IBM SPSS Categories 21

Jacqueline J. Meulman

Willem J. Heiser

Note: Before using this information and the product it supports, read the general information under Notices on p. 301.

This edition applies to IBM® SPSS® Statistics 21 and to all subsequent releases and modifications until otherwise indicated in new editions.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Preface

IBM® SPSS® Statistics is a comprehensive system for analyzing data. The Categories optional add-on module provides the additional analytic techniques described in this manual. The Categories add-on module must be used with the SPSS Statistics Core system and is completely integrated into that system.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative visit *http://www.ibm.com/spss*.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using IBM Corp. products or for installation help for one of the supported hardware environments. To reach Technical Support, see the IBM Corp. web site at *http://www.ibm.com/support*. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Technical Support for Students

If you're a student using a student, academic or grad pack version of any IBM SPSS software product, please see our special online Solutions for Education (*http://www.ibm.com/spss/rd/students/*) pages for students. If you're a student using a university-supplied copy of the IBM SPSS software, please contact the IBM SPSS product coordinator at your university.

Customer Service

If you have any questions concerning your shipment or account, contact your local office. Please have your serial number ready for identification.

Training Seminars

IBM Corp. provides both public and onsite training seminars. All seminars feature hands-on workshops. Seminars will be offered in major cities on a regular basis. For more information on these seminars, go to *http://www.ibm.com/software/analytics/spss/training*.

Acknowledgements

The optimal scaling procedures and their implementation in IBM® SPSS® Statistics were developed by the Data Theory Scaling System Group (DTSS), consisting of members of the departments of Education and Psychology of the Faculty of Social and Behavioral Sciences at Leiden University.

Willem Heiser, Jacqueline Meulman, Gerda van den Berg, and Patrick Groenen were involved with the original 1990 procedures. Jacqueline Meulman and Peter Neufeglise participated in the development of procedures for categorical regression, correspondence analysis, categorical principal components analysis, and multidimensional scaling. In addition, Anita van der Kooij contributed especially to CATREG, CORRESPONDENCE, and CATPCA. Willem Heiser, Jacques Commandeur, Frank Busing, Gerda van den Berg, and Patrick Groenen participated in the development of the PROXSCAL procedure. Frank Busing, Willem Heiser, Patrick Groenen, and Peter Neufeglise participated in the development of the PREFSCAL procedure.

Contents

Part I: User's Guide

Part II: Examples

9 Categorical Regression 94

10 Categorical Principal Components Analysis 137

11 Nonlinear Canonical Correlation Analysis 186

15 Multidimensional Unfolding 258

Appendices

Part I:
User's Guide

Chapter

1

Introduction to Optimal Scaling Procedures for Categorical Data

Categories procedures use optimal scaling to analyze data that are difficult or impossible for standard statistical procedures to analyze. This chapter describes what each procedure does, the situations in which each procedure is most appropriate, the relationships between the procedures, and the relationships of these procedures to their standard statistical counterparts.

Note: These procedures and their implementation in IBM® SPSS® Statistics were developed by the Data Theory Scaling System Group (DTSS), consisting of members of the departments of Education and Psychology, Faculty of Social and Behavioral Sciences, Leiden University.

What Is Optimal Scaling?

The idea behind optimal scaling is to assign numerical quantifications to the categories of each variable, thus allowing standard procedures to be used to obtain a solution on the quantified variables.

The optimal scale values are assigned to categories of each variable based on the optimizing criterion of the procedure in use. Unlike the original labels of the nominal or ordinal variables in the analysis, these scale values have metric properties.

In most Categories procedures, the optimal quantification for each scaled variable is obtained through an iterative method called **alternating least squares** in which, after the current quantifications are used to find a solution, the quantifications are updated using that solution. The updated quantifications are then used to find a new solution, which is used to update the quantifications, and so on, until some criterion is reached that signals the process to stop.

Why Use Optimal Scaling?

Categorical data are often found in marketing research, survey research, and research in the social and behavioral sciences. In fact, many researchers deal almost exclusively with categorical data.

While adaptations of most standard models exist specifically to analyze categorical data, they often do not perform well for datasets that feature:

- Too few observations
- Too many variables
- Too many values per variable

By quantifying categories, optimal scaling techniques avoid problems in these situations. Moreover, they are useful even when specialized techniques are appropriate.

Rather than interpreting parameter estimates, the interpretation of optimal scaling output is often based on graphical displays. Optimal scaling techniques offer excellent exploratory analyses, which complement other IBM® SPSS® Statistics models well. By narrowing the focus of your investigation, visualizing your data through optimal scaling can form the basis of an analysis that centers on interpretation of model parameters.

Optimal Scaling Level and Measurement Level

This can be a very confusing concept when you first use Categories procedures. When specifying the level, you specify not the level at which variables are *measured* but the level at which they are *scaled*. The idea is that the variables to be quantified may have nonlinear relations regardless of how they are measured.

For Categories purposes, there are three basic levels of measurement:

- The **nominal** level implies that a variable's values represent unordered categories. Examples of variables that might be nominal are region, zip code area, religious affiliation, and multiple choice categories.
- The **ordinal** level implies that a variable's values represent ordered categories. Examples include attitude scales representing degree of satisfaction or confidence and preference rating scores.
- The **numerical** level implies that a variable's values represent ordered categories with a meaningful metric so that distance comparisons between categories are appropriate. Examples include age in years and income in thousands of dollars.

For example, suppose that the variables *region*, *job*, and *age* are coded as shown in the following table.

Table 1-1
Coding scheme for region, job, and age

Region code	Region value	Job code	Job value	Age
1	North	1	intern	20
2	South	2	sales rep	22
3	East	3	manager	25
4	West			27

The values shown represent the categories of each variable. *Region* would be a nominal variable. There are four categories of *region*, with no intrinsic ordering. Values 1 through 4 simply represent the four categories; the coding scheme is completely arbitrary. *Job*, on the other hand, could be assumed to be an ordinal variable. The original categories form a progression from intern to manager. Larger codes represent a job higher on the corporate ladder. However, only the order information is known—nothing can be said about the distance between adjacent categories. In contrast, *age* could be assumed to be a numerical variable. In the case of *age*, the distances between the values are intrinsically meaningful. The distance between 20 and 22 is the same as the distance between 25 and 27, while the distance between 22 and 25 is greater than either of these.

Selecting the Optimal Scaling Level

It is important to understand that there are no intrinsic properties of a variable that automatically predefine what optimal scaling level you should specify for it. You can explore your data in any way that makes sense and makes interpretation easier. By analyzing a numerical-level variable at the ordinal level, for example, the use of a nonlinear transformation may allow a solution in fewer dimensions.

The following two examples illustrate how the "obvious" level of measurement might not be the best optimal scaling level. Suppose that a variable sorts objects into age groups. Although age can be scaled as a numerical variable, it may be true that for people younger than 25 safety has a positive relation with age, whereas for people older than 60 safety has a negative relation with age. In this case, it might be better to treat age as a nominal variable.

As another example, a variable that sorts persons by political preference appears to be essentially nominal. However, if you order the parties from political left to political right, you might want the quantification of parties to respect this order by using an ordinal level of analysis.

