

www.statsoft.com

- Products
- Solutions
- Buy
- Trials
- Support

TextbookCorrespondence Analysis

What can we help you find?

Search

Looking for info about statistics?

We wrote the book on it.
And you can read it for free!

- Elementary Concepts
- Statistics Glossary
- Basic Statistics
- ANOVA / MANOVA
- Association Rules
- Boosting Trees
- Canonical Analysis
- CHAID Analysis
- C & R Trees
- Classification Trees
- Cluster Analysis
- Correspondence Analysis
- Data Mining Techniques
- Discriminant Analysis
- Distribution Fitting
- Experimental Design
- Factor Analysis
- General Discrim. Analysis
- General Linear Models
- Generalized Additive Mod.
- Generalized Linear Mod.
- General Regression Mod.
- Graphical Techniques
- Ind.Components Analysis
- Linear Regression
- Log-Linear Analysis
- MARSplines
- Machine Learning
- Multidimensional Scaling
- Neural Networks
- Nonlinear Estimation
- Nonparametric Statistics
- Partial Least Squares
- Power Analysis
- Process Analysis
- Quality Control Charts
- Reliability / Item Analysis
- SEPATH (Structural eq.)
- Survival Analysis
- Text Mining
- Time Series / Forecasting

How To Analyze Simple Two-Way and Multi-Way Table, Correspondence Analysis









- [General Purpose](#)
- [Supplementary Points](#)
- [Multiple Correspondence Analysis](#)
- [Burt Tables](#)

General Purpose

Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information which is similar in nature to those produced by [Factor Analysis](#) techniques, and they allow you to explore the structure of categorical variables included in the table. The most common kind of table of this type is the two-way frequency crosstabulation table (see, for example, [Basic Statistics](#) or [Log-Linear](#)).

In a typical correspondence analysis, a crosstabulation table of frequencies is first standardized, so that the relative frequencies across all cells sum to 1.0. One way to state the goal of a typical analysis is to represent the entries in the table of relative frequencies in terms of the distances between individual rows and/or columns in a low-dimensional space. This is best illustrated by a simple example, which will be described below. There are several parallels in interpretation between correspondence analysis and Factor Analysis, and some similar concepts will also be pointed out below.

For a comprehensive description of this method, computational details, and its applications (in the English language), refer to the classic text by Greenacre (1984). These methods were originally developed primarily in France by Jean-Paul Benzérci in the early 1960's and 1970's (e.g., see Benzérci, 1973; see also Lebart, Morineau, and Tabard, 1977), but have only more recently gained increasing popularity in English-speaking countries (see, for example, Carrol, Green, and Schaffer, 1986; Hoffman and Franke, 1986). (Note that similar techniques were developed independently in several countries, where they were known as optimal scaling, reciprocal averaging, optimal scoring, quantification method, or homogeneity analysis). In the following paragraphs, a general introduction to correspondence analysis will be presented.

-  [Variance Components](#)
-  [Statistical Advisor](#)
-  [Distribution Tables](#)
-  [References Cited](#)
-  [Send Comments](#)
-  [Business Solutions](#)
-  [Free Resources](#)
-  [About Textbook](#)

Overview. Suppose you collected data on the smoking habits of different employees in a company. The following data set is presented in Greenacre (1984, p. 55).

	Smoking Category				
Staff Group	(1) None	(2) Light	(3) Medium	(4) Heavy	Row Totals
(1) Senior Managers	4	2	3	2	11
(2) Junior Managers	4	3	7	4	18
(3) Senior Employees	25	10	12	4	51
(4) Junior Employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
Column Totals	61	45	62	25	193

You can think of the 4 column values in each row of the table as coordinates in a 4-dimensional space, and you could compute the (Euclidean) distances between the 5 row points in the 4-dimensional space. The distances between the points in the 4-dimensional space summarize all information about the similarities between the rows in the table above. Now suppose you could find a lower-dimensional space, in which to position the row points in a manner that retains all, or almost all, of the information about the differences between the rows. You could then present all information about the similarities between the rows (types of employees in this case) in a simple 1, 2, or 3-dimensional graph. While this may not appear to be particularly useful for small tables such as the one shown above, you can easily imagine how the presentation and interpretation of very large tables (e.g., differential preference for 10 consumer items among 100 groups of respondents in a consumer survey) could greatly benefit from the simplification that can be achieved via correspondence analysis (e.g., represent the 10 consumer items in a two-dimensional space).

