and difference components are thus, respectively:

$$\text{sum}: \quad 89.349^2 + 44.548^2 + 15.394^2 + 12.634^2 = 10364$$

$$\text{difference}: \quad 36.393^2 + 10.297^2 + 7.5997^2 + 5.046^2 = 1514$$

which are the components we obtained before by direct calculation. Thus separate maps of dimensions 1 and 2 together and dimensions 3 and 6 together visualize the sum and difference components in their respective optimal planes (Figures 2 and 3).

Figure 2 gives the overall view of the countries, irrespective of the gender differences, and here we see the general picture we saw in Figure 1 of Germany and Austria separating out to the right and a vertical spread of the other countries on the left with Italy at the top, generally more in favour of women staying at home in the early and later periods of marriage, and Great Britain below in favour of women staying at home after the birth of the first child. Notice the similarity of the biplot vectors in this Figure compared to Figure 1. The percentage of variance explained of the total variance (11878) is 83.9%, which if expressed as a percentage of the sum component (10364) rises to 96.2%.

Figure 3 shows us a more accurate map of the differences, which we did not see before. Notice that the biplot vectors are in different positions now, all positive on the horizontal axis, but opposing the third period mainly against the first two along the vertical axis[2]. Since we are looking at male–female differences all the positive scalar products formed by the countries with the vectors show us that all the differences are positive. Here Ireland is maybe the sole exception since it is practically orthogonal to the first two variables, which indicates very little male–female difference here, already notices in Figure 1 in the case of the second variable. We also seee again that Irish males show a strong difference on the third period. The biggest male–female differences are for Italy and Hungary, since these countries are furthest from the centre. Hungarian males are particularly different from females with respect to the first two periods. This map explains 94.5% of the sum-of-squared differences, which is 12.0% of the total variance.

[2] The three figures in this study would all benefit from axis rotation. After rotation in Figure 3 the first two variables would coincide with one axis, the third variable with the other, with the fourth variable more or less equally loaded on both axes.

In summary, the two maps explain a total of 95.9% of the variance, or 96.2% of the sum component and 94.5% of the difference component. Notice that interpreting more than two dimensions in a PCA is normally a complicated exercise because the later dimensions need to be interpreted remembering what has already been accounted for in the previous dimensions. But in this case, by separating out the variance into sum and difference components, the interpretation is compartmentalized and thus easier to manage and comprehend, even though the sum and difference components are not orthogonal in multidimensional space.

4 Special case: square matrices

Previous work by Greenacre (2000) is a special case of the present one. Let N be a square asymmetric matrix. The SVD of of the symmetric part of N, $\frac{1}{2}(N + N^T)$ and the skew-symmetric part, $\frac{1}{2}(N - N^T)$, can be recovered in the SVD of the block matrix:

$$A = \begin{bmatrix} N & N^T \\ N^T & N \end{bmatrix}$$

In this case, matrices A and B are the square matrix N and its transpose N^T respectively.

5 Doubling of preferences in correspondence analysis

Greenacre and Torres (1999) have investigated different ways of analyzing preferences, or ranking data, and other types of so-called “dominance data”, using correspondence analysis. Suppose that K is a table of rankings, where the rows are respondents and the columns are the n ranked items, and the ranks are from 1 to n where 1 indicates most highly preferred item. Greenacre and Torres have shown that the correspondence analysis of the “doubled” matrix:

$$\begin{bmatrix} L \\ M \end{bmatrix}$$

where

$$L_{ij} = K_{ij} - 1 \quad \text{and} \quad M_{ij} = n - 1 - L_{ij};$$

gives a solution equivalent to that of dual scaling (Nishisato, 1980). The equivalence lies in the fact that dual scaling is the analysis of the difference $M - L$, called the \dominance table" by Nishisato. This analysis is in turn equivalent to a row-centred principal component analysis of the matrix H with general element $H_{ij} = (n+1) - K_{ij}$, so that the item with value n is the one most preferred, and 1 is the least preferred. Note that H and K differ from M and L respectively by 1, the former ranging from 1 to n and the latter from 0 to $n - 1$. It is the latter form which is used in correspondence analysis (see Greenacre and Torres 1999).

