J. DOUGLAS CARROLL, PAUL E. GREEN, and CATHERINE M. SCHAFFER*

Correspondence analysis is a metric technique for finding a spatial representation of data that has particular applicability to the analysis of cross tabulations (or contingency tables). The authors focus on some of the geometric underpinnings of the procedure and the kinds of statements that can be made about interobject similarity based on a distance model. In particular, a rationale is developed, along with a scaling of coordinates, that enables one to compare both between-set and within-set squared distances. This scaling differs from the conventional scaling currently used in correspondence analysis, where only one set of distances usually are compared.

Interpoint Distance Comparisons in Correspondence Analysis

Correspondence analysis is a metric multidimensional scaling (MDS) method that has been associated mainly with the French school (Benzécri 1969, 1973) of data analysis. It has been used primarily to analyze two-way (and sometimes higher way) contingency tables of frequency data, so that squared distances between certain sets of points in the derived space bear simple relationships to the original tabular entries.[1] Correspondence analysis and its various aliases—dual scaling, reciprocal averaging, homogeneity analysis, and canonical scoring, to name a few—go back to at least the early 1930s (Nishisato 1980, Ch. 1).

Recently there has been a marked renewal of interest in the technique and an accompanying diffusion of information about it. Several books and monographs (Bouroche and Saporta 1980; de Leeuw 1973; Gifi 1981; Greenacre 1984; Heiser 1981; Lebart, Morineau, and Warwick 1984; Meulman 1982) have appeared and curiosity about the methodology is increasing among marketing researchers (Hoffman and Franke 1986). Each of the books/monographs listed provides an expository account of correspondence analysis. It is not the purpose of our technical note to duplicate these efforts; rather, we focus on the geometric bases of correspondence analysis. To the best of our knowledge, this aspect of the methodology has not been discussed in a way that specifies the conditions under which squared interpoint distances (particularly between-set distances) can be compared.

THE BASICS OF CORRESPONDENCE ANALYSIS

As background for the ensuing discussion, consider the 4 × 3 table of frequency data (Table 1). In the usual analysis of two-way cross tabulations, the data of Table 1 are submitted to a chi square test. In this case the chi square value is 21.1 which, with 6 degrees of freedom, is significant beyond the .01 level.[2] If we assume that evidence of association between rows and columns is found, how can row-column association be represented?

The objective of (two-way) correspondence analysis is to portray data (like those of Table 1) geometrically as a set of row and column points in, say, two-dimensional space for ease of visualization. The questions of interest for our discussion are:

[1]Though mainly applied to cross-tabulated frequencies, this method can be used to analyze data other than frequencies (see Nishisato 1980, Ch. 1).

*J. Douglas Carroll is a Distinguished Member of the Technical Staff, AT&T Bell Laboratories. Paul E. Green is S. S. Kresge Professor of Marketing and Catherine M. Schaffer is a doctoral candidate in marketing, Wharton School, The University of Pennsylvania.

[2]It is not necessary to precede this application of correspondence analysis with a chi square test.

Journal of Marketing Research
Vol. XXIII (August 1986), 271–80

Table 1
A 4 × 3 CROSS TABULATION OF (HYPOTHETICAL) FREQUENCIES

Rows	Columns 5	6	7	
1	1	5	1	7
2	5	1	3	9
3	2	10	2	14
4	1	1	7	9
	9	17	13	39

1. What is the relationship between two row (column) points that are "close" in the geometric configuration to the input data?
2. What is the relationship (if any) between row and column points that are close in the geometric configuration to the input data?

We could summarize these two queries by asking what comparisons can be made of *distances* (between row points, column points, and/or row and column points) in the geometric representation obtained from correspondence analysis.

As Lebart, Morineau, and Warwick (1984, p. 46) point out, when talking about the X and Y coordinate sets,

> . . . it is legitimate to interpret distances among elements of one set of points. . . . It is also legitimate to interpret the relative positions of one point of one set with respect to *all the points* of the other set. Except in special cases, it is extremely dangerous to interpret the proximity of *two* points corresponding to different sets of points.

However, they do not delve into detail about what the special conditions are and why dangers are associated with comparing between-set relationships. (This issue has, in large part, motivated our note.)

Application of Two-Way Correspondence Analysis

In the conventional application of correspondence analysis to the 4 × 3 frequency matrix **F** of Table 1, one would first normalize **F** (with general entry f_{ij}; $i = 1, I$; $j = 1, J$) as follows.

$$\mathbf{H} = \mathbf{R}^{-1/2}\mathbf{F}\mathbf{C}^{-1/2} \quad (1)$$

where $\mathbf{R}^{-1/2}$ and $\mathbf{C}^{-1/2}$ are diagonal matrices whose entries consist of reciprocals of the square roots of the row marginals and column marginals, respectively. Numerical results for this normalization are given in Table 2.

