



The Geometric Interpretation of Correspondence Analysis
Author(s): Michael Greenacre and Trevor Hastie
Source: *Journal of the American Statistical Association*, Vol. 82, No. 398 (Jun., 1987), pp. 437-447
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: https://www.jstor.org/stable/2289445
Accessed: 26-01-2020 22:24 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms



Taylor & Francis, Ltd., American Statistical Association are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

The Geometric Interpretation of Correspondence Analysis

MICHAEL GREENACRE and TREVOR HASTIE*

Correspondence analysis is an exploratory multivariate technique that converts a data matrix into a particular type of graphical display in which the rows and columns are depicted as points. The method has a long and varied history and has appeared in different forms in the psychometric and ecological literature, among others. In this article we review the geometry of correspondence analysis and its geometric interpretation. We also discuss various extensions of correspondence analysis to multivariate categorical data (multiple correspondence analysis) and a variety of other data types.

KEY WORDS: Graphical display; Singular value decomposition; Principal components; Contingency tables.

1. INTRODUCTION

Correspondence analysis is an exploratory multivariate technique that converts a matrix of nonnegative data into a particular type of graphical display in which the rows and columns of the matrix are depicted as points. It is a method that, algebraically at least, has been known for more than 50 years, the first mathematical account being by Hirschfeld (1935). Since then the same algebraic and numerical procedure has been rediscovered in different contexts, notably in ecology (reciprocal averaging) and psychology (dual scaling). The method was rediscovered in France in the early 1960s and has been used extensively in that country as a method of graphical data analysis (Benzécri 1973; Lebart, Morineau, and Tabard 1977). Detailed descriptions of the evolution of the various forms of correspondence analysis can be found in Benzécri (1977), Nash (1978), Nishisato (1980, sec. 1.2), Greenacre (1984, sec. 1.3), and Tenenhaus and Young (1985). A bibliography by Nishisato (1986) lists more than 1,000 references in the period 1975–1986 that are directly or indirectly relevant to the topic of correspondence analysis.

In this article we focus specifically on the geometry of correspondence analysis and its geometric interpretation. This will be seen to be a variant of principal components analysis, but tailored to categorical rather than continuous data. The most basic form of correspondence analysis is its application to a two-way contingency table, which we shall call *simple correspondence analysis*. The geometry of this leading case, discussed in Section 2, provides the basic rules of interpretation. All other forms of correspondence analysis are the application of the same algorithm to other types of data matrices, with a consequent adaptation of the interpretation. In Section 3 we treat the case of a multiway contingency table that is coded as a matrix of indicator (or "dummy") variables. The correspondence analysis of this indicator matrix is known as *multiple correspondence analysis* or *homogeneity analysis* (Gifi 1981). In Section 4 we show how other data types may be reexpressed in order to be geometrically interpretable using correspondence analysis.

2. GEOMETRY OF SIMPLE CORRESPONDENCE ANALYSIS

2.1 Basic Geometry

The correspondence analysis of a two-way contingency table $\mathbf{N}(I \times J)$ provides a leading case of the method's geometry and interpretation. We shall denote the row and column totals of $\mathbf{N}$ by n_{i+} $(i = 1 \cdots I)$ and n_{+j} $(j = 1 \cdots J)$, respectively, and the overall total simply by n. We shall use the 5×3 contingency table of Table 1 throughout this section as an illustration. It is a cross-tabulation of 312 people, all identified as readers of a particular newspaper in a readership survey, according to five educational groups and three categories of readership of the newspaper. This example has the advantage that its geometry is precisely three-dimensional so that we can visually depict the technique's concepts and mechanism without abstraction.

We suppose initially that we are interested in comparing the rows of the table. The proportions of readership types in each education group are given in parentheses in Table 1. Each of these vectors is known as a row *profile*, denoted by $\mathbf{a}_i = [n_{i1}/n_{i+} \cdots n_{iJ}/n_{i+}]^T$, for example, $\mathbf{a}_2 = [.214\ .548\ .238]^T$. Each row in this example can be represented by its profile as a point vector in three-dimensional Euclidean space. The fact that there is a linear dependency among the coordinates of the profile vectors (they each sum to 1) means geometrically that the five points are contained exactly in a two-dimensional regular simplex, namely, the triangle with vertices at the unit points along each of the three coordinate axes. The points may be plotted directly in this triangle in what is generally known as *triangular* (or *barycentric*) *coordinates*.

Apart from serving as the vertices of the triangle bounding the row points, the unit points may be considered as the most extreme, or most polarized, types of row profile observable. For example, the unit point $\mathbf{e}_1 = [1\ 0\ 0]^T$ represents a row profile in which the total frequency is concentrated into the "casual" reading category. In this sense the jth column of the data matrix is represented by the vertex $\mathbf{e}_j$ of the triangle.

For reasons that will soon become apparent, we wish to define distances between profiles not by the usual Eu-

* Michael Greenacre is Professor, Department of Statistics, University of South Africa, Pretoria 0001, South Africa. Trevor Hastie is a member of the Technical Staff, Statistics and Data Analysis, Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974. This article was inspired in part by a series of seminars on correspondence analysis, organized by Perci Diaconis, at Stanford University in 1984. The authors thank the participants of the seminars for their contributions and acknowledge the constructive suggestions of the referees and the editor of the Statistical Graphics special section.

© 1987 American Statistical Association

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

Table 1. Contingency Table From Readership Survey

Education group	Category of readership C1	C2	C3	Totals
E1	5 (.357)	7 (.500)	2 (.143)	14
E2	18 (.214)	46 (.548)	20 (.238)	84
E3	19 (.218)	29 (.333)	39 (.448)	87
E4	12 (.119)	40 (.396)	49 (.485)	101
E5	3 (.115)	7 (.269)	16 (.615)	26
Totals	57 (.183)	129 (.413)	126 (.404)	312

NOTE: Education groups—E1, some primary; E2, primary completed; E3, some secondary; E4, secondary completed; E5, some tertiary. Readership categories—C1, glance; C2, fairly thorough; C3, very thorough. Row profiles are given in parentheses (i.e., row frequencies as a proportion of row totals).

clidean metric but rather by the following weighted Euclidean metric, called the *chi-squared metric*:

$$d_c(\mathbf{a}_i, \mathbf{a}_{i'}) = (\mathbf{a}_i - \mathbf{a}_{i'})^T \mathbf{D}_c^{-1} (\mathbf{a}_i - \mathbf{a}_{i'}) = \sum_{j=1}^{J} \frac{(n_{ij}/n_{i+} - n_{i'j}/n_{i'+})^2}{(n_{+j}/n)}, \quad (2.1)$$

where $\mathbf{D}_c$ is the diagonal matrix of elements $c_j = n_{+j}/n$ ($j = 1 \cdots J$). The vector $\mathbf{c} = [c_1 \cdots c_J]^T$, in this case the proportions of all respondents in the readership categories, is called the *average row profile*. To observe chi-squared distances between points, we can rescale the row profiles as follows: $\tilde{\mathbf{a}}_i = \mathbf{D}_c^{-1/2}\mathbf{a}_i$, so chi-squared distances are transformed to ordinary Euclidean distances, since $(\mathbf{a}_i - \mathbf{a}_{i'})^T\mathbf{D}_c^{-1}(\mathbf{a}_i - \mathbf{a}_{i'}) = (\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_{i'})^T(\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_{i'})$. Equivalently, we can stretch the coordinate axes in proportion to the values $c_j^{-1/2}$ so that each axis has a different scale. In our example this results in a stretched triangle, no longer equilateral, which bounds the cloud of points (Figs. 1 and 2). In this metric two vertices $\mathbf{e}_j$ and $\mathbf{e}_{j'}$ are at a squared distance apart of $1/c_j + 1/c_{j'}$, so a vertex corresponding to a column with low c_j is pushed away from the other vertices. The profile points are accordingly stretched out in the direction of this vertex.

Figures 1 and 2 also show the position of the average row profile $\mathbf{c} = [.183\ .413\ .404]^T$. This vector is also called the row *centroid* because it is the weighted average of the row profiles: $\mathbf{c} = \sum_{i=1}^{I} r_i\mathbf{a}_i$, where $r_i = n_{i+}/n$. Each row profile receives a weight proportional to the respective row total in the original data. The relative weights r_i are called the *row masses*.

