CORRESPONDENCE ANALYSIS AS A TOOL FOR SURVEY STATISTICIANS

Margaret J. Krier and J. Edward Jackson
Eastman Kodak Company

Introduction

One almost certain result of a survey will be the presentation of one or more cross-tabulations of frequencies. These cross-tabs may then be subjected to contingency table analysis to determine whether or not the factors represented by the rows are independent of those represented by the columns. In the past few years, great strides have been made in facilitating the analysis of multidimensional contingency tables with such software packages as BMDP4F and C-TAB. If a particular table exhibits independence, then that table may be replaced in a report with only the marginals, making the report shorter and easier to assimilate. (It is our observation that in market research, at least, these tests for multidimensional tables are seldom carried out and the client is still being buried with pounds of cross-tabs which may be unnecessary.)

Now, suppose one has a contingency table which contains one or more significant interactions. How are these conditions to be described? If the number of levels is fairly small, a glance at the table itself, or the table of residuals, may be sufficient to determine the nature of the lack of independence. If the table is larger, and in particular has more than two dimensions, this may not be so easy to do. There have been a number of analytical techniques proposed to aid in this process but most of them still require some insight into the nature of the problem. In this day and age, graphical procedures are becoming quite useful in many fields of statistical diagnoses and and we would like to describe one such technique, Correspondence Analysis, for use with contingency tables.

CORRESPONDENCE ANALYSIS

Correspondence Analysis (CA) is a multidimensional scaling technique. The concept goes back to the thirties (Hill, 1974) but it is only in the past ten years or so that much headway has been made towards making the technique operational. The principal activity in this area has come from the French who, more properly, refer to this as the Analysis of Correspondence although it has also been called the method of reciprocal averages (see for example, Lebart et al., 1977, 1984). There are many alternate motivations and formulations for CA in the literature [Fisher (1940), Hill (1974), Nishisato (1980) and Lebart et al (1977, 1984) etc.] Due to space considerations, we will not consider these various formulations but will look at the characterization of CA from the singular value decomposition viewpoint. For those familiar with other multi-dimensional scaling techniques, it is similar in many ways to such point-vector (linear compensatory) models such as Biplots (Gabriel, 1981) or MDPREF (Green and Rao, 1972). Both of these can be used to represent dominance data (e.g. preference scores or ranks) for t stimuli by n respondents such that the stimuli and respondents may be represented on the same graph. Essentially, these methods decompose an nxt data matrix X into the product of two matrices ($X=AB$) where A is an nxk matrix representing the respondents, B is an kxt matrix representing the stimuli and k is the dimensionality required to adequately represent this set of data [$k \leq \min(n,t)$]. A and B are obtained by the singular value (Eckart-Young) decomposition of X which essentially amounts to obtaining the eigenvectors of both $X'X$ and XX'. For preference data, the stimuli are customarily represented as points and respondents as vectors. The locations of the projections of the individual stimuli on each vector should be correlated with the stimulus ratings for the respondent. While often used for preference data, these techniques may in theory, be used to represent any two-dimensional array.

Our application of CA is a singular value decomposition of a contingency table with r rows and c columns into an rxk matrix representing the row characteristics and a kxc matrix representing the columns. In addition to the starting matrix being a matrix of counts rather than pre-

ferences or ranks, CA is differentiated from other similar techniques in that a certain amount of preprocessing of the contingency table must be carried out before the singular value decomposition. Also, CA is based on a chi-square metric rather than the usual euclidean metric. While CA was designed with contingency tables in mind, it could be used for any other data matrix where the chi-square metric might be appropriate.

The discussion to follow shows for CA both its relationship to the chi-square test for independence in a contingency table and its relationship to principal component and biplot analysis.

Consider the matrix $Z=D_r^{-1}PD_c^{-1} - J$ where:

P = an rxc contingency table of frequencies f_{ij} divided by n, the sum of entries in the table,

D_r = a diagonal matrix of row sums, r_i, of P,

D_c = a diagonal matrix of column sums, c_j, of P, and

J = an rxc matrix of ones.