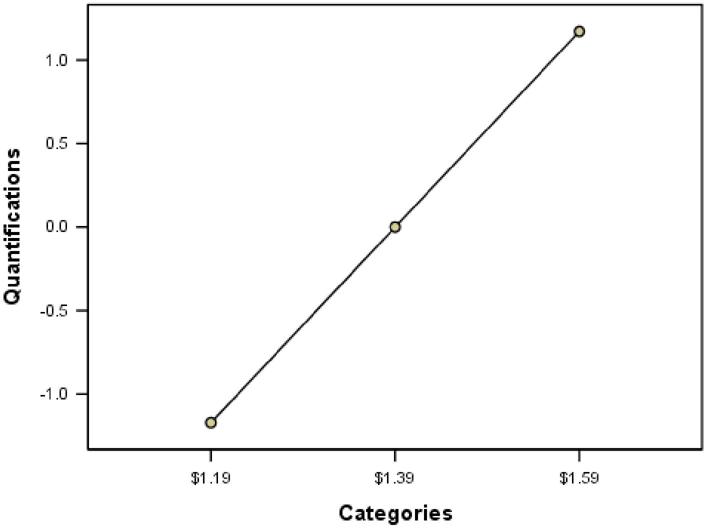
Even though there are no predefined properties of a variable that make it exclusively one level or another, there are some general guidelines to help the novice user. With single-nominal quantification, you don't usually know the order of the categories but you want the analysis to impose one. If the order of the categories is known, you should try ordinal quantification. If the categories are unorderable, you might try multiple-nominal quantification.

Transformation Plots

The different levels at which each variable can be scaled impose different restrictions on the quantifications. Transformation plots illustrate the relationship between the quantifications and the original categories resulting from the selected optimal scaling level. For example, a linear transformation plot results when a variable is treated as numerical. Variables treated as ordinal result in a nondecreasing transformation plot. Transformation plots for variables treated nominally that are U-shaped (or the inverse) display a quadratic relationship. Nominal variables could also yield transformation plots without apparent trends by changing the order of the categories completely. The following figure displays a sample transformation plot.

Transformation plots are particularly suited to determining how well the selected optimal scaling level performs. If several categories receive similar quantifications, collapsing these categories into one category may be warranted. Alternatively, if a variable treated as nominal receives quantifications that display an increasing trend, an ordinal transformation may result in a similar fit. If that trend is linear, numerical treatment may be appropriate. However, if collapsing categories or changing scaling levels is warranted, the analysis will not change significantly.

Figure 1-1
Transformation plot of price (numerical)



Category Codes

Some care should be taken when coding categorical variables because some coding schemes may yield unwanted output or incomplete analyses. Possible coding schemes for *job* are displayed in the following table.

Table 1-2
Alternative coding schemes for job

Category	**A**	**B**	**C**	**D**
intern	1	1	5	1
sales rep	2	2	6	5
manager	3	7	7	3

Some Categories procedures require that the range of every variable used be defined. Any value outside this range is treated as a missing value. The minimum category value is always 1. The maximum category value is supplied by the user. This value is not the *number* of categories for a variable—it is the *largest* category value. For example, in the table, scheme A has a maximum category value of 3 and scheme B has a maximum category value of 7, yet both schemes code the same three categories.

The variable range determines which categories will be omitted from the analysis. Any categories with codes outside the defined range are omitted from the analysis. This is a simple method for omitting categories but can result in unwanted analyses. An incorrectly defined maximum category can omit *valid* categories from the analysis. For example, for scheme B, defining the maximum category value to be 3 indicates that *job* has categories coded from 1 to 3; the *manager* category is treated as missing. Because no category has actually been coded 3, the third category in the analysis contains no cases. If you wanted to omit all manager categories, this analysis would be appropriate. However, if managers are to be included, the maximum category must be defined as 7, and missing values must be coded with values above 7 or below 1.

For variables treated as nominal or ordinal, the range of the categories does not affect the results. For nominal variables, only the label and not the value associated with that label is important. For ordinal variables, the order of the categories is preserved in the quantifications; the category values themselves are not important. All coding schemes resulting in the same category ordering will have identical results. For example, the first three schemes in the table are functionally equivalent if *job* is analyzed at an ordinal level. The order of the categories is identical in these schemes. Scheme D, on the other hand, inverts the second and third categories and will yield different results than the other schemes.

Although many coding schemes for a variable are functionally equivalent, schemes with small differences between codes are preferred because the codes have an impact on the amount of output produced by a procedure. All categories coded with values between 1 and the user-defined maximum are valid. If any of these categories are empty, the corresponding quantifications will be either system-missing or 0, depending on the procedure. Although neither of these assignments affect the analyses, output is produced for these categories. Thus, for scheme B, *job* has four categories that receive system-missing values. For scheme C, there are also four categories receiving system-missing indicators. In contrast, for scheme A there are no system-missing quantifications. Using consecutive integers as codes for variables treated as nominal or ordinal results in much less output without affecting the results.

Coding schemes for variables treated as numerical are more restricted than the ordinal case. For these variables, the differences between consecutive categories are important. The following table displays three coding schemes for *age*.

Table 1-3
Alternative coding schemes for age

Category	A	B	C
20	20	1	1
22	22	3	2
25	25	6	3
27	27	8	4

Any recoding of numerical variables must preserve the differences between the categories. Using the original values is one method for ensuring preservation of differences. However, this can result in many categories having system-missing indicators. For example, scheme A employs the original observed values. For all Categories procedures except for Correspondence Analysis, the maximum category value is 27 and the minimum category value is set to 1. The first 19 categories are empty and receive system-missing indicators. The output can quickly become rather cumbersome if the maximum category is much greater than 1 and there are many empty categories between 1 and the maximum.

To reduce the amount of output, recoding can be done. However, in the numerical case, the Automatic Recode facility should not be used. Coding to consecutive integers results in differences of 1 between all consecutive categories, and, as a result, all quantifications will be equally spaced. The metric characteristics deemed important when treating a variable as numerical are destroyed by recoding to consecutive integers. For example, scheme C in the table corresponds to automatically recoding *age*. The difference between categories 22 and 25 has changed from three to one, and the quantifications will reflect the latter difference.

An alternative recoding scheme that preserves the differences between categories is to subtract the smallest category value from every category and add 1 to each difference. Scheme B results from this transformation. The smallest category value, 20, has been subtracted from each category, and 1 was added to each result. The transformed codes have a minimum of 1, and all differences are identical to the original data. The maximum category value is now 8, and the zero quantifications before the first nonzero quantification are all eliminated. Yet, the nonzero quantifications corresponding to each category resulting from scheme B are identical to the quantifications from scheme A.

Which Procedure Is Best for Your Application?

The techniques embodied in four of these procedures (Correspondence Analysis, Multiple Correspondence Analysis, Categorical Principal Components Analysis, and Nonlinear Canonical Correlation Analysis) fall into the general area of multivariate data analysis known as **dimension reduction**. That is, relationships between variables are represented in a few dimensions—say two or three—as often as possible. This enables you to describe structures or patterns in the relationships that would be too difficult to fathom in their original richness and complexity. In market research applications, these techniques can be a form of **perceptual mapping**. A major advantage of these procedures is that they accommodate data with different levels of optimal scaling.

Categorical Regression describes the relationship between a categorical response variable and a combination of categorical predictor variables. The influence of each predictor variable on the response variable is described by the corresponding regression weight. As in the other procedures, data can be analyzed with different levels of optimal scaling.

Multidimensional Scaling and Multidimensional Unfolding describe relationships between objects in a low-dimensional space, using the proximities between the objects.

