



Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research

Author(s): Donna L. Hoffman and George R. Franke

Source: *Journal of Marketing Research*, Vol. 23, No. 3 (Aug., 1986), pp. 213-227

Published by: American Marketing Association

Stable URL: <http://www.jstor.org/stable/3151480>

Accessed: 01/09/2009 16:23

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=ama>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.



American Marketing Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of Marketing Research*.

<http://www.jstor.org>

Correspondence analysis is an exploratory data analysis technique for the graphical display of contingency tables and multivariate categorical data. Its history can be traced back at least 50 years under a variety of names, but it has received little attention in the marketing literature. Correspondence analysis scales the rows and columns of a rectangular data matrix in corresponding units so that each can be displayed graphically in the same low-dimensional space. The authors present the theory behind the method, illustrate its use and interpretation with an example representing soft drink consumption, and discuss its relationship to other approaches that jointly represent the rows and columns of a rectangular data matrix.

Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research

Marketing researchers often need to detect and interpret relationships among the variables in a rectangular data matrix. To facilitate this task, multidimensional scaling and unfolding, discriminant analysis, canonical correlation analysis, factor analysis, and principal components analysis all have been used to represent graphically the rows and/or columns of a data matrix. However, these methods have little applicability to the categorical data that arise in many marketing research applications. The purpose of our article is to direct the attention of the marketing community to correspondence analysis, a multivariate descriptive statistical method that represents graphically the rows and columns of a categorical data matrix in the same low-dimensional space.

In correspondence analysis, numerical scores are assigned to the rows and columns of a data matrix so as to maximize their interrelationship. The scores are in

corresponding units, allowing all the variables to be plotted in the same space for ease of interpretation. This representation then can be used to reveal the structure and patterns inherent in the data. In this sense, correspondence analysis is in that class of methods known as "exploratory data analysis" (cf. de Leeuw 1973; Heiser 1981; Tukey 1977).

Correspondence analysis has several features that contribute to its usefulness to marketing researchers. Much of its value relates to its multivariate treatment of the data through the simultaneous consideration of multiple categorical variables. The multivariate nature of correspondence analysis can reveal relationships that would not be detected in a series of pairwise comparisons of variables. Correspondence analysis also helps to show *how* variables are related, not just that a relationship exists. The joint graphical display obtained from a correspondence analysis can help in detecting structural relationships among the variable categories. Finally, correspondence analysis has highly flexible data requirements. The only strict data requirement for a correspondence analysis is a rectangular data matrix with non-negative entries. Thus, the researcher can gather suitable data quickly and easily.

A distinct advantage of correspondence analysis over other methods yielding joint graphical displays is that it produces two *dual* displays whose row and column geometries have similar interpretations, facilitating analysis and detection of relationships. In other multivariate

*Donna L. Hoffman is Assistant Professor, Graduate School of Business, Columbia University. George R. Franke is Assistant Professor, Department of Advertising, The University of Texas, Austin.

The authors thank William Moore, Donald Morrison, Thomas Novak, William Perreault, and two anonymous *JMR* reviewers for their helpful comments on a draft of this article. A longer version of the article with a technical appendix is available from the first author. Professor Hoffman gratefully acknowledges support from the Faculty Research Fund of the Graduate School of Business, Columbia University.

approaches to graphical data representation, this duality is not present.

Correspondence analysis as a geometric approach to multivariate descriptive data analysis originated in France; Benzécri (1969, 1973a, b) and his colleagues have done much to popularize the technique. The term "correspondence analysis" is a translation of the French "analyse factorielle des correspondances." The technique has received considerable attention in the statistical and psychometric literature under a variety of names, including dual scaling, method of reciprocal averages, optimal scaling, canonical analysis of contingency tables, categorical discriminant analysis, homogeneity analysis, quantification of qualitative data, and simultaneous linear regression. Complete histories of correspondence analysis are given by de Leeuw (1973), Greenacre (1984), and Nishisato (1980).

Though very few applications of correspondence analysis have been reported in the marketing literature, interest is increasing. Levine's (1979) procedure for the analysis of "pick-any" data, which is related closely to correspondence analysis, has been discussed by Holbrook, Moore, and Winer (1982). Green et al. (1983) use correspondence analysis in a cross-national examination of family purchasing roles. Franke (1983) illustrates the use of "dual scaling" with a reanalysis of data from a study by Belk, Painter, and Semenik (1981) on perceived causes of and preferred solutions to the energy crisis. Franke (1985) also discusses the use of dual scaling in examining measurement-level assumptions and interpreting responses to a measure. Additionally, Benzécri (1973b) describes two marketing-oriented applications of correspondence analysis, one evaluating competing cigarette brands and the other selecting a name for a new brand of cigarettes.

There is virtually no limit to the number of marketing applications for correspondence analysis. In the development of market segments, for example, correspondence analysis could be used to detect relatively homogeneous groupings of individuals. Correspondence analysis also can aid in product positioning studies. For example, suppose interest centers on consumer perceptions of brands as a basis for positioning a particular brand. Correspondence analysis of the categorical brands by attributes matrix gives information on the positioning of each brand *vis-à-vis* the attributes selected to describe them.

Correspondence analysis has been used to monitor the efficiency of advertising campaigns in France (Marc 1973). Before the ad campaign, a study is carried out to monitor advertising efficiency. After the campaign, another study is conducted. Together, the results of these studies reveal movement in product positioning attributable to the advertising campaign.

The method also may prove useful in the design phase of the new-product development process. Suppose a new-product manager gathers (binary) endorsements of consumers on a variety of proposed features of a new of-

fering. Correspondence analysis of this consumers by product features matrix affords guidelines for appropriate segmentation bases and potential marketing mix strategies. The method can be applied also in the concept-testing phase when several concepts are competing for developmental funds. Analysis of the concepts by attributes matrix can indicate those concepts that have the most favorable profiles and, consequently, should be developed further.

In the next two sections we use an artificial example to describe the theory behind the method of correspondence analysis. Appropriate types of data for its use and guidelines for interpretations also are discussed. We then illustrate correspondence analysis with an example that empirically demonstrates practical data considerations and issues of interpretation. The relationship of correspondence analysis to other multivariate methods is examined. In the concluding section we discuss the issues of supplementary variables and outliers, provide some cautions to the researcher, and comment on implementation.

THE METHOD OF CORRESPONDENCE ANALYSIS

An Artificial Example

In many marketing research applications, the data collected are categorical, mainly because of the limitations and constraints imposed on the data collection process. For example, a researcher may be interested in the relationship between several brands in a product class and a variety of attributes believed to describe the brands. Frequently the researcher gives consumers a list of brands and asks them to check off the attributes that describe the brands, rather than asking them to rate each brand on a scale. The advantages of this common data collection process are that it is quicker, easier, and less expensive than obtaining rating scale (i.e., interval-level) data.

As an example, suppose data were collected from 100 consumers on three brands, and six attributes were hypothesized to describe those brands. For each brand, respondents indicate whether the attribute describes the brand. The data generated from such a procedure might be arrayed as in Table 1. We have calculated, by subtraction, the "no" category for each attribute (we explain why subsequently). Twenty-nine percent of the respondents indicated that attribute 1 described brand A, 20% said that attribute 2 described brand A, and so on. These data, originally zeros and ones, have been aggregated over individuals and proportions calculated.

Suppose we are interested in the following questions.

1. What are the similarities and differences among the three brands with respect to the six attributes?
2. What are the similarities and differences among the six attributes with respect to the three brands?
3. What is the relationship among the brands and attributes?
4. Can these relationships be represented graphically in a joint low-dimensional space?