The usual chi-squared statistic χ^2 that tests the null hypothesis of row–column independence can now be reexpressed as follows:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{(n_{i+}n_{+j}/n)} = n\sum_{i=1}^{I} r_i(\mathbf{a}_i - \mathbf{c})^T\mathbf{D}_c^{-1}(\mathbf{a}_i - \mathbf{c}). \quad (2.2)$$

In other words, χ^2/n can be defined geometrically as the weighted average of the squared (chi-squared) distances of the row profiles to their centroid. The quantity χ^2/n crops up frequently in correspondence analysis and is called the *total inertia* of the data matrix.

The null hypothesis of row–column independence, $n_{ij} = n_{i+}n_{+j}/n$ ($i = 1 \cdots I, j = 1 \cdots J$), is equivalent to the hypothesis of homogeneity of the rows: $n_{1j}/n_{1+} = n_{2j}/n_{2+} = \cdots = n_{Ij}/n_{I+}$ ($j = 1 \cdots J$). Each row of $\mathbf{N}$ may be viewed as the realization of a multinomial distribution, conditional on the respective row total. Under the homogeneity assumption, this distribution is common to all of the rows up to a rescaling by the row totals. This scaled multinomial distribution is completely described by a set of probabilities whose maximum likelihood estimates are the elements c_j of the row centroid. Thus a significant χ^2 can be interpreted geometrically as a significant deviation of the row profiles from their centroid, that is, from the homogeneity hypothesis. Figures 1 and 2 show the deviations

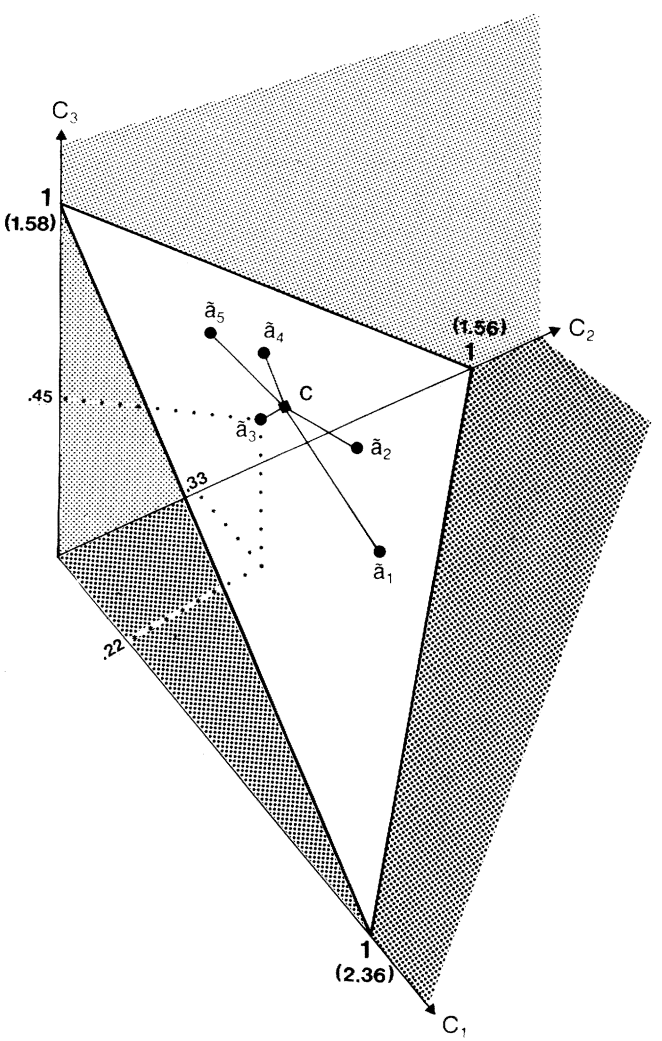


Figure 1. A Geometric View of the Profiles of the Five Education Groups in the "Stretched" Coordinate System. Each of the three axes has a different scale, with the unit points at $1/c_1^{1/2} = 2.36$, $1/c_2^{1/2} = 1.56$, and $1/c_3^{1/2} = 1.58$, respectively. The transformed row profiles $\tilde{\mathbf{a}}_i$ ($i = 1, \ldots, 5$) all lie in the simplex joining the unit points. The lines joining the row profiles to the centroid profile $\mathbf{c}$ show the deviations from the "independence" point.

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

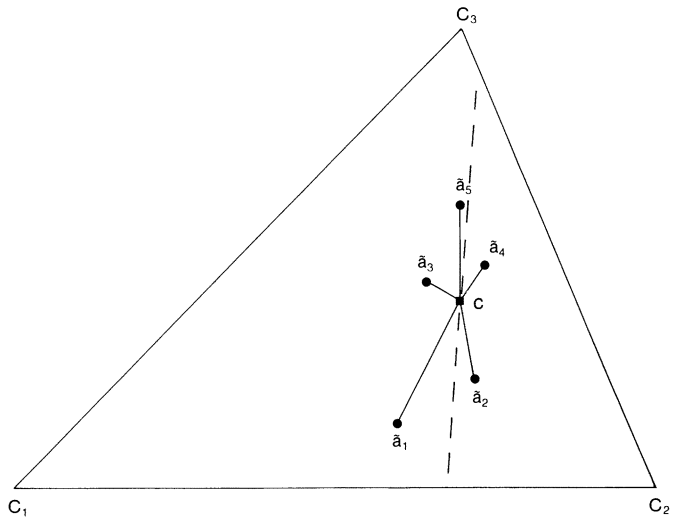


Figure 2. The Triangular Simplex of Figure 1 Laid Flat, Showing the Profiles in "Stretched" Triangular Coordinates (i.e., the chi-squared metric). Each profile receives a weight proportional to its respective row total. The weighted average of the squared distances between the row profiles and their centroid **c** *is equal to the total inertia χ^2/n, where χ^2 is the chi-squared statistic that tests row–column independence on Table 1. The best-fitting principal axis is indicated by the dashed line. This axis minimizes the weighted sum of squared distances from the points to the axis.*

of each row profile from the centroid. The χ^2 statistic for these data is 26.0 (with 8 df, $p = .001$). Apart from leading to this geometric interpretation of χ^2, a further justification for the chi-squared metric is that it is the Mahalanobis metric between row profiles based on their estimated covariance matrix under the homogeneity assumption (Greenacre 1984, sec. 4.4).

Correspondence analysis provides a low-dimensional explanation for the lack of homogeneity in the row profiles or, equivalently, for the interaction or dependence between the rows and columns of the contingency table. For illustration we first concentrate on a one-dimensional explanation, so the problem may now be posed of finding a line that in some sense best fits the cloud of profile points. The problem is analogous to finding the largest principal component of a set of I observations on J variables, with simple generalizations to accommodate the weighting of the observations and the weighted metric. A natural choice of fit is a weighted least squares one using the row masses r_i as weights, as this will then provide a decomposition of the chi-squared statistic into components along the line and orthogonal to it. With this choice of objective function it is easy to show that the best line passes through the centroid **c**, since **c** is the best zero-dimensional (point) summary. If the origin of the display is transferred to **c**, then the best-fitting line can be shown to be the principal eigenvector of the nonsymmetric matrix

$$\sum_{i=1}^{I} r_i(\mathbf{a}_i - \mathbf{c})(\mathbf{a}_i - \mathbf{c})^T\mathbf{D}_\mathbf{c}^{-1} = (\mathbf{A} - \mathbf{1}\mathbf{c}^T)^T\mathbf{D}_\mathbf{r}(\mathbf{A} - \mathbf{1}\mathbf{c}^T)\mathbf{D}_\mathbf{c}^{-1}, \quad (2.3)$$

where $\mathbf{D}_\mathbf{r}$ is the diagonal matrix of row masses and **A** is the matrix with the row profiles as rows ($\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2 \cdots \mathbf{a}_I]^T$) (for a proof see, e.g., Greenacre 1984, sec. 4.1). The eigenvector, denoted by $\mathbf{v}_1$, defines the first principal axis of the row profiles and is normalized in the metric $\mathbf{D}_\mathbf{c}^{-1}$: $\mathbf{v}_1^T\mathbf{D}_\mathbf{c}^{-1}\mathbf{v}_1 = 1$. Figure 2 shows the best-fitting line in our example. Remember that this display has been stretched, so chi-squared distances are represented as Euclidean distances.