The second step in correspondence analysis is to find the Eckart-Young (1936) or singular value decomposition of **H**, which can be written as

$$\mathbf{H} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}' \quad (2)$$

with $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}$ and $\boldsymbol{\Delta}$ is diagonal.[3] Eckart-Young analysis is a type of factoring (Green with Carroll 1976) in which T "dimensions" are extracted and **H** is approximated as a product of three matrices, where $\mathbf{P}_T'\mathbf{P}_T = \mathbf{I}_T$ and $\mathbf{Q}_T'\mathbf{Q}_T = \mathbf{I}_T$ are orthonormal sections and $\boldsymbol{\Delta}_T$ is diagonal with ordered non-negative entries (the "singular values" of $\mathbf{H}_T$).[4] Often (but not necessarily) T is set at two dimensions.[5] The full Eckart-Young or singular value decomposition corresponds to the case in which T corresponds to "full rank" (generally the minimum of I and J). In the full rank case, $\mathbf{H}_T$ approximates **H** perfectly.

The last step in correspondence analysis is to define the configuration's row and column coordinates, **X** and **Y** (after dropping the dimensionality subscript), as follows.

$$\mathbf{X} = \mathbf{R}^{-1/2}\mathbf{P}\boldsymbol{\Delta} \quad (3)$$

$$\mathbf{Y} = \mathbf{C}^{-1/2}\mathbf{Q} \quad (4)$$

It is the two sets of coordinates, **X** and **Y**, that represent the final set of outputs of interest to the researcher. They usually are plotted together in a space of two (or more) dimensions.

Figure 1 is a two-dimensional plot obtained from a correspondence analysis of **X** and **Y** for the data of Table 1. Points 1 through 4 are the four row points and points 5 through 7 are the three column points.[6] As noted from Figure 1, row points 1 and 3 are close (actually superimposed) and are relatively far from row points 2 and 4 (which are also relatively far from each other). Column points 5, 6, and 7 are relatively far from each other and not particularly close to any of the row points.

As we point out next, however, only the interpoint distances of the four row points are comparable in the usual application of correspondence analysis. That is,

Table 2
THE 4 × 3 MATRIX H OBTAINED FROM $\mathbf{H} = \mathbf{R}^{-1/2}\mathbf{F}\mathbf{C}^{-1/2}$

Rows	Columns 5	6	7
1	.1260	.4583	.1049
2	.5554	.0808	.2774
3	.1782	.6482	.1483
4	.1111	.0808	.6472

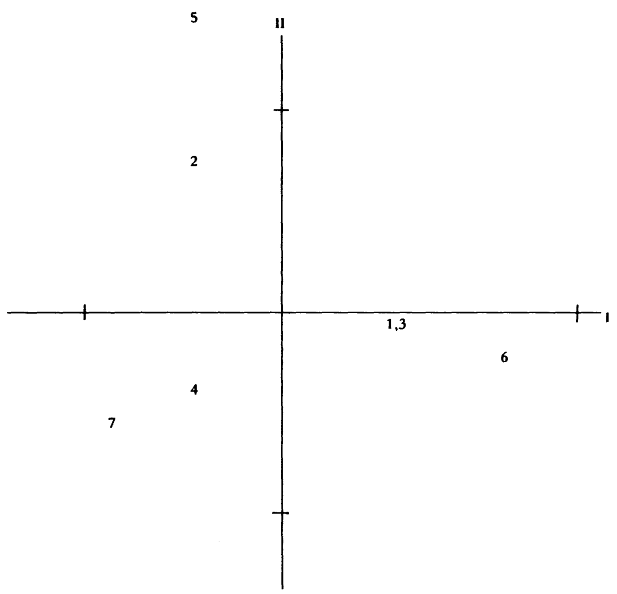
[3]As described by many authors (e.g., Gower and Digby 1981), the first eigenvector consists of constants (the so-called "trivial" eigenvector) and is ignored in solving for $\mathbf{P}_t$, $\boldsymbol{\Delta}_t$, and $\mathbf{Q}_t'$. Alternatively, by a preliminary transformation of **H** (see Gifi 1981, p. 136), the trivial vector can be removed at the outset.

[4]In a previous version of this note (obtainable from the authors) we show how correspondence analysis also can be formally represented as a *weighted* Eckart-Young analysis of a transformation of **F**.

[5]Another issue of importance is how well a low-dimensional space (e.g., two dimensions) represents the data. As Benzécri et al. (1973, Vol. 2, p. 46–7) indicate, two points close to each other in the two-dimensional space may be very far apart in higher dimensions.

[6]The two-dimensional fit accounts for all of the variance in the input matrix **H**.

Figure 1
TWO-DIMENSIONAL PLOT OF CORRESPONDENCE ANALYSIS OF TABLE 1 DATA



though all distances look comparable in Figure 1, the traditional scaling of the **X** and **Y** coordinates does not permit this interpretation. In the next section we explain why only one set of distances are comparable.

SOME GEOMETRIC ASPECTS OF CORRESPONDENCE ANALYSIS

A central objective of correspondence analysis is to find a set of coordinates (denoted by **X**), representing the *rows* of the two-way contingency table **F**, so that squared Euclidean distances between the rows of **X** correspond in a straightforward way to squared distances between rows in **F**. Benzécri has referred to these distances as "chi square" distances, and a rationale for this term is given by Gifi (1981, p. 136).