To answer these questions, we present the method of

Table 1
ARTIFICIAL DATA ON THREE BRANDS AND SIX ATTRIBUTES FROM 100 INDIVIDUALS

Brand	Attribute												Row mass
	1		2		3		4		5		6		
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
A	29 ^a (016 ^b)	71 039	20 011	80 044	18 010	82 045	24 013	76 042	20 011	80 044	13 007	87 048)	333
B	26 (014)	74 041	15 008	85 047	25 014	75 041	30 016	70 039	10 005	90 050	34 019	66 036)	333
C	25 (014)	75 041	26 014	74 041	31 017	69 038	21 011	79 044	15 008	85 047	24 013	76 042)	333
Column mass	044	122	034	132	042	124	041	125	024	142	040	126	

^aAll entries are proportions. Decimal points are omitted.

^bFigures in parentheses are rescaled so that their sum equals unity (before rounding).

correspondence analysis using as an example the artificial data of Table 1. Notation and general data concepts are introduced first. Correspondence analysis involves terminology that may be unfamiliar to marketing researchers. We maintain this terminology in our exposition for consistency with the psychometric and statistical literature. In the following discussion, boldface capital letters represent matrices, boldface lowercase letters represent vectors, and lowercase italic letters represent scalars.

Notation and Data Doubling

Let \mathbf{X} represent the 3×12 brands by attribute categories categorical data matrix displayed in Table 1. In general, the matrix is "objects by variable categories." The term "objects" is used to represent the extensive variety of products, commodities, goods, and consumers investigated in marketing research studies. Hence, objects may be brands, individuals, product classes, segments of consumers, etc. The term "variables" is used in the broadest sense possible and refers, in general, to characteristics of the objects being studied. These characteristics may be attributes, store locations, marketing mix variables, attitude statements, etc.

The general q -variate categorical data matrix \mathbf{X} is $n \times p$, where the q variables (e.g., attributes) are represented by sets of columns and categorical measurements of objects (e.g., brands) on these variables are represented by rows. Each variable has P_r categories (columns), with $r = 1, \dots, q$ and $P_1 + \dots + P_r + \dots + P_q = p$. The general entry x_{ij} is some categorical measure of the j^{th} variable category,¹ $j = 1, \dots, p$, on the i^{th} object, $i = 1, \dots, n$.

¹Actually, j indexes the l^{th} category of the r^{th} variable, $l = 1, \dots, P_r$, but this level of precision in notation is not required for the exposition we present.

When the q variables have only two possible responses (e.g., yes/no, endorse/do not endorse, purchase/do not purchase, etc.), only two categories are possible for each variable and $P_r = 2$ for all r . In practice, the researcher typically obtains only the positive endorsements and infers the negative by subtraction. In applications of correspondence analysis to data other than contingency tables the data matrix can be "doubled" to obtain this full set of responses. Doubling creates a symmetry between the two "poles" of each binary variable and renders the correspondence analysis invariant with respect to the direction in which we choose to scale the data (Greenacre 1984). The artificial example in this and the following section is based on such a doubled data matrix.

Algebraic Considerations in Correspondence Analysis

A variety of approaches lead to the equations of correspondence analysis (Tenenhaus and Young 1985). As a theoretical basis for developing the logic of correspondence analysis, we use the notion of the singular value decomposition (SVD) of a matrix (Eckart and Young 1936; Green with Carroll 1978). This "principal components analysis" approach, due largely to Greenacre (1978, 1984), is useful because it emphasizes the geometric properties of correspondence analysis and illuminates the practical implications of the data analysis. The singular value decomposition embodies the idea of the basic structure of a matrix, consisting of basic values and basic vectors. The eigenstructure (eigenvalues and eigenvectors) of a symmetric matrix is a special case of the SVD.

The philosophy behind correspondence analysis is to obtain a graphical representation of both the rows and columns of the original data matrix in terms of as few dimensions as possible. In correspondence analysis, each row of \mathbf{X} represents a point *profile* in p -dimensional space and each column represents a point profile in n -dimen-

sional space. Attention is directed to the profiles of the frequency distributions rather than their raw occurrence, because the raw frequencies in Table 1 do not yield a meaningful interpretation of distances between row points and between column points.

In terms of the n brands, say, in p -dimensional attribute space, it is clear that some brands will "occur" frequently and consequently some attribute categories will be endorsed frequently for those brands. Other brands will have small frequencies of occurrence and hence the attribute categories attributed to them will appear less frequently. The brand profiles are *conditional frequencies* of attribute category j given brand i . Similarly for the p attributes in n -dimensional brand space, the conditional frequencies of brand i given attribute category j are the quantities of interest.

To perform a correspondence analysis, one rescales the original data matrix \mathbf{X} so that the sum of the elements equals 1.

$$(1) \quad \mathbf{P} = \mathbf{X}/\mathbf{1}'\mathbf{X}\mathbf{1}, \text{ with } \mathbf{1}'\mathbf{P}\mathbf{1} = 1,$$

where $\mathbf{1}' = (1 \dots 1)'$, either $1 \times n$ or $1 \times p$, depending on the context. \mathbf{P} is the *correspondence matrix* whose elements are the relative frequencies and if \mathbf{X} is a contingency table, \mathbf{P} is the probability density on the cells of \mathbf{X} . The row sums of \mathbf{P} are written into \mathbf{D}_r , an $n \times n$ diagonal matrix,

$$(2) \quad \mathbf{D}_r = \text{diag}(\mathbf{r})$$

where $\mathbf{r} = \mathbf{P}\mathbf{1}$, and the column sums of \mathbf{P} are written into \mathbf{D}_c , a $p \times p$ diagonal matrix,

$$(3) \quad \mathbf{D}_c = \text{diag}(\mathbf{c})$$

where $\mathbf{c} = \mathbf{P}'\mathbf{1}$. These row and column sums are referred to as *masses* in correspondence analysis. The masses enable us to *weight* each profile point in proportion to its frequency. Again, if we are working with a contingency table, \mathbf{r} and \mathbf{c} are the marginal densities. These densities are only analogies when \mathbf{X} is not a contingency table. The entries of \mathbf{P} , with row and column masses \mathbf{r} and \mathbf{c} , respectively, are in parentheses in Table 1 below the corresponding entries of \mathbf{X} .

Note that the brand masses are all equal (.33) and that each attribute also has equal mass (.16), though this quantity is distributed differently for each attribute between the yes and no categories, depending on the frequency of responses in each category. In this example, the masses are equal because each row sums to the same constant value and each pair of columns sums to the same constant value, by design. In other situations, such as in contingency tables, the masses will not necessarily be equal.

The row and column profiles of \mathbf{P} are defined as the vectors of row and column elements of \mathbf{P} divided by their respective masses. The n row profiles in p -dimensional space are written in the rows of \mathbf{R} and the p col-

umn profiles in n -dimensional space are written in the rows of \mathbf{C} .

$$(4) \quad \mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P}$$

and

$$(5) \quad \mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}'$$

Note that a profile (row or column) sums to unity. The correspondence analysis problem is to find a low-rank approximation to the original data matrix that optimally represents both these row and column profiles in k -dimensional subspaces, where k is generally much smaller than either n or p .² These two k -dimensional subspaces (one for the row profiles and one for the column profiles) have a geometric correspondence, which we examine hereafter, that enables us to represent both in one joint display.

Because we wish to represent graphically the distances between row (or column) profiles, we orient the configuration of points at the "center of gravity" of both sets. The centroid of the set of row points in its space is \mathbf{c} , the vector of column masses. This defines the "average" row profile. The centroid of the set of column points in its space is \mathbf{r} , the vector of row masses. This is the average column profile. To perform the analysis relative to the center of gravity, \mathbf{P} is centered "symmetrically" by rows and columns, that is, $\mathbf{P} - \mathbf{rc}'$, so that the origin corresponds to the average profile of both sets of points.