The trace of matrix (2.3) is equal to χ^2/n, so its set of eigenvalues $\lambda_1, \lambda_2, \ldots$, or *principal inertias*, is a decomposition of the total inertia. There is an equivalent decomposition of χ^2 into components $n\lambda_1, n\lambda_2, \ldots$. In the example the total inertia is .0833, and the first principal axis corresponds to an eigenvalue $\lambda_1 = .0704$, which accounts for 84.5% of the total inertia.

As in principal components analysis, the eigenvectors corresponding to the two largest eigenvalues define the plane closest to the row profiles. In general, there are I row profiles lying exactly in a $(J - 1)$-dimensional simplex with J vertices. The K eigenvectors corresponding to the K largest eigenvalues of (2.3) define the K-dimensional affine subspace closest to the profiles in the sense of weighted least squares. The exact dimensionality of a set of profiles—that is, the rank of matrix (2.3)—is at most $\min\{I, J\} - 1$.

2.2 Interpretation

In our example the row profiles are contained exactly in a plane. Hence their display with respect to the first two principal axes is an exact representation of the profiles and merely a rotation of the display of Figure 2. The geometrical issues are easily illustrated in such a simple example; in general the data lie in higher dimensions, and all that we see graphically are the projections. This display is given in Figure 3 as well as the projections of the profiles onto the first principal axis. These projections may be interpreted as approximations to the profiles' actual two-dimensional positions. Similarly, in the case of larger data matrices a two-dimensional display, say, would approximate the high-dimensional configuration of points.

To assist in the interpretation of the axis, the three vertices of the triangular coordinate system may also be projected onto it. This is analogous to factor analysis, where the variables are correlated with a factor to get factor loadings that are used in naming the factor. This is achieved by projecting the unit coordinate vectors onto the axis, which is precisely what we do here. Interpretation consists of looking for groupings and contrasts in the configuration of projected vertices and in the configuration of projected profiles. Thus the proximity of the two points representing columns C1 and C2 compared with their distance from C3 indicates that the axis reflects a contrast between the third readership category ("very thorough" reading) and the first two categories. In other words, the heterogeneity within the contingency table is concentrated into the contrast between column C3 and the pair of relatively homogeneous columns C1 and C2. Strictly speaking, there is no distance interpretation implied between the positions of the projected vertices, since the vertices are always fixed in space. It is merely what lies to a greater or lesser extent

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

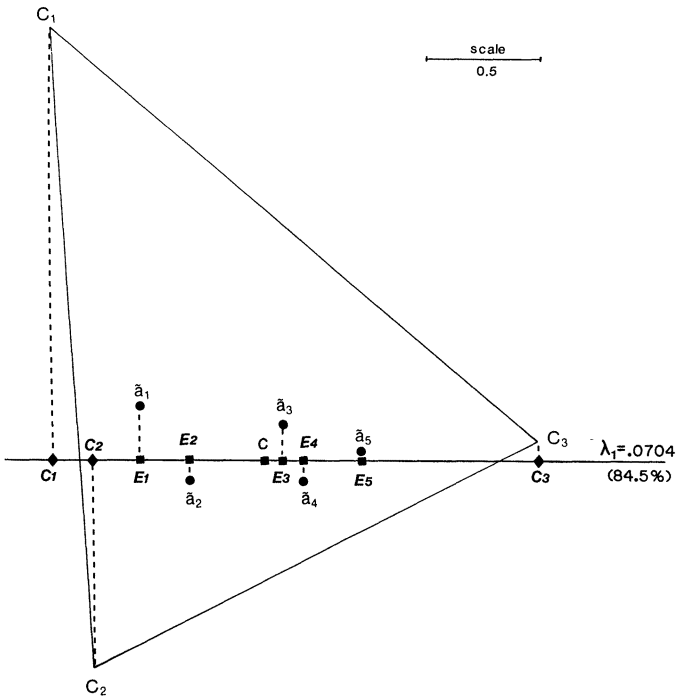


Figure 3. The Best One-Dimensional Summary of the Row Profiles in the Stretched Coordinate System, Accounting for 84.5% of the Inertia (or equivalently, of χ^2). The profile points are connected to their one-dimensional estimates, which are the projections onto this line. The "corner" profiles, or vertices, are also projected onto the principal axis.

on one side of the axis compared with what lies on the other side that provides the axial interpretation. Back in the original space, the axis shows us the directions along which the profiles vary the most, and the positions of the column "corners" let us add a descriptive label to this direction of spread.

On the other hand, the row points along the axis do have a specific metric interpretation. Distances between projected row points are approximate chi-squared distances between the row profiles, and these may be compared among one another. For example, there appears to be more difference between educational groups E1 and E2 than between E3 and E4.

Finally, there is a geometric relationship between the row points and the column points along the axis. In the original space (Fig. 1) the vertices are the basis vectors of the space and the profile vectors consist of nonnegative elements that sum to 1. Thus each row point is a weighted average of the vertices, where the weights are the proportions of each readership category in the respective education group. We can deduce that education groups on the left side of the axis, particularly group E1, have relatively high proportions of the readership categories C1 and C2, whereas those on the right side, particularly E5, have a relatively high proportion of C3. The other education groups are strung out along this axis between these two extremes, with the average row profile at the origin of the new display. The positions of the row profiles along the principal axis reflect their differences on the specific feature characterized by the projected column vertices.

As in principal component analysis, the first principal axis may be regarded as optimal in two senses that are made equivalent thanks to the Pythagorean theorem. On the one hand, it is the best-fitting line in a weighted least squares sense (see Pearson 1901). On the other hand, it is inertia maximizing, which in the present case of a contingency table is the same as variance maximizing (see Hotelling 1933). In our example the projections of the vertices can be considered an optimal quantification, or scaling, of the three readership categories. Each education group is positioned on this scale according to the average value of its members—that is, the projected row profile. The variance of these averages, weighted by the respective group sample sizes, is maximized by the scaling provided by the first principal axis.

Geometrically, the principal inertia is the weighted average of squared (chi-squared) distances from the centroid to the projections of the row profiles on the respective principal axis. It is an absolute measure of dispersion of the row profiles in the direction of this axis. Its maximum value is 1, when all projected row profiles coincide with projected column vertices. The "significance" of a principal axis may be judged in two different ways. First, in the example the axis is successful in that it recovers a meaningful ordering of the education groups and readership groups. Second, if the data do arise from multinomial sampling, the null hypothesis of random dispersion along the first axis may be tested using asymptotic results by O'Neill (1978, 1980), summarized by Greenacre (1984, sec. 8.1).

Each principal inertia can be further decomposed into components due to each row profile. The study of these components, or *contributions to inertia*, is another important feature of the geometric interpretation. The rows that contribute highly to a principal axis have, in effect, largely determined the orientation and thus the identity of the corresponding principal axis. Also worth studying in conjunction with the display are the cosines of the angles between the row profiles' deviation vectors from the centroid (e.g., shown in Fig. 2) and the principal axis. These permit a description of how closely each profile vector lines up, or "correlates," with a principal axis; thus they measure how well the display approximates the profiles' true positions (Greenacre 1984, sec. 3.3; Lebart, Morineau, and Warwick 1984, pp. 46–49).

2.3 The Dual Problem

The methodology just described may be applied in an equivalent, symmetric fashion to the columns of the contingency table $\mathbf{N}$—that is, by repeating the preceding on the transposed table $\mathbf{N}^T$ ($J \times I$). We now look for the principal axes of the column profiles, weighted by masses that are the elements of $\mathbf{c}$, in a space with a chi-squared metric defined by the diagonal matrix $\mathbf{D}_\mathbf{r}^{-1}$. Thus the elements of $\mathbf{r}$ and $\mathbf{c}$ play dual roles, weighting the profiles on one hand and rescaling the dimensions on the other.