The notion of a distance (actually *squared* distance) between row profiles i and k in the original frequency matrix $\mathbf{F} \equiv (f_{ij})$ can be defined as

$$d_{ik}^2 = n \sum_{j=1}^{J} \frac{1}{c_j} \left(\frac{f_{ij}}{r_i} - \frac{f_{kj}}{r_k} \right)^2 \tag{5}$$

where:

$$c_j \equiv f_{\cdot j} \equiv \sum_{i=1}^{I} f_{ij}$$

$$r_i \equiv f_{i\cdot} \equiv \sum_{j=1}^{J} f_{ij}$$

and

$$n = \sum_i \sum_j f_{ij} = \sum_j c_j = \sum_i r_i.$$

Indeed, d_{ik}^2 is in fact n times the squared Euclidean distance between profiles $\mathbf{g}_i$ and $\mathbf{g}_k$, with $\mathbf{g}_i$ defined by the entries

$$g_{ij} = \frac{f_{ij}}{r_i\sqrt{c_j}}. \tag{6}$$

Thus, to obtain the g_{ij}'s one normalizes each original frequency f_{ij} by dividing it by the product of that entry's row marginal and the *square root* of its column marginal. This procedure has the effect of adjusting for "size" effects.

Because

$$d_{ik}^2 = n \sum_j (g_{ij} - g_{kj})^2 \tag{7}$$

and

$$g_{ij} = \frac{f_{ij}}{r_i\sqrt{c_j}} \tag{8}$$

when we approximate $\mathbf{G} = (g_{ij})$, we also are approximating any function of **G**, in particular the matrix of squared Euclidean distances between rows of **G**. Table 3 shows **G**, as obtained from the **F** matrix of Table 1 by the normalization described.

Though the procedure is not shown here, one could just as easily start from the *columns* of **F** and compute squared profile distances between columns across rows. The reasoning is analogous, but of course the numerical results will differ.

It can be shown that a weighted Eckart-Young analysis of **G** (with weights given by $\mathbf{R}^{1/2}$) is equivalent for rows (but not for columns) to a correspondence analysis of **F** (details can be obtained from the authors). Thus, we can make squared distance comparisons for rows (but not columns) or vice versa, depending on which set is to be emphasized. That is, the development can proceed from the computation of squared distances between row profiles or column profiles but not both, simultaneously.

Therefore, if we reexamine Figure 1, showing the normalization

$$\mathbf{X} = \mathbf{R}^{-1/2}\mathbf{P}\boldsymbol{\Delta} \tag{9}$$

Table 3
THE 4 × 3 MATRIX G OBTAINED FROM A NORMALIZATION OF F

	Columns		
Rows	*5*	*6*	*7*
1	.0476	.1732	.0396
2	.1852	.0269	.0924
3	.0476	.1732	.0396
4	.0370	.0269	.2157

$$\mathbf{Y} = \mathbf{C}^{-1/2}\mathbf{Q}, \tag{10}$$

only the placement of the **X** (row) coordinates follows from the preceding argument. Practitioners of correspondence analysis know this, and hence interpret **Y** as the set of (column) coordinates such that each row point is represented as a weighted centroid of the full set of **Y** coordinates, with weights given by the frequency with which each column appears in a given row. (This is what Lebart, Morineau, and Warwick mean by, "It is also legitimate to interpret the relative positions of one point of one set with respect to *all the points* of the other set.")

Had the development proceeded the other way around, the following definitions would apply

$$\mathring{\mathbf{Y}} = \mathbf{C}^{-1/2}\mathbf{Q}\boldsymbol{\Delta} \tag{11}$$

$$\mathring{\mathbf{X}} = \mathbf{R}^{-1/2}\mathbf{P} \tag{12}$$

and it would be (only) the column coordinates **Y** that provide a squared distance relationship. The configuration, of course, would differ from that shown in Figure 1.

One "solution" to the situation is to plot only one set of points—rows or columns, depending on whether the diagonal matrix $\boldsymbol{\Delta}$ is associated with **X** or **Y**. Another alternative (actually the one used by most applications researchers) is to apply the squared distance idea to (say) *only* the row points and the weighted centroid idea to the relationship between row and column points. A third approach (the one pursued here) is to define a new problem representation, metric, and apportionment of $\boldsymbol{\Delta}$, the matrix of singular values, so that *all* squared interpoint distances (within set and between sets) are comparable.

ANOTHER VIEW OF CORRESPONDENCE ANALYSIS

At this point we seek a new formulation of correspondence analysis that enables us to make squared distance comparisons of *all* interpoint relationships, both within set and between sets. This objective can be accomplished by taking a different view of the original frequency table.

For example, let us reconsider the 4 × 3 contingency table (Table 1). This table (whose matrix is now denoted **E**) can be re-expressed as a 7 × 39 pseudocontingency table (whose matrix is denoted $\tilde{\mathbf{E}}$) made up of 0, 1 dummy-variable coding (Table 4) without any loss of information. Given this table, the contingency table **E** can be constructed uniquely. Alternatively, given the contingency table **E**, the pseudocontingency table $\tilde{\mathbf{E}}$ can be constructed uniquely up to an arbitrary permutation of the columns, which can be viewed as corresponding to individual observations on each of the two attributes. (The same idea can be extended to multiway contingency tables.)