Following are brief guidelines for each of the procedures:

- Use Categorical Regression to predict the values of a categorical dependent variable from a combination of categorical independent variables.
- Use Categorical Principal Components Analysis to account for patterns of variation in a single set of variables of mixed optimal scaling levels.
- Use Nonlinear Canonical Correlation Analysis to assess the extent to which two or more sets of variables of mixed optimal scaling levels are correlated.
- Use Correspondence Analysis to analyze two-way contingency tables or data that can be expressed as a two-way table, such as brand preference or sociometric choice data.
- Use Multiple Correspondence Analysis to analyze a categorical multivariate data matrix when you are willing to make no stronger assumption that all variables are analyzed at the nominal level.
- Use Multidimensional Scaling to analyze proximity data to find a least-squares representation of a single set of objects in a low-dimensional space.
- Use Multidimensional Unfolding to analyze proximity data to find a least-squares representation of two sets of objects in a low-dimensional space.

Categorical Regression

The use of Categorical Regression is most appropriate when the goal of your analysis is to predict a dependent (response) variable from a set of independent (predictor) variables. As with all optimal scaling procedures, scale values are assigned to each category of every variable such that these values are optimal with respect to the regression. The solution of a categorical regression maximizes the squared correlation between the transformed response and the weighted combination of transformed predictors.

Relation to other Categories procedures. Categorical regression with optimal scaling is comparable to optimal scaling canonical correlation analysis with two sets, one of which contains only the dependent variable. In the latter technique, similarity of sets is derived by comparing each set to an unknown variable that lies somewhere between all of the sets. In categorical regression, similarity of the transformed response and the linear combination of transformed predictors is assessed directly.

Relation to standard techniques. In standard linear regression, categorical variables can either be recoded as indicator variables or be treated in the same fashion as interval level variables. In the first approach, the model contains a separate intercept and slope for each combination of the levels of the categorical variables. This results in a large number of parameters to interpret. In the second approach, only one parameter is estimated for each variable. However, the arbitrary nature of the category codings makes generalizations impossible.

If some of the variables are not continuous, alternative analyses are available. If the response is continuous and the predictors are categorical, analysis of variance is often employed. If the response is categorical and the predictors are continuous, logistic regression or discriminant analysis may be appropriate. If the response and the predictors are both categorical, loglinear models are often used.

Regression with optimal scaling offers three scaling levels for each variable. Combinations of these levels can account for a wide range of nonlinear relationships for which any single "standard" method is ill-suited. Consequently, optimal scaling offers greater flexibility than the standard approaches with minimal added complexity.

In addition, nonlinear transformations of the predictors usually reduce the dependencies among the predictors. If you compare the eigenvalues of the correlation matrix for the predictors with the eigenvalues of the correlation matrix for the optimally scaled predictors, the latter set will usually be less variable than the former. In other words, in categorical regression, optimal scaling makes the larger eigenvalues of the predictor correlation matrix smaller and the smaller eigenvalues larger.

Categorical Principal Components Analysis

The use of Categorical Principal Components Analysis is most appropriate when you want to account for patterns of variation in a single set of variables of mixed optimal scaling levels. This technique attempts to reduce the dimensionality of a set of variables while accounting for as much of the variation as possible. Scale values are assigned to each category of every variable so that these values are optimal with respect to the principal components solution. Objects in the analysis receive component scores based on the quantified data. Plots of the component scores reveal patterns among the objects in the analysis and can reveal unusual objects in the data. The solution of a categorical principal components analysis maximizes the correlations of the object scores with each of the quantified variables for the number of components (dimensions) specified.

An important application of categorical principal components is to examine preference data, in which respondents rank or rate a number of items with respect to preference. In the usual IBM® SPSS® Statistics data configuration, rows are individuals, columns are measurements for the items, and the scores across rows are preference scores (on a 0 to 10 scale, for example), making the data row-conditional. For preference data, you may want to treat the individuals as variables. Using the Transpose procedure, you can transpose the data. The raters become the variables, and all variables are declared ordinal. There is no objection to using more variables than objects in CATPCA.

Relation to other Categories procedures. If all variables are declared multiple nominal, categorical principal components analysis produces an analysis equivalent to a multiple correspondence analysis run on the same variables. Thus, categorical principal components analysis can be seen as a type of multiple correspondence analysis in which some of the variables are declared ordinal or numerical.

Relation to standard techniques. If all variables are scaled on the numerical level, categorical principal components analysis is equivalent to standard principal components analysis.

More generally, categorical principal components analysis is an alternative to computing the correlations between non-numerical scales and analyzing them using a standard principal components or factor-analysis approach. Naive use of the usual Pearson correlation coefficient as a measure of association for ordinal data can lead to nontrivial bias in estimation of the correlations.

Nonlinear Canonical Correlation Analysis

Nonlinear Canonical Correlation Analysis is a very general procedure with many different applications. The goal of nonlinear canonical correlation analysis is to analyze the relationships between two or more sets of variables instead of between the variables themselves, as in principal components analysis. For example, you may have two sets of variables, where one set of variables might be demographic background items on a set of respondents and a second set might be responses to a set of attitude items. The scaling levels in the analysis can be any mix of nominal, ordinal, and numerical. Optimal scaling canonical correlation analysis determines the similarity among the sets by simultaneously comparing the canonical variables from each set to a compromise set of scores assigned to the objects.

Relation to other Categories procedures. If there are two or more sets of variables with only one variable per set, optimal scaling canonical correlation analysis is equivalent to optimal scaling principal components analysis. If all variables in a one-variable-per-set analysis are multiple nominal, optimal scaling canonical correlation analysis is equivalent to multiple correspondence analysis. If there are two sets of variables, one of which contains only one variable, optimal scaling canonical correlation analysis is equivalent to categorical regression with optimal scaling.

Relation to standard techniques. Standard canonical correlation analysis is a statistical technique that finds a linear combination of one set of variables and a linear combination of a second set of variables that are maximally correlated. Given this set of linear combinations, canonical correlation analysis can find subsequent independent sets of linear combinations, referred to as canonical variables, up to a maximum number equal to the number of variables in the smaller set.

If there are two sets of variables in the analysis and all variables are defined to be numerical, optimal scaling canonical correlation analysis is equivalent to a standard canonical correlation analysis. Although IBM® SPSS® Statistics does not have a canonical correlation analysis procedure, many of the relevant statistics can be obtained from multivariate analysis of variance.

Optimal scaling canonical correlation analysis has various other applications. If you have two sets of variables and one of the sets contains a nominal variable declared as single nominal, optimal scaling canonical correlation analysis results can be interpreted in a similar fashion to regression analysis. If you consider the variable to be multiple nominal, the optimal scaling analysis is an alternative to discriminant analysis. Grouping the variables in more than two sets provides a variety of ways to analyze your data.

Correspondence Analysis

The goal of correspondence analysis is to make biplots for correspondence tables. In a correspondence table, the row and column variables are assumed to represent unordered categories; therefore, the nominal optimal scaling level is always used. Both variables are inspected for their nominal information only. That is, the only consideration is the fact that some objects are in the same category while others are not. Nothing is assumed about the distance or order between categories of the same variable.

One specific use of correspondence analysis is the analysis of two-way contingency tables. If a table has *r* active rows and *c* active columns, the number of dimensions in the correspondence analysis solution is the minimum of *r* minus 1 or *c* minus 1, whichever is less. In other words, you could perfectly represent the row categories or the column categories of a contingency table in a space of dimensions. Practically speaking, however, you would like to represent the row and column categories of a two-way table in a low-dimensional space, say two dimensions, for the reason that two-dimensional plots are more easily comprehensible than multidimensional spatial representations.