There is no need to recompute the dual solution, since it may be obtained from the first problem (Greenacre 1984, sec. 4.1). The total inertia and its decomposition into principal inertias are exactly the same in the two prob-

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

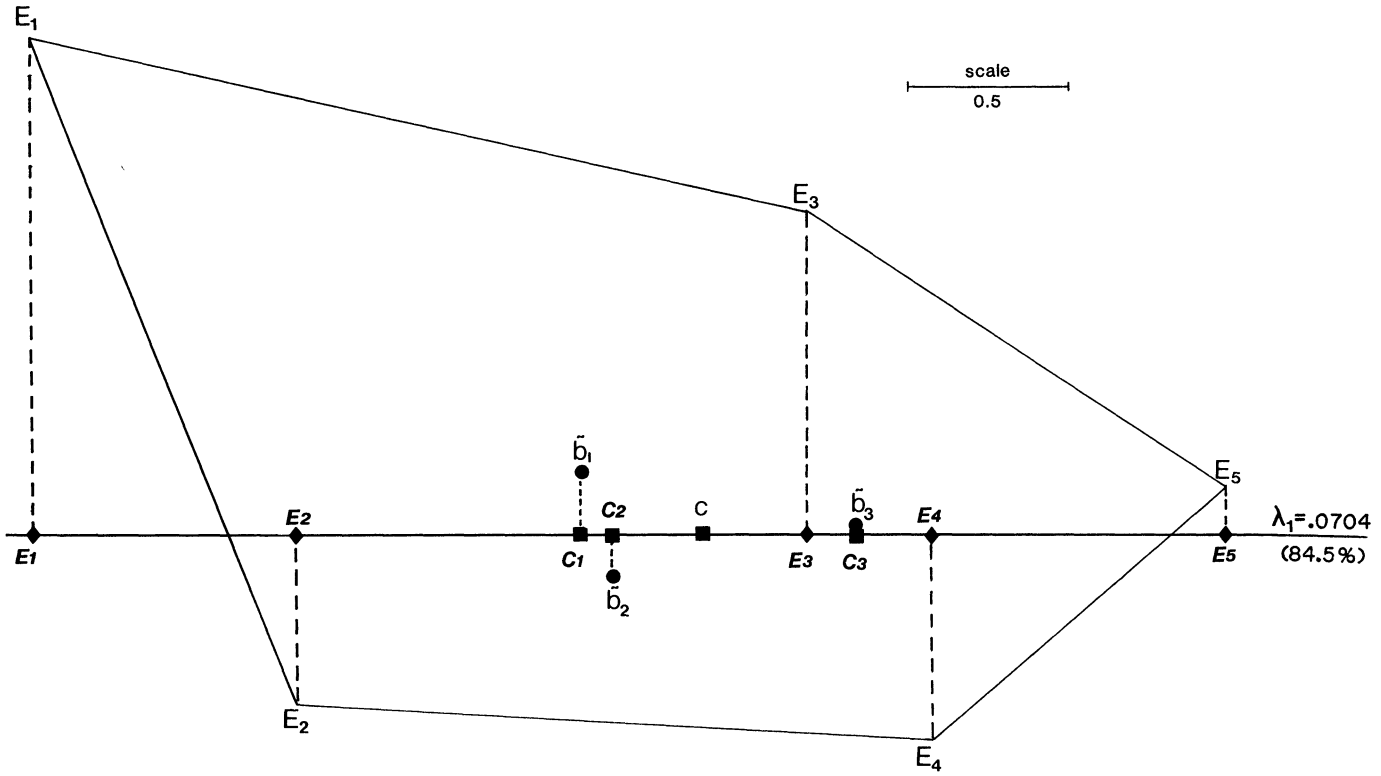


Figure 4. The Display Dual to Figure 3 of the Stretched Column Profiles $\tilde{\mathbf{b}}_j$, Where the Row Points Are the Projections of the Vertices of the Five-Cornered Simplex. The first principal axis is again the horizontal axis, and the projections of the profiles and vertices onto the axis are shown. Again, 84.5% of the total inertia is accounted for by this axis. Notice that the positions of the projected column profiles are a uniform contraction of the positions of the column vertices in Figure 3.

lems. In each problem the projections of profiles onto their kth principal axis can be obtained from the projections of their respective vertices in the dual problem, rescaled by a factor equal to $\lambda_k^{1/2}$, the square root of the kth common principal inertia. A similar duality does not exist in principal component analysis, unless the rows and columns of the data matrix sum to 0.

Figure 4 shows the dual solution of our illustrative example, where the chi-squared distances between the column profiles are displayed exactly and the five vertices representing the row points are projected onto the display. (This differs slightly from the previous situation in which the vertices were also displayed exactly in two-dimensional space. Here the plane cuts through the four-dimensional simplex defined by the five vertices.)

This close relationship between the two problems prompts most users of correspondence analysis to overlay the respective plots of the row and column profiles in a joint display. Figure 5 represents the joint display for our illustrative example. Here we have plotted the first two principal axes (and only two for this problem). Thus the projection of the points onto the λ_1 axis is the joint display for the one-dimensional solutions in Figures 3 and 4. In this display the interrow and intercolumn distances may be interpreted as approximate chi-squared distances, but row-to-column distances are meaningless. A practical advantage here is that the dispersions of row and column profiles are more or less the same, whereas in the plot of profiles and vertices, the profiles are often a very tight bunch of points in the display, needing magnification to be able to interpret their relative positions.

2.4 Supplementary Points

Once the principal axes of a cloud of profiles have been established, it is possible to display additional points that are defined in the profile space. Such points may be projected onto individual principal axes or onto any subspace spanned by principal axes. The ability to display such supplementary points is useful for enhancing the interpretation of the principal axes and the patterns observed in the displayed points. We have already illustrated this idea by projecting the unit profiles, or vertices, onto a principal axis (Figs. 3 and 4).

As a further example, we might have reason to subdivide education group E2 into two subgroups—say, private schools with frequencies E2a : [8 15 5] and public schools with frequencies E2b : [10 31 15]. Their profiles may be added to the display computed from the original data and projected onto the first principal axis (Fig. 5). Alternatively, one could combine the frequencies in an observed cluster in order to plot a point that represents the cluster.

Another use of supplementary points by Greenacre (1984, sec. 8.1) is reminiscent of bootstrapping. Replicated sets of row frequencies may be randomly generated by multinomial sampling. The new profiles can be displayed as supplementary points in order to gauge the sampling variability of the profile points (Fig. 5).