The solution to finding a representation of both row and column profiles in a low-dimensional space involves the generalized singular value decomposition (GSVD) and low-rank matrix approximation theory (Seber 1984). The GSVD of the symmetrically centered correspondence matrix \mathbf{P} defines the theoretical correspondence analysis problem.

$$(6) \quad \mathbf{P} - \mathbf{rc}' = \tilde{\mathbf{M}}\mathbf{D}_{\tilde{\mu}}\tilde{\mathbf{N}}'$$

where $\tilde{\mathbf{M}}'\mathbf{D}_r^{-1}\tilde{\mathbf{M}} = \tilde{\mathbf{N}}'\mathbf{D}_c^{-1}\tilde{\mathbf{N}} = \mathbf{I}$, with $\tilde{\mathbf{M}}$ $n \times k$, $\tilde{\mathbf{N}}$ $p \times k$, and $\mathbf{D}_{\tilde{\mu}}$ $k \times k$, and $\tilde{\mu}_1 \geq \dots \geq \tilde{\mu}_k \geq 0$.

The columns of $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{N}}$ hold the first k left and right generalized basic vectors of $\mathbf{P} - \mathbf{rc}'$, in the metrics \mathbf{D}_r^{-1} and \mathbf{D}_c^{-1} , corresponding to the k largest basic values and define the optimal weighted Euclidean k -dimensional subspaces in terms of weighted sum of squared distances. $\mathbf{D}_{\tilde{\mu}}$ is a diagonal matrix holding the generalized basic values $\tilde{\mu}_1, \dots, \tilde{\mu}_k$, in descending order, corresponding to the generalized basic vectors. In other words, the principal axes of the attribute category (column) set of points are defined by the columns of $\tilde{\mathbf{M}}$ and the principal axes of the brand (row) set of points are defined by the columns of $\tilde{\mathbf{N}}$. The weighted centers of

²Correspondence analysis optimizes several criteria simultaneously. See Tenenhaus and Young (1985) for a detailed discussion.

gravity of each set of points are both at the origin of the principal axes.³

The *principal coordinates* (cf. Gower 1966) of the brand and attribute category profiles, with respect to their principal axes, are written in the rows of \mathbf{F} and \mathbf{G} , respectively.

$$(7) \quad \mathbf{F} = (\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1c}')\mathbf{D}_c^{-1}\tilde{\mathbf{N}}$$

and

$$(8) \quad \mathbf{G} = (\mathbf{D}_c^{-1}\mathbf{P}' - \mathbf{1r}')\mathbf{D}_r^{-1}\tilde{\mathbf{M}}$$

The set of points defined in equation 7 are the n row profiles in weighted Euclidean k -dimensional space, with masses defined by the n elements of \mathbf{r} and principal axis weights defined by the inverses of the elements of \mathbf{c} , that is, \mathbf{D}_c^{-1} . A similar definition holds for the column set of points defined in equation 8. These are the p column profiles in weighted Euclidean k -dimensional space, with masses defined by the p elements of \mathbf{c} and principal axis weights defined by the inverses of the elements of \mathbf{r} , that is, \mathbf{D}_r^{-1} . Thus, the principal axes are weighted inversely by the elements of the average profile.

Each set of points can be related to the principal axes of the *other* set of profile points through rescalings by the basic values.

$$(9) \quad \mathbf{F} = \mathbf{D}_r^{-1}\tilde{\mathbf{M}}\mathbf{D}_\mu$$

and

$$(10) \quad \mathbf{G} = \mathbf{D}_c^{-1}\tilde{\mathbf{N}}\mathbf{D}_\mu$$

In practice, the correspondence analysis problem is restated in an equivalent form in terms of the SVD for computational convenience.

$$(11) \quad \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}')\mathbf{D}_c^{-1/2} = \mathbf{M}\mathbf{D}_\mu\mathbf{N}'$$

where $\mathbf{M}'\mathbf{M} = \mathbf{N}'\mathbf{N} = \mathbf{I}$, with \mathbf{M} $n \times k$, \mathbf{N} $p \times k$, \mathbf{D}_μ $k \times k$, and $\mu_1 \geq \dots \geq \mu_r \geq \dots \geq \mu_k > 0$. Then,

$$(12) \quad \mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{M}\mathbf{D}_\mu, \text{ where } \mathbf{M} = \mathbf{D}_r^{-1}\tilde{\mathbf{M}},$$

and

$$(13) \quad \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{N}\mathbf{D}_\mu, \text{ where } \mathbf{N} = \mathbf{D}_c^{-1}\tilde{\mathbf{N}},$$

and plotting the rows of \mathbf{F} and \mathbf{G} in the same space results in a k -dimensional correspondence analysis.

Correspondence analysis can be considered a dual generalized principal components analysis (Greenacre 1984). The columns of \mathbf{F} are the eigenvectors of \mathbf{RC} and

the columns of \mathbf{G} are the eigenvectors of \mathbf{CR} .

$$(14) \quad (\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}\mathbf{P}')\mathbf{F} = \mathbf{F}\mathbf{D}_\lambda$$

and

$$(15) \quad (\mathbf{D}_c^{-1}\mathbf{P}'\mathbf{D}_r^{-1}\mathbf{P})\mathbf{G} = \mathbf{G}\mathbf{D}_\lambda$$

The eigenvalues, λ_r , are the *weighted variances* of each principal axis (the weighted sums of squares of the points' coordinates along the r^{th} principal axis in each set) and are equal to the corresponding squared basic values from the SVD in equation 11.

$$(16) \quad \mathbf{F}'\mathbf{D}_r\mathbf{F} = \mathbf{D}_\mu^2 = \mathbf{D}_\lambda$$

and

$$(17) \quad \mathbf{G}'\mathbf{D}_c\mathbf{G} = \mathbf{D}_\mu^2 = \mathbf{D}_\lambda$$

The axes are orthogonal, though the metric is "chi square" and not ordinary Euclidean as in principal components analysis.

The *transition formulas* relate the brand and attribute category *coordinates* to each other.

$$(18) \quad \mathbf{F} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{G}\mathbf{D}_\mu^{-1} = \mathbf{R}\mathbf{G}\mathbf{D}_\mu^{-1}$$

and

$$(19) \quad \mathbf{G} = \mathbf{D}_c^{-1}\mathbf{P}'\mathbf{F}\mathbf{D}_\mu^{-1} = \mathbf{C}\mathbf{F}\mathbf{D}_\mu^{-1}$$

Hill (1974) considers these the defining formulas for a correspondence analysis.

The transition formulas in equations 18 and 19 are important because they provide the mechanism for obtaining one set of coordinates from the other set. To see the geometric importance of these formulas, consider the i^{th} row of \mathbf{F} , \mathbf{f}'_i .

$$(20) \quad \mathbf{f}'_i = (\tilde{\mathbf{r}}'_i\mathbf{G})\mathbf{D}_\mu^{-1}$$

where $\tilde{\mathbf{r}}'_i$ is the i^{th} row profile from \mathbf{R} in equation 4. Equation 20 defines a "barycenter," actually a *center of mass*, of the p column profile points in \mathbf{G} , because the sum of the elements of $\tilde{\mathbf{r}}'_i$ equals unity. Postmultiplication by \mathbf{D}_μ^{-1} divides the coordinates of the centroid by the singular values. Geometrically, a particular row profile will be "attracted" to a position in its subspace that corresponds to the column variable categories prominent in that row profile. A corresponding definition and interpretation holds for the rows of \mathbf{G} .

Distances (squared) between points in the same set are given by

$$(21) \quad d_{ii'}^2 = \sum_j 1/c_j(p_{ij}/r_i - p_{i'j}/r_{i'})^2$$

for row points i and i' and

$$(22) \quad d_{jj'}^2 = \sum_i 1/r_i(p_{ij}/c_j - p_{ij'}/c_{j'})^2$$

for column points j and j' . These are similar to ordinary

³The centering operation has the effect of removing the *trivial axes* with corresponding basic values of unity. That is, without centering, the first axes (columns) extracted from the left and right generalized basic vectors would correspond to \mathbf{r} and \mathbf{c} , respectively, and the first diagonal element of \mathbf{D}_μ would equal 1. In this case, the analysis is performed relative to the origin, rather than from the "center of gravity." Because the GSVD of $\mathbf{P} - \mathbf{rc}'$ is "contained" in the GSVD of \mathbf{P} (Greenacre 1978, 1984), attention is restricted to $\mathbf{P} - \mathbf{rc}'$.