Mass. To continue with the simpler example of the two-way table presented above, computationally, the program will first compute the relative frequencies for the frequency table, so that the sum of all table entries is equal to 1.0 (each element will be divided by the total, i.e., 193). You could say that this table now shows how one unit of *mass* is distributed across the cells. In the terminology of correspondence analysis, the row and column totals of the matrix of relative frequencies are called the row mass and column mass, respectively.

Inertia. The term *inertia* in correspondence analysis is used by analogy with the definition in applied mathematics of "moment of inertia," which stands for the integral of mass times the squared distance to the centroid (e.g., Greenacre, 1984, p. 35). Inertia is defined as the total Pearson [Chi-square](#) for the two-way divided by the total sum (193 in the present example)..

Inertia and row and column profiles. If the rows and columns in a table are completely independent of each other, the entries in the table (distribution of mass) can be reproduced from the row and column totals alone, or row and column *profiles* in the terminology of correspondence analysis. According to the well-known formula for computing the Chi-square statistic for two-way tables, the expected frequencies in a table, where the column and rows are independent of each other, are equal to the respective column total times the row total, divided by the grand total. Any deviations from the expected values (expected under the hypothesis of complete independence of the row and column variables) will contribute to the overall *Chi-square*. Thus, another way of looking at correspondence analysis is to consider it a method for decomposing the overall *Chi-square* statistic (or $Inertia = Chi\text{-}square / Total\ N$) by identifying a small number of dimensions in which the deviations from the expected values can be represented. This is similar to the goal of [Factor Analysis](#), where the total variance is decomposed, so as to arrive at a lower-dimensional representation of the variables that allows you to reconstruct most of the variance/covariance matrix of variables.

Analyzing rows and columns. This simple example began with a discussion of the row-points in the table shown above. However, you may rather be interested in the column totals, in which case you could plot the column points in a small-dimensional space, which satisfactorily reproduces the similarity (and distances) between the relative frequencies for the columns, across the rows, in the table shown above. In fact it is customary to simultaneously plot the

column points and the row points in a single graph, to summarize the information contained in a two-way table.

Reviewing results. Let's now look at some of the results for the table shown above. First, shown below are the so-called *singular values*, *eigenvalues*, *percentages of inertia explained*, *cumulative percentages*, and the contribution to the overall *Chi-square*.

Eigenvalues and Inertia for all Dimensions					
Input Table (Rows x Columns): 5 x 4					
Total Inertia = .08519 Chi ² = 16.442					
No. of Dims	Singular Values	Eigen-Values	Perc. of Inertia	Cumulativ Percent	Chi Squares
1	.273421	.074759	87.75587	87.7559	14.42851
2	.100086	.010017	11.75865	99.5145	1.93332
3	.020337	.000414	.48547	100.0000	.07982

Note that the dimensions are "extracted" so as to maximize the distances between the row or column points, and successive dimensions (which are independent of or orthogonal to each other) will "explain" less and less of the overall *Chi-square* value (and, thus, inertia). Thus, the extraction of the dimensions is similar to the extraction of *principal components* in *Factor Analysis*.

First, it appears that, with a single dimension, 87.76% of the inertia can be "explained," that is, the relative frequency values that can be reconstructed from a single dimension can reproduce 87.76% of the total *Chi-square* value (and, thus, of the inertia) for this two-way table; two dimensions allow you to explain 99.51%.

Maximum number of dimensions. Since the sums of the frequencies across the columns must be equal to the row totals, and the sums across the rows equal to the column totals, there are in a sense only (*no. of columns*-1) independent entries in each row, and (*no. of rows*-1) independent entries in each column of the table (once you know what these entries are, you can fill in the rest based on your knowledge of the column and row marginal totals). Thus, the maximum number of eigenvalues that can be extracted from a two-way table is equal to the minimum of the number of columns minus 1, and the number of rows minus 1. If you choose to extract (i.e., interpret) the maximum number of dimensions that can be extracted, then you can reproduce exactly all information contained in the table.

