Graphical Methods for Marketing Research

Warren F. Kuhfeld

Abstract

Correspondence analysis, multiple correspondence analysis, preference mapping, and multidimensional preference analysis are descriptive statistical methods that generate graphical displays from data matrices. These methods are used by marketing researchers to investigate relationships among products and individual differences in preferences for those products. The end result is a two- or three-dimensional scatter plot that shows the most salient information in the data matrix. This chapter describes these methods, shows examples of the graphical displays, and discusses marketing research applications.*

Introduction

Correspondence analysis (CA), multiple correspondence analysis (MCA), preference mapping (PREFMAP), and multidimensional preference analysis (MDPREF) are descriptive statistical methods that generate graphical displays from data matrices. These methods are sometimes referred to as perceptual mapping methods. They simultaneously locate two or more sets of points in a single plot, and all emphasize presenting the geometry of the data. CA simultaneously displays in a scatter plot the row and column labels from a two-way contingency table or crosstabulation constructed from two categorical variables. MCA simultaneously displays in a scatter plot the category labels from more than two categorical variables. MDPREF displays both the row labels (products) and column labels (people) from a data matrix of continuous variables. PREFMAP shows rating scale data projected into a plot of row labels—for example, from an MDPREF analysis. These methods are used by marketing researchers to investigate relationships among products and individual differences in preferences for those products.

This chapter will only discuss these techniques as methods of generating two-dimensional scatter plots. However, three-dimensional and higher-dimensional results can also be generated and displayed with modern interactive graphics software and with scatter plot matrices.

*This chapter is a revision of a paper that was published in the 1992 National Computer Graphics Association Conference Proceedings. Copies of this chapter (TS-722L) and all of the macros are available on the web http://support.sas.com/techsup/tnote/tnote_stat.html#market.

Methods

This section presents the algebra and example plots for MDPREF, PREFMAP, CA, and MCA. These methods are all similar in spirit to the biplot, which is discussed first to provide a foundation for the other methods.

The Biplot. A *biplot* (Gabriel 1981) simultaneously displays the rows and columns of a data matrix in a low-dimensional (typically two-dimensional) plot. The "bi" in "biplot" refers to the *joint* display of rows and columns, not to the dimensionality of the plot. Consider an $(n \times m)$ data matrix $\mathbf{Y}$, an $(n \times q)$ matrix $\mathbf{A}$ with row vectors $\mathbf{a}'_1$, $\mathbf{a}'_2$, ..., $\mathbf{a}'_n$, and an $(m \times q)$ matrix $\mathbf{B}$ with row vectors $\mathbf{b}'_1$, $\mathbf{b}'_2$, ..., $\mathbf{b}'_m$. The n rows of $\mathbf{A}$ correspond to the rows of $\mathbf{Y}$, and the m columns of $\mathbf{B}'$ correspond to the columns of $\mathbf{Y}$. The rank of $\mathbf{Y}$ is $q \leq \text{MIN}(n, m)$. $\mathbf{A}$ and $\mathbf{B}$ are chosen such that $y_{ij} = \mathbf{a}'_i\mathbf{b}_j$. If $q = 2$ and the rows of $\mathbf{A}$ and $\mathbf{B}$ are plotted in a two-dimensional scatter plot, the scalar product of the coordinates $\mathbf{a}'_i$ and $\mathbf{b}'_j$ *exactly* equals the data value y_{ij}. This kind of scatter plot is a biplot; it geometrically shows the algebraic relationship $\mathbf{AB}' = \mathbf{Y}$. Typically, the row coordinates are plotted as points, and the column coordinates are plotted as vectors.

When $q > 2$ and two dimensions are plotted, then $\mathbf{a}'_i\mathbf{b}_j$ is *approximately* equal to y_{ij}, and the display is an *approximate biplot.*[*] The approximate biplot geometrically shows the algebraic relationship $\mathbf{AB}' \approx \mathbf{Y}$. The best values for $\mathbf{A}$ and $\mathbf{B}$, in terms of minimum squared error in approximating $\mathbf{Y}$, are found using a singular value decomposition (SVD),[†] $\mathbf{Y} = \mathbf{AB}' = \mathbf{UDV}'$, where $\mathbf{D}$ is a $(q \times q)$ diagonal matrix and $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_q$, a $(q \times q)$ identity matrix. Solutions for $\mathbf{A}$ and $\mathbf{B}$ include $\mathbf{A} = \mathbf{U}$ and $\mathbf{B} = \mathbf{VD}$, or $\mathbf{A} = \mathbf{UD}$ and $\mathbf{B} = \mathbf{V}$, or more generally $\mathbf{A} = \mathbf{UD}^r$ and $\mathbf{B} = \mathbf{VD}^{(1-r)}$, for $0 \leq r \leq 1$. See Gabriel (1981) for more information on the biplot.

Multidimensional Preference Analysis. Multidimensional Preference Analysis (Carroll 1972) or MDPREF is a biplot analysis for preference data. Data are collected by asking respondents to rate their preference for a set of objects. Typically in marketing research, the objects are products−the client's products and the competitors'. Questions that can be addressed with MDPREF analyses include: Who are my customers? Who else should be my customers? Who are my competitors' customers? Where is my product positioned relative to my competitors' products? What new products should I create? What audience should I target for my new products?

For example, consumers can be asked to rate their preference for a group of automobiles on a 0 to 9 scale, where 0 means no preference and 9 means high preference. $\mathbf{Y}$ is an $(n \times m)$ matrix that contains ratings of the n products by the m consumers. The data are stored as the transpose of the typical data matrix, since the columns are the people. The goal is to produce a plot with the cars represented as points and the consumers represented as vectors. Each person's vector points in *approximately* the direction of the cars that the person most preferred and away from the cars that are least preferred.