Euclidean distances except that each squared term is weighted by the inverse of the relative frequency (mass) corresponding to the term.

These distances, approximated in the k -dimensional subspaces by

$$(23) \quad d_{ii'}^2 \approx (\mathbf{f}_i - \mathbf{f}_{i'})'(\mathbf{f}_i - \mathbf{f}_{i'})$$

and

$$(24) \quad d_{jj'}^2 \approx (\mathbf{g}_j - \mathbf{g}_{j'})'(\mathbf{g}_j - \mathbf{g}_{j'})$$

for row and column points, respectively, are defined as *chi square* distances. This distance measure is chosen because it guarantees invariance according to the property of *distributional equivalence*:

- If two rows having identical column profiles are aggregated, the distances between columns remain unchanged.
- If two columns having identical row profiles are aggregated, the distances between rows remain unchanged.

Clearly, identical profiles imply equal or proportional raw data.

The Correspondence Analysis Model

The correspondence analysis "model" on \mathbf{P} in k dimensions reveals how an element of \mathbf{P} is approximated in the k -dimensional weighted Euclidean subspace.

$$(25) \quad \mathbf{P} \approx \mathbf{rc}' + \mathbf{D}_r \mathbf{F} \mathbf{D}_c^{-1} \mathbf{G}' \mathbf{D}_c$$

From equation 25 it is clear that the model treats rows and columns symmetrically, as nothing changes if we begin with \mathbf{X}' instead of \mathbf{X} .

Data Considerations

Because of the inherent symmetry of correspondence analysis, the implied data matrix for analysis is a contingency table. However, the method can be applied to almost any matrix of categorical data, as long as the entries are non-negative. Excellent data classifications for correspondence analysis are given by Benzécri (1973b), Nishisato (1980), and Greenacre (1984).

Many situations in marketing research lead to data at the nominal or ordinal level of measurement (Perreault and Young 1980). Such data are often intractable with traditional analytical methods. A common source of this type of data is the evaluation of objects (e.g., retail outlets, competing products, individuals) on attributes (e.g., product features, attitude statements) with binary judgments rather than 5- or 7-point rating scales. Binary judgments are useful when the researcher has many objects or attributes to measure, when respondent cooperation is difficult to obtain, when it is difficult to make fine distinctions between objects on the attributes, and whenever rating scales are difficult to use.

Another source of data common in marketing research is the open-ended elicitation of attributes, brands, stores, and so on, from respondents (i.e., "pick-any" data). With an unconstrained set of alternatives, failure to mention an alternative does not necessarily imply rejection of it.

Correspondence analysis is appropriate for such data, whereas standard multidimensional scaling methods are not (Holbrook, Moore, and Winer 1982).

Though correspondence analysis is ideally suited to those research situations in which categorical measurements are the most reasonably obtained, it also can be applied to ordered categories and "discretized" quantitative variables (see Jambu and Lebeaux 1983), but the original ordering may not be maintained after scaling unless the solution is constrained (Nishisato and Sheu 1984). This type of application allows investigation of possible nonlinearities among the categories with respect to the principal axes. It can lead to the discovery of relationships between scale value categories that are obscured if the data are dichotomized, or if methods are used that recognize only the metric properties of the data. Thus, a "loss of information" in ignoring the ordered or interval nature of the data yields a meaningful gain in understanding (Lebart, Morineau, Warwick 1984).

Correspondence analysis of other forms of data, such as rank-order data, sorting data, paired comparison data, and successive categories data, is discussed by Jambu and Lebeaux (1983), Nishisato (1980), Nishisato and Nishisato (1983), and Nishisato and Sheu (1984). Applications of correspondence analysis are virtually unlimited, but Lebart, Morineau, and Warwick (1984) suggest three conditions that should be satisfied if correspondence analysis is to be most effective.

1. The data matrix must be large enough that visual inspection or simple statistical analysis cannot reveal its structure.
2. The variables must be "homogeneous," so that it makes sense to calculate a statistical distance between rows and columns and so that distances can be interpreted meaningfully.
3. The data matrix must be "amorphous, *a priori*." In other words, the method is most fruitfully applied to data whose structure is either unknown or only poorly understood.

INTERPRETING A CORRESPONDENCE ANALYSIS

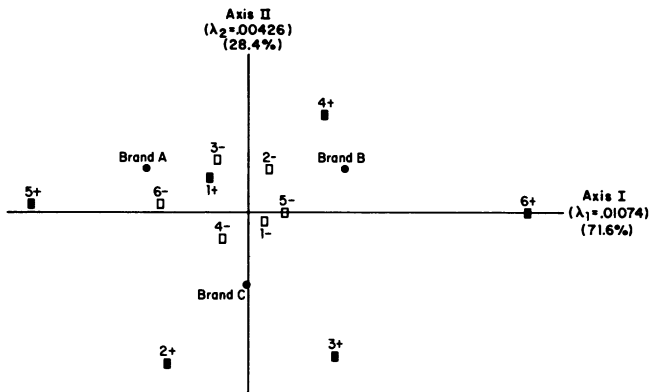
The principal coordinates of the brand and attribute category profile points from the correspondence analysis in two dimensions of the artificial data of Table 1 are plotted in Figure 1. The plots are merged into one joint display for ease of interpretation.

The overall spatial variation in each set of points can be quantified and assists in interpretation. This variation, the *total inertia*, is defined as the weighted sum of squared distances from the points to their respective centroids and is equivalent for both sets of points.

$$(26) \quad \text{Inertia (Total)} = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} =$$

$$(27) \quad \text{Inertia (rows)} = \sum_i r_i \left[\sum_j 1/c_j (p_{ij}/r_i - c_j)^2 \right] \\ = \sum_i r_i (\bar{\mathbf{r}}_i - \mathbf{c})' \mathbf{D}_c^{-1} (\bar{\mathbf{r}}_i - \mathbf{c}) =$$

Figure 1
TWO-DIMENSIONAL CORRESPONDENCE ANALYSIS OF
THE DOUBLED DATA MATRIX IN TABLE 1



$$(28) \quad \text{Inertia (columns)} = \sum_j c_j \left[\sum_i 1/r_i (p_{ij}/c_j - r_i)^2 \right] \\ = \sum_j c_j (\bar{c}_j - \mathbf{r})' \mathbf{D}_r^{-1} (\bar{c}_j - \mathbf{r}).$$

It is because of the geometric correspondence of the two sets of points, in position and inertia, that we can merge the two displays into one joint display. The advantage of this merger is that a concise graphical display representing varied features of the data is obtained in a single picture. The geometric display of each set of points reveals the nature of similarities and variation within the set, and the joint display shows the correspondence between sets. However, distances between points from different sets *cannot* be interpreted because these distances do not approximate any defined quantity. Distances between points in the same set are equal to the relevant chi square distances in equations 23 and 24, whereas the between-set correspondence is influenced by the barycentric nature of the transition formulas in equations 18 and 19.

The total inertia also can be decomposed along the principal axes. Each eigenvalue, λ_i , indicates the weighted variance (inertia) explained by the i^{th} principal axis of the display. Summed over all k principal axes, these eigenvalues represent the total inertia of the spatial representation.

The first principal axis in the artificial example accounts for 71.6% of the spatial variation in the data ($\lambda_1 = .01074$). The second principal axis accounts for the remaining 28.4% ($\lambda_2 = .00426$). In this artificial example, two dimensions recover exactly the original data matrix because with three brands there are at most two mutually exclusive dimensions. Real applications involving reduced dimensionalities of larger data matrices will necessarily be approximations.

From Figure 1 we see that brands A, B, and C are relatively far from each other in terms of the attributes that describe them. Their relative positions as points in the two-dimensional space indicate the similarities and differences among them with respect to the attributes. The first dimension separates attribute 6+ on the right from attribute 5+ on the left, and also separates brand B on the right from brand A on the left. The second dimension separates attributes 2+ and 3+ on the bottom from attribute 4+ on the top. This dimension, in addition, differentiates brand C on the bottom from brands A and B on the top.