Row and column coordinates. Next look at the coordinates for the two-dimensional solution.

Row Name	Dim. 1	Dim. 2
(1) Senior Managers	-.065768	.193737
(2) Junior Managers	.258958	.243305
(3) Senior Employees	-.380595	.010660
(4) Junior Employees	.232952	-.057744
(5) Secretaries	-.201089	-.078911

Of course, you can plot these coordinates in a two-dimensional scatterplot. Remember that the purpose of correspondence analysis is to reproduce the distances between the row and/or column points in a two-way table in a lower-dimensional display; note that, as in *Factor Analysis*, the actual rotational orientation of the axes is arbitrarily chosen so that successive dimensions "explain" less and less of the overall *Chi-square* value (or *inertia*). You could, for example, reverse the signs in each column in the table shown above, thereby effectively rotating the respective axis in the plot by 180 degrees.

What is important are the distances of the points in the two-dimensional display, which are informative in that row points that are close to each other are similar with regard to the pattern of relative frequencies across the columns. If you have produced this plot you will see that, along the most important first axis in the plot, the *Senior employees* and *Secretaries* are relatively close together on the left side of the origin (scale position 0). If you looked at the table of relative row frequencies (i.e., frequencies standardized, so that their sum in each row is equal to 100%), you will see that these two groups of employees indeed show very similar patterns of

relative frequencies across the categories of smoking intensity.

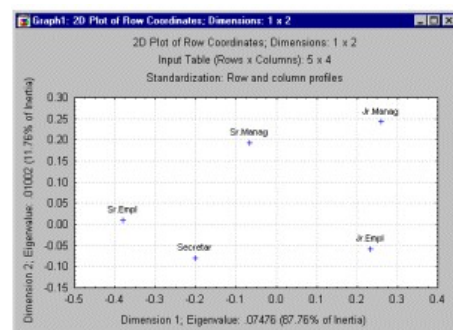
Percentages of Row Totals					
	Smoking Category				
Staff Group	(1) None	(2) Light	(3) Medium	(4) Heavy	Row Totals
(1) Senior Managers	36.36	18.18	27.27	18.18	100.00
(2) Junior Managers	22.22	16.67	38.89	22.22	100.00
(3) Senior Employees	49.02	19.61	23.53	7.84	100.00
(4) Junior Employees	20.45	27.27	37.50	14.77	100.00
(5) Secretaries	40.00	24.00	28.00	8.00	100.00

Obviously the final goal of correspondence analysis is to find theoretical interpretations (i.e., meaning) for the extracted dimensions. One method that may aid in interpreting extracted dimensions is to plot the column points. Shown below are the column coordinates for the first and second dimension.

Smoking category	Dim. 1	Dim. 2
None	-.393308	.030492
Light	.099456	-.141064
Medium	.196321	-.007359
Heavy	.293776	.197766

It appears that the first dimension distinguishes mostly between the different degrees of smoking, and in particular between category None and the others. Thus, you can interpret the greater similarity of *Senior Managers* with *Secretaries*, with regard to their position on the first axis, as mostly deriving from the relatively large numbers of *None* smokers in these two groups of employees.

Compatibility of row and column coordinates. It is customary to summarize the row and column coordinates in a single plot. However, it is important to remember that in such plots, you can only interpret the distances between row points, and the distances between column points, but not the distances between row points and column points.



To continue with this example, it would not be appropriate to say that the category *None* is similar to *Senior Employees* (the two points are very close in the simultaneous plot of row and column coordinates). However, as was indicated earlier, it is appropriate to make general statements about the nature of the dimensions, based on which side of the origin particular points fall. For example, because category *None* is the only column point on the left side of the origin for the first axis, and since employee group *Senior Employees* also falls onto that side of the first axis, you may conclude that the first axis separates *None* smokers from the other categories of smokers, and that *Senior Employees* are different from, for example, *Junior Employees*, in that there are relatively more non-smoking *Senior Employees*.