This matrix has entry

$$z_{ij} = \frac{f_{ij} - \frac{r_i c_i}{n}}{\frac{r_i c_i}{n}} = \frac{nf_{ij}}{r_i c_i} - 1$$

which is the ratio of the observed frequency to the expected frequency under chi-square independence minus 1. Thus, the matrix Z displays the deviations from the independence assumption.

The generalized singular value decomposition of Z is

$$Z = ND_\alpha M',$$

where the rows of M and N are orthonormal with respect to D_r and D_c respectively, i.e.

$$N'D_rN=M'D_cM=I$$

and D_α is the matrix of generalized eigenvalues associated with Z(Greenacre, 1982).

A generalized least squares rank k approximation to the matrix Z is obtained by using the k largest eigenvalues and their associated vectors. One measure of the goodness of fit of this approximation is:

$$\sum_{i=1}^{k} \alpha_i^2 / \sum_{i=1}^{q} \alpha_i^2$$

where α_i are the eigenvalues of Z and q is the rank of Z.

For the graphical representation of the rows and columns of Z, consider the matrices F of row point coordinates and G of column point coordinates where:

$$F=N_{(k)}\, D^a_{\alpha(k)} \text{ and } G=M_{(k)}\, D^b_{\alpha(k)}.$$

The subscript k indicates a rank k approximation to Z is being used. The different characterizations of CA correspond to the different choices of a and b in F and G respectively.

In the classical French characterization of correspondence analysis, a=1 and b=1. In the corresponding CA plot, the row point configuration approximates the weighted chi-square distances between them in the original P-matrix. The chi-square distance between the row points i and ℓ is

$$d_{\chi^2}(i,\ell) = n \Sigma (1/c_j)(f_{ij}/r_i - f_{\ell j}/r_\ell)^2.$$

It should be noted that in these chi-square distances, the $1/c_j$ weighting tends to equalize the contributions to the structure of the space of the low and high frequency columns. A similar interpretation is given to the column points. In this characterization, only global relationships between column and row points relative to the principal axes can be made. Essentially, only within row points (or within column points) relationships can be inferred from the plot.

When a+b=1, the biplot interpretation of Gabriel (1981) is applicable, i.e. the inner product of the ith row vector of F and the ith row vector of G approximate the datum entry z_{ij}. This characterization will permit interset comparisons to be made as well as intraset comparisons. The reciprocal averaging characterization of CA has a=1 and b=0, or a=0 and b=1. In particular, the symmetric characterization a=b=(1/2), used in the examples discussed below, produce row and column points on the same scale. (See Greenacre, 1981, 1982; Heiser and Muelman, 1983; and Kester and Schriever, 1982 for further interpretive information on these characterizations.)

NUMERICAL EXAMPLES

The numerical examples which we will use to illustrate correspondence analysis are taken from Discrete Multivariate Analysis Theory and Practice (Bishop, Fienberg and Holland, 1975).

1. Cancer data. The first example deals with some breast cancer displayed in Table I (Morrison et al. 1973). A loglinear analysis of these data shows a number of significant two-way intractions. In particular, all of the column variables interact with location. In addition, the malignant-benign breakdown interacts both with the size of the inflammation and the probability of survival. Age and location also interact. There are no three-way interactions. The fact that some of the column and row variables interact suggest that a CA plot might be useful and this is shown in Figure 1. The first two dimensions account for 72% of the variability represented in Table I. The various physical conditions (the columns of Table 1) are represented by points. The interpoint distances between these points represent dissimilarities among these conditions. The three geographic locations: Tokyo, Boston and Glamorgan (Wales) broken down by the three age groups are shown as vectors. The choice of points and vectors is arbitrary depending on which one can best be used to describe the information in Table I. The interpoint distances between the end-points of these vectors also represent dissimilarities among these rows of data. One cannot say anything about the interpoint distances between the column points and the endpoints of the vectors. This is NOT a point-point diagram. The relationship between the points and vectors is one of projection. The projection of the points perpendicularly on the vectors (extended in both directions) should be related to the values of the chi-square deviates (the Z-matrix described in the previous section) which are given in Table II. If one projects all eight points on the TY vector, the order of these projections are the same as for these metrics with the exception of SMM and DGM which are reversed and which would be resolved by the inclusion of the third dimension.