Figure 1 contains an example in which 25 consumers rated their preference for 17 new (at the time) 1980 automobiles. This plot is based on a principal component model. It differs from a proper biplot of $\mathbf{Y}$ due to scaling factors. In principal components, the columns in data matrix $\mathbf{Y}$ are standardized to mean zero and variance one. The SVD is $\mathbf{Y} = \mathbf{UDV}'$, and the principal component model is $\mathbf{Y} = ((n-1)^{1/2}\mathbf{U})\,((n-1)^{-1/2}\mathbf{D})\,(\mathbf{V}')$. The standardized principal component scores matrix, $\mathbf{A} = (n-1)^{1/2}\mathbf{U}$, and the component structure matrix, $(n-1)^{-1/2}\mathbf{DV}'$, are plotted. The advantage of creating a biplot based on $(n-1)^{1/2}\mathbf{U}$ and $(n-1)^{-1/2}\mathbf{DV}'$ instead of $\mathbf{U}$ and $\mathbf{DV}'$ is that the coordinates

[*]In practice, the term biplot is sometimes used without qualification to refer to an approximate biplot.

[†]SVD is sometimes referred to in the psychometric literature as an Eckart-Young (1936) decomposition.

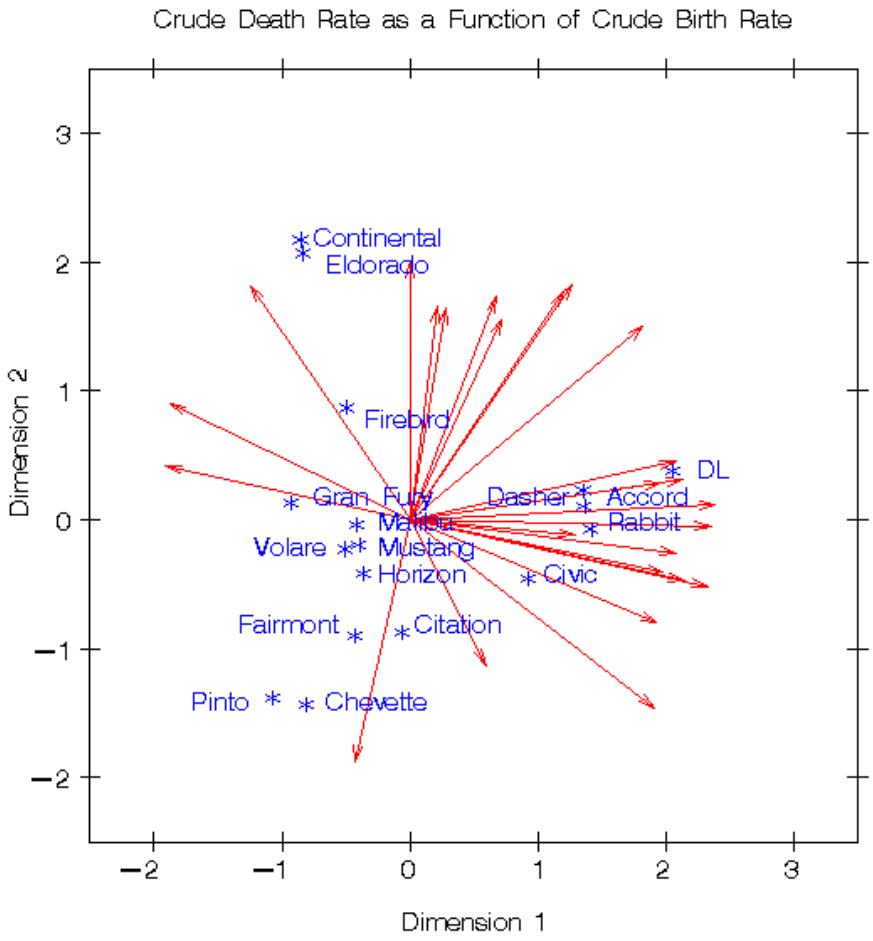

Figure 1. Multidimensional Preference Analysis

do not get smaller as sample size increases. The fit, or proportion of the variance in the data accounted for by the first two dimensions, is the sum of squares of the first two elements of $(n-1)^{-1/2}\mathbf{D}$ divided by the sum of squares of all of the elements of $(n-1)^{-1/2}\mathbf{D}$.

The dimensions of the MDPREF biplot are the first two principal components. The first principal component represents the information that is most salient to the preference judgments. At one end of the plot of the first principal component are the most preferred automobiles; the least preferred automobiles are at the other end of the plot. The second principal component represents the direction that is most salient to the preference judgments that is orthogonal to the first principal component. The automobile point coordinates are the scores of the automobile on the first two principal components. The judge vectors point in *approximately* the direction of judges most preferred cars, with preference increasing as the vector moves from the origin.