Scaling of the coordinates (standardization options). Another important decision that the analyst must make concerns the scaling of the coordinates. The nature of the choice pertains to whether or not you want to analyze the relative row percentages, column percentages, or both. In the context of the example described above, the row percentages were shown to illustrate

how the patterns of those percentages across the columns are similar for points which appear more closely together in the graphical display of the row coordinates. Put another way, the coordinates are based on the analysis of the *row profile matrix*, where the sum of the table entries in a row, across all columns, is equal to 1.0 (each entry r_{ij} in the row profile matrix can be interpreted as the conditional probability that a case belongs to column j , given its membership in row i). Thus, the coordinates are computed so as to maximize the differences between the points with respect to the row profiles (row percentages). The row coordinates are computed from the row profile matrix, the column coordinates are computed from the column profile matrix.

A fourth option, *Canonical* standardization (see Gifi, 1981), is also provided, and it amounts to a standardization of the columns and rows of the matrix of relative frequencies. This standardization amounts to a rescaling of the coordinates based on the row profile standardization and the column profile standardization, and this type of standardization is not widely used. Note also that a variety of other custom standardizations can be easily performed if you have the raw *eigenvalues* and *eigenvector* matrices.

Metric of coordinate system. In several places in this introduction, the term *distance* was (loosely) used to refer to the differences between the pattern of relative frequencies for the rows across the columns, and columns across the rows, which are to be reproduced in a lower-dimensional solution as a result of the correspondence analysis. Actually, these distances represented by the coordinates in the respective space are not simple Euclidean distances computed from the relative row or column frequencies, but rather, they are weighted distances. Specifically, the weighting that is applied is such that the metric in the lower-dimensional space is a *Chi-square* metric, provided that (1) you are comparing row points, and chose either row-profile standardization or both row- and column-profile standardization, or (2) you are comparing column points, and chose either column-profile standardization or both row- and column-profile standardization.

In that case (but *not* if you chose the canonical standardization), the squared Euclidean distance between, for example, two row points i and i' in the respective coordinate system of a given number of dimensions actually approximates a weighted (i.e., *Chi-square*) distance between the relative frequencies (see Hoffman and Franke, 1986, formula 21):

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

In this formula, $d_{ii'}$ stands for the squared distance between the two points, c_j stands for the column total for the j 'th column of the standardized frequency table (where the sum of all entries or mass is equal to 1.0), p_{ij} stands for the individual cell entries in the standardized frequency table (row i , column j), r_i stands for the row total for the i 'th row of the relative frequency table, and the summation \sum is over the columns of the table. To reiterate, only the distances between row points, and correspondingly, between column points are interpretable in this manner; the distances between row points and column points cannot be interpreted.

Judging the quality of a solution. A number of auxiliary statistics are reported, to aid in the evaluation of the *quality* of the respective chosen numbers of dimensions. The general concern here is that all (or at least most) points are properly represented by the respective solution, that is, that their distances to other points can be approximated to a satisfactory degree. Shown below are all statistics reported for the row coordinates for the example table discussed so far, based on a one-dimensional solution only (i.e., only one dimension is used to reconstruct the patterns of relative frequencies across the columns).

Row Coordinates and Contributions to Inertia						
Staff Group	Coordin. Dim. 1	Mass	Quality	Relative Inertia	Inertia Dim. 1	Cosine ² Dim. 1

(1) Senior Managers	-.065768	.056995	.092232	.031376	.003298	.092232
(2) Junior Managers	.258958	.093264	.526400	.139467	.083659	.526400
(3) Senior Employees	-.380595	.264249	.999033	.449750	.512006	.999033
(4) Junior Employees	.232952	.455959	.941934	.308354	.330974	.941934
(5) Secretaries	-.201089	.129534	.865346	.071053	.070064	.865346

Coordinates. The first numeric column shown in the table above contains the coordinates, as discussed in the previous paragraphs. To reiterate, the specific interpretation of these coordinates depends on the standardization chosen for the solution (see above). The number of dimensions is chosen by the user (in this case we chose only one dimension), and coordinate values will be shown for each dimension (i.e., there will be one column with coordinate values for each dimension).

Mass. The *Mass* column contains the row totals (since these are the row coordinates) for the table of relative frequencies (i.e., for the table where each entry is the respective *mass*, as discussed earlier in this section). Remember that the coordinates are computed based on the matrix of conditional probabilities shown in the *Mass* column.