It is clear that we could obtain the same solution by applying the usual column-centred principal component analysis to either of the following block matrices:

$$\begin{bmatrix} H & K \\ K & H \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} M & L \\ L & M \end{bmatrix}$$

In both cases the sum part will be constant (yielding zero eigenvalues) and the difference part will be exactly the dominance matrix.

6 More than two matched matrices: two sets of matched pairs

The above ideas can be generalized to two sets of matched matrix pairs. Consider the case where a matched pair of matrices is itself matched with another matched pair. This could be, for example, two matrices summarizing results in a survey for males and females, and both of these repeated at another time point. Let A_1 and B_1 be the data for males and females at time point 1 and A_2 and B_2 the corresponding results at time point 2. We can set up the four data matrices in a 2×2 block format with each block itself a 2×2 block matrix { we call this a nested block matrix:

$$\begin{bmatrix} A_1 & B_1 & A_2 & B_2 \\ B_1 & A_1 & B_2 & A_2 \\ A_2 & B_2 & A_1 & B_1 \\ B_2 & A_2 & B_1 & A_1 \end{bmatrix}$$

Applying the result of Secftion 2 to the \major" blocks, we recover the SVDs of the sum and difference matrices

$$\begin{bmatrix} A_1 + A_2 & B_1 + B_2 \\ B_1 + B_2 & A_1 + A_2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A_1 - A_2 & B_1 - B_2 \\ B_1 - B_2 & A_1 - A_2 \end{bmatrix}$$

And then , applying the result a second time to each of these \minor" block matrices, we recover the SVDs of four different matrices:

(1) $A_1 + A_2 + B_1 + B_2$

(2) $(A_1 + A_2) - (B_1 + B_2)$

(3) $(A_1 + B_1) - (A_2 + B_2)$

(4) $(A_1 - B_1) - (A_2 - B_2)$

The first matrix consists of the overall results irrespective of sex or time of the survey. The second matrix is the difference between males and females irrespective of time points (data from the two points are aggregated for each sex). The third matrix is the difference between the time points irrespective of sex (the male and female data are aggregated at each time point). The fourth matrix is the change in the male–female difference from time point 1 to time point 2. In this way the different components of variance, a main effect, a marginal effect for time and a marginal effect for sex, and finally a time–sex interaction effect, are separated in the analysis.

The special case of square matrices has been dealt with by Greenacre and Clavel (1998). If we have two square matrices N and M, for example, two transition matrices between different time points, or two transition matrices for two groups (e.g., males and females), then we can analyze the 4×4 block matrix:

$$\begin{bmatrix} N & N^T & M & M^T \\ N^T & N & M^T & M \\ M & M^T & N & N^T \\ M^T & M & N^T & N \end{bmatrix}$$

which is equivalent to the analysis of the two block matrices:

$$\begin{bmatrix} N + M & N^T + M^T \\ N^T + M^T & N + M \end{bmatrix} \text{ and } \begin{bmatrix} N - M & N^T - M^T \\ N^T - M^T & N - M \end{bmatrix}$$

which, in turn, is equivalent to the analysis of the four matrices:

(1) $N + M + N^T + M^T$

(2) $(N + M) - (N^T + M^T)$

(3) $(N + N^T) - (M + M^T)$

(4) $(N - N^T) - (M - M^T)$

The first matrix can be called the average symmetric component; the second, the average skew-symmetric component; the third, the difference between symmetric components; and the fourth, the difference between skew-symmetric components.

7 Further results

We can also look at generalizations of the idea to the case of three or more matched matrices.

For example, consider the case of three matched matrices: A_1, A_2 and A_3. We set up the following 3×3 block matrix:

$$A = \begin{bmatrix} A_1 & A_2 & A_3 \\ A_3 & A_1 & A_2 \\ A_2 & A_3 & A_1 \end{bmatrix}$$

which is called a block circulant matrix.