The transition formulas in equations 18 and 19 make clear that, geometrically, a particular brand will tend to a position in its space corresponding to the attribute categories prominent in that brand profile. Similarly, given the display of brand profiles, a particular attribute category will tend along the principal axes in the direction of the brands that are relatively substantial in that category. For example, the attribute category 3+ point is on the negative side of the second principal axis, and brand C, which is relatively high on attribute 3+ (see Table 1), is on the negative side of its second principal axis. Points near the center of the display have undifferentiated profile distributions as a consequence of the origin placed at the center of gravity. Notice that we have been careful *not* to interpret between-set distances.

The interpretation of the correspondence analysis is not yet complete. The two-dimensional display in Figure 1 shows the projections of the point profiles onto the plane, but does not indicate which points have had the most impact in determining the orientation of the axes. For a complete and correct interpretation of the graphical display, we must use additional information.

Because the total inertia of each set of points is decomposed along the principal axes and among the points in similar and symmetric fashion, the inertia for each set of points can be decomposed in a manner analogous to the decomposition of variance. These various decompositions are used to assist in the interpretation of the graphical display.

Table 2 is the numerical representation of the correspondence analysis depicted in Figure 1. Each column represents a particular decomposition of the variation in each set of points and is discussed in turn. The two columns headed "Coordinate" contain the coordinates of the points on the first and second principal axes, respectively. The weights for each point (column headed "Mass") are repeated from Table 1 for completeness.

Inertia of the Points

The inertia of the i^{th} brand point is equal to

$$(29) \quad r_i \left[\sum_j 1/c_j (p_{ij}/r_i - c_j)^2 \right] = r_i \sum_i f_{ii}^2.$$

Equation 29 represents the contribution of the i^{th} brand to the total inertia, with r_i the mass of that brand and the

Table 2
NUMERICAL RESULTS OF CORRESPONDENCE ANALYSIS OF TABLE 1 DATA^a

Name	Quality	Mass	Inertia ^b	Axis 1			Axis 2		
				Coordinate	Squared correlation	Contribution ^b	Coordinate	Squared correlation	Contribution ^b
Brand A	1000	333	405	-130	883	500	50	117	167
Brand B	1000	333	405	130	883	500	50	117	167
Brand C	1000	333	190	0	0	0	-90	1000	666
1+ (yes)	1000	44	10	-50	519	9	40	481	21
1- (no)	1000	122	1	20	519	3	-20	481	7
2+	1000	34	110	-100	206	32	-20	794	307
2-	1000	132	31	30	206	8	50	794	78
3+	1000	42	130	120	289	51	-180	711	316
3-	1000	124	44	-40	289	17	60	711	104
4+	1000	41	60	100	429	37	110	571	124
4-	1000	125	20	-30	429	12	-40	571	41
5+	1000	24	123	-270	1000	172	0	0	0
5-	1000	142	21	50	1000	30	0	0	0
6+	1000	40	340	360	999	480	-10	1	1
6-	1000	126	110	-110	999	149	0	1	1

^aAll values are multiplied by 1000 and decimal points are omitted.

^bScaled (before multiplication by 1000) to sum to unity.

quantity in brackets the squared chi squared distance of the brand profile to the center of gravity c in the brand space (i.e., $\sum_i f_{ii}^2$). A similar definition holds for each attribute category point. These contributions, summed over all brands (or attribute categories), equal the total inertia.

The inertias for each point are in the column headed "Inertia" in Table 2. Brand A's inertia in the set of brand points is 40.5% of the total inertia, as is brand B's. Brand C accounts for 19% of the total inertia in this set. Attribute category 6+ has an inertia that is 34% of the total inertia in the attribute set of points and accounts for by far the largest proportion.

Absolute Contributions to Inertia

The inertia along the t^{th} axis, λ_t , consists of the weighted sum of squared distances to the origin of the displayed row (or column) profiles, where the weights are the masses for each row (or column) point. For the brand profiles, this inertia can be expressed as

$$(30) \quad \lambda_t = \sum_i r_i f_{ii}^2.$$

A similar definition holds for the attribute category profiles. Thus, each eigenvalue also represents the inertia of the projections of the brand set (or attribute category set) of points on each axis.

If each term in the summation is expressed as a percentage relative to the inertia "explained" by each axis, that is,

$$(31) \quad r_i f_{ii}^2 / \lambda_t,$$

the absolute contribution of the i^{th} brand to the t^{th} principal axis is obtained. The absolute contributions quantify the importance of each point in determining the direction of the principal axes and serve as guides to interpretation of each axis. They are interpreted as the percentage of (weighted) variance explained by each point in relation to each axis.

It is clear from the decomposition that a point can contribute to a principal axis (i.e., make a high contribution to the inertia of that axis) in two ways: when it has a large mass and/or when it is a large distance from the centroid, even if it has relatively low mass.

Because all the brands have equal mass, it is their distance from the centroid that determines their contributions to the inertia of each axis. The absolute contributions, in the columns headed "Contribution" in Table 2, indicate that brands A and B contribute equally and solely to the direction of axis 1 and brand C contributes primarily to axis 2.

Similarly for the attributes, categories 6+, 5+, and 6- define the first principal axis whereas categories 3+, 2+, 4+, and 3- define the second principal axis. Attribute categories 1-, 5-, and to a lesser extent 1+ contribute essentially nothing to the inertia of each axis and consequently are near the origin (note that their profiles are virtually identical to the average column profile \bar{r}).

Relative Contributions to Inertia

After the dimensional interpretation, the next step in a correspondence analysis is to determine the "quality"

of the representation of each point in the display. The quantity

$$(32) \quad f_{ii}^2 / \sum_t f_{it}^2$$

gives the *relative contribution* of the t^{th} principal axis to the inertia of the i^{th} brand. A similar definition holds for the relative contributions of the attribute categories. These values are independent of the point's mass and indicate how well each point is "fit" by the representation.

A relative contribution is actually a squared correlation, because it is equal to the \cos^2 of the angle θ between the point and the t^{th} principal axis. High values of $\cos^2\theta$ indicate that the axis explains the point's inertia very well; θ is low and the profile point lies in the direction of the axis and correlates highly with it. Summed over all the axes of interest (in this case two), the relative contributions give the *quality* of the representation. This is just the \cos^2 of the angle the point makes with the subspace. Thus, the relative contribution gives that part of the variance of a point explained by an axis, and the quality gives the goodness of fit of each point's representation in the subspace. The sum of the relative contributions over *all* axes (not just those used for the display) equals unity (as in Table 2).

The relative contributions are in the columns headed "Correlation" in Table 2. The first axis explains 88.3% of the inertia of brands A and B and nothing of brand C, whereas the second axis explains 11.7% of the inertia of brands A and B and 100% of that of brand C. Similarly, the first axis explains 51.9% of attribute 1 and the second axis explains the remaining 48.1%. The relative contributions are equal for each attribute category pair because the doubling procedure gives each pair of attribute categories equal mass. The qualities of each point in the two-dimensional space (all equal to unity) are in the column headed "Quality."

The various decompositions of the total inertia, in conjunction with the principal coordinate values for the brands and attribute categories, make possible a complete interpretation of the correspondence analysis of the data in Table 1. As discussed hereafter, external information can be fit into the display through the transition formulas in equations 18 and 19 and also can be helpful in interpreting correspondence analysis results. Another aid is cluster analysis, which with large data matrices may be useful in detecting homogeneous groups and in presenting results (Jambu and Lebeaux 1983). Whatever aids are used, we emphasize that visual inspection of the graphical display is a key step in interpreting the results.