Quality. The *Quality* column contains information concerning the quality of representation of the respective row point in the coordinate system defined by the respective numbers of dimensions, as chosen by the user. In the table shown above, only one dimension was chosen, and the numbers in the *Quality* column pertain to the quality of representation in the one-dimensional space. To reiterate, computationally, the goal of the correspondence analysis is to reproduce the distances between points in a low-dimensional space. If you extracted (i.e., interpreted) the maximum number of dimensions (which is equal to the minimum of the number of rows and the number of columns, minus 1), you could reconstruct all distances exactly. The *Quality* of a point is defined as the ratio of the squared distance of the point from the origin in the chosen number of dimensions, over the squared distance from the origin in the space defined by the maximum number of dimensions (remember that the metric here is *Chi-square*, as described earlier). By analogy to [Factor Analysis](#), the quality of a point is similar in its interpretation to the communality for a variable in factor analysis. .

Note that the *Quality* measure reported is independent of the chosen method of standardization, and always pertains to the default standardization (i.e., the distance metric is *Chi-square*, and the quality measure can be interpreted as the "proportion of *Chi-square* accounted for" for the respective row, given the respective number of dimensions). A low quality means that the current number of dimensions does not well represent the respective row (or column). In the table shown above, the quality for the first row (*Senior Managers*) is less than .1, indicating that this row point is not well represented by the one-dimensional representation of the points.

Relative inertia. The *Quality* of a point (see above) represents the proportion of the contribution of that point to the overall *inertia* (*Chi-square*) that can be accounted for by the chosen number of dimensions. However, it does not indicate whether or not, and to what extent, the respective point does in fact contribute to the overall inertia (*Chi-square* value). The relative inertia represents the proportion of the total inertia accounted for by the respective point, and it is independent of the number of dimensions chosen by the user. Note that a particular solution may represent a point very well (high *Quality*), but the same point may not contribute much to the overall inertia (e.g., a row point with a pattern of relative frequencies across the columns that is similar to the average pattern across all rows).

Relative inertia for each dimension. This column contains the relative contribution of the respective (row) point to the *inertia* "accounted for" by the respective dimension. Thus, this value will be reported for each (row or column) point, for each dimension.

Cosine² (quality or squared correlations with each dimension). This column contains the quality for each point, by dimension. The sum of the values in these columns across the

dimensions is equal to the total *Quality* value discussed above (since in the example table above, only one dimension was chose, the values in this column are identical to the values in the overall *Quality* column). This value may also be interpreted as the "correlation" of the respective point with the respective dimension. The term *Cosine*² refers to the fact that this value is also the squared cosine value of the angle the point makes with the respective dimension (refer to Greenacre, 1984, for details concerning the geometric aspects of correspondence analysis).

A note about "statistical significance." It should be noted at this point that correspondence analysis is an exploratory technique. Actually, the method was developed based on a philosophical orientation that emphasizes the development of models that fit the data, rather than the rejection of hypotheses based on the lack of fit (Benzecri's "second principle" states that "The model must fit the data, not vice versa;" see Greenacre, 1984, p. 10). Therefore, there are no statistical significance tests that are customarily applied to the results of a correspondence analysis; the primary purpose of the technique is to produce a simplified (low- dimensional) representation of the information in a large frequency table (or tables with similar measures of correspondence).

[To index](#)

Supplementary Points

The introductory section provides an overview of how to interpret the coordinates and related statistics computed in a correspondence analysis. An important aid in the interpretation of the results from a correspondence analysis is to include supplementary row or column points, that were not used to perform the original analyses. For example, consider the following results which are based on the example given in the introductory (based on Greenacre, 1984).

Row Name	Dim. 1	Dim. 2
(1) Senior Managers	-.065768	.193737
(2) Junior Managers	.258958	.243305
(3) Senior Employees	-.380595	.010660
(4) Junior Employees	.232952	-.057744
(5) Secretaries	-.201089	-.078911
National Average	-.258368	-.117648

The table above shows the coordinate values (for two dimensions) computed for a frequency table of different types of employees by type of smoking habit. The row labeled *National Average* contains the coordinate values for the supplementary point, which is the national average (percentages) for the different smoking categories (which make up the columns of the table; those fictitious percentages reported in Greenacre (1984) are: Nonsmokers: 42%, light smokers: 29%, medium smokers, 20%; heavy smokers: 9%). If you plotted these coordinates in a two-dimensional scatterplot, along with the column coordinates, it would be apparent that the *National Average* supplementary row point is plotted close to the point representing the Secretaries group, and on the same side of the horizontal axis (first dimension) as the *Nonsmokers* column point. If you refer back to the original two-way table shown in the [introductory section](#), this finding is consistent with the entries in the table of row frequencies, that is, there are relatively more nonsmokers among the *Secretaries*, and in the *National Average*. Put another way, the sample represented in the original frequency table contains more smokers than the national average.