Then it can be shown[3] that the analysis of A yields a component corresponding to the sum $A_1 + A_2 + A_3$ and the remaining components corresponding to the matrix of differences:

$$\begin{bmatrix} A_1 - \frac{1}{3}(A_1 + A_2 + A_3) & A_2 - \frac{1}{3}(A_1 + A_2 + A_3) & A_3 - \frac{1}{3}(A_1 + A_2 + A_3) \\ A_3 - \frac{1}{3}(A_1 + A_2 + A_3) & A_1 - \frac{1}{3}(A_1 + A_2 + A_3) & A_2 - \frac{1}{3}(A_1 + A_2 + A_3) \\ A_2 - \frac{1}{3}(A_1 + A_2 + A_3) & A_3 - \frac{1}{3}(A_1 + A_2 + A_3) & A_1 - \frac{1}{3}(A_1 + A_2 + A_3) \end{bmatrix}$$

which can also be written as:

$$\frac{2}{3}\begin{bmatrix} A_1 - \frac{1}{2}(A_2 + A_3) & A_2 - \frac{1}{2}(A_1 + A_3) & A_3 - \frac{1}{2}(A_1 + A_2) \\ A_3 - \frac{1}{2}(A_1 + A_2) & A_1 - \frac{1}{2}(A_2 + A_3) & A_2 - \frac{1}{2}(A_1 + A_3) \\ A_2 - \frac{1}{2}(A_1 + A_3) & A_3 - \frac{1}{2}(A_1 + A_2) & A_1 - \frac{1}{2}(A_2 + A_3) \end{bmatrix}$$

that is, the block circulant matrix of differences between each submatrix and the average of the other two submatrices.

Furthermore we can show that the dimensions of the difference components occur in “bimension” pairs, i.e. pairs of dimensions with equal singular values, and the coordinates of the rwo and column points in the form of equilateral triangles.

This result generalizes to four and more matrices in a surprising way. One might expect that for four matrices, the difference components would be “trimensions” with the points forming tetrahedra, but they still occur in bimension pairs, with the points forming squares. For five matrices, they form pentagons still in two dimensions, and so on.

[3] This is the subject of another research report in preparation.

References

Aitchison, J. (1986), The Statistical Analysis of Compositional Data, Chapman & Hall, London.

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

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

Gabriel, K.R. Odoro®, C. (1990), Biplots in Medical Research, Statistics in Medecine,9, 469{485.

Gower, J.C. and Hand, D.J. (1996), Biplots. Chapman and Hall, London.

Greenacre, M.J. & Underhill, L.G. (1982), Scaling a data matrix in low-dimensional Euclidean space, in Hawkins, D.M. (ed.), Topics in Applied Multivariate Analysis, Cambridge University Press, Cambridge.

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

Greenacre, M.J. (1992), Biplots in correspondence analysis, Journal of Applied Statistics, 20, 251{269.

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

Greenacre, M.J. (2000), Correspondence analysis of square asymmetric matrices, Applied Statistics, 49, 297{310.

Greenacre, M.J. and Clavel, J.G. (1998), Analysis of a pair of transition matrices, Economics Working Paper 298, Universitat Pompeu Fabra.

Greenacre, M.J. and Torres, A, (1999), A note on the dual scaling of dominance data and its relationship to correspondence analysis, Economics Working Paper 430, Universitat Pompeu Fabra.

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

van der Heijden, P.G.M. and Mooijaart, A. (1995), Some new log bilinear models for the analysis of asymmetry in a square contingency table, Sociological Methods and Research.