In addition to deciding what should be represented by points and what by vectors, the analyst has a second problem in examples such as this one in that the original data form a five-way contingency table yet this must be reformatted into two dimensions to carry out the correspondence analysis. There again, this choice is dictated by the specific problem but this will have an important bearing on the results. For this example, we let all of the physical conditions be represented in the columns and the demographics by the rows. This example could also be analyzed using Multiple Correspondence Analysis, a generalization of CA to more than two-way tables (Lebart et al, 1977).

Now, how may this display be of use? First, it can be seen that all of the benign conditions are in the top half of this display and the malignant conditions in the lower half. Considering combinations of conditions with the size of inflammation, the pair DGB and SGB are farther apart from SMB and DMB than are their counterparts for the malignant conditions. When considering probability of survival, one notes that SMM and DMM are located in the southeast quadrant while the other pairs, e.g. SGM and DGM are related primarily to the vertical axis alone. While it may appear from this plot that a three-way interaction may exist among these three factors, we already know from the contingency table analysis that this is not so. Always run this analysis first. If the interactions are not significant, there is no need for CA (or cross-tabs as we have already noted). There may be some concern about PGB since it represents only seven individuals and six empty cells. The deletion of this column had little effect on the location of the other points or on the orientation of the vectors.

The interaction between these factors and location is fairly easy to see. The age vectors for each location are grouped together but the locations themselves do not overlap at all. Boston and Glamorgan tend to have higher death rates than Tokyo. Boston also tends to have a higher benign rate than the other two. Glamorgan and Tokyo differ in the size of inflammation. There is also an interaction between location and

age. Both Glamorgan and Tokyo vectors sweep out an arc of roughly 45 degrees and are in order by age group. The Boston vectors, on the other hand, sweep out about 90 degrees. This seems to be due to the 50-69 year old group which may be pulled out of the normal pattern because it contained two of the seven DGB cases as well as large number of both SMB and DMB. Not shown in this display are the correlations between the projections of each point on each vector with their corresponding deviations in the Z-matrix. These correlations ranged from r=.69 to r=.98 except for the over 70 segment in Tokyo which had a correlation of only r=.12 for this two-dimensional case and will require a higher order representation to explain it.

2. Father-son data. Our other example, displayed in Table III, has the distribution of 637 men by various occupations crossed by the occupations of their fathers (Pearson, 1904). (The original data consisted of 775 men but we deleted teaching, agriculture, and the navy because of the paucity of data.) These data differ from the last set in that one would not expect the columns to be independent of the rows; rather, one would expect a heavy concentration of frequencies along the diagonals reflecting the notion (at that point in time) that sons generally followed the same occupations as their fathers. In this example, only 56% of the variability is explained by two dimensions and it was necessary to go to three dimensions to get up to 70%. (The fourth dimension adds 12%, the fifth, 7%.) Most multidimensional scaling articles show examples that can be represented in two dimensions but the world is not always flat. Once one gets beyond two dimensions, one needs to be a bit creative with one's graphics. In Figure 2, we have represented the third dimension of the points (the son's occupations) by circles of varying radii. For the vectors (the fathers) which depart markedly from the plane of the first two dimensions, we have indicated by a sign the nature of the departure and a double sign (i.e. ++ or --) for those who extend the most along the third axis. Some procedures, such as MDPREF, avoid some of these problems by obtaining vectors of unit length; departure from a unit circle then indicates the amount of the third dimension involved in each vector.)