While this type of information could have been easily gleaned from the original frequency table (that was used as the input to the analysis), in the case of very large tables, such conclusions may not be as obvious.

Quality. Another interesting result for supplementary points concerns the [quality](#) of their representation in the chosen number of dimensions (see the [introductory section](#) for a more detailed discussion of the concept of *quality*). To reiterate, the goal of the correspondence analysis is to reproduce the distances between the row or column coordinates (patterns of relative frequencies across the columns or rows, respectively) in a low-dimensional solution.

Given such a solution, you may ask whether particular supplementary points of interest can be represented equally well in the final space, that is, whether or not their distances from the other points in the table can also be represented in the chosen numbers of dimensions. Shown below are the summary statistics for the original points, and the supplementary row point *National Average*, for the two-dimensional solution of representation of representation of supplementary points.

Staff Group	Quality	Cosine ² Dim. 1	Cosine ² Dim. 2
(1) Senior Managers	.892568	.092232	.800336
(2) Junior Managers	.991082	.526400	.464682
(3) Senior Employees	.999817	.999033	.000784
(4) Junior Employees	.999810	.941934	.057876
(5) Secretaries	.998603	.865346	.133257
National Average	.761324	.630578	.130746

The statistics reported in the table above are discussed in the [introductory section](#). In short, the [Quality](#) of a row or column point is defined as the ratio of the squared distance of the point from the origin in the chosen number of dimensions, over the squared distance from the origin in the space defined by the maximum number of dimensions (remember that the metric here is [Chi-square](#), as described in the introductory section). In a sense, the overall quality is the "proportion of squared distance-from-the-overall-centroid accounted for." The supplementary row point *National Average* has a quality of .76, indicating that it is reasonably well represented in the two-dimensional solution. The *Cosine²* statistic is the quality "accounted for" by the respective row point, by the respective dimension (the sum of the *Cosine²* values over the respective number of dimensions is equal to the total *Quality*, see also the [introductory section](#)).

[To index](#)

Multiple Correspondence Analysis (MCA)

Multiple correspondence analysis (MCA) may be considered to be an extension of simple correspondence analysis to more than two variables. For an introductory overview of simple correspondence analysis, refer to the [introductory section](#). Multiple correspondence analysis is a simple correspondence analysis carried out on an indicator (or design) matrix with cases as rows and categories of variables as columns. Actually, one usually analyzes the inner product of such a matrix, called the *Burt Table* in an MCA; this will be discussed later. However, to clarify the interpretation of the results from a multiple correspondence analysis, it is easier to discuss the simple correspondence analysis of an indicator or design matrix.

Indicator or design matrix. Consider again the simple two-way table presented in the [introductory section](#):

	Smoking Category				
Staff Group	(1) None	(2) Light	(3) Medium	(4) Heavy	Row Totals
(1) Senior Managers	4	2	3	2	11
(2) Junior Managers	4	3	7	4	18
(3) Senior Employees	25	10	12	4	51
(4) Junior Employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
Column Totals	61	45	62	25	193

Suppose you had entered the data for this table in the following manner, as an *indicator or design* matrix:

	Staff Group					Smoking			
Case Number	Senior Manager	Junior Manager	Senior Employee	Junior Employee	Secretary	None	Light	Medium	Heavy
1	1	0	0	0	0	1	0	0	0
2	1	0	0	0	0	1	0	0	0
3	1	0	0	0	0	1	0	0	0
4	1	0	0	0	0	1	0	0	0

5	1	0	0	0	0	0	1	0	0
...
...
...
191	0	0	0	0	1	0	0	1	0
192	0	0	0	0	1	0	0	0	1
193	0	0	0	0	1	0	0	0	1

Each one of the 193 total cases in the table is represented by one case in this data file. For each case a 1 is entered into the category where the respective case "belongs," and a 0 otherwise. For example, case 1 represents a *Senior Manager* who is a *None* smoker. As can be seen in the table above, there are a total of 4 such cases in the two-way table, and thus there will be four cases like this in the indicator matrix. In all, there will be 193 cases in the indicator or design matrix.