Next, suppose the "chi square" metric is applied to the pseudocontingency matrix, to get

$$d_{ik}^2 = n \sum \frac{1}{c_j}\left(\frac{\tilde{e}_{ij}}{r_i} - \frac{\tilde{e}_{kj}}{r_k}\right)^2 \tag{13}$$

$$= \frac{n}{m} \sum (\overset{*}{\tilde{e}}_{ij} - \overset{*}{\tilde{e}}_{kj})^2$$

where:

$$\overset{*}{\tilde{e}}_{ij} = \frac{\tilde{e}_{ij}}{r_i} = \frac{\tilde{e}_{ij}}{n_i}$$

and

$$n_i = \sum_j \tilde{e}_{ij} = \tilde{e}_{i\cdot} = r_i$$

so that

$$\sum_j \overset{*}{\tilde{e}}_{ij} = 1 \text{ for all } i.$$

Note also that m denotes the number of "ways" of the original contingency table; in the preceding example $m = 2$.

Thus, except for a scale factor of n/m, the metric in this case is simply a squared Euclidean metric between *rows* of $\tilde{\mathbf{E}} = (\tilde{e}_{ij}/\tilde{e}_{i\cdot})$. Applying correspondence analysis to $\tilde{\mathbf{E}}$ is *equivalent* to applying correspondence analysis to the two-way contingency table **E** with which it is as-

Table 4
A 7 × 39 PSEUDOCONTINGENCY TABLE

																																								Sum
	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	14
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	9
$\tilde{\mathbf{E}}$ =	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
	1	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	9
	0	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	17
	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	13
Sum	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	

sociated (given satisfaction of the conditions to follow). The following points should be noted.

1. In the analysis of $\tilde{\mathbf{E}}$ one obtains, as an added feature, a representation of the individuals (or other observational units) as points. A distance interpretation does *not* hold for these points, however. Rather, they are interpretable in terms of the weighted centroid principle. That is, the point for an individual (or other observational unit) is the weighted centroid of the points for the row and column entities (attribute levels) present for, or chosen by, that individual.
2. If correspondence analysis is done on the "standard" contingency table (**E**), the diagonal matrix of singular values $\mathbf{\Delta}$ must be modified in the following ways.

   a. First, an identity matrix is added to $\mathbf{\Delta}$.
   b. Then the resulting matrix sum $(\mathbf{\Delta} + \mathbf{I})$ must be split *equally* between the two matrices **P** and **Q** in the Eckart-Young decomposition, rather than being absorbed into only one (say **P**).

   The result is that the overall correspondence analysis of the standard table **E** is now of the form

$$\tilde{\mathbf{X}} = \mathbf{R}^{-1/2}\mathbf{P}(\mathbf{\Delta} + \mathbf{I})^{1/2} \tag{14}$$

$$\tilde{\mathbf{Y}} = \mathbf{C}^{-1/2}\mathbf{Q}(\mathbf{\Delta} + \mathbf{I})^{1/2} \tag{15}$$

   instead of the representation of **X** and **Y**, as defined in equations 3 and 4.
3. Comparing these two sets of equations makes it clear that equations 14 and 15 simply divide the $\mathbf{\Delta} + \mathbf{I}$ matrix symmetrically between **P** and **Q** (or, more precisely, between $\mathbf{R}^{-1/2}\mathbf{P}$ and $\mathbf{C}^{-1/2}\mathbf{Q}$) by multiplying each by $(\mathbf{\Delta} + \mathbf{I})^{1/2}$ rather than multiplying one (say **P**) by $\mathbf{\Delta}$. The reasons are based on the relationship of correspondence analysis to generalized canonical correlation (Carroll 1968), as first pointed out by Saporta (1980).
4. There is another scale factor to be considered, namely a factor of $1/\sqrt{m}$ (or $1/\sqrt{2}$ in the two-way case), by which both $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ should be multiplied if correspondence analysis is applied to the pseudocontingency table $\tilde{\mathbf{E}}$. Because an overall scale factor (which uniformly rescales the overall configuration) makes no essential difference, it can be safely ignored. However, if the pseudocontingency table uses a more general matrix without the constraint that all columns sum to a constant, there would be no simple scale factor. In the latter case the pseudocontingency table would not correspond to a "standard" contingency table (such as **E**).
5. Given our alternative scaling of the matrices for the row and column representations ($\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ instead of **X** and **Y**), *all* squared distances are interpretable, both within each attribute and between attributes. As indicated before, they are interpretable as approximations to (squared) distances between the corresponding rows of the matrix $\tilde{\mathbf{E}}$, which also can be viewed as the chi square or Benzécri metric applied to the pseudocontingency table $\tilde{\mathbf{E}}$.