When fewer than the maximum number of possible dimensions is used, the statistics produced in the analysis describe how well the row and column categories are represented in the low-dimensional representation. Provided that the quality of representation of the two-dimensional solution is good, you can examine plots of the row points and the column points to learn which categories of the row variable are similar, which categories of the column variable are similar, and which row and column categories are similar to each other.

Relation to other Categories procedures. Simple correspondence analysis is limited to two-way tables. If there are more than two variables of interest, you can combine variables to create interaction variables. For example, for the variables *region*, *job*, and *age*, you can combine *region* and *job* to create a new variable *rejob* with the 12 categories shown in the following table. This new variable forms a two-way table with *age* (12 rows, 4 columns), which can be analyzed in correspondence analysis.

Table 1-4
Combinations of region and job

Category code	Category definition	Category code	Category definition
1	North, intern	7	East, intern
2	North, sales rep	8	East, sales rep
3	North, manager	9	East, manager

Category code	Category definition	Category code	Category definition
4	South, intern	10	West, intern
5	South, sales rep	11	West, sales rep
6	South, manager	12	West, manager

One shortcoming of this approach is that any pair of variables can be combined. We can combine *job* and *age*, yielding another 12-category variable. Or we can combine *region* and *age*, which results in a new 16-category variable. Each of these interaction variables forms a two-way table with the remaining variable. Correspondence analyses of these three tables will not yield identical results, yet each is a valid approach. Furthermore, if there are four or more variables, two-way tables comparing an interaction variable with another interaction variable can be constructed. The number of possible tables to analyze can get quite large, even for a few variables. You can select one of these tables to analyze, or you can analyze all of them. Alternatively, the Multiple Correspondence Analysis procedure can be used to examine all of the variables simultaneously without the need to construct interaction variables.

Relation to standard techniques. The Crosstabs procedure can also be used to analyze contingency tables, with independence as a common focus in the analyses. However, even in small tables, detecting the cause of departures from independence may be difficult. The utility of correspondence analysis lies in displaying such patterns for two-way tables of any size. If there is an association between the row and column variables—that is, if the chi-square value is significant—correspondence analysis may help reveal the nature of the relationship.

Multiple Correspondence Analysis

Multiple Correspondence Analysis tries to produce a solution in which objects within the same category are plotted close together and objects in different categories are plotted far apart. Each object is as close as possible to the category points of categories that apply to the object. In this way, the categories divide the objects into homogeneous subgroups. Variables are considered homogeneous when they classify objects in the same categories into the same subgroups.

For a one-dimensional solution, multiple correspondence analysis assigns optimal scale values (category quantifications) to each category of each variable in such a way that overall, on average, the categories have maximum spread. For a two-dimensional solution, multiple correspondence analysis finds a second set of quantifications of the categories of each variable unrelated to the first set, attempting again to maximize spread, and so on. Because categories of a variable receive as many scorings as there are dimensions, the variables in the analysis are assumed to be multiple nominal in optimal scaling level.

Multiple correspondence analysis also assigns scores to the objects in the analysis in such a way that the category quantifications are the averages, or centroids, of the object scores of objects in that category.

Relation to other Categories procedures. Multiple correspondence analysis is also known as homogeneity analysis or dual scaling. It gives comparable, but not identical, results to correspondence analysis when there are only two variables. Correspondence analysis produces unique output summarizing the fit and quality of representation of the solution, including stability information. Thus, correspondence analysis is usually preferable to multiple correspondence analysis in the two-variable case. Another difference between the two procedures is that the input to multiple correspondence analysis is a data matrix, where the rows are objects and the columns are variables, while the input to correspondence analysis can be the same data matrix, a general proximity matrix, or a joint contingency table, which is an aggregated matrix in which both the rows and columns represent categories of variables. Multiple correspondence analysis can also be thought of as principal components analysis of data scaled at the multiple nominal level.

Relation to standard techniques. Multiple correspondence analysis can be thought of as the analysis of a multiway contingency table. Multiway contingency tables can also be analyzed with the Crosstabs procedure, but Crosstabs gives separate summary statistics for each category of each control variable. With multiple correspondence analysis, it is often possible to summarize the relationship between all of the variables with a single two-dimensional plot. An advanced use of multiple correspondence analysis is to replace the original category values with the optimal scale values from the first dimension and perform a secondary multivariate analysis. Since multiple correspondence analysis replaces category labels with numerical scale values, many different procedures that require numerical data can be applied after the multiple correspondence analysis. For example, the Factor Analysis procedure produces a first principal component that is equivalent to the first dimension of multiple correspondence analysis. The component scores in the first dimension are equal to the object scores, and the squared component loadings are equal to the discrimination measures. The second multiple correspondence analysis dimension, however, is not equal to the second dimension of factor analysis.

Multidimensional Scaling

The use of Multidimensional Scaling is most appropriate when the goal of your analysis is to find the structure in a set of distance measures between a single set of objects or cases. This is accomplished by assigning observations to specific locations in a conceptual low-dimensional space so that the distances between points in the space match the given (dis)similarities as closely as possible. The result is a least-squares representation of the objects in that low-dimensional space, which, in many cases, will help you further understand your data.

Relation to other Categories procedures. When you have multivariate data from which you create distances and which you then analyze with multidimensional scaling, the results are similar to analyzing the data using categorical principal components analysis with object principal normalization. This kind of PCA is also known as principal coordinates analysis.

Relation to standard techniques. The Categories Multidimensional Scaling procedure (PROXSCAL) offers several improvements upon the scaling procedure available in the Statistics Base option (ALSCAL). PROXSCAL offers an accelerated algorithm for certain models and allows you to put restrictions on the common space. Moreover, PROXSCAL attempts to minimize normalized raw stress rather than S-stress (also referred to as **strain**). The normalized raw stress is generally preferred because it is a measure based on the distances, while the S-stress is based on the squared distances.

Multidimensional Unfolding

The use of Multidimensional Unfolding is most appropriate when the goal of your analysis is to find the structure in a set of distance measures between two sets of objects (referred to as the row and column objects). This is accomplished by assigning observations to specific locations in a conceptual low-dimensional space so that the distances between points in the space match the given (dis)similarities as closely as possible. The result is a least-squares representation of the row and column objects in that low-dimensional space, which, in many cases, will help you further understand your data.

Relation to other Categories procedures. If your data consist of distances between a single set of objects (a square, symmetrical matrix), use Multidimensional Scaling.

Relation to standard techniques. The Categories Multidimensional Unfolding procedure (PREFSCAL) offers several improvements upon the unfolding functionality available in the Statistics Base option (through ALSCAL). PREFSCAL allows you to put restrictions on the common space; moreover, PREFSCAL attempts to minimize a penalized stress measure that helps it to avoid degenerate solutions (to which older algorithms are prone).

Aspect Ratio in Optimal Scaling Charts

Aspect ratio in optimal scaling plots is isotropic. In a two-dimensional plot, the distance representing one unit in dimension 1 is equal to the distance representing one unit in dimension 2. If you change the range of a dimension in a two-dimensional plot, the system changes the size of the other dimension to keep the physical distances equal. Isotropic aspect ratio cannot be overridden for the optimal scaling procedures.

Recommended Readings

See the following texts for general information on optimal scaling techniques:

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, and H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: *Methodologies of Pattern Recognition,* S. Watanabe, ed. New York: Academic Press, 35–74.

Bishop, Y. M., S. E. Feinberg, and P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. In: *Data Analysis and Informatics III,* E. Diday, et al., ed., 415–424.