As one would expect, most of the vectors track their corresponding points quite well indicating that the "null hypothesis" of following in the father's footsteps had some validity. The first two dimensions indicate the status quo and also indicate some clustering of occupations. One cluster consists of the army, politics, law and landowners. Another cluster consists of commerce, medicine, literature, scholarship and science, and divinity. Arts and crafts appear by themselves. The third dimension is required to resolve medicine which does not follow the trend as well; true, most sons whose fathers were in medicine also went into that activity but so did some others, particularly those whose fathers were in divinity. (The fourth and fifth dimensions attempted to resolve similar situations involving the army, crafts, law and politics.) In this example, we associated the fathers with the vectors, but we could have just as well done it the other way around.

CAUTION

One requirement of multidimensional scaling procedures in general is that they require a reasonable number of entries to be worthwhile. This precludes the use of CA for many contingency tables which have a small number of rows and/or columns. In these cases, however, any interactions which do exist should be easy enough to interpret without the need for CA. The principal advantage of CA lies in its ability to graphically portray large tables which otherwise may be difficult to diagnose.

REFERENCES

1. Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. The MIT Press.

2. Fisher, R. A. (1940). The precision of discriminant functions. Ann. Eugen. Lond. 10 422-429.

3. Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. Interpreting Multivariate Data (V. Barnett, Editor). John Wiley and Sons, Inc., 147-173.

4. Green, P. E. and Rao, V. R. (1972). Applied Multidimensional Scaling. Dryden Press.

5. Greenacre, M. J. (1981). Practical correspondence analysis. Interpreting Multivariate Data (V. Barnett, Editor). John Wiley and Sons, Inc., 119-146.

6. Greenacre, M. J. (1982). Basic structure (Singular value decomposition) and multivariate analysis. (Preprint)

7. Heiser, W. J. and Meulman, J. (1983). Analyzing rectangular tables by joint and constrained multidimensional scaling. J. Econometrics 22, 139-167.

8. Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. Applied Statistics 23, 340-354.

9. Kester, N. K. and Schriever, B. F. (1982). Analysis of Association of Categorical Variables by Numerical Scores and Graphical Representations. Mathematical Centrum.

10. Lebart, L., Morineau, A. and Tabard, N. (1977). Techniques de la Description Statistique. Dunod.

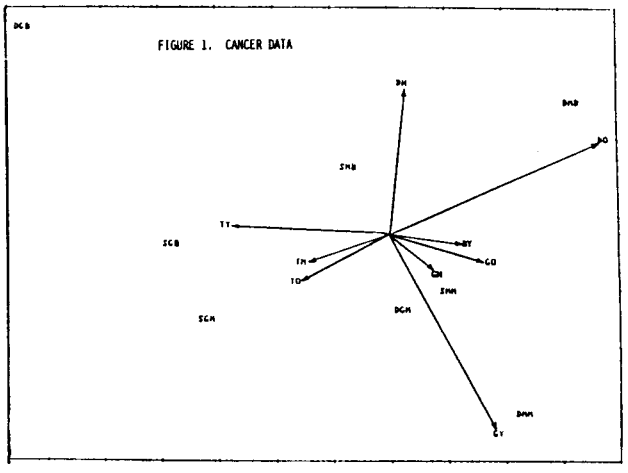
11. Lebart, L., Morineau, A. and Warwick, K. M. (1984). Multivariate Descriptive Statistical Analysis. John Wiley and Sons, Inc.

12. Morrison, A. S., Block, M. M., Lowe, C. R., MacMahon, B. and Youse, S. (1973). Some international differences in histology and survival in breast cancer. Int. J. Cancer 11, 261-267.

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

14. Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. Draper's Co. Res. Mem. Biometric Ser. 1. Reprinted (1948) in Karl Pearson's Early Papers. Cambridge University Press.

15. Ries, P. N. and Smith, H. (1963). The use of chi-square for preference in multidimensional problems. Chem. Eng. Progress 59, 39-43.