   Because the factor $1/m$ is a constant, the chi squared distance, except for an arbitrary scale factor of n/m, is equivalent to

$$\tilde{d}_{ik}^2 = \sum_{j=1}^{J}\left(\frac{\tilde{e}_{ij}}{\tilde{e}_{i\cdot}} - \frac{\tilde{e}_{kj}}{\tilde{e}_{k\cdot}}\right) \tag{16}$$

   or, in words, equals the squared Euclidean distance between rows i and k of $\tilde{\mathbf{E}}$, after each row is normalized to unit length so that the sum of squares of each row is equal to one.
6. A somewhat curious result is that, within attributes (either rows or columns), the squared Euclidean distance is proportional to

$$\tilde{d}_{ii'}^2 = \frac{1}{n_i} + \frac{1}{n_{i'}} \tag{17}$$

where:

$$n_i = \tilde{e}_{i\cdot} = \sum_j \tilde{e}_{ij}.$$

   In words, the within-attribute squared distances for two different levels of the same attribute are proportional to the sums of the reciprocals of the two associated marginal frequencies. Stated another way, the within-attribute squared distance is proportional to the reciprocal of the harmonic mean of the marginal frequencies.
7. The pseudocontingency table approach generalizes readily to the multiway contingency table case (what Benzécri calls multiple correspondence analysis), as well as leading to an interpretation of both between- and within-set squared distances among attribute categories. Hence, we advocate the adoption of this approach. Moreover, this formulation of multiple correspondence analysis can be shown to be equivalent to a generalized canonical correlation method called CANCOR (Carroll 1968), as applied to submatrices of dummy variables corresponding to the levels of m attributes.

Solving the Numerical Example

We first apply conventional correspondence analysis to the 7×39 pseudocontingency table, $\tilde{\mathbf{E}}$ (Table 4). We obtain the singular value matrix associated with the first two dimensions,

$$\mathbf{\Delta}_p = \begin{bmatrix} 0.8983 & 0 \\ 0 & 0.8379 \end{bmatrix} \tag{18}$$

which is shown in Table 5. Also in Table 5 are the two-dimensional coordinates, where each singular value is

Table 5

CORRESPONDENCE ANALYSIS OF 4 × 3 CONTINGENCY AND PSEUDOCONTINGENCY TABLES

Singular values		*Contingency table*		*Pseudocontingency table*	
Δ_1		.6148		.8983	
Δ_2		.4037		.8379	
Scaled coordinates		*I*	*II*	*I*	*II*
Rows	1	1.164	−0.160	0.823	−0.113
	2	−1.085	1.911	−0.767	1.352
	3	1.164	−0.160	0.823	−0.113
	4	−1.630	−1.539	−1.152	−1.088
Columns	5	−0.644	2.078	−0.456	1.469
	6	1.410	−0.295	0.997	−0.209
	7	−1.398	−1.052	−0.988	−0.744

computed and allocated to the rows (representing attribute levels) of the 7×39 pseudocontingency table.

Next, the contingency table **E** is analyzed, leading to singular values of

$$\Delta_c = \begin{bmatrix} 0.6148 & 0 \\ 0 & 0.4037 \end{bmatrix} \quad (19)$$

In the two-attribute case, the relationship between sets of singular values is

$$\delta_p^2 = \frac{(\delta_c + 1)}{2} \quad (20)$$

where δ_p, δ_c denote singular values based on the pseudocontingency and contingency tables, respectively. Thus, from Table 5 we note that

$$(0.8983)^2 = \frac{(0.6148 + 1)}{2}; \quad (0.8379)^2 = \frac{(0.4037 + 1)}{2} \quad (21)$$

$$\approx 0.807 \qquad\qquad \approx 0.702.$$

Next, the coordinates are shown for the contingency table analysis, based on equations 14 and 15.

We note from Table 5 that, except for a constant scale factor ($\approx \sqrt{2} = 1.414$), the coordinates of the two solutions are the same. In terms of the aforementioned profile metric, all squared interpoint distances are comparable. Thus, in practice we do not actually apply correspondence analysis to the pseudocontingency table; however, through the use of the scaling shown in equations 14 and 15, we have a representation that one *would* obtain if conventional correspondence analysis were applied to the pseudocontingency table.

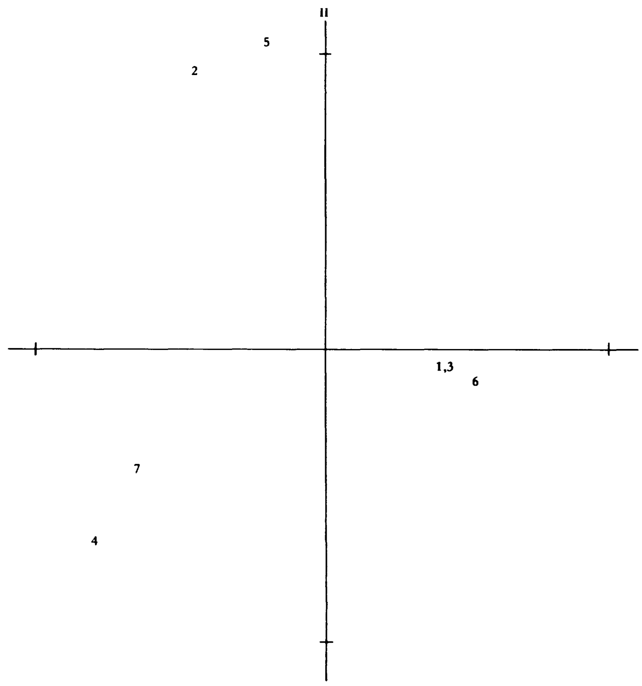
Again, we return to the initial 4×3 frequency table (Table 1). The conventional correspondence analysis produced the plot of Figure 1. For comparison, Figure 2 shows the appropriate plot when scaled in terms of the "profile metric," based on the pseudocontingency table analog.