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. In: *Progress in Multivariate Analysis,* S. Das Gupta, and J. Sethuraman, eds. Calcutta: Indian Statistical Institute.

De Leeuw, J., and J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. In: *Data Analysis and Informatics,* E. Diday,et al., ed. Amsterdam: North-Holland, 231–242.

De Leeuw, J., F. W. Young, and Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 471–503.

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Heiser, W. J., and J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. In: *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski, ed. Oxford: Oxford UniversityPress, 51–89.

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Krzanowski, W. J., and F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. London: Edward Arnold.

Lebart, L., A. Morineau, and K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, 7–12.

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, 539–565.

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In: *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah, ed. Amsterdam: North-Holland, 3–22.

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, 287–315.

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, 357–387.

Chapter

2

Categorical Regression (CATREG)

Categorical regression quantifies categorical data by assigning numerical values to the categories, resulting in an optimal linear regression equation for the transformed variables. Categorical regression is also known by the acronym CATREG, for *cat*egorical *reg*ression.

Standard linear regression analysis involves minimizing the sum of squared differences between a response (dependent) variable and a weighted combination of predictor (independent) variables. Variables are typically quantitative, with (nominal) categorical data recoded to binary or contrast variables. As a result, categorical variables serve to separate groups of cases, and the technique estimates separate sets of parameters for each group. The estimated coefficients reflect how changes in the predictors affect the response. Prediction of the response is possible for any combination of predictor values.

An alternative approach involves regressing the response on the categorical predictor values themselves. Consequently, one coefficient is estimated for each variable. However, for categorical variables, the category values are arbitrary. Coding the categories in different ways yield different coefficients, making comparisons across analyses of the same variables difficult.

CATREG extends the standard approach by simultaneously scaling nominal, ordinal, and numerical variables. The procedure quantifies categorical variables so that the quantifications reflect characteristics of the original categories. The procedure treats quantified categorical variables in the same way as numerical variables. Using nonlinear transformations allow variables to be analyzed at a variety of levels to find the best-fitting model.

Example. Categorical regression could be used to describe how job satisfaction depends on job category, geographic region, and amount of travel. You might find that high levels of satisfaction correspond to managers and low travel. The resulting regression equation could be used to predict job satisfaction for any combination of the three independent variables.

Statistics and plots. Frequencies, regression coefficients, ANOVA table, iteration history, category quantifications, correlations between untransformed predictors, correlations between transformed predictors, residual plots, and transformation plots.

Data. CATREG operates on category indicator variables. The category indicators should be positive integers. You can use the Discretization dialog box to convert fractional-value variables and string variables into positive integers.

Assumptions. Only one response variable is allowed, but the maximum number of predictor variables is 200. The data must contain at least three valid cases, and the number of valid cases must exceed the number of predictor variables plus one.

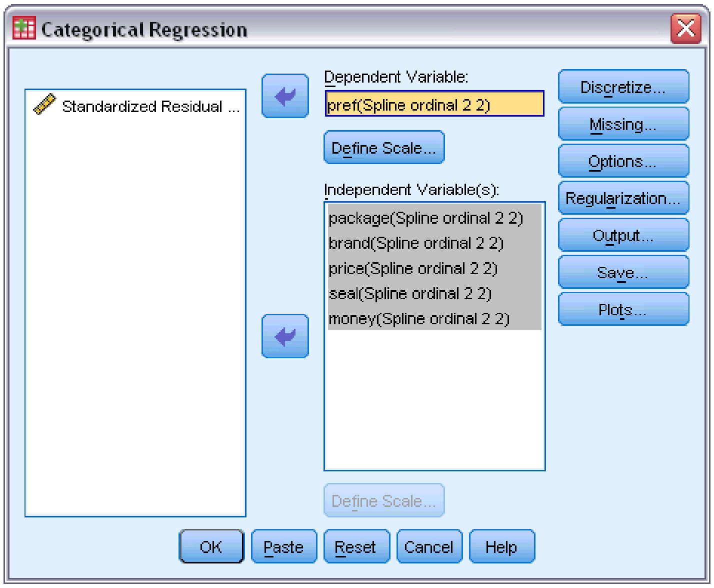
Related procedures. CATREG is equivalent to categorical canonical correlation analysis with optimal scaling (OVERALS) with two sets, one of which contains only one variable. Scaling all variables at the numerical level corresponds to standard multiple regression analysis.

To Obtain a Categorical Regression

- From the menus choose:

  Analyze > Regression > Optimal Scaling (CATREG)...

Figure 2-1
Categorical Regression dialog box



Dependent Variable: pref(Spline ordinal 2 2)

Independent Variable(s):
package(Spline ordinal 2 2)
brand(Spline ordinal 2 2)
price(Spline ordinal 2 2)
seal(Spline ordinal 2 2)
money(Spline ordinal 2 2)

- Select the dependent variable and independent variable(s).

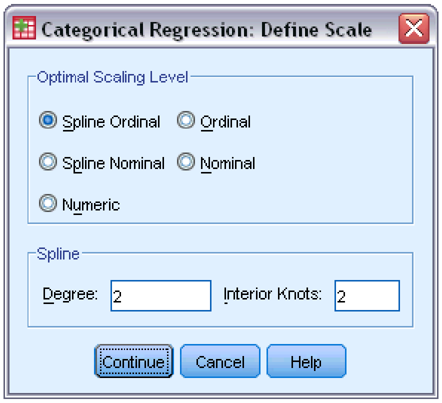
- Click OK.

  Optionally, change the scaling level for each variable.

Define Scale in Categorical Regression

You can set the optimal scaling level for the dependent and independent variables. By default, they are scaled as second-degree monotonic splines (ordinal) with two interior knots. Additionally, you can set the weight for analysis variables.

Figure 2-2
Define Scale dialog box



Categorical Regression: Define Scale

Optimal Scaling Level

● Spline Ordinal ○ Ordinal

○ Spline Nominal ○ Nominal

○ Numeric

Spline

Degree: 2 Interior Knots: 2

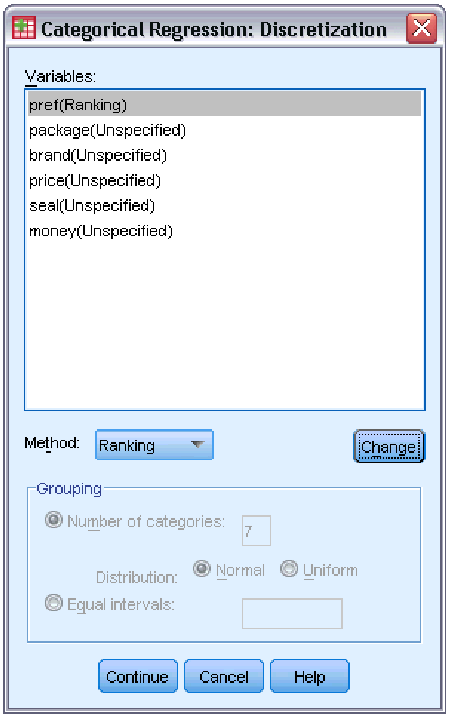
Optimal Scaling Level. You can also select the scaling level for quantifying each variable.