ILLUSTRATING CORRESPONDENCE ANALYSIS

Empirical Example: Beverage Purchase and Consumption

A group of male and female MBA students from Columbia University were asked to indicate, for a variety of popular soft drinks, the frequency with which they purchased and consumed the soft drinks in a 1-month

period. For illustrative purposes, the scale used to collect the information was coded 1 to indicate purchase and consumption at least every other week and 0 to indicate purchase and consumption less than every other week. The data about eight soft drinks from 34 of the students were used for our example. The soft drinks are Coke, Diet Coke, Diet Pepsi, Diet 7Up, Pepsi, Sprite, Tab, and 7Up. The 34×8 binary indicator matrix is displayed in Table 3.

A correspondence analysis was performed on the 34×16 matrix obtained by doubling Table 3 about the columns. The resulting eigenvalues equal their corresponding proportions of inertia because the total inertia equals unity (in this example). The eigenvalues for the first three principal axes are .482, .151, and .099, with cumulative proportions of inertia equaling .482, .633, and .732. Eight dimensions recover perfectly the 34×16 doubled data matrix (i.e., $\sum \lambda_i = 1$), hence the other five axes account for the remaining 26.8% of the inertia.

In a doubled binary data matrix with q variables, the row sums equal a constant value (8 in this case) and the

Table 3
THE 34×8 BINARY INDICATOR MATRIX OF BEVERAGE PURCHASE AND CONSUMPTION

Individual	Soft drink						
	Coke	Diet Coke	Diet Pepsi	Diet 7Up	Pepsi	Sprite	Tab 7Up
1	1	0	0	0	1	1	0 1
2	1	0	0	0	1	0	0 0
3	1	0	0	0	1	0	0 0
4	0	1	0	1	0	0	1 0
5	1	0	0	0	1	0	0 0
6	1	0	0	0	1	1	0 0
7	0	1	1	1	0	0	1 0
8	1	1	0	0	1	1	0 1
9	1	1	0	0	0	1	1 1
10	1	0	0	0	1	0	0 1
11	1	0	0	0	1	1	0 0
12	0	1	0	0	0	0	1 0
13	0	0	1	1	0	1	0 1
14	1	0	0	0	0	1	0 0
15	0	1	1	0	0	0	1 0
16	0	0	0	0	1	1	0 0
17	0	1	0	0	0	1	0 0
18	1	1	0	0	1	0	0 0
19	1	0	0	0	0	0	0 1
20	1	1	1	0	1	0	0 0
21	1	0	0	0	1	0	0 0
22	1	0	0	0	1	0	0 0
23	0	1	0	1	0	0	1 0
24	1	1	0	0	1	0	0 0
25	0	1	1	1	0	0	0 0
26	0	1	0	1	0	0	1 0
27	0	1	0	0	0	0	1 0
28	1	0	0	0	0	1	0 1
29	1	0	0	0	0	1	0 0
30	0	1	1	0	0	0	1 0
31	1	0	0	0	1	0	0 1
32	0	1	1	0	0	0	1 0
33	1	0	0	0	1	0	0 1
34	0	1	1	1	0	0	1 0

Table 4
DECOMPOSITION OF INERTIA AMONG THE SOFT DRINKS FOR THE FIRST TWO PRINCIPAL AXES^a

<i>Soft drink</i>	<i>Quality</i> ^b	<i>Mass</i>	<i>Inertia</i> ^c	<i>Axis 1</i>		<i>Axis 2</i>	
				<i>Squared correlation</i>	<i>Contribution</i> ^c	<i>Squared correlation</i>	<i>Contribution</i> ^c
Coke+	809	70	55	785	84	18	6
Coke-	809	55	70	785	120	18	9
Diet Coke+	664	62	62	614	80	19	8
Diet Coke-	664	62	62	614	80	19	8
Diet Pepsi+	643	25	100	404	80	1	1
Diet Pepsi-	643	100	25	404	25	1	1
Diet 7Up+	626	25	100	458	94	49	32
Diet 7Up-	626	100	25	458	24	49	8
Pepsi+	811	62	62	550	76	227	98
Pepsi-	811	62	62	550	67	227	88
Sprite+	826	40	85	149	26	533	298
Sprite-	826	85	40	149	13	533	142
Tab+	751	40	85	712	125	0	0
Tab-	751	85	40	712	60	0	0
7Up+	727	35	90	181	34	364	221
7Up-	727	90	35	181	12	364	80

^aAll values are multiplied by 1000 and decimal points are omitted.

^bMeasured over the first three principal axes.

^cScaled (before multiplication by 1000) to sum to unity.

Table 5
DECOMPOSITION OF INERTIA AMONG THE INDIVIDUALS FOR THE FIRST TWO PRINCIPAL AXES^a

<i>Individual</i>	<i>Quality</i> ^b	<i>Mass</i>	<i>Inertia</i> ^c	<i>Axis 1</i>		<i>Axis 2</i>	
				<i>Squared correlation</i>	<i>Contribution</i> ^c	<i>Squared correlation</i>	<i>Contribution</i> ^c
1	911	30	27	701	47	198	42
2,3,5,21,22	909	30	18	520	19	384	44
4,23,26	702	30	40	690	55	0	0
6,11	716	30	18	623	30	1	1
7,34	955	30	49	888	90	1	1
8	453	30	27	322	21	130	27
9	463	30	40	0	0	371	91
10,31,33	808	30	27	593	32	1	1
12,27	776	30	27	490	25	34	6
13	864	30	60	42	5	572	219
14,29	621	30	18	248	12	145	21
15,30,32	704	30	40	688	51	16	3
16	300	30	27	162	9	14	2
17	611	30	27	35	2	129	21
18,24	678	30	18	124	4	541	61
19	428	30	27	245	13	85	14
20	444	30	27	0	0	301	55
25	715	30	40	543	49	1	1
28	908	30	27	347	23	560	116

^aAll values are multiplied by 1000 and decimal points are omitted.

^bMeasured over the first three principal axes.

^cScaled (before multiplication by 1000) to sum to unity.

ysis, and discriminant analysis, through the generalized singular value decomposition (see Greenacre 1984, Appendix A.2). The differences among methods are determined by the type of transformation applied to the original data matrix, the metrics in which the principal axes are defined, and how the basic values are assigned to the left and right basic vectors. In correspondence analysis, the transformation is defined by equation 11, the metrics are the chi square metrics defined by the inverses of equations 2 and 3, and the basic values are assigned according to equations 12 and 13.

A distinct advantage of correspondence analysis over other methods, in terms of obtaining a joint graphical display, is that correspondence analysis produces two *dual* displays whose row and column geometries have similar interpretations. In other multivariate approaches as they are commonly employed, this duality does not exist.

Tenenhaus and Young (1985) have shown that four broad data analytic approaches lead to the equations of correspondence analysis, the "method of reciprocal averages" (Fisher 1940; Hirschfeld 1935; Horst 1935; Richardson and Kuder 1933), the analysis of variance approach (Bock 1960; de Leeuw 1973; Guttman 1941; Hayashi 1950, 1952, 1954; Nishisato 1980; van Rijckevorsel and de Leeuw 1978), the principal components analysis (PCA) approach (Benzécri 1969, 1973a, b; Burt 1950; Greenacre 1978, 1984), and the generalized canonical analysis approach (McKeon 1966). We use the PCA approach to demonstrate correspondence analysis because it illustrates clearly the geometric aspects of the method. However, the equivalences yield additional interpretations of the results of a correspondence analysis (involving the meaning of the eigenvalues), which illuminate other aspects of the method.

The method of reciprocal averages is defined by the transition formulas, where an individual's (row's) scale value (principal coordinate) is the mean of the scale values of the categories (columns) chosen by that individual, and the scale value of a category is the mean of the scale values of the individuals in that category. This renders more intuitive the "barycentric" nature of the transition formulas. The internal consistency of the scale values is maximized and each eigenvalue is a measure of the internal consistency of each scaling (i.e., dimension) induced on the rows and columns.

In the analysis of variance approach, the ratio of the sum of squares between rows (columns) is maximized and the ratio of the sum of squares within rows (columns) is minimized relative to the total sum of squares. The successive squared correlation ratios (the between sum of squares relative to the total sum of squares) are equivalent to the eigenvalues.