In contrast to Figure 1's plot, Figure 2 enables one to compare row, column, *and* row-column distances. As noted, row points 1 and 3 plot close to column point 6, row point 2 is grouped with column point 5, and row point 4 is near column point 7. In short, the points break clearly into three clusters. We also note that points 4 and 7 project on axes I and II in reversed order when Figures 1 and 2 are compared. Thus, the proposed scaling can lead to rankings of point projections (involving between-set comparisons) that differ from those obtained by the conventional scaling procedure.

AN EMPIRICAL EXAMPLE

Though the hypothetical data of Table 1 serve expository purposes, the primary motivation for proposing the "profile metric" scaling procedure is its value in empirical problems. To this end, the two types of scaling were applied to a much larger two-way table. The data in Table 6 were obtained from a survey of 252 customers of

Figure 2
TWO-DIMENSIONAL PLOT OF CORRESPONDENCE ANALYSIS OF TABLE 1 DATA, BASED ON PSEUDOCONTINGENCY TABLE ANALOG



overnight air delivery services. Each respondent, previously classified by favorite shipper (Alpha, Beta, Gamma, and Delta) and by size of business, was asked to pick those attributes that best described his or her favorite supplier. Table 7 lists the attribute names and respondent category descriptions.

Figure 3 shows the two-dimensional configuration entailing conventional correspondence analysis scaling, in which only row-based (squared) distances are comparable.[7] As noted, the 10 supplier-respondent categories break down into a cluster of Gamma users and Delta users, a cluster of Alpha users (in which the very small businesses are somewhat separated from the rest), and the Beta user group, plotted near Alpha. Apparently, in terms of supplier perceptions, size of business has less influence on Gamma users than it does on Alpha users.

Figure 4 shows the counterpart two-dimensional scaling based on the proposed "profile metric" in which all interpoint distances are comparable. Gamma and Delta users perceive their suppliers' attributes to be (1) always delivering packages by the promised time, (2) easy to trace a package that has gone astray, (3) friendly and congenial employees, and (4) sensitive to customer needs

[7]The variance accounted for by the two-dimensional solution (in both Figures 3 and 4) is 88.5%.

Table 6
CROSS TABULATION OF SUPPLIER FAVORITE/SIZE CATEGORY BY ATTRIBUTE MATCHING
(see Table 7 for row and column legends)

Row	*Column* 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	4.00	5.00	5.00	20.00	1.00	21.00	2.00	1.00	22.00	8.00	3.00	13.00	1.00	4.00	9.00	119.00
2	2.00	5.00	7.00	13.00	1.00	16.00	4.00	2.00	16.00	10.00	6.00	10.00	7.00	6.00	5.00	110.00
3	9.00	9.00	8.00	12.00	2.00	17.00	5.00	3.00	16.00	11.00	6.00	13.00	10.00	7.00	11.00	139.00
4	5.00	11.00	10.00	20.00	1.00	15.00	3.00	2.00	25.00	10.00	6.00	14.00	10.00	6.00	8.00	146.00
5	25.00	25.00	19.00	19.00	20.00	11.00	11.00	8.00	19.00	19.00	13.00	2.00	15.00	15.00	10.00	231.00
6	25.00	25.00	21.00	20.00	19.00	14.00	12.00	10.00	18.00	20.00	15.00	3.00	15.00	18.00	14.00	249.00
7	22.00	23.00	18.00	15.00	20.00	12.00	14.00	8.00	15.00	22.00	19.00	2.00	14.00	16.00	11.00	231.00
8	17.00	17.00	18.00	11.00	11.00	10.00	11.00	3.00	10.00	14.00	12.00	1.00	13.00	11.00	12.00	171.00
9	7.00	11.00	8.00	7.00	3.00	7.00	3.00	2.00	10.00	7.00	8.00	13.00	6.00	3.00	5.00	100.00
10	12.00	13.00	13.00	9.00	11.00	8.00	12.00	9.00	9.00	14.00	10.00	1.00	10.00	11.00	7.00	149.00
	128.00	144.00	127.00	146.00	89.00	131.00	77.00	48.00	160.00	135.00	98.00	72.00	101.00	97.00	92.00	1645.00