Let $\mathbf{a}_i'$ be row i of $\mathbf{A} = (n-1)^{1/2}\mathbf{U}$, $\mathbf{b}_j'$ be row j of $\mathbf{B} = (n-1)^{-1/2}\mathbf{VD}$, $\|\mathbf{a}_i\|$ be the length of $\mathbf{a}_i$, $\|\mathbf{b}_j\|$ be the length of $\mathbf{b}_j$, and θ be the angle between the vectors $\mathbf{a}_i$ and $\mathbf{b}_j$. The predicted degree of (scaled) preference that an individual judge has for an automobile is $\mathbf{a}_i'\mathbf{b}_j = \|\mathbf{a}_i\|\,\|\mathbf{b}_j\| cos\theta$. Each car point can be orthogonally projected onto each judge's vector. The projection of the ith car on the jth judge vector is $\mathbf{b}_j((\mathbf{a}_i'\mathbf{b}_j)/(\mathbf{b}_j'\mathbf{b}_j))$, and the length of this projection is $\|\mathbf{a}_i\| cos\theta$. The automobile that

projects farthest along a judge vector has the highest predicted preference. The length of this projection, $\|\mathbf{a}_i\| cos\theta$, differs from the predicted preference, $\|\mathbf{a}_i\| \, \|\mathbf{b}_j\| cos\theta$, only by $\|\mathbf{b}_j\|$, which is constant within each judge. Since the goal is to look at projections of points onto the vectors, the absolute length of a judge's vector is unimportant. The relative lengths of the vectors indicate fit, with longer vectors indicating better fit. The coordinates for the endpoints of the vectors were multiplied by 2.5 to extend the vectors and create a better graphical display. The direction of the preference scale is important. The vectors point in the direction of increasing values of the data values. If the data had been ranks, with 1 the most preferred and n the least preferred, then the vectors would point in the direction of the least preferred automobiles.

The people in the top left portion of the plot most prefer the large American cars. Other people, with vectors pointing up and nearly vertical, also show this pattern of preference. There is a large cluster of people who prefer the Japanese and European cars. A few people, most notably the person whose vector passes through the "e" in "Chevette", prefer the small and inexpensive American Cars. There are no vectors pointing through the bottom left portion of the of the plot, which suggests that the smaller American cars are generally not preferred by anyone within this group.

The first dimension, which is a measure of overall evaluation, discriminates between the American cars on the left and the Japanese and European cars on the right. The second dimension seems to reflect the sizes of the automobiles. Some cars have a similar pattern of preference, most notably Continental and Eldorado, which share a symbol in the plot. Marketers of Continental or Eldorado may want to try to distinguish their car from the competition. Dasher, Accord, and Rabbit were rated similarly, as were Malibu, Mustang, Volare, and Horizon.

This 1980 example is quite prophetic even though it is based on a small nonrandom sample. Very few vectors point toward the smaller American cars, and Mustang is the only one of them that is still being made. Many vectors are pointing toward the European and Japanese cars, and they are still doing quite well in the market place. Many vectors are pointing in the one to two o'clock range where there are no cars in the plot. One can speculate that these people would prefer Japanese and European luxury cars such as Accura, Lexus, Infinity, BMW, and Mercedes.

Preference Mapping. Preference mapping[‡] (Carroll 1972) or PREFMAP plots resemble biplots, but are based on a different model. The goal in PREFMAP is to take a set of coordinates for a set of objects, such as the MDPREF car coordinates in example in Figure 1, and project in external information that can aid in interpreting the configuration of points. Questions that can be addressed with PREFMAP analyses include: Where is my product positioned relative to my competitors' products? Why is my product positioned there? How can I reposition my existing products? What new products should I create?

The Preference Mapping Vector Model. Figure 2 contains an example in which three attribute variables (ride, reliability, and miles per gallon) are displayed in the plot of the first two principal components of the car preference data. Each of the automobiles was rated on these three dimensions on a 1 to 5 scale, where 1 is poor and 5 is good. Figure 2 is based on the simplest version of PREFMAP−the *vector model*. The vector model assumes that some is good and more is *always* better. This model is appropriate for miles per gallon and reliability−the more miles a motorist can travel without refueling or breaking down, the better. The end points for the attribute vectors are obtained by projecting the attribute variables into the car space. If the attribute ratings are stored in matrix $\mathbf{R}$, then the coordinates for the end points are in the matrix $\boldsymbol{\beta}$ from the multivariate linear regression model $\mathbf{R} =$

[‡]Preference mapping is sometimes referred to as external unfolding.