- **Spline Ordinal.** The order of the categories of the observed variable is preserved in the optimally scaled variable. Category points will be on a straight line (vector) through the origin. The resulting transformation is a smooth monotonic piecewise polynomial of the chosen degree. The pieces are specified by the user-specified number and procedure-determined placement of the interior knots.
- **Spline Nominal.** The only information in the observed variable that is preserved in the optimally scaled variable is the grouping of objects in categories. The order of the categories of the observed variable is not preserved. Category points will be on a straight line (vector) through the origin. The resulting transformation is a smooth, possibly nonmonotonic, piecewise polynomial of the chosen degree. The pieces are specified by the user-specified number and procedure-determined placement of the interior knots.
- **Ordinal.** The order of the categories of the observed variable is preserved in the optimally scaled variable. Category points will be on a straight line (vector) through the origin. The resulting transformation fits better than the spline ordinal transformation but is less smooth.
- **Nominal.** The only information in the observed variable that is preserved in the optimally scaled variable is the grouping of objects in categories. The order of the categories of the observed variable is not preserved. Category points will be on a straight line (vector) through the origin. The resulting transformation fits better than the spline nominal transformation but is less smooth.
- **Numeric.** Categories are treated as ordered and equally spaced (interval level). The order of the categories and the equal distances between category numbers of the observed variable are preserved in the optimally scaled variable. Category points will be on a straight line (vector) through the origin. When all variables are at the numeric level, the analysis is analogous to standard principal components analysis.

Categorical Regression Discretization

The Discretization dialog box allows you to select a method of recoding your variables. Fractional-value variables are grouped into seven categories (or into the number of distinct values of the variable if this number is less than seven) with an approximately normal distribution unless otherwise specified. String variables are always converted into positive integers by assigning category indicators according to ascending alphanumeric order. Discretization for string variables applies to these integers. Other variables are left alone by default. The discretized variables are then used in the analysis.

Figure 2-3
Discretization dialog box



Variables:
pref(Ranking)
package(Unspecified)
brand(Unspecified)
price(Unspecified)
seal(Unspecified)
money(Unspecified)

Method: Ranking

Grouping
● Number of categories: 7
Distribution: ● Normal ○ Uniform
○ Equal intervals:

Method. Choose between grouping, ranking, and multiplying.

- **Grouping.** Recode into a specified number of categories or recode by interval.
- **Ranking.** The variable is discretized by ranking the cases.
- **Multiplying.** The current values of the variable are standardized, multiplied by 10, rounded, and have a constant added so that the lowest discretized value is 1.

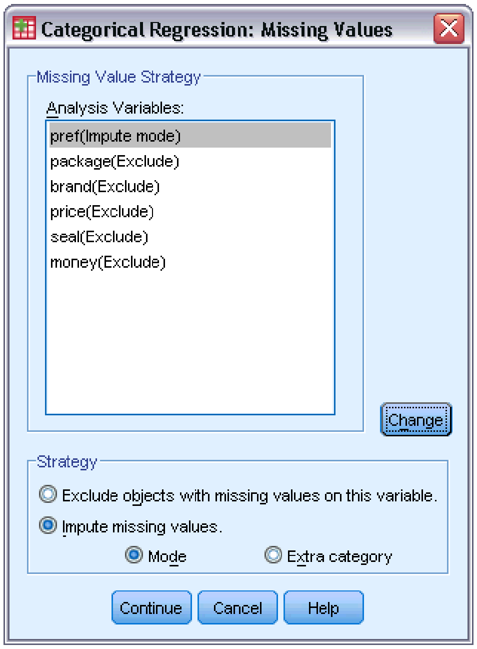
Grouping. The following options are available when discretizing variables by grouping:

- **Number of categories.** Specify a number of categories and whether the values of the variable should follow an approximately normal or uniform distribution across those categories.
- **Equal intervals.** Variables are recoded into categories defined by these equally sized intervals. You must specify the length of the intervals.

Categorical Regression Missing Values

The Missing Values dialog box allows you to choose the strategy for handling missing values in analysis variables and supplementary variables.

Figure 2-4
Missing Values dialog box



Categorical Regression: Missing Values

Missing Value Strategy

Analysis Variables:

pref(Impute mode)
package(Exclude)
brand(Exclude)
price(Exclude)
seal(Exclude)
money(Exclude)

Strategy

○ Exclude objects with missing values on this variable.
● Impute missing values.
● Mode ○ Extra category

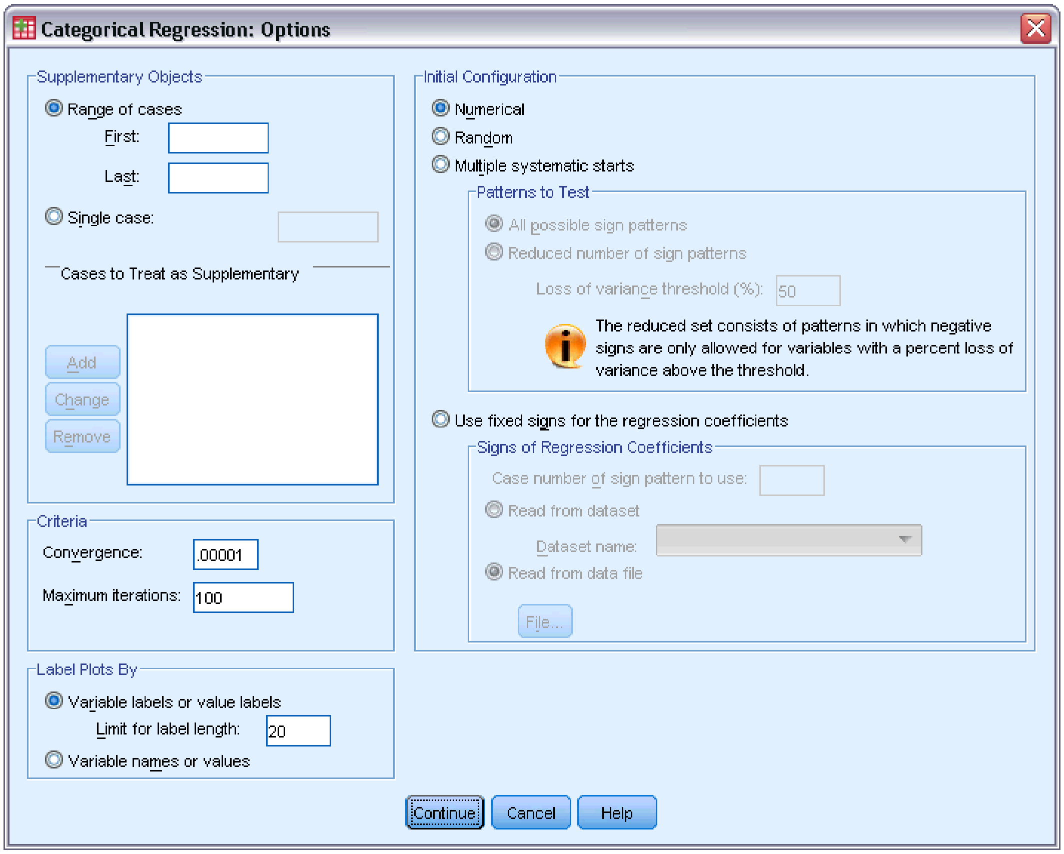
Strategy. Choose to exclude objects with missing values (listwise deletion) or impute missing values (active treatment).

- **Exclude objects with missing values on this variable.** Objects with missing values on the selected variable are excluded from the analysis. This strategy is not available for supplementary variables.
- **Impute missing values.** Objects with missing values on the selected variable have those values imputed. You can choose the method of imputation. Select Mode to replace missing values with the most frequent category. When there are multiple modes, the one with the smallest category indicator is used. Select Extra category to replace missing values with the same quantification of an extra category. This implies that objects with a missing value on this variable are considered to belong to the same (extra) category.