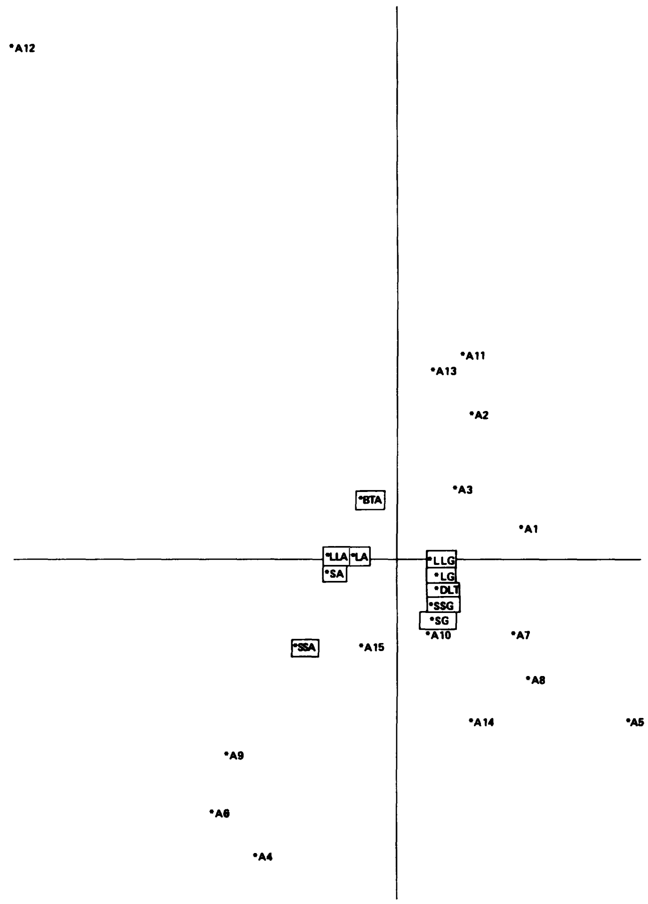
Table 7
ROW AND COLUMN DESCRIPTIONS FOR TABLE 6

Rows (*supplier and buyer size category*)		*Columns* (*supplier attributes*)
1. Alpha—very small	(SSA)	1. Almost always delivers the package by the promised time
2. Alpha—small	(SA)	
3. Alpha—large	(LA)	2. Packages are almost always delivered in good condition
4. Alpha—very large	(LLA)	
5. Gamma—very small	(SSG)	3. Almost never late in pickup of packages to be shipped
6. Gamma—small	(SG)	
7. Gamma—large	(LG)	4. Little paperwork is required in preparing packages for shipment
8. Gamma—very large	(LLG)	
9. Beta	(BTA)	5. Will make special trips to pick up packages if needed
10. Dela	(DLT)	
		6. Easy to calculate the shipping costs
		7. Easy to trace a package that has gone astray
		8. Sensitive to customer needs in settling claims
		9. Capable of shipping packages almost everywhere I want
		10. Friendly and congenial employees
		11. Extremely efficient in all their business dealings
		12. Less expensive than most
		13. Responsible and dependable in customer relations
		14. Uses the most advanced technology
		15. Really interested in the smaller customer

in settling claims. Beta is perceived primarily as less expensive. Very small customers of Alpha perceive its attributes to be (1) requiring little paperwork and (2) easy to calculate shipping costs, whereas larger customers perceive them to be (1) capable of shipping anywhere I want and (2) really interested in the smaller customer.

Perhaps the most striking difference in supplier-business interpoint distances between Figures 3 and 4 is that Beta is widely separated from the rest of the points in Figure 4. Moreover, very small users of Alpha are more widely separated from the rest of the Alpha points in Figure 4. In effect, the "profile metric" scaling has pulled the row points apart along the vertical axis. The column points are much less affected.

Figure 3
TWO-DIMENSIONAL PLOT OF CORRESPONDENCE ANALYSIS OF TABLE 6 DATA



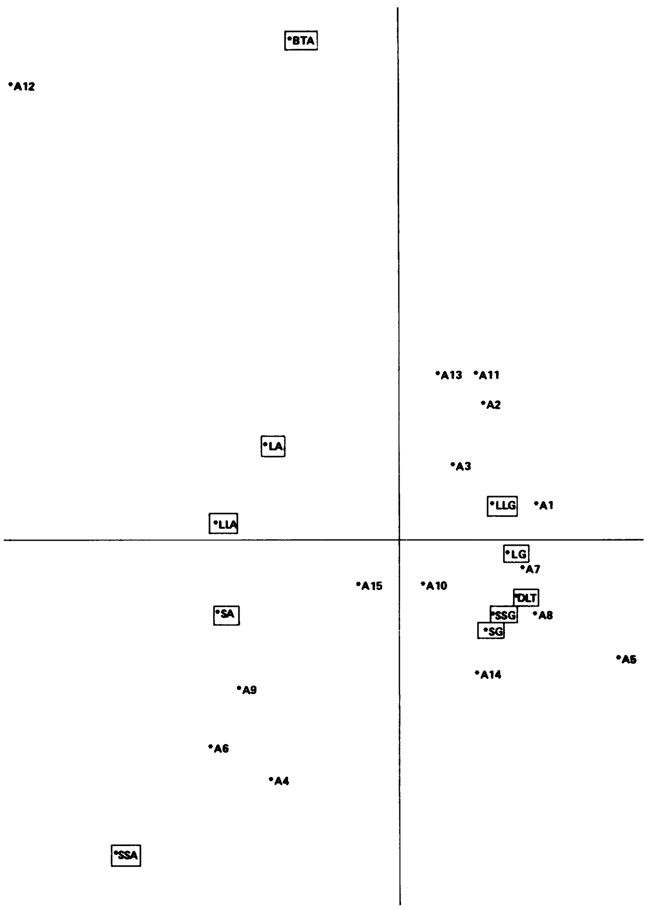
SUMMARY AND CONCLUSIONS

Our technical note has had two main objectives:

1. To describe the conditions (and associated scaling of the singular values) under which a squared distance interpretation is appropriate for both within-set and between-set relationships.
2. To illustrate the proposed scaling in the context of an actual data set.