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

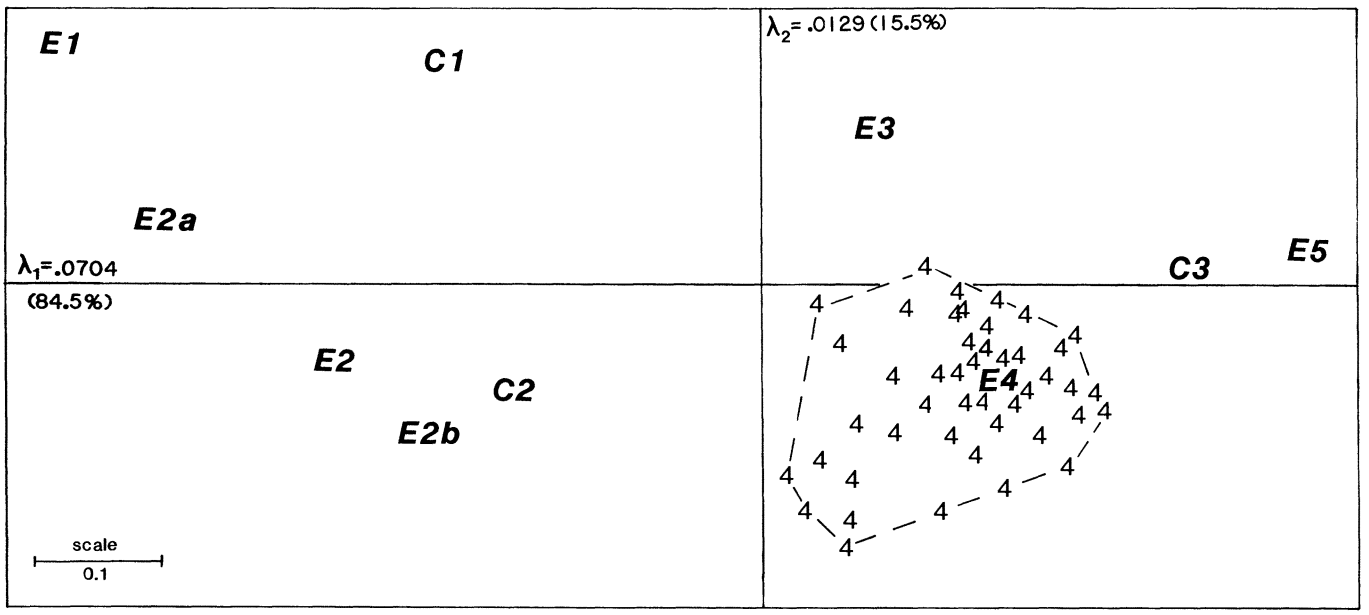


Figure 5. Joint Display of the Row and Column Profiles, Showing Examples of Supplementary Points. Points E2a and E2b are subgroups of E2. Their profiles are plotted with respect to the existing principal axes. The points labeled "4" are replicated profiles of group E4, generated by multinomial sampling and displayed as supplementary points.

2.5 Relationship to the Singular Value Decomposition and the Biplot

The row and column coordinates with respect to principal axes may be obtained from the singular value decomposition (SVD) of the double centered and standardized matrix:

$$\mathbf{D}_r^{-1/2}[(1/n)\mathbf{N} - \mathbf{r}\mathbf{c}^T]\mathbf{D}_c^{-1/2} = \mathbf{X}\mathbf{D}_\alpha\mathbf{Y}^T, \qquad (2.4)$$

where $\mathbf{X}^T\mathbf{X} = \mathbf{Y}^T\mathbf{Y} = \mathbf{I}$. The singular values are the square roots of the principal inertias: $\mathbf{D}_\alpha = \mathbf{D}_\lambda^{1/2}$. The principal axes of the row and column profiles are the column vectors of $\mathbf{D}_r^{1/2}\mathbf{Y}$ and $\mathbf{D}_c^{1/2}\mathbf{X}$, respectively. The K-dimensional coordinates of profile points and vertex points in the dual problems are the rows of the first K columns of the following matrices:

Row problem
row profiles, $\mathbf{D}_r^{-1/2}\mathbf{X}\mathbf{D}_\alpha$; column vertices, $\mathbf{D}_c^{-1/2}\mathbf{Y}$

Column problem
column profiles, $\mathbf{D}_c^{-1/2}\mathbf{Y}\mathbf{D}_\alpha$; row vertices, $\mathbf{D}_r^{-1/2}\mathbf{X}$.

These results show that correspondence analysis is not a biplot (Gabriel 1971, 1981) of the matrix of standardized residuals. A biplot would typically plot the rows of the first K columns of $\mathbf{X}\mathbf{D}_\alpha^{1/2}$ and $\mathbf{Y}\mathbf{D}_\alpha^{1/2}$ (or, e.g., $\mathbf{X}$ and $\mathbf{Y}\mathbf{D}_\alpha$) in a joint display. The geometric interpretation of this biplot is that the scalar product between the ith row point and jth column point (with respect to the origin of the display) is a least squares approximation to the (i, j)th standardized residual.

3. MULTIPLE CORRESPONDENCE ANALYSIS

Multiple correspondence analysis is concerned with displaying the categories of more than two discrete variables. This is achieved by defining indicator variables (or "dummy variables") for each category and reexpressing the data in the form of a cases-by-variables indicator matrix. Simple correspondence analysis of a cross-table of two discrete variables is not a natural special case of multiple correspondence analysis. As a stepping-stone to multiple correspondence analysis, we explain how it would apply to the special case of two discrete variables and how this compares with the simple correspondence analysis of the same data.

3.1 The Bivariate Case

The indicator matrix $\mathbf{Z}$ corresponding to the bivariate example in Table 1 is of order 312×8. The columns refer to the eight categories of the two discrete variables, education and readership, and each row refers to a respondent in the survey. The data in $\mathbf{Z}$ are zeros except for ones that indicate to which categories the cases belong. For example, there are five cases in category E1 of education and C1 of readership; hence there are five rows of $\mathbf{Z}$ that have elements [1 0 0 0 0 1 0 0]. Apart from their case identifications in the survey, each row can be labeled by its respective pair of categories. The five rows that we have just described can thus each be labeled "11." Clearly there are only 15 different types of row in $\mathbf{Z}$, corresponding to the 15 cells of the contingency table. If $\mathbf{Z}$ is partitioned as $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2]$ in terms of the two sets of categories, then the contingency table that cross-tabulates the two variables is $\mathbf{N} = \mathbf{Z}_1^T\mathbf{Z}_2$.

Multiple correspondence analysis is essentially the application of the same correspondence analysis algorithm described in Section 2 to the indicator matrix $\mathbf{Z}$, resulting in the display of the rows (cases) and the columns (eight