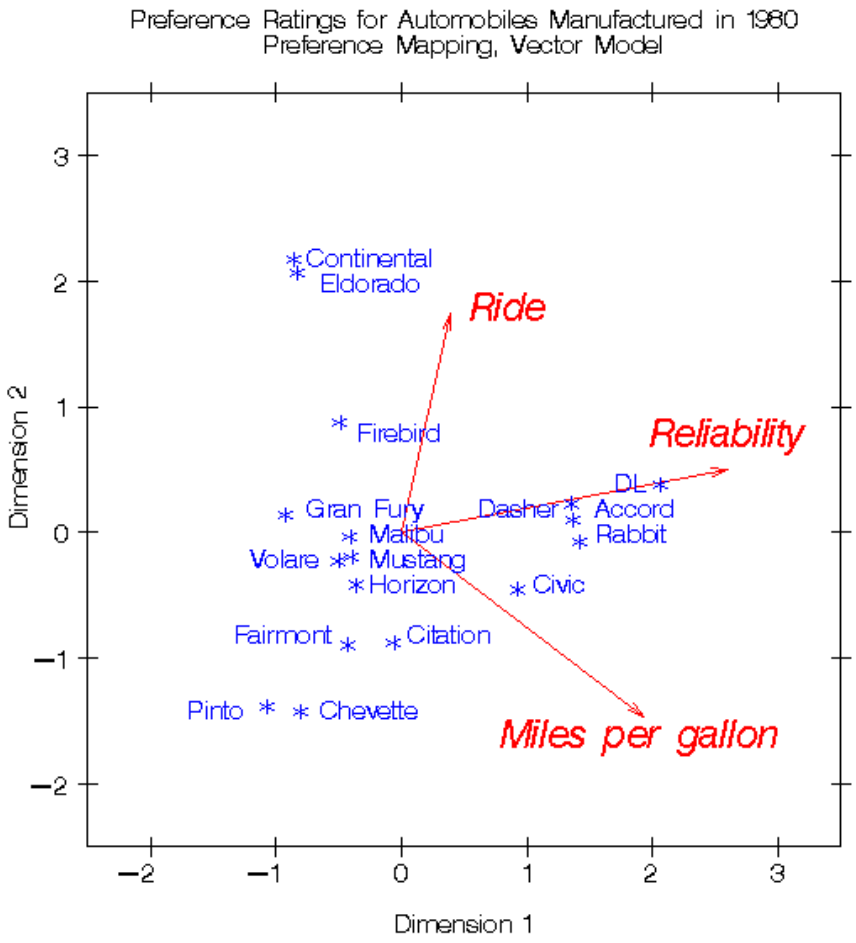

Figure 2. Preference Mapping, Vector Model

$\mathbf{A}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. $\mathbf{A}$ is the matrix of standardized principal component scores, or $\mathbf{A}$ could be the coordinates from a multidimensional scaling analysis. The relative lengths of the vectors indicate fit, which is given by the R^2. As with MDPREF, the lengths of all vectors can be scaled by the same constant to make a better graphical display.

PREFMAP analyses can help in the interpretation of principal component, multidimensional scaling, and MDPREF analyses by projecting in external information that helps explain the configuration. Orthogonal projections of the product points on an attribute vector give an *approximate* ordering of the products on the attribute. The ride vector points almost straight up showing that the larger cars, such as the Eldorado and Continental, have the best ride. In Figure 1, it was shown that most people preferred the DL, Japanese cars, and larger American cars. Figure 2 shows that the DL and Japanese cars were rated as the most reliable and have the best fuel economy. The small American cars are not rated highly on any of the three dimensions, although some are on the positive end of miles per gallon.

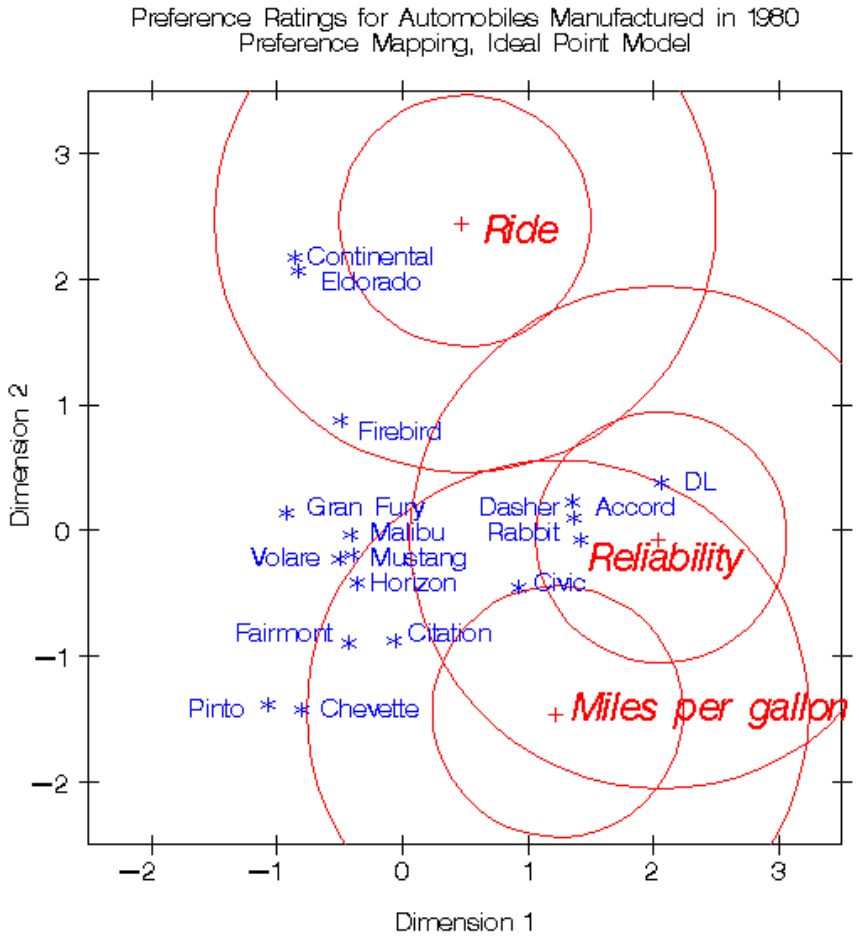

Figure 3. Preference Mapping, Ideal Point Model

The Preference Mapping Ideal Point Model. The *ideal point* model differs from the vector model in that the ideal point model does not assume that more is better, *ad infinitum*. Consider the sugar content of cake. There is an ideal amount of sugar that cake should contain—not enough sugar is not good, and too much sugar is also not good. In the current example, the ideal number of miles per gallon and the ideal reliability are unachievable. It makes sense to consider a vector model, because the ideal point is infinitely far away. This argument is less compelling for ride; the point for a car with smooth, quiet ride may not be infinitely far away. Figure 3 shows the results of fitting an ideal point model for the three attributes. In the vector model, results are interpreted by orthogonally projecting the car points on the attribute vectors. In the ideal point model, Euclidean distances between car points and ideal points are compared. Eldorado and Continental have the best predicted ride, because they are closest to the ride ideal point. The concentric circles drawn around the ideal points help to show distances between the cars and the ideal points. The numbers of circles and their radii are arbitrary. The overall interpretations of Figures 2 and 3 are the same. All three ideal points are at the edge of the car points, which suggests the simpler vector model is sufficient for these data.