Figures

Figure 1

Principal Component Analysis of Table 1 stacked

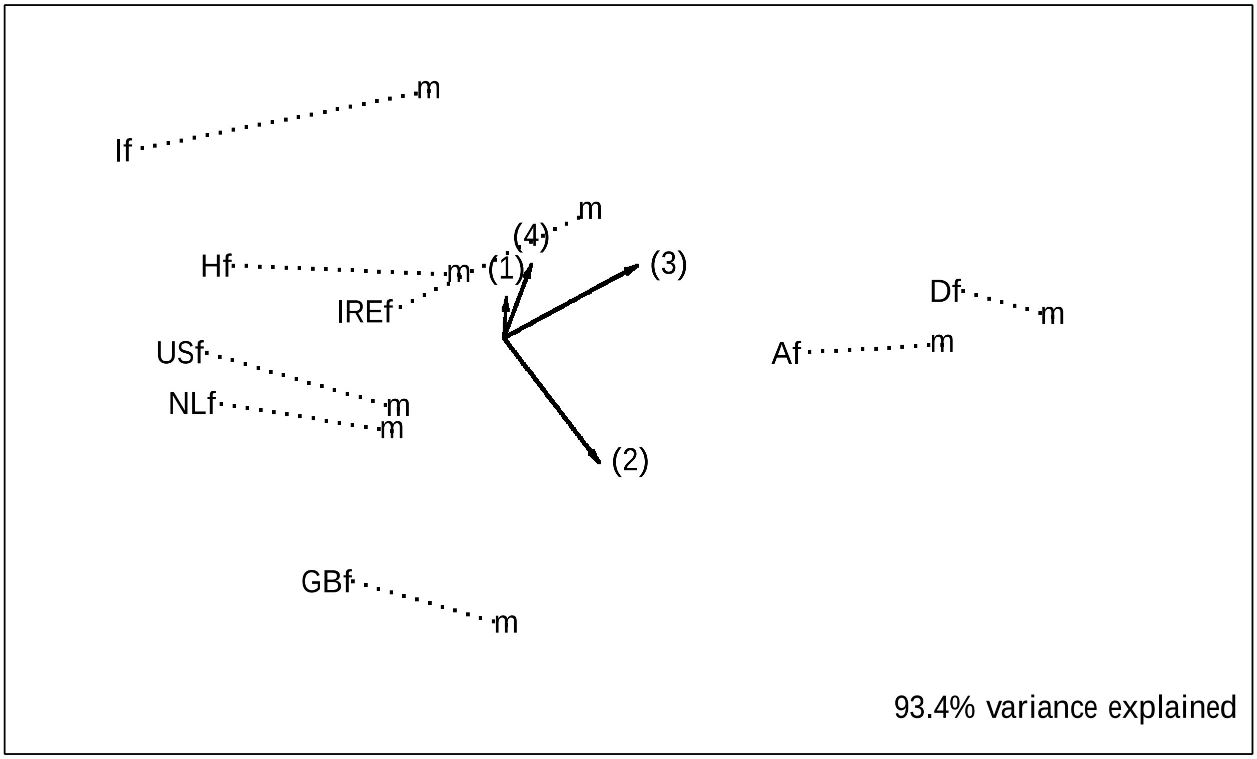


Figure 2
Principal Component Analysis of Sum Component

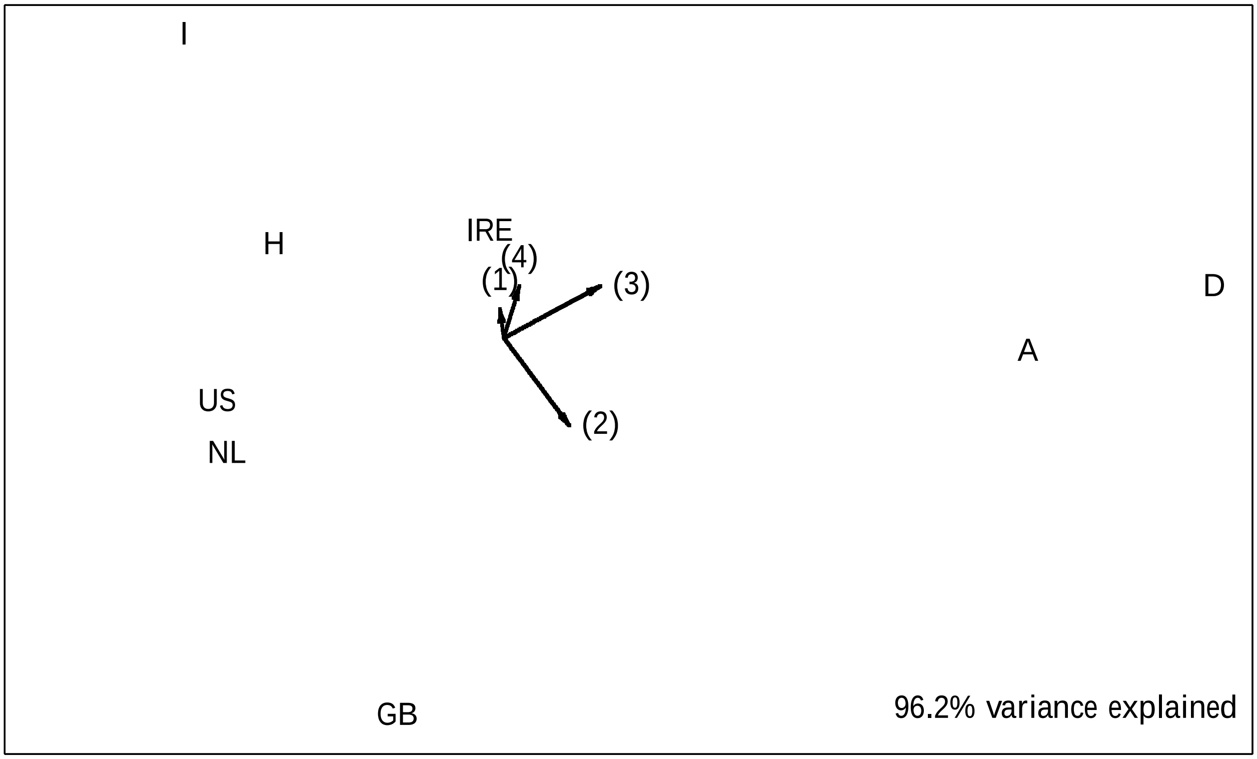