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

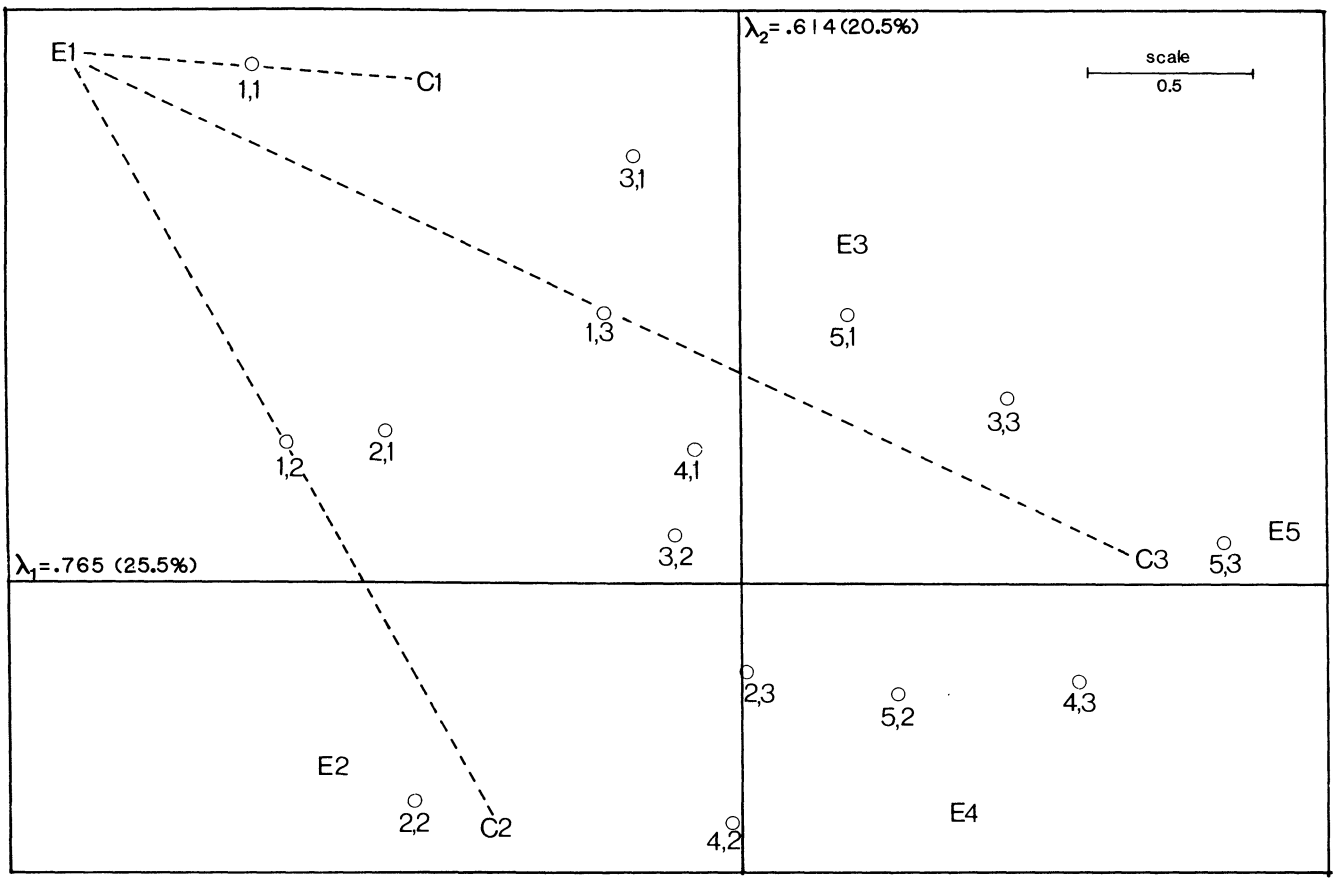


Figure 6. Multiple Correspondence Analysis of the Bivariate Indicator Matrix Derived From the Contingency Table of Table 1. The row points lie in 15 distinct positions, labeled according to their two categories in each case. Thus all rows characterized by the first categories of the two variables lie halfway between category points E1 and C1 at the position labeled "1, 1." The percentages of inertia are artificially low and may be adjusted to yield the same percentages (84.5% and 15.5%, respectively) of the equivalent simple correspondence analysis of Figure 5 (Benzécri 1979; Greenacre 1986).

categories). The two-dimensional principal-axes display of the row profiles and the column vertices is given in Figure 6. Applying the same rules of interpretation of simple correspondence analysis to this situation leads us to the following geometric description of this display.

First, the eight vertices of the simplex within which the row profiles lie are projected onto the plane in the positions indicated. Contrary to simple correspondence analysis, the vertices, or unit points, are now artificial extreme responses that cannot be achieved in the data set. The subset of five points for education and the subset of three points for reading level each has the same centroid, namely, the origin of the display. Each category point is again weighted by the marginal frequency of the respective category.

Second, the row profiles are now of a very special form. Each row of $\mathbf{Z}$ consists of zeros, except for two ones in the respective category positions, so each row profile has values of $\frac{1}{2}$ in these same positions and 0 elsewhere. In this situation the barycentric relationship implies that each row point will lie midway between its two respective category points. For example, the five points labeled "11" fall midway between projected vertices E1 and C1, the seven points labeled "12" fall midway between E1 and C2, and so on for each of the 15 clumps of row points. The principal axes are the axes of maximum variance of the row points, that is, of the 15 clumps of points, weighted by the size of the respective clumps.

The positions of the category points may be recovered from the analysis of the contingency table $\mathbf{N}$. In fact, the positions of the education and readership category points are exactly those of the row and column vertices in Figures 4 and 3, respectively [for a proof of this result, see Greenacre (1984, pp. 130–131) or Lebart et al. (1984, pp. 86–87)]. Carroll, Green and Schaffer (1986) discussed this equivalence at length and interpreted the distances between vertices.

There is also a direct relationship between the principal inertias of the two analyses: If λ_k is the kth largest principal inertia in the analysis of $\mathbf{N}$, then $\frac{1}{2}(1 + \lambda_k^{1/2})$ is the kth largest principal inertia in the analysis of $\mathbf{Z}$. This does not fully account for all of the eigenvalues in the analysis of the 312 × 8 indicator matrix, but the remaining eigenvalues and their associated eigenvectors are irrelevant to the geometric interpretation. These remaining eigenvalues that can effectively be ignored are less than or equal to $\frac{1}{2}$. The presence of these redundant eigenvalues is to lower the percentages of inertia (as seen in Fig. 6). We know that the first principal axis accounts for 84.5% of the total inertia of $\mathbf{N}$, whereas only 25.5% is accounted for by the

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

first principal axis of $\mathbf{Z}$. The latter percentage is a highly pessimistic measure of explained variance, and more realistic alternatives were given by Benzécri (1979) and Greenacre (1986).

3.2 The Multivariate Case

The coding of categorical data in an indicator matrix provides a natural extension to more than two variables. Suppose there are Q categorical variables so that the indicator matrix is of the form $\mathbf{Z}$ $(I \times J) = [\mathbf{Z}_1 \cdots \mathbf{Z}_Q]$. Suppose that the qth variable has J_q categories and hence that $\mathbf{Z}_q$ is $I \times J_q$ and that $J = \sum_{q=1}^{Q} J_q$ is the total number of categories. There are $J_1 \times \cdots \times J_q$ types of responses possible.

Again the same correspondence analysis algorithm may be applied to $\mathbf{Z}$ to obtain a graphical display of the J categories and, if necessary, the I row points. An example will demonstrate the generalization of the geometric interpretation in this case.

Figure 7 shows the display of a set of data from another readership survey. Three hundred fifteen people are categorized according to their readership of 19 publications ($Q = 19$). For each publication there are five categories of readership ($J_q = 5$ for all q): "don't read," "glance," "read some," "read most," and "read all." The indicator matrix analyzed is thus of order 315×95. For any given configuration of column points representing the 95 categories, each respondent (row) point lies at the average