The ideal point model is fit with a multiple regression model and some pre- and post-processing. First the $\mathbf{A}$ matrix is augmented by a variable that is the sum of squares of columns of $\mathbf{A}$ creating $\mathbf{A}^*$. Then solve for $\boldsymbol{\beta}$ from $\mathbf{R} = \mathbf{A}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}$. For a two-dimensional scatter plot, the ideal point coordinates are

given by dividing each coefficient for the two axes by the coefficient for the sum-of-squares variable, then multiplying the resulting values by −0.5. The coordinates are $-0.5\,\boldsymbol{\beta}\,\mathrm{diag}(\boldsymbol{\beta}_3)^{-1}$, where $\mathrm{diag}(\boldsymbol{\beta}_3)$ is a diagonal matrix constructed from the third row of $\boldsymbol{\beta}$. This is a constrained response-surface model. The fit is given by the R^2. See Carroll (1972) for the justification for the formula.

The results in Figure 3 were modified from the raw results to eliminate *anti-ideal points*. The ideal point model is a distance model. The rating data are interpreted as distances between attribute ideal points and the products. In this example, each of the automobiles was rated on these three dimensions, on a 1 to 5 scale, where 1 is poor and 5 is good. The data are the reverse of what they should be−a ride rating of 1 should mean this car is similar to a car with a good ride, and a rating of 5 should mean this car is different from a car with a good ride. So the raw coordinates must be multiplied by −1 to get ideal points. Even if the scoring had been reversed, anti-ideal points can occur. If the coefficient for the sum-of-squares variable is negative, the point is an anti-ideal point. In this example, there is the possibility of *anti-anti-ideal points*. When the coefficient for the sum-of-squares variable is negative, the two multiplications by −1 cancel, and the coordinates are ideal points. When the coefficient for the sum-of-squares variable is positive, the coordinates are multiplied by −1 to get an ideal point.

Other PREFMAP Models. The ideal point model presented here is based on an ordinary Euclidean distance model. All points falling on a circle centered around an ideal point are an equal distance from the ideal point. Two more PREFMAP models are sometimes used. The more general models allow for differential weighting of the axes and rotations, so ellipses, not circles, show equal weighted distances. All three ideal point models are response surface models. See Carroll (1972) for more information.

Correspondence Analysis. Correspondence analysis (CA) is a weighted SVD of a contingency table. It is used to find a low-dimensional graphical representation of the association between rows and columns of a table. Each row and column is represented by a point in a Euclidean space determined from cell frequencies. Like MDPREF, CA is based on a singular value decomposition, but ordinary SVD of a contingency table does not portray a desirable geometry.

Questions that can be addressed with CA and MCA include: Who are my customers? Who else should be my customers? Who are my competitors' customers? Where is my product positioned relative to my competitors' products? Why is my product positioned there? How can I reposition my existing products? What new products should I create? What audience should I target for my new products?

CA is a popular data analysis method in France and Japan. In France, CA was developed under the strong influence of Jean-Paul Benzécri; in Japan, under Chikio Hayashi. CA is described in Lebart, Morineau, and Warwick (1984); Greenacre (1984); Nishisato (1980); Tenenhaus and Young (1985); Gifi (1990); Greenacre and Hastie (1987); and many other sources. Hoffman and Franke (1986) provide a good introductory treatment using examples from marketing research.

Simple CA. This section is primarily based on the theory of CA found in Greenacre (1984). Let $\mathbf{N}$ be an $(n_r \times n_c)$ contingency table of rank $q \leq MIN(n_r, n_c)$. Let $\mathbf{1}$ be a vector of ones of the appropriate order, $\mathbf{I}$ be an identity matrix, and diag() be a matrix-valued function that creates a diagonal matrix from a vector. Let $f = \mathbf{1'N1}$, $\mathbf{P} = (1/f)\mathbf{N}$, $\mathbf{r} = \mathbf{P1}$, $\mathbf{c} = \mathbf{P'1}$, $\mathbf{D_r} = \mathrm{diag}(\mathbf{r})$, and $\mathbf{D_c} = \mathrm{diag}(\mathbf{c})$. The scalar f is the sum of all elements in $\mathbf{N}$. $\mathbf{P}$ is a matrix of relative frequencies. The vector $\mathbf{r}$ contains row marginal proportions or row *masses*. The vector $\mathbf{c}$ contains column marginal proportions or column masses. $\mathbf{D_r}$ and $\mathbf{D_c}$ are diagonal matrices of marginals. The coordinates of the CA are based on the generalized singular value decomposition of $\mathbf{P}$, $\mathbf{P} = \mathbf{AD_uB'}$, where $\mathbf{A'D_r}^{-1}\mathbf{A} = \mathbf{B'D_c}^{-1}\mathbf{B} = \mathbf{I}$. $\mathbf{A}$

is an $(n_r \times q)$ matrix of left generalized singular vectors, $\mathbf{D_u}$ is a $(q \times q)$ diagonal matrix of singular values, and $\mathbf{B}$ is an $(n_c \times q)$ matrix of right generalized singular vectors. The first (trivial) column of $\mathbf{A}$ and $\mathbf{B}$ and the first singular value in $\mathbf{D_u}$ are discarded before any results are displayed. This step centers the table and is analogous to centering the data in ordinary principal component analysis. In practice, this centering is done by subtracting ordinary chi-square expected values from $\mathbf{P}$ before the SVD. The columns of $\mathbf{A}$ and $\mathbf{B}$ define the principal axes of the column and row point clouds, respectively. The fit, or proportion of the *inertia* (analogous to variance) in the data accounted for by the first two dimensions, is the sum of squares of the first two singular values, divided by the sum of squares of all of the singular values. Three sets of coordinates are typically available from CA, one based on rows, one based on columns, and the usual set is based on both rows and columns.