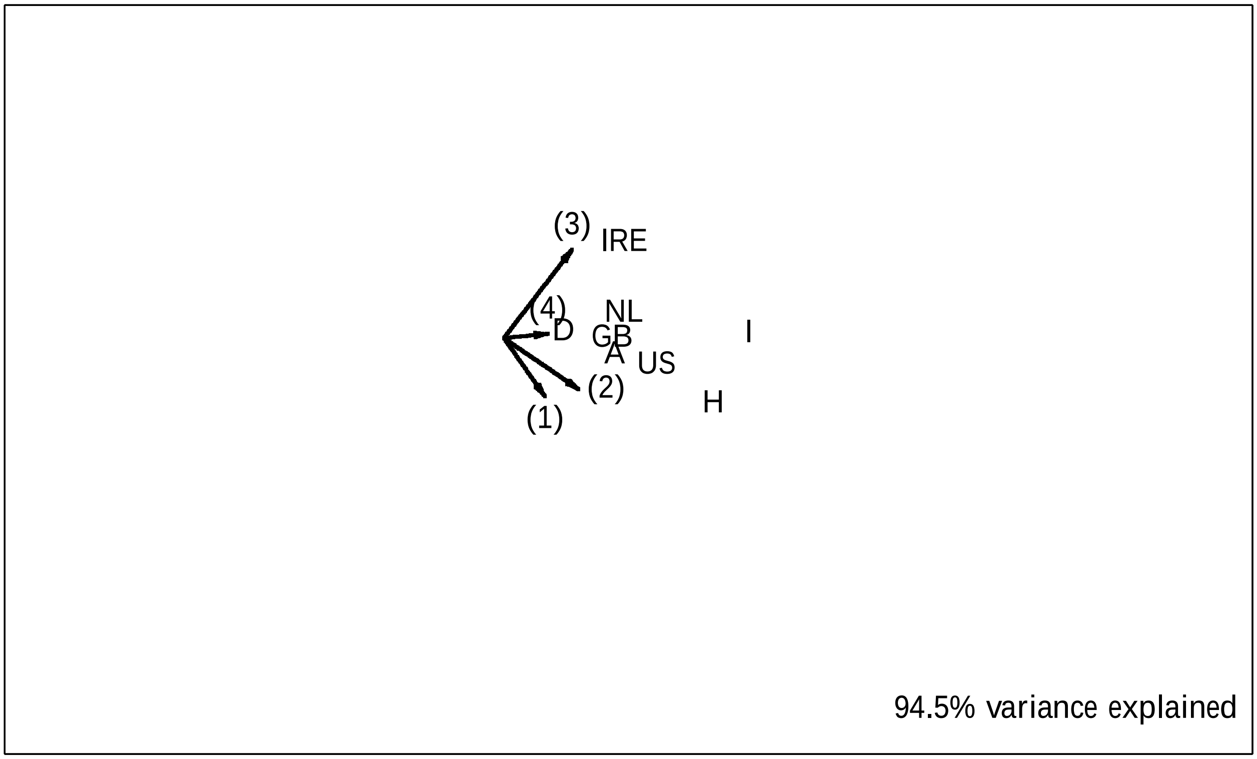


Figure 3

Principal Component Analysis of Di®erence Component


(3)
IRE
(4)
NL
D
GB
I
A
US
(2)
H
(1)
94.5% variance explained