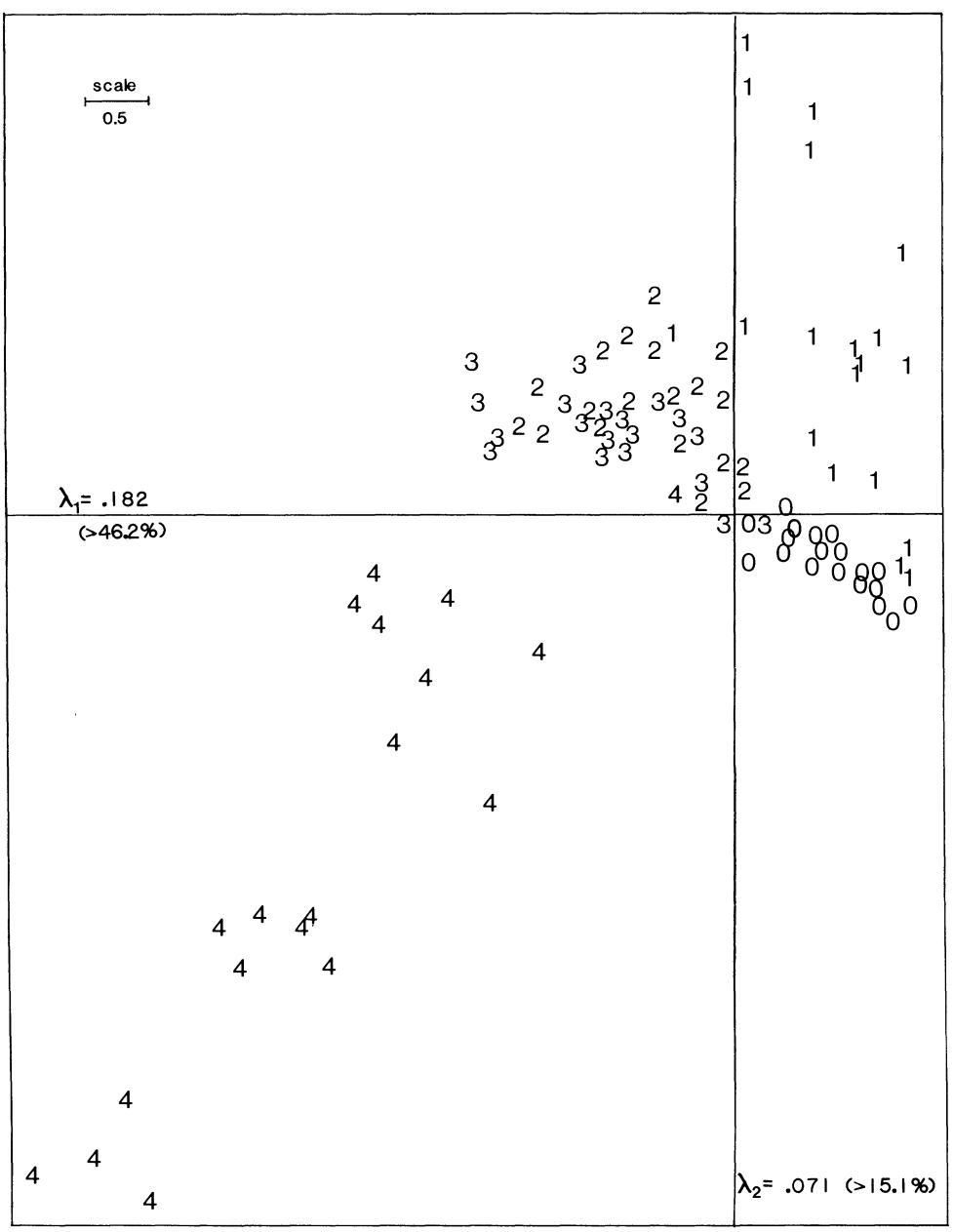


Figure 7. Multiple Correspondence Analysis of the 19-Variate Indicator Matrix of Readership Data. For each of 19 publications there are five category points: 0, do not read; 1, glance; 2, read some; 3, read most; 4, read all. Only the projected vertices of the category points are displayed. The computed percentages of inertia for the two axes are 34.3% and 13.4%, respectively, but we show the more realistic lower bounds according to results by Greenacre (1986).

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

position of the 19 category points that characterize his or her response vector. Figure 7 is that configuration of category points that maximizes the dispersion of the respondent points in a planar display.

Notice that points spread and contrast along diagonal axes rather than the principal axes. This is of no consequence to our final conclusions. The "don't read" categories of all of the media are in a tight bunch on the lower right side, opposing all of the "reading" categories. There are a number of "read most" points and a "read all" point on the right near the "don't read" categories, which means that some people read a few specific publications intensively, but very little else.

The vertical spread of the reading categories from bottom left to top right shows the spread of the reading thoroughness categories. The "glance" categories are well separated. There is a large overlap of the middle two categories of thoroughness. Finally, the "read all" categories are generally well separated from the other categories. Remember again that the patterns in the category points reflect the patterns among the respondents. Where the responses "read some" or "read most" occur, there is usually a mixture of these and the distinction between these two categories is rather blurred, compared with the quite separate "glance" and "read all" categories.

This display reveals a pattern that might be unexpected beforehand. It is tempting to assign unidimensional scale values such as 0, 1, 2, 3, 4 to the reading categories and use these values in conventional multivariate analysis. But the assumption that the "read some" and "read most" categories are separate in the data is contradicted by the display. It also appears that at least two dimensions should be considered in the reduction of these data: one that measures the read/no-read dichotomy by counting the number of publications read, say, and one that quantifies the thoroughness of actual reading, not necessarily on an equi-spaced scale. The optimal scaling interpretation, mentioned in Section 2.2 for simple correspondence analysis, carries over to the multivariate case as follows. Any set of scale values for all of the categories (i.e., columns of $\mathbf{Z}$) implies a score for each respondent (i.e., row of $\mathbf{Z}$), where the score is the average scale value of the categories into which the respondent falls. The scale values provided by the positions of the category points on the first principal axis yield row scores with maximum variance.

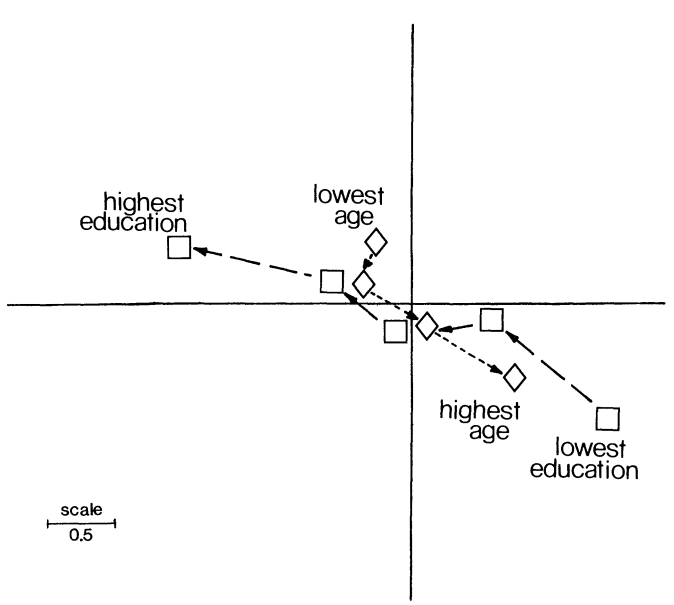


Figure 8. Positions of Supplementary Education and Age Categories, to Be Overlaid on Figure 7.

In this survey we had additional biographical information for each respondent. Figure 8 shows the positions of five categories of education and four categories of age as supplementary column points (these would normally be overlaid on Fig. 7). Education seems to line up more or less in agreement with the read/no-read dichotomy, with the lower education groups reading the least. Age moves in the opposite direction, with the youngest group reading the most, the oldest the least. Education shows a higher association with the display than age, because its category points are spread out more.

Although all 315 respondent points may be displayed, attention in such analyses shifts to groups of row points—for example, row points grouped by biographical categories. This is an alternative interpretation of the display of the biographical categories. A point representing the lowest category of education can be defined as the centroid of all of the respondent points who fall into this category. The fact that this centroid is on the right side of the display indicates that this group of respondents has a relatively high number of "don't read" responses.

4. MORE GENERAL CODING SCHEMES

Many different types of data may be recoded into a form suitable for correspondence analysis. The most commonly used coding scheme is called *fuzzy coding* (*codage flou* in the French literature), a generalization of the strict logical coding of the indicator matrix. Instead of a 1 indicating a specific category, with zeros elsewhere, we can assign a set of nonnegative values that add up to 1. These can even be considered probabilities that the case lies in the respective categories. Fuzzy coding can be used to recode continuous data into ordered categories. When a data value lies near the boundary between categories, it may be allocated to both categories in appropriate amounts. Various ways of handling missing data can be explored using different coding schemes (Greenacre 1984, sec. 5.3; Meulman 1982).

The special case of two categories per question (i.e., $J_q = 2$ for all q) is often encountered, especially when dealing with bipolar rating scales. For example, Kosslyn (1985, table 1) presented ratings of five books on graphics according to 14 criteria. The books are by Bertin (1983), Chambers, Cleveland, Kleiner, and Tukey (1983), Tufte (1983), Fisher (1982), and Schmid (1983), and they will be denoted by B, C, T, F, and S, respectively. Each rating scale is 10 points, with 1 indicating "very poor" and 10 "excellent." In a large survey one might consider treating each of the criteria as a discrete variable with 10 categories,