The *row profile* (conditional probability) matrix is defined as $\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{P} = \mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u}\mathbf{B}'$. Each (i, j) element of $\mathbf{R}$ contains the observed probability of being in column j given membership in row i. The values in each row of $\mathbf{R}$ sum to one. The row coordinates, $\mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u}$, and column coordinates, $\mathbf{D_c}^{-1}\mathbf{B}$, provide a CA based on the row profile matrix. The *principal* row coordinates, $\mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u}$, and *standard* column coordinates, $\mathbf{D_c}^{-1}\mathbf{B}$, provide a decomposition of $\mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u}\mathbf{B}'\mathbf{D_c}^{-1} = \mathbf{D_r}^{-1}\mathbf{P}\mathbf{D_c}^{-1} = \mathbf{R}\mathbf{D_c}^{-1}$. Since $\mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u} = \mathbf{R}\mathbf{D_c}^{-1}\mathbf{B}$, the row coordinates are weighted centroids of the column coordinates. Each column point, with coordinates scaled to standard coordinates, defines a vertex in $(n_c - 1)$-dimensional space. All of the principal row coordinates are located in the space defined by the standard column coordinates. Distances among row points have meaning, but distances among column points and distances between row and column points are not interpretable.

The formulas for the analysis of the *column profile* matrix can easily be derived by applying the row profile formulas to the transpose of $\mathbf{P}$. The principal column coordinates $\mathbf{D_c}^{-1}\mathbf{B}\mathbf{D_u}$ are weighted centroids of the standard row coordinates $\mathbf{D_r}^{-1}\mathbf{A}$. Each row point, with coordinates scaled to standard coordinates, defines a vertex in $(n_r - 1)$-dimensional space. All of the principal column coordinates are located in the space defined by the standard row coordinates. Distances among column points have meaning, but distances among row points and distances between row and column points are not interpretable.

The usual sets of coordinates[§] are given by $\mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u}$ and $\mathbf{D_c}^{-1}\mathbf{B}\mathbf{D_u}$. One advantage of using these coordinates is that both sets are postmultiplied by the diagonal matrix $\mathbf{D_u}$, whose diagonal values are all less than or equal to one. When $\mathbf{D_u}$ is a part of the definition of only one set of coordinates, that set forms a tight cluster near the centroid, whereas the other set of points is more widely dispersed. Including $\mathbf{D_u}$ in both sets makes a better graphical display. However, care must be taken in interpreting such a plot. No correct interpretation of distances between row points and column points can be made. Less specific statements, such as "two points are on the same side of the plot" have meaning.

Another property of this choice of coordinates concerns the geometry of distances between points within each set. Distances between row (or column) profiles are computed using a *chi-square metric*. The rationale for computing distances between row profiles using the non-Euclidean chi-square metric is as follows. Each row of the contingency table may be viewed as a realization of a multinomial distribution conditional on its row marginal frequency. The null hypothesis of row and column independence is equivalent to the hypothesis of homogeneity of the row profiles. A significant chi-square statistic is geometrically interpreted as a significant deviation of the row profiles from their centroid, $\mathbf{c}'$. The chi-square metric is the Mahalanobis metric between row profiles based on their estimated covariance matrix under the homogeneity assumption (Greenacre and Hastie 1987). A parallel argument can be made for the column profiles.

[§]This set is often referred to as the French standardization due to its popularity in France.