Analyzing the design matrix. If you now analyzed this data file (design or indicator matrix) shown above as if it were a two-way frequency table, the results of the correspondence analysis would provide column coordinates that would allow you to relate the different categories to each other, based on the distances between the row points, i.e., between the individual cases. In fact, the two-dimensional display you would obtain for the column coordinates would look very similar to the combined display for row and column coordinates, if you had performed the simple correspondence analysis on the two-way frequency table (note that the metric will be different, but the relative positions of the points will be very similar).

More than two variables. The approach to analyzing categorical data outlined above can easily be extended to more than two categorical variables. For example, the indicator or design matrix could contain two additional variables *Male* and *Female*, again coded 0 and 1, to indicate the subjects' gender; and three variables could be added to indicate to which one of three age groups a case belongs. Thus, in the final display, you could represent the relationships (similarities) between *Gender*, *Age*, *Smoking habits*, and *Occupation (Staff Groups)*.

Fuzzy coding. It is not necessary that each case is assigned exclusively to only one category of each categorical variable. Rather than the 0-or-1 coding scheme, you could enter probabilities for membership in a category, or some other measure that represents a fuzzy rule for group membership. Greenacre (1984) discusses different types of coding schemes of this kind. For example, suppose in the example design matrix shown earlier, you had missing data for a few cases regarding their smoking habits. Instead of discarding those cases entirely from the analysis (or creating a new category *Missing data*), you could assign to the different smoking categories proportions (which should add to 1.0) to represent the probabilities that the respective case belongs to the respective category (e.g., you could enter proportions based on your knowledge about estimates for the national averages for the different categories).

Interpretation of coordinates and other results. To reiterate, the results of a multiple correspondence analysis are identical to the results you would obtain for the column coordinates from a simple correspondence analysis of the design or indicator matrix. Therefore, the interpretation of coordinate values, [quality](#) values, \cos^2 's and other statistics reported as the results from a multiple correspondence analysis can be interpreted in the same manner as described in the context of the simple correspondence analysis (see [introductory section](#)), however, these statistics pertain to the total [inertia](#) associated with the entire design matrix.

Supplementary column points and "multiple regression" for categorical variables. Another application of the analysis of design matrices via correspondence analysis techniques is that it allows you to perform the equivalent of a [Multiple Regression](#) for categorical variables, by adding supplementary columns to the design matrix. For example, suppose you added to the design matrix shown earlier two columns to indicate whether or not the respective subject had or had not been ill over the past year (i.e., you could add one column *Ill* and another column *Not ill*, and again enter 0's and 1's to indicate each subject's health status). If, in a simple

correspondence analysis of the design matrix, you added those columns as supplementary columns to the analysis, then (1) the summary statistics for the quality of representation (see the [introductory section](#)) for those columns would give you an indication of how well you can "explain" illness as a function of the other variables in the design matrix, and (2) the display of the column points in the final coordinate system would provide an indication of the nature (e.g., direction) of the relationships between the columns in the design matrix and the column points indicating illness; this technique (adding supplementary points to an MCA analysis) is also sometimes called [predictive mapping](#).

The Burt table. The actual computations in multiple correspondence analysis are not performed on a design or indicator matrix (which, potentially, may be very large if there are many cases), but on the inner product of this matrix; this matrix is also called the *Burt* matrix. With frequency tables, this amounts to tabulating the stacked categories against each other; for example the *Burt* for the two-way frequency table presented earlier would look like this.