Categorical Regression Options

The Options dialog box allows you to select the initial configuration style, specify iteration and convergence criteria, select supplementary objects, and set the labeling of plots.

Figure 2-5
Options dialog box



Categorical Regression: Options

Supplementary Objects

● Range of cases
First:
Last:
○ Single case:

Cases to Treat as Supplementary

Criteria
Convergence: .00001
Maximum iterations: 100

Label Plots By
● Variable labels or value labels
Limit for label length: 20
○ Variable names or values

Initial Configuration
● Numerical
○ Random
○ Multiple systematic starts

Patterns to Test
○ All possible sign patterns
○ Reduced number of sign patterns
Loss of variance threshold (%): 50

The reduced set consists of patterns in which negative signs are only allowed for variables with a percent loss of variance above the threshold.

○ Use fixed signs for the regression coefficients

Signs of Regression Coefficients
Case number of sign pattern to use:
○ Read from dataset
Dataset name:
○ Read from data file

Supplementary Objects. This allows you to specify the objects that you want to treat as supplementary. Simply type the number of a supplementary object (or specify a range of cases) and click Add. You cannot weight supplementary objects (specified weights are ignored).

Initial Configuration. If no variables are treated as nominal, select the Numerical configuration. If at least one variable is treated as nominal, select the Random configuration.

Alternatively, if at least one variable has an ordinal or spline ordinal scaling level, the usual model-fitting algorithm can result in a suboptimal solution. Choosing Multiple systematic starts with all possible sign patterns to test will always find the optimal solution, but the necessary processing time rapidly increases as the number of ordinal and spline ordinal variables in the dataset increase. You can reduce the number of test patterns by specifying a percentage of loss of variance threshold, where the higher the threshold, the more sign patterns will be excluded. With this option, obtaining the optimal solution is not garantueed, but the chance of obtaining a suboptimal solution is diminished. Also, if the optimal solution is not found, the chance that the suboptimal solution is very different from the optimal solution is diminished. When multiple systematic starts are requested, the signs of the regression coefficients for each start are written to an external IBM® SPSS® Statistics data file or dataset in the current session. For more information, see the topic Categorical Regression Save on p. 23.

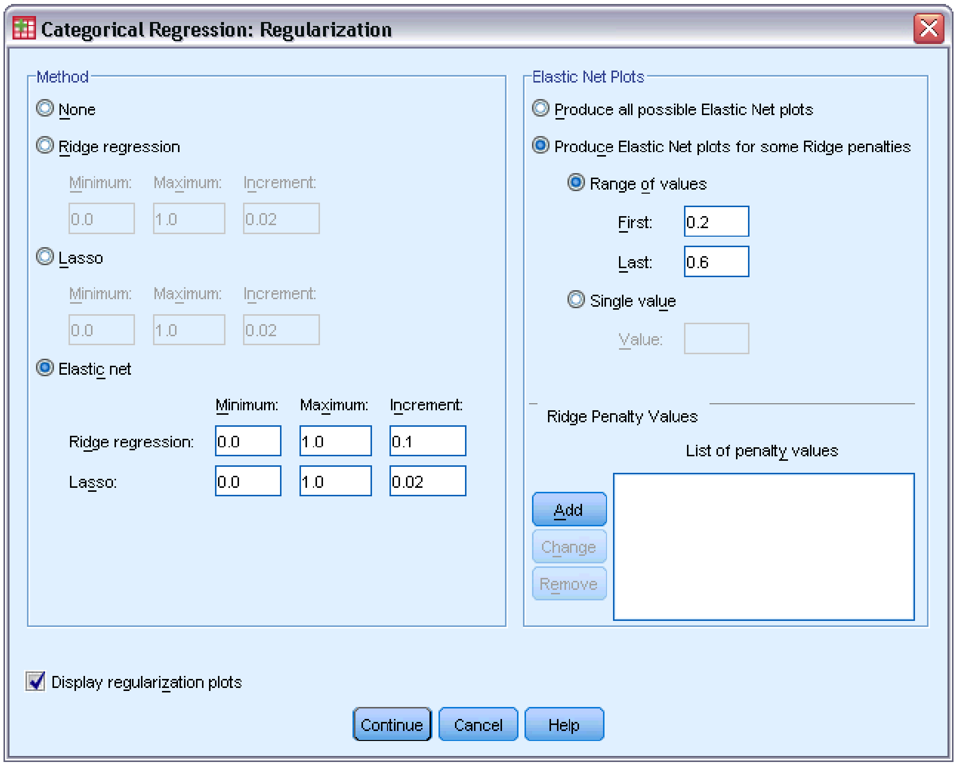
The results of a previous run with multiple systematic starts allows you to Use fixed signs for the regression coefficients. The signs (indicated by 1 and −1) need to be in a row of the specified dataset or file. The integer-valued starting number is the case number of the row in this file that contains the signs to be used.

Criteria. You can specify the maximum number of iterations that the regression may go through in its computations. You can also select a convergence criterion value. The regression stops iterating if the difference in total fit between the last two iterations is less than the convergence value or if the maximum number of iterations is reached.

Label Plots By. Allows you to specify whether variables and value labels or variable names and values will be used in the plots. You can also specify a maximum length for labels.

Categorical Regression Regularization

Figure 2-6
Regularization dialog box



Categorical Regression: Regularization

Method

○ None

○ Ridge regression

Minimum: 0.0 Maximum: 1.0 Increment: 0.02

○ Lasso

Minimum: 0.0 Maximum: 1.0 Increment: 0.02

● Elastic net

	Minimum:	Maximum:	Increment:
Ridge regression:	0.0	1.0	0.1
Lasso:	0.0	1.0	0.02

Elastic Net Plots

○ Produce all possible Elastic Net plots

● Produce Elastic Net plots for some Ridge penalties

● Range of values

First: 0.2

Last: 0.6

○ Single value

Value:

Ridge Penalty Values

List of penalty values

☑ Display regularization plots

Method. Regularization methods can improve the predictive error of the model by reducing the variability in the estimates of regression coefficient by shrinking the estimates toward 0. The Lasso and Elastic Net will shrink some coefficient estimates to exactly 0, thus providing a form of variable selection. When a regularization method is requested, the regularized model and coefficients for each penalty coefficient value are written to an external IBM® SPSS® Statistics data file or dataset in the current session. For more information, see the topic Categorical Regression Save on p. 23.

- **Ridge regression.** Ridge regression shrinks coefficients by introducing a penalty term equal to the sum of squared coefficients times a **penalty coefficient**. This coefficient can range from 0 (no penalty) to 1; the procedure will search for the "best" value of the penalty if you specify a range and increment.
- **Lasso.** The Lasso's penalty term is based on the sum of absolute coefficients, and the specification of a penalty coefficient is similar to that of Ridge regression; however, the Lasso is more computationally intensive.
- **Elastic net.** The Elastic Net simply combines the Lasso and Ridge regression penalties, and will search over the grid of values specified to find the "best" Lasso and Ridge regression penalty coefficients. For a given pair of Lasso and Ridge regression penalties, the Elastic Net is not much more computationally expensive than the Lasso.