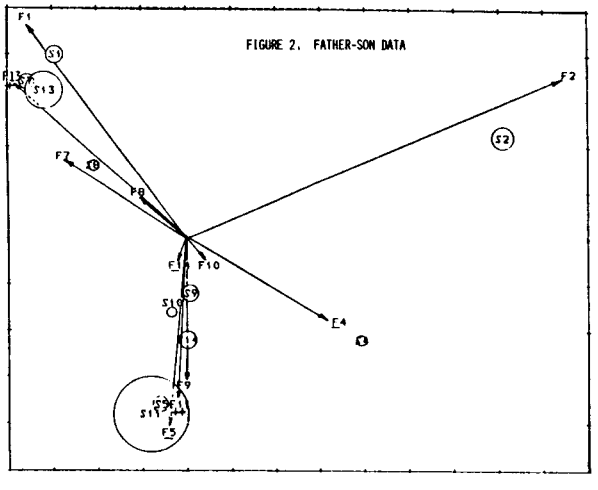


TABLE I

CANCER DATA

DIAGNOSTIC CENTER		SURV. INFLM. DIAGN. AGE	NO MINOR MALIG. DMM	NO MINOR BENIG. DMB	NO MAJOR MALIG. DGM	NO MAJOR BENIG. DGB	YES MINOR MALIG. SMM	YES MINOR BENIG. SMB	YES MAJOR MALIG. SGM	YES MAJOR BENIG. SGB
Tokyo	TY	< 50	9	7	4	3	26	68	25	9
	TM	50 - 69	9	9	11	2	20	46	18	5
	TO	≥ 70	2	3	1	0	1	6	5	1
Boston	BY	< 50	6	7	6	0	11	24	4	0
	BM	50 - 69	8	20	3	2	18	58	10	3
	BO	≥ 70	9	18	3	0	15	26	1	1
Glamorgan	GY	≤ 50	16	7	3	0	16	20	8	1
	GM	50 - 69	14	12	3	0	27	39	10	4
	GO	≥ 70	3	7	3	0	12	11	4	1

TABLE II

Z-MATRIX FOR CANCER DATA

	DMM	DMB	DGM	DGB	SMM	SMB	SGM	SGB
TY	- .40	- .61	- .45	1.17	- .10	.15	.49	.82
TM	- .25	- .36	.89	.82	- .13	- .02	.35	.27
TO	.06	.34	.09	-1.00	- .72	- .19	1.37	.61
BY	.04	.02	1.14	-1.00	- .01	.06	- .38	-1.00
BM	- .34	.39	- .49	.79	- .23	.22	- .26	- .25
BO	.24	1.09	- .15	-1.00	.08	- .09	- .88	- .58
GY	1.27	- .16	- .13	-1.00	.18	- .28	.01	- .57
GM	.29	- .07	- .43	-1.00	.30	- .08	- .18	.12
GO	- .26	.45	.51	-1.00	.53	- .31	- .12	- .25

TABLE III

PEARSON'S FATHER-SON OCCUPATIONAL DATA

OCCUPATIONS OF FATHERS	OCCUPATIONS OF SONS 1	2	4	5	7	8	9	10	11	13	14
1	28	0	0	0	1	3	3	0	3	5	2
2	2	51	1	2	0	1	2	0	0	1	1
4	0	12	6	5	0	1	7	1	2	0	10
5	5	5	1	54	0	6	9	4	12	1	13
7	17	1	0	14	6	11	4	1	3	17	7
8	3	5	0	6	2	18	13	1	1	8	5
9	0	1	0	4	0	1	4	0	2	1	4
10	12	16	1	15	0	5	13	11	6	7	15
11	0	4	0	1	0	0	3	0	20	5	6
13	5	0	0	3	1	8	1	2	2	23	1
14	5	3	2	6	1	3	1	0	0	1	9

OCC. CODES: 1-Army, 2-Art, 4-Crafts, 5-Divinity, 7-Landowners, 8-Law, 9-Literature, 10-Commerce, 11-Medicine, 13-Politics, 14-Scholarship and Science