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

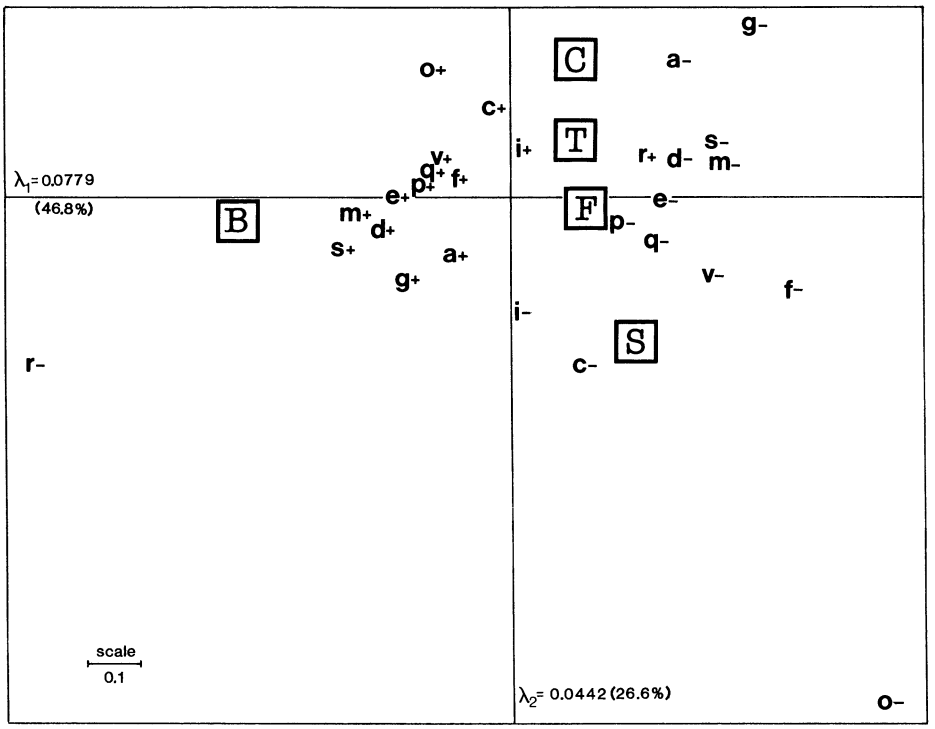


Figure 9. Correspondence Analysis of "Doubled" Table of Book Evaluation Data (Kosslyn 1985). The books are by Bertin (B), Chambers et al. (C), Tufte (T), Fisher (F), and Schmid (S). The criteria are readability (r), originality (o), generality (g), discriminability (d), visual properties (v), processing priorities (p), perceptual distortion (c), perceptual grouping (e), memory limits (m), ambiguity (a), inferences (i), purposes (s), questions (q), and data formats (f). For each criterion the favorable pole of the scale is labeled "+" and the opposite pole "−." Notice that this is the joint display of the row and column profiles.

then exploring the data by multiple correspondence analysis. This is clearly not possible with a sample size of 5, so we treat each criterion as a fuzzy categorical variable with just two categories. These two categories, labeled by "+" and "−," are the two extreme poles of the scale. The data are coded in terms of how much weight is allocated to the respective poles. For example, a rating of 9 on readability for book C is coded as [.9 .1] for the respective categories, showing that a weight of .9 is allocated to the positive readability pole and .1 to the negative readability pole. The process is called *doubling* (*dédoublement* in French) because the number of columns is effectively doubled. In this example the original 5 × 14 data matrix is converted into a 5 × 28 fuzzy indicator matrix.

The correspondence analysis of this matrix is given in Figure 9. Each criterion is represented by its two polar points and these lie on a line through the origin. The analysis can be thought of as a multiple correspondence analysis in which the category points are constrained to be regularly ordered and spaced on a straight line. Each pair of + and − points has the usual property of having its centroid at the origin, so the origin lies exactly at the mean rating for the particular criterion. All of the + points are clustered on the left side, opposing the − points on the right side, apart from the criterion "readability" that lies in exactly the opposite direction. This horizontal axis separates book B (on the left) from the other four books on the right. B generally receives the highest ratings, except for "readability." The other books separate out along the second vertical axis. S seems to be the least liked, opposing C, which rates highly on "originality" and its treatment of "perceptual distortion" and "drawing inferences."

Continuous variables may also be coded into a bipolar form. Escofier (1979) proposed that a standardized variable z (with mean 0 and variance 1) be reexpressed as two fuzzy variables: $y+ = (1 + z)/2$ and $y- = (1 - z)/2$. The centroid of each pair of + and − points is still at the origin, and their masses are the same, so they are equidistant from the origin. The coding reflects to what extent an observation lies above or below the mean. This scheme is useful when analyzing continuous and discrete data at the same time.

The general principle of such coding schemes is that a constant unit is spread across two or more categories of the recoded variable, possibly in negative amounts as in the last example. The geometry is a simple generalization of that for indicator matrices and usually amounts to talking about weighted averages of category points instead of simple averages.

5. DISCUSSION

Because correspondence analysis is really the principal components analysis of categorical data, it is surprising that the technique still remains relatively unknown outside the fields of psychology and ecology. French statisticians have elevated it to a jack-of-all-trades technique of data analysis. They place large emphasis on how data, often at

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms

different measurement levels, are reexpressed prior to analysis. The same algorithm and geometric framework can handle a multitude of different data types and structures.

Finally, although we have motivated simple correspondence analysis from a geometric point of view, the geometry of the indicator matrix in multiple correspondence analysis is admittedly not as convincing. Distances between the row profiles of an indicator matrix and projections of artificial column vertices have less intuitive appeal. However, the scaling interpretation remains attractive in this case; the displays are graphical representations of optimal scale values for the categories.

[*Received June 1986. Revised October 1986.*]

REFERENCES

Benzécri, J.-P. (1973), *L'Analyse des Données, Tome 2: L'Analyse des Correspondances*, Paris: Dunod.

——— (1977), "Histoire et Préhistoire de l'Analyse des Données, 5—L'Analyse des Correspondances," *Cahiers de l'Analyse des Données*, 2, 9–40.

——— (1979), "Sur le Calcul des Taux d'Inertie Dans l'Analyse d'un Questionnaire," *Cahiers de l'Analyse des Données*, 4, 377–378.

Bertin, J. (1983), *Semiology of Graphs* (translated by W. Berg), Madison: University of Wisconsin Press.

Carroll, J. D., Green, P. E., and Schaffer, C. M. (1986), "Interpoint Distance Comparisons in Correspondence Analysis," *Journal of Market Research*, 23, 271–280.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth.

Escofier, B. (1979), "Traitement Simultané de Variables Qualitatives et Quantitatives en Analyse Factorielle," *Cahiers de l'Analyse des Données*, 4, 137–146.

Fisher, H. T. (1982), *Mapping Information*, Cambridge, MA: Abt Books.

Gabriel, K. R. (1971), "The Biplot-Graphic Display of Matrices With Application to Principal Component Analysis," *Biometrika*, 58, 453–467.

——— (1981), "Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis," in *Interpreting Multivariate Data*, ed. V. Barnett, Chichester, U.K.: John Wiley, pp. 147–174.

Gifi, A. (1981), *Nonlinear Multivariate Analysis*, University of Leiden, Dept. of Data Theory.

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

——— (1986), "Multiple Correspondence Analysis and the Least-Squares Fitting of All Two-Way Tables of a Set of Categorical Variables," Research Report 86/8, University of South Africa, Dept. of Statistics.

Hirschfeld, H. O. (1935), "A Connection Between Correlation and Contingency," in *Proceedings of the Cambridge Philosophical Society* (*Mathematical Proceedings*), 31, 520–524.

Hotelling, H. O. (1933), "Analysis of a Complex of Statistical Variables Into Principal Components," *Journal of Educational Psychology*, 24, 417–441 & 498–520.

Kosslyn, S. M. (1985), "Graphics and Human Information Processing," *Journal of the American Statistical Association*, 80, 499–512.

Lebart, L., Morineau, A., and Tabard, N. (1977), *Techniques de la Description Statistique: Méthodes et Logiciels Pour l'Analyse des Grands Tableaux*, Paris: Dunod.

Lebart, L., Morineau, A., and Warwick, K. (1984), *Multivariate Descriptive Statistical Analysis*, New York: John Wiley.

Meulman, J. (1982), *Homogeneity Analysis of Incomplete Data*, Leiden: DSWO Press.

Nash, S. W. (1978), Review of *Techniques de la Description Statistique: Méthodes et Logiciels Pour l'Analyse des Grands Tableaux*, by L. Lebart, A. Morineau, and N. Tabard, *Journal of the American Statistical Association*, 74, 254–255.

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

——— (1986), *Quantification of Categorical Data: A Bibliography 1975–1986*, Toronto: Microstats.

O'Neill, M. E. (1978), "Asymptotic Distributions of the Canonical Correlations From Contingency Tables," *Australian Journal of Statistics*, 20, 75–82.

——— (1980), "The Distribution of Higher-Order Interactions in Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 42, 357–365.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to a System of Points in Space," *Philosophical Magazine*, 2, 559–572.

Schmid, C. F. (1983), *Statistical Graphics*, New York: John Wiley.

Tenenhaus, M., and Young, F. W. (1985), "An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Data," *Psychometrika*, 50, 91–119.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

This content downloaded from 171.66.12.245 on Sun, 26 Jan 2020 22:24:02 UTC
All use subject to https://about.jstor.org/terms