Both illustrations described are based on two-way tables and, hence, use two-way correspondence analysis. Fortunately, however, the squared interpoint distance comparison can be extended readily to multiple corre-

Figure 4
TWO-DIMENSIONAL PLOT OF CORRESPONDENCE ANALYSIS OF TABLE 6 DATA BASED ON PSEUDOCONTINGENCY TABLE ANALOG



spondence analysis (Carroll 1968; Greenacre 1984) without major modification. In multiple correspondence analysis the data are arranged as three-way (or higher) tables. This type of data also can be expressed in dummy-variable form, similar to the pseudocontingency formulation described before. Hence, all of the results described for the two-way case generalize readily to higher way tables.

In sum, the proposed scaling provides marketing researchers with a metric joint-space model in which all within-set and between-set distances are comparable. The basic model is a type of internal "unfolding" analysis (Carroll 1980) that can be used for preference-type data or other kinds of cross-tabulation data, either two-way or higher way. (See Heiser 1981 for a discussion of the use of correspondence analysis in unfolding; however, Heiser uses the "standard" scaling of dimensions rather than the alternate scaling proposed here.) In our opinion, the approach can be a highly useful way for the analyst to understand row and column relationships in large contingency tables (such as Table 6); such visual representations could augment the more conventional examination of numerical frequencies alone.

The future looks promising for the application of correspondence analysis to the graphic interpretation of contingency tables and other kinds of cross-tabulation data. The virtues of correspondence analysis include its ease of application and the portrayal of data in a joint-space fashion so that both row and column relationships are comparable in terms of a (squared) distance model. As such it provides a (metric) point-point analog to Carroll's popular MDPREF (point-vector) model.[8]

REFERENCES

Benzécri, J. P. (1969), "Statistical Analysis as a Tool to Make Patterns Emerge from Data," in *Methodologies of Pattern Recognition,* S. Watanabe, ed. New York: Academic Press, Inc., 35–74.

——— et al. (1973), *L'Analyse des Données, Vol. I, La Taxinomie; Vol. II, L'Analyse des Correspondances.* Paris: Dunod.

Bouroche, J. M. and G. Saporta (1980), *L'Analyse des Données.* Paris: Presses Universitaires de France.

Carroll, J. Douglas (1968), "Generalization of Canonical Correlation Analysis to Three or More Sets of Variables," in *Proceedings of 76th Annual Conference of the American Psychological Association,* 3, 227–8.

——— (1980), "Models and Methods for Multidimensional Analysis of Preferential Choice (or Other Dominance) Data," in *Similarity and Choice,* E. D. Lantermann and H. Feger, eds. Vienna: Hans Huber Publishers.

de Leeuw, Jan (1973), *Canonical Analysis of Categorical Data,* unpublished doctoral dissertation, Psychological Institute, University of Leiden, The Netherlands.

Eckart, C. and Gale Young (1936), "The Approximation of One Matrix by Another of Lower Rank," *Psychometrika,* 1 (September), 211–8.

Gifi, A. (1981), *Non-Linear Multivariate Analysis.* Leiden, The Netherlands: Department of Data Theory, University of Leiden.

Gower, J. C. and P. G. N. Digby (1981), "Expressing Complex Relationships in Two Dimensions," in *Interpreting Multivariate Data,* V. Barnett, ed. Chichester, UK: John Wiley & Sons, Inc., 83–118.

Green, Paul E. with J. Douglas Carroll (1976), *Mathematical Tools for Applied Multivariate Analysis.* New York: Academic Press, Inc.

Greenacre, Michael J. (1981), "Practical Correspondence Analysis," in *Interpreting Multivariate Data,* V. Barnett, ed. Chichester, UK: John Wiley & Sons, Inc., 119–46.

——— (1984), *Theory and Application of Correspondence Analysis.* London: Academic Press, Inc.

Heiser, Willem J. (1981), *Unfolding Analysis of Proximity Data.* Leiden, The Netherlands: Department of Data Theory, University of Leiden.

Hill, M. O. (1974), "Correspondence Analysis: A Neglected

[8]Additional material on correspondence analysis is given by Greenacre (1981) and Hill (1974, 1982).

Multivariate Method," *Applied Statistics*, 23 (3), 340–54.
——— (1982), "Correspondence Analysis," in *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz and N. L. Johnson, eds. New York: John Wiley & Sons, Inc., 204–10.
Hoffman, Donna L. and George R. Franke (1986), "Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research," *Journal of Marketing Research*, 23 (August), 213–27.
Lebart, Ludovic, Alain Morineau, and Kenneth M. Warwick (1984), *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: John Wiley & Sons, Inc.
Meulman, Jacqueline (1982), *Homogeneity Analysis of Incomplete Data*. Leiden, The Netherlands: DSWO Press.
Nishisato, Shizuhiko (1980), *Analysis of Categorical Data: Dual Scaling and Its Applications*. Toronto: University of Toronto Press.
Saporta, G. (1980), "About Some Remarkable Properties of Generalized Canonical Analysis," paper presented at the Second European Meeting of the Psychometric Society, Groningen, The Netherlands (June).
Young, G. (1941), "Maximum Likelihood Estimation and Factor Analysis," *Psychometrika*, 6, 49–53.