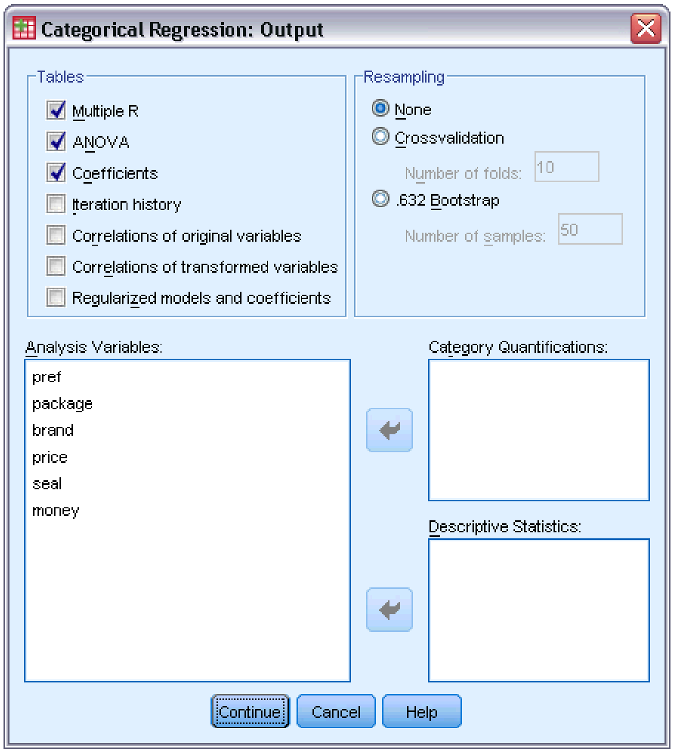
Display regularization plots. These are plots of the regression coefficients versus the regularization penalty. When searching a range of values for the "best" penalty coefficient, it provides a view of how the regression coefficients change over that range.

Elastic Net Plots. For the Elastic Net method, separate regularization plots are produced by values of the Ridge regression penalty. All possible plots uses every value in the range determined by the minimum and maximum Ridge regression penalty values specified. For some Ridge penalties allows you to specify a subset of the values in the range determined by the minimum and maximum. Simply type the number of a penalty value (or specify a range of values) and click Add.

Categorical Regression Output

The Output dialog box allows you to select the statistics to display in the output.

Figure 2-7
Output dialog box



Categorical Regression: Output

Tables

☑ Multiple R
☑ ANOVA
☑ Coefficients
□ Iteration history
□ Correlations of original variables
□ Correlations of transformed variables
□ Regularized models and coefficients

Resampling

● None
○ Crossvalidation
Number of folds: 10
○ .632 Bootstrap
Number of samples: 50

Analysis Variables:
pref
package
brand
price
seal
money

Category Quantifications:

Descriptive Statistics:

Tables. Produces tables for:

- **Multiple R.** Includes R^2, adjusted R^2, and adjusted R^2 taking the optimal scaling into account.
- **ANOVA.** This option includes regression and residual sums of squares, mean squares, and F. Two ANOVA tables are displayed: one with degrees of freedom for the regression equal to the number of predictor variables and one with degrees of freedom for the regression taking the optimal scaling into account.
- **Coefficients.** This option gives three tables: a Coefficients table that includes betas, standard error of the betas, t values, and significance; a Coefficients-Optimal Scaling table with the standard error of the betas taking the optimal scaling degrees of freedom into account; and a table with the zero-order, part, and partial correlation, Pratt's relative importance measure for the transformed predictors, and the tolerance before and after transformation.
- **Iteration history.** For each iteration, including the starting values for the algorithm, the multiple R and regression error are shown. The increase in multiple R is listed starting from the first iteration.
- **Correlations of original variables.** A matrix showing the correlations between the untransformed variables is displayed.

- **Correlations of transformed variables.** A matrix showing the correlations between the transformed variables is displayed.
- **Regularized models and coefficients.** Displays penalty values, R-square, and the regression coefficients for each regularized model. If a resampling method is specified or if supplementary objects (test cases) are specified, it also displays the prediction error or test MSE.

Resampling. Resampling methods give you an estimate of the prediction error of the model.

- **Crossvalidation.** Crossvalidation divides the sample into a number of subsamples, or folds. Categorical regression models are then generated, excluding the data from each subsample in turn. The first model is based on all of the cases except those in the first sample fold, the second model is based on all of the cases except those in the second sample fold, and so on. For each model, the prediction error is estimated by applying the model to the subsample excluded in generating it.
- **.632 Bootstrap.** With the bootstrap, observations are drawn randomly from the data with replacement, repeating this process a number of times to obtain a number bootstrap samples. A model is fit for each bootstrap sample, and the prediction error for each model is estimated by this fitted model is then applied to the cases not in the bootstrap sample.

Category Quantifications. Tables showing the transformed values of the selected variables are displayed.

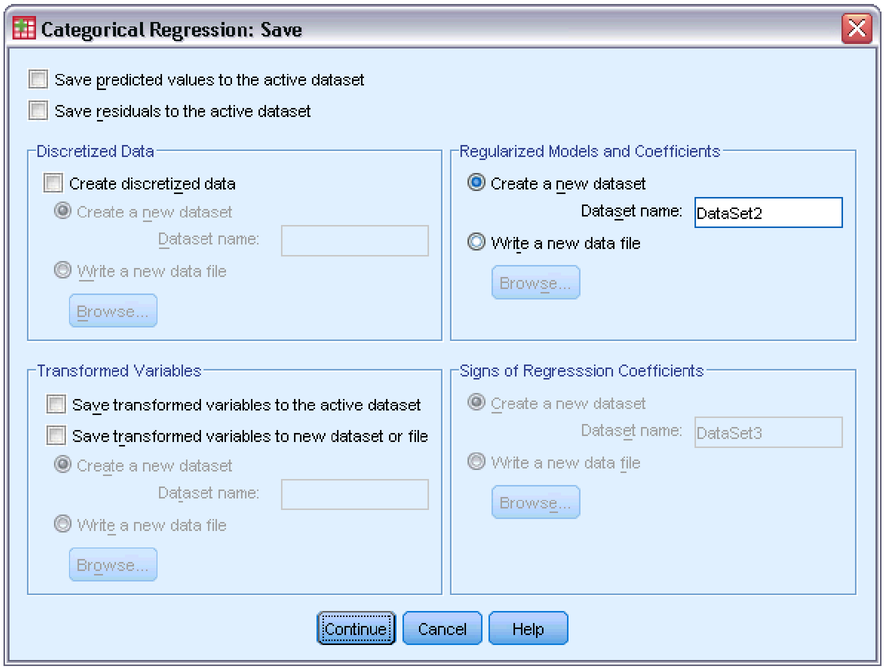
Descriptive Statistics. Tables showing the frequencies, missing values, and modes of the selected variables are displayed.

Categorical Regression Save

The Save dialog box allows you to save predicted values, residuals, and transformed values to the active dataset and/or save discretized data, transformed values, regularized models and coefficients, and signs of regression coefficients to an external IBM® SPSS® Statistics data file or dataset in the current session.

- Datasets are available during the current session but are not available in subsequent sessions unless you explicitly save them as data files. Dataset names must adhere to variable naming rules.
- Filenames or dataset names must be different for each type of data saved.

Figure 2-8
Save dialog box



Categorical Regression: Save

□ Save predicted values to the active dataset

□ Save residuals to the active dataset

Discretized Data

□ Create discretized data

● Create a new dataset

Dataset name:

○ Write a new data file

Regularized Models and Coefficients

● Create a new dataset

Dataset name: DataSet2

○ Write a new data file

Transformed Variables

□ Save transformed variables to the active dataset

□ Save transformed variables to new dataset or file

● Create a new dataset

Dataset name:

○ Write a new data file

Signs of Regresssion Coefficients

● Create a new dataset

Dataset name: DataSet3

○ Write a new data file