In the generalized canonical analysis approach, the sum of the squared correlations between the scaled individuals and scaled variable categories is maximized. This maximized value equals the sum of the eigenvalues and each eigenvalue is the canonical correlation between each successive joint scaling of the rows and columns.

It is also useful to contrast correspondence analysis with multidimensional unfolding, another approach for the joint display of a data matrix. Correspondence analysis displays the positions of the rows (or columns) of the data matrix *relative* to the set of rows (or columns) included in the analysis. This is a consequence of using profiles, rather than absolute frequencies. Multidimensional unfolding methods, however, *directly* approximate the entries in the data matrix, which are assumed to be row-to-column distances (dissimilarities). In this case, there is a direct interpretation of the graphical representation in terms of interpoint distances. If the data can be considered as row-to-column distances, multidimensional unfolding is an appropriate technique. If the data cannot be considered as such, correspondence analysis may be the more appropriate method for constructing joint representations.

DISCUSSION

Supplementary Points: Fitting External Information Into the Display

The transition formulas in equations 18 and 19 provide the means for fitting external information into the graphical display from a correspondence analysis. These "supplementary points" enrich interpretation of the display, in much the same way that regression procedures assist in the interpretation of multidimensional scaling solutions (Kruskal and Wish 1978; Schiffman, Reynolds, and Young 1981).

Suppose we have information about physical characteristics of the eight soft drinks in the preceding example and array these data in a characteristics by soft drinks matrix. It is possible to consider each row of the matrix as defining a point in the space of the row (individual) profiles of the individuals by soft drinks matrix. Through the use of transition formula 18 we can make a transition from columns (the soft drinks) to rows (the physical characteristics) to obtain point locations for each characteristic. Each physical characteristic profile then can be projected onto the plane defined by the first two principal axes to see which characteristics are associated with which soft drinks. If we had information on the individuals, such as demographic data, we could use transition formula 19 and go from rows to columns. In this case, each column of the individuals by demographics matrix defines a column profile in the same space as the profiles of the soft drinks across the individuals. The transition from rows to columns yields a set of points that can be displayed in the original space, thereby providing information on the demographic characteristics of the individuals.

The fitting of supplementary points also can serve as a validity check (Lebart, Morineau, and Warwick 1984, p. 163). Because a supplementary variable makes no contribution to the axis, its squared correlation (relative contribution) with each principal axis can be examined. High values indicate good fit into the previously defined

display and imply validation of the variables being investigated.

Handling Outliers

Outlier points plague correspondence analysis solutions. Occasionally, a row (or column) profile point is so "rare" (in profile) in its set of points that it has a major role in determining the higher order principal axes. This situation is easily discerned by examining the points' contributions to the axes. When a point has a very large absolute contribution and a large principal coordinate on a major principal axis, it can be considered an outlier.

Two such points are individuals 13 and 28 in the empirical example. These points consume nearly 34% of the inertia on the second principal axis, determining its orientation to a large degree. The solution lies in redefining these points as supplementary and performing the analysis again without them, permitting them no influence on the direction of the principal axes. Then the points can be fit *a posteriori* on the axes calculated for the remaining points with transition formula 18.

Inspection of the data matrix provides information about the nature of the "rarity" of an outlying point. Treating the point as supplementary allows more detailed study of the structure of the remaining points whose multivariate association is not as readily determined by inspection.

A Caveat

Correspondence analysis does have limitations. It is a multivariate *descriptive* statistical method and is not appropriate for hypothesis testing. Other approaches are better suited to searching for parsimonious models that can account for most of the variance in the data, such as weighted least squares (Grizzle, Starmer, and Koch 1969) and loglinear modeling (Bishop, Fienberg, and Holland 1975). Recently, van der Heijden and de Leeuw (1985) showed that, under certain conditions, correspondence analysis can be interpreted in terms of specific loglinear models. However, statistical tests for correspondence analysis are still being developed; earlier tests were shown either theoretically or through simulations to be unjustified (Lebart 1976). Nonetheless, correspondence analysis may be helpful in detecting models that merit further consideration by other methods.

As discussed before, an important caveat for interpreting correspondence analysis results is that the between-set distances cannot be interpreted. The joint display of coordinates shows the relationship between a point from one set and *all* the points of the other set, not between individual points from each set. (See Carroll, Green, and Schaffer 1986 for an alternative scaling of the coordinates that provides for comparability of all within-set and between-set distances.) When the correspondence analysis solution has more than two dimensions, proximity with one pair of axes may disappear when other pairs are plotted.

Correspondence analysis also suffers from the "curse

of dimensionality." There is no method for conclusively determining the appropriate number of and what combinations of dimensions to plot and inspect. As with other multivariate methods, the researcher must balance parsimony against interpretability in determining the number of dimensions to use.

Finally, it must be recognized that in many ways correspondence analysis is a subjective technique. Many different portrayals of a data set often are possible, leading to different analysis categories and solutions. By its flexibility, correspondence analysis can lead to greater insight into the phenomena being studied because it affords several different views of the same data set. Subjectivity of analysis is part of the price of this flexibility.

Implementation

A variety of computer programs are available for carrying out a correspondence analysis. The SPAD system of FORTRAN programs written by Lebart and Morineau (1982) for mainframe computer systems is particularly applicable to large data sets. A specialized version of this program is described by Lebart, Morineau, and Warwick (1984). Nishisato and Nishisato (1983) have prepared a program that performs correspondence analysis ("dual scaling") on the IBM PC. The program accepts as input up to six different types of data. Greenacre (1984) presents a simple program to do correspondence analysis using the high-level programming language GENSTAT. An extensive collection of computer programs for correspondence analysis and related techniques is provided by Jambu and Lebeaux (1983). If the researcher has access to a matrix subroutine that performs a singular value decomposition, he or she has the tools necessary to implement the method. For example, correspondence analysis is programmed easily with the MATRIX procedure in SAS (SAS Institute 1982).

Concluding Remarks

As we present it, correspondence analysis is a method of exploratory data analysis that (1) quantifies multivariate categorical data, (2) affords a graphical representation of the structure in the data, and (3) does not pose stringent measurement requirements. For many applications, its use is straightforward and unambiguous. When complex multivariate relationships are examined, correspondence analysis is limited only by the researcher's ingenuity in interpreting the derived spatial map. As a graphical method of data analysis, correspondence analysis is applied best as a multivariate descriptive statistical technique supplemental to other forms of analysis.

Correspondence analysis is very flexible. Not only is it flexible in terms of data requirements, but it also allows for the incorporation of marketing knowledge. In studying a product class, say, the researcher can set masses of brands equal to the market share or dollar sales of each, or perhaps to the percentage of consumers who use the product in the population. The technique of fitting

supplementary points in the display is an interesting and virtually limitless way to incorporate external information into the analysis. It is also useful as a check on data validity and as a tool for handling troublesome outliers. Though correspondence analysis has limitations, the most important being that between-set distances in the graphical display are not interpretable, its flexibility may render it more suitable than other methods for marketing research applications in many situations.

Categorical data are common products of marketing research. However, the analysis of such data often is hindered by the requirements and limitations of many familiar research tools. Correspondence analysis is a versatile and easily implemented analytical method that can do much to assist researchers in detecting and explaining relationships among complex marketing phenomena.