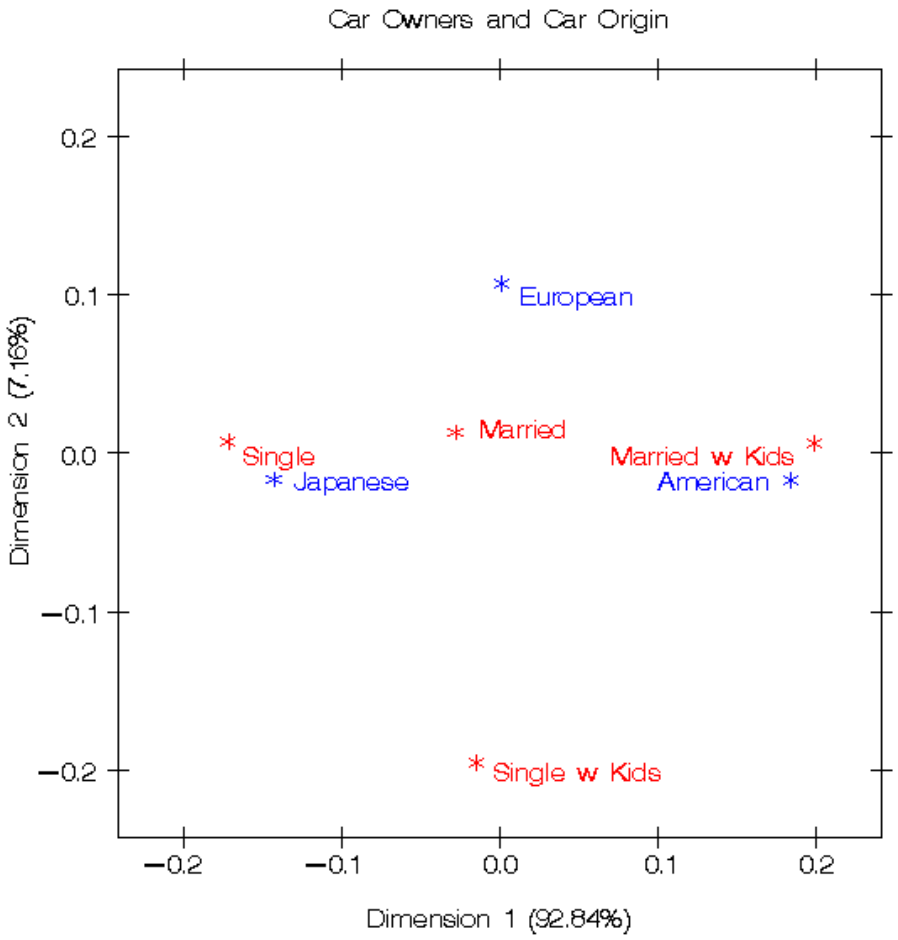

Figure 4. Simple Correspondence Analysis

The row coordinates are $\mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u} = \mathbf{D_r}^{-1}\mathbf{A}\mathbf{D_u}\mathbf{B}'\mathbf{D_c}^{-1}\mathbf{B} = (\mathbf{D_r}^{-1}\mathbf{P})(\mathbf{D_c}^{-1/2})(\mathbf{D_c}^{-1/2}\mathbf{B})$. They are row profiles $\mathbf{D_r}^{-1}\mathbf{P}$ rescaled by $\mathbf{D_c}^{-1/2}$ (rescaled so that distances between profiles are transformed from a chi-square metric to a Euclidean metric), then orthogonally rotated with $\mathbf{D_c}^{-1/2}\mathbf{B}$ to a principal axes orientation. Similarly, the column coordinates are column profiles rescaled to a Euclidean metric and orthogonally rotated to a principal axes orientation.

CA Example. Figure 4 contains a plot of the results of a simple CA of a survey of car owners. The questions included origin of the car (American, Japanese, European), and marital/family status (single, married, single and living with children, and married living with children). Both variables are categorical. Table 1 contains the crosstabulation and the observed minus expected frequencies. It can be seen from the observed minus expected frequencies that four cells have values appreciably different from zero (Married w Kids/American, Single/American, Married w Kids/Japanese, Single/Japanese). More people who are married with children drive American cars than would be expected if the rows and columns are independent, and more people who are single with no children drive Japanese cars than would be expected if the rows and columns are independent.

Table 1
Simple Correspondence Example Input

	Contingency Table			Observed Minus Expected Values		
	American	European	Japanese	American	European	Japanese
Married	37	14	51	-1.5133	0.4602	1.0531
Married w Kids	52	15	44	10.0885	0.2655	-10.3540
Single	33	15	63	-8.9115	0.2655	8.6460
Single w Kids	6	1	8	0.3363	-0.9912	0.6549

CA graphically shows the information in the observed minus expected frequencies. The right side of Figure 4 shows the association between being married with children and owning an American Car. The left side of the plot shows the association between being single and owning a Japanese Car. This interpretation is based on points being located in approximately the same direction from the origin and in approximately the same region of the space. Distances between row points and column points are not defined.

Multiple Correspondence Analysis. Multiple correspondence analysis (MCA) is a generalization of simple CA for more than two variables. The input is a *Burt table*, which is a partitioned symmetric matrix containing all pairs of crosstabulations among a set of categorical variables. Each diagonal partition is a diagonal matrix containing marginal frequencies (a crosstabulation of a variable with itself). Each off-diagonal partition is an ordinary contingency table. Each contingency table above the diagonal has a transposed counterpart below the diagonal. A Burt table is the inner product of a partitioned design matrix. There is one partition per categorical variable, and each partition is a binary design matrix. Each design matrix has one column per category, and a single 1 in each row. The partitioned design matrix has exactly m ones in each row, where m is the number of categorical variables. The results of a MCA of a Burt table, $\mathbf{N}$, are the same as the column results from a simple CA of the design matrix whose inner product is the Burt table. MCA is not a simple CA of the Burt table. The coordinates for MCA are $\mathbf{D_c}^{-1}\mathbf{B}\mathbf{D_u}$, from $(1/f)\mathbf{N} = \mathbf{P} = \mathbf{P}' = \mathbf{B}\mathbf{D_u}^2\mathbf{B}'$, where $\mathbf{B}'\mathbf{D_c}^{-1}\mathbf{B} = \mathbf{I}$.

MCA Example. Figure 5 contains a plot of the results of an MCA of a survey of car owners. The questions included origin of the car (American, Japanese, European), size of car (small, medium, large), type of car (family, sporty, work vehicle), home ownership (owns, rents), marital/family status (single, married, single and living with children, and married living with children), and sex (male, female). The variables are all categorical.