	Employee					Smoking			
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)
(1) Senior Managers	11	0	0	0	0	4	2	3	2
(2) Junior Managers	0	18	0	0	0	4	3	7	4
(3) Senior Employees	0	0	51	0	0	25	10	12	4
(4) Junior Employees	0	0	0	88	0	18	24	33	13
(5) Secretaries	0	0	0	0	25	10	6	7	2
(1) Smoking:None	4	4	25	18	10	61	0	0	0
(2) Smoking:Light	2	3	10	24	6	0	45	0	0
(3) Smoking:Medium	3	7	12	33	7	0	0	62	0
(4) Smoking:Heavy	2	4	4	13	2	0	0	0	25

The *Burt* has a clearly defined structure. In the case of two categorical variables (shown above), it consists of 4 partitions: (1) the crosstabulation of variable *Employee* against itself, (2) the crosstabulation of variable *Employee* against variable *Smoking*, (3), the crosstabulation of variable *Smoking* against variable *Employee*, and (4) the crosstabulation of variable *Smoking* against itself. Note that the matrix is symmetrical, and that the sum of the diagonal elements in each partition representing the crosstabulation of a variable against itself must be the same (e.g., there were a total of 193 observations in the present example, and hence, the diagonal elements in the crosstabulation tables of variable *Employee* against itself, and *Smoking* against itself must also be equal to 193).

Note that the off-diagonal elements in the partitions representing the crosstabulations of a variable against itself are equal to 0 in the table shown above. However, this is not necessarily always the case, for example, when the *Burt* was derived from a design or indicator matrix that included fuzzy coding of category membership (see above).

[To index](#)

Burt Tables

The Burt table is the result of the inner product of a design or indicator matrix, and the multiple correspondence analysis results are identical to the results one would obtain for the column points from a simple correspondence analysis of the indicator or design matrix (see also [MCA](#)).

For example, suppose you had entered data concerning the *Survival* for different *Age* groups in different *Locations* like this:

Case No.	SURVIVAL		AGE			LOCATION		
	NO	YES	LESST50	A50TO69	OVER69	TOKYO	BOSTON	GLAMORGAN
1	0	1	0	1	0	0	0	1
2	1	0	1	0	0	1	0	0
3	0	1	0	1	0	0	1	0
4	0	1	0	0	1	0	0	1
...
...

...
762	1	0	0	1	0	1	0	0
763	0	1	1	0	0	0	1	0
764	0	1	0	1	0	0	0	1

In this data arrangement, for each case a 1 was entered to indicate to which category, of a particular set of categories, a case belongs (e.g., *Survival*, with the categories *No* and *Yes*). For example, case 1 survived (a 0 was entered for variable *No*, and a 1 was entered for variable *Yes*), case 1 is between age 50 and 69 (a 1 was entered for variable *A50to69*), and was observed in *Glamorgn*). Overall there are 764 observations in the data set.

If you denote the data (design or indicator matrix) shown above as matrix **X**, then matrix product **X'X** is a *Burt* table); shown below is an example of a *Burt* table that one might obtain in this manner.

	SURVIVAL		AGE			LOCATION		
	NO	YES	<50	50-69	69+	TOKYO	BOSTON	GLAMORGN
SURVIVAL:NO	210	0	68	93	49	60	82	68
SURVIVAL:YES	0	554	212	258	84	230	171	153
AGE:UNDER_50	68	212	280	0	0	151	58	71
AGE:A_50TO69	93	258	0	351	0	120	122	109
AGE:OVER_69	49	84	0	0	133	19	73	41
LOCATION:TOKYO	60	230	151	120	19	290	0	0
LOCATION:BOSTON	82	171	58	122	73	0	253	0
LOCATION:GLAMORGN	68	153	71	109	41	0	0	221

The *Burt* table has a clearly defined structure. Overall, the data matrix is symmetrical. In the case of 3 categorical variables (as shown above), the data matrix consists $3 \times 3 = 9$ partitions, created by each variable being tabulated against itself, and against the categories of all other variables. Note that the sum of the diagonal elements in each diagonal partition (i.e., where the respective variables are tabulated against themselves) is constant (equal to 764 in this case).

The off-diagonal elements in each diagonal partition in this example are all 0. If the cases in the design or indicator matrix are assigned to categories via fuzzy coding (i.e., if probabilities are used to indicate likelihood of membership in a category, rather than 0/1 coding to indicate actual membership), then the off-diagonal elements of the diagonal partitions are not necessarily equal to 0.

[To index](#)