REFERENCES

- Belk, Russell, John Painter, and Richard Semenik (1981), "Preferred Solutions to the Energy Crisis as a Function of Causal Attributions," *Journal of Consumer Research*, 8 (December), 306-12.
- 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. (1973a), *L'Analyse des Données. Vol. I, La Taxinomie*. Paris: Dunod.
- et al. (1973b), *L'Analyse des Données. Vol. II, L'Analyse des Correspondances*. Paris: Dunod.
- Bishop, Yvonne M. M., S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Bock, R. Darrell (1960), "Methods and Applications of Optimal Scaling," Laboratory Report No. 25, L. L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill.
- Burt, C. (1950), "The Factorial Analysis of Qualitative Data," *British Journal of Psychology (Statistical Section)*, 3 (November), 166-85.
- Carroll, J. Douglas, Paul E. Green, and Catherine M. Schaffer (1986), "Interpoint Distance Comparisons in Correspondence Analysis," *Journal of Marketing Research*, 23 (August), 271-80.
- 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), 211-18.
- Fisher, Ronald A. (1940), "The Precision of Discriminant Functions," *Annals of Eugenics*, 10 (December), 422-9.
- Franke, George R. (1983), "Dual Scaling: A Model for Interpreting and Quantifying Categorical Data," in *Research Methods and Causal Modeling in Marketing*, W. R. Darden, K. B. Monroe, and W. R. Dillon, eds. Chicago: American Marketing Association, 111-4.
- (1985), "Evaluating Measures Through Data Quantification: Applying Dual Scaling to an Advertising Copytest," *Journal of Business Research*, 13 (February), 61-9.
- Gabriel, K. R. (1971), "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis," *Biometrika*, 58 (December), 453-67.
- (1981), "Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis," in *Interpreting Multivariate Data*, V. Barnett, ed. Chichester: John Wiley & Sons, Inc., 147-73.
- Gower, J. C. (1966), "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis," *Biometrika*, 53 (December), 325-38.
- Green, Paul E., with J. Douglas Carroll (1978), *Mathematical Tools for Applied Multivariate Analysis*. New York: Academic Press, Inc.
- , Vithala R. Rao, and Wayne S. DeSarbo (1978), "Incorporating Group-Level Similarity Judgments in Conjoint Analysis," *Journal of Consumer Research*, 5 (December), 187-93.
- Green, Robert T., Jean-Paul Leonardi, Jean-Louis Chandon, Isabella C. M. Cunningham, Bronis Verhage, and Alain Strazzeri (1983), "Societal Development and Family Purchasing Roles: A Cross-National Study," *Journal of Consumer Research*, 9 (March), 436-42.
- Greenacre, Michael J. (1978), "Some Objective Methods of Graphical Display of a Data Matrix" (English translation of 1978 doctoral thesis), Department of Statistics and Operations Research, University of South Africa.
- (1984), *Theory and Application of Correspondence Analysis*. London: Academic Press, Inc.
- Grizzle, J. E., C. F. Starmer, and G. G. Koch (1969), "Analysis of Categorical Data by Linear Models," *Biometrics*, 25 (September), 489-504.
- Guttman, Louis (1941), "The Quantification of a Class of Attributes: A Theory and Method of Scale Construction," in *Prediction of Personal Adjustment*, The Committee on Social Adjustment, ed. New York: Social Science Research Council, 319-48.
- Hayashi, C. (1950), "On the Quantification of Qualitative Data from the Mathematico-Statistical Point of View," *Annals of the Institute of Statistical Mathematics*, 2 (1), 35-47.
- (1952), "On the Prediction of Phenomena from Qualitative Data and the Quantification of Qualitative Data from the Mathematico-Statistical Point of View," *Annals of the Institute of Statistical Mathematics*, 3 (2), 69-98.
- (1954), "Multidimensional Quantification—with the Applications to Analysis of Social Phenomena," *Annals of the Institute of Statistical Mathematics*, 5 (2), 121-43.
- 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 Multivariate Method," *Applied Statistics*, 23 (3), 340-54.
- Hirschfeld, H. O. (1935), "A Connection Between Correlation and Contingency," *Proceedings of the Cambridge Philosophical Society*, 31 (October), 520-4.
- Holbrook, Morris B., William L. Moore, and Russell S. Winer (1982), "Constructing Joint Spaces from Pick-Any Data: A New Tool for Consumer Analysis," *Journal of Consumer Research*, 9 (June), 99-105.
- Horst, Paul (1935), "Measuring Complex Attitudes," *Journal of Social Psychology*, 6 (3), 369-74.
- Jambu, M. and M-O. Lebeaux (1983), *Cluster Analysis and Data Analysis*. Amsterdam: North Holland Publishing Company.

- Kruskal, Joseph B. and Myron Wish (1978), *Multidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-011. Beverly Hills, CA: Sage Publications, Inc.
- Lebart, Ludovic (1976), "The Significance of Eigenvalues Issued from Correspondence Analysis," *Proceedings in Computational Statistics (COMPSTAT)*. Vienna: Physica Verlag, 38-45.
- and Alain Morineau (1982), "SPAD: A System of FORTRAN Programs for Correspondence Analysis," *Journal of Marketing Research*, 19 (November), 608-9.
- , ———, and Kenneth M. Warwick (1984), *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: John Wiley & Sons, Inc.
- Levine, Joel H. (1979), "Joint-Space Analysis of 'Pick-Any' Data: Analysis of Choices from an Unconstrained Set of Alternatives," *Psychometrika*, 44 (March), 85-92.
- Marc, Marcel (1973), "Some Practical Uses of 'The Factorial Analysis of Correspondence,'" *European Research*, 1 (July), 2-8.
- McKeon, J. J. (1966), "Canonical Analysis: Some Relations Between Canonical Correlation, Factor Analysis, Discriminant Function Analysis and Scaling Theory," *Psychometrika*, Monograph No. 13.
- Nishisato, Shizuhiko (1980), *Analysis of Categorical Data: Dual Scaling and Its Applications*. Toronto: University of Toronto Press.
- and Ira Nishisato (1983), *An Introduction to Dual Scaling*, 1st ed. Islington, Ontario: MicroStats.
- and Wen-Jenn Sheu (1984), "A Note on Dual Scaling of Successive Categories Data," *Psychometrika*, 49 (December), 493-500.
- Perreault, William D., Jr., and Forrest W. Young (1980), "Alternating Least Squares Optimal Scaling: Analysis of Non-metric Data in Marketing Research," *Journal of Marketing Research*, 17 (February), 1-13.
- Richardson, M. and G. F. Kuder (1933), "Making a Rating Scale That Measures," *Personnel Journal*, 12 (June), 36-40.
- SAS Institute (1982), *SAS User's Guide: Statistics*. Cary, NC: SAS Institute Inc.
- Schiffman, Susan S., M. Lance Reynolds, and Forrest W. Young (1981), *Introduction to Multidimensional Scaling*. New York: Academic Press, Inc.
- Seber, G. A. F. (1984), *Multivariate Observations*. New York: John Wiley & Sons, Inc.
- Tenenhaus, Michel and Forrest W. Young (1985), "An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data," *Psychometrika*, 50 (March), 91-119.
- Tukey, John W. (1977), *Exploratory Data Analysis*. Reading, MA: Addison-Wesley Publishing Company, Inc.
- van der Heijden, Peter G. M. and Jan de Leeuw (1985), "Correspondence Analysis Used Complementary to Loglinear Analysis," *Psychometrika*, 50 (December), 429-47.
- van Rijkevorsel, Jan and Jan de Leeuw (1978), "An Outline to HOMALS-1," Department of Data Theory, Faculty of Social Sciences, University of Leiden, The Netherlands.

JMR REPRINT POLICY AND PROCEDURE

All reprints are sold according to the following price schedule, in minimum orders of 50 copies.

50 copies with covers	\$ 75.00
100 copies with covers	125.00
(multiples of 100 may be ordered)	

There is NO RETURN and NO EXCHANGE on reprints.

Under the "fair use" provision of the new copyright law taking effect January 1978, anyone may make a photocopy of a copyrighted article for his or her own use without seeking permission. Also, a single copy reprint or an order of less than 50 copies may be obtained from University Microfilms International, 300 N. Zeeb Road, Ann Arbor, MI 48106. Articles are priced *prepaid* at \$6.00 plus \$1.00 for each additional copy of the same article. Complete issues are obtainable at 10¢ per page, minimum order \$10.00.

To obtain permission to reproduce one's own reprints in quantity, please contact the Permissions Department, American Marketing Association, 250 S. Wacker Drive, Chicago, IL 60606.