The top-right quadrant of the plot shows that the categories single, single with kids, 1 income, and renting a home are associated. Proceeding clockwise, the categories sporty, small, and Japanese are associated. In the bottom-left quadrant we see the association between being married, owning your own home, and having two incomes. Having children is associated with owning a large American family car. Such information could be used in marketing research to identify target audiences for advertisements. This interpretation is based on points being located in approximately the same direction from the origin

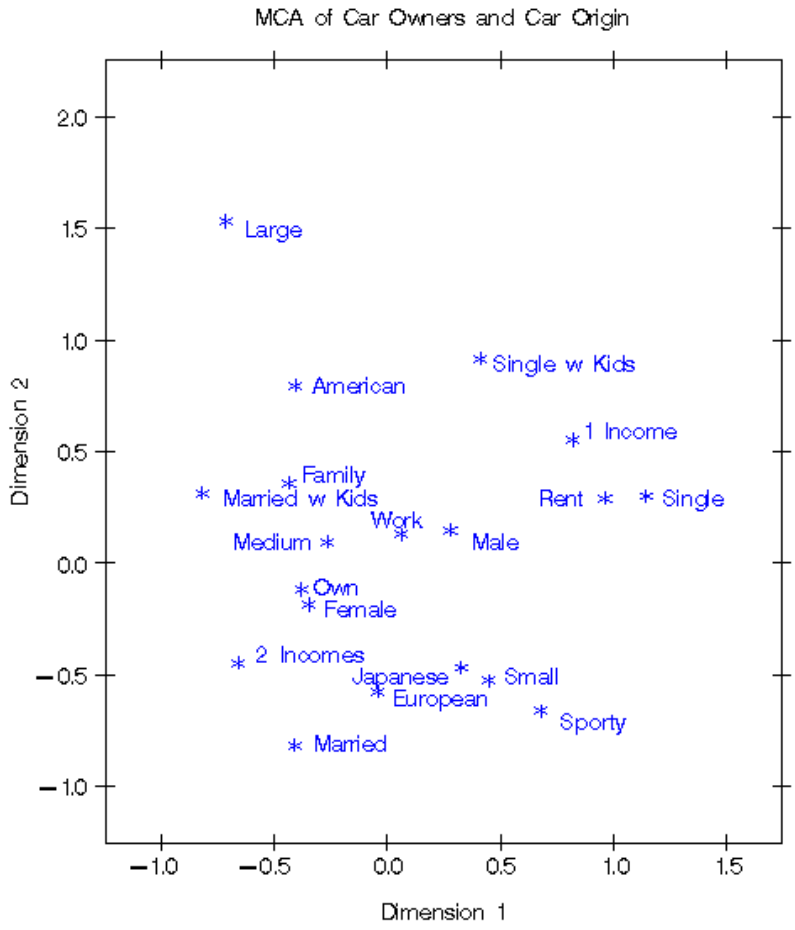

Figure 5. Multiple Correspondence Analysis

and in approximately the same region of the space. Distances between points are not interpretable in MCA.

Other CA Standardizations. Other standardizations have been proposed for CA by several authors. The usual goal is to provide a standardization that avoids the problem of row and column distances being undefined. Unfortunately, this problem remains unsolved. Carroll, Green, and Schaffer (1986) proposed that simple CA coordinates should be transformed to MCA coordinates before they are plotted. They argued that all distances are then comparable, but Greenacre (1989) showed that their assertion was incorrect. Others have also claimed to have discovered a method of defining the between row and column differences, but so far no method has been demonstrated to be correct.

Notes

The Geometry of the Scatter Plots. All of the scatterplots in this chapter were created with the axes equated so that a centimeter on the y-axis represents the same data range as a centimeter on the x-axis. *This is important.* Distances, angles between vectors, and projections are evaluated to interpret the plots. When the axes are equated, distances and angles are correctly represented in the plot. When axes are scaled independently, for example to fill the page, then the correct geometry is not presented. The important step of equating the axes is often overlooked in practice.

In a true biplot, $\mathbf{A} = \mathbf{U}\mathbf{D}^r$ and $\mathbf{B} = \mathbf{V}\mathbf{D}^{(1-r)}$ are plotted, and the elements of $\mathbf{Y}$ can be approximated from $y_{ij} \approx \mathbf{a}_i'\mathbf{b}_j$. For MDPREF and PREFMAP, the absolute lengths of the vectors are not important since the goal is to project points on vectors, not look at scalar products of row points and column vectors. It is often necessary to change the lengths of *all* of the vectors to improve the graphical display. If all of the vectors are relatively short with end points clustered near the origin, the display will not look good. To avoid this problem in Figure 1, *both* the x-axis and y-axis coordinates were multiplied by the constant 2.5, to lengthen all vectors by the same relative amount. The coordinates must not be scaled independently.

Software. All data analyses were performed with Release 8.00 of the SAS System. MDPREF is performed with PROC PRINQUAL, simple and multiple correspondence analysis are performed with PROC CORRESP, and PREFMAP is performed with PROC TRANSREG. The plots are prepared with the SAS `%PlotIt` autocall macro. If your site has installed the autocall libraries supplied by SAS Institute and uses the standard configuration of SAS software supplied by the Institute, you need only to ensure that the SAS system option `mautosource` is in effect to begin using the autocall macros. See pages 597–599 and pages 753–783.

Conclusions

Marketing research helps marketing decision makers understand their customers and their competition. Correspondence analysis compactly displays survey data to aid in determining what kinds of consumers are buying certain products. Multidimensional preference analysis shows product positioning, group preferences, and individual preferences. The plots may suggest how to reposition products to appeal to a broader audience. They may also suggest new groups of customers to target. Preference mapping is used as an aid in understanding MDPREF and multidimensional scaling results. PREFMAP displays product attributes in the same plot as the products. The insight gained from perceptual mapping methods can be a valuable asset in marketing decision making. These techniques can help marketers gain insight into their products, their customers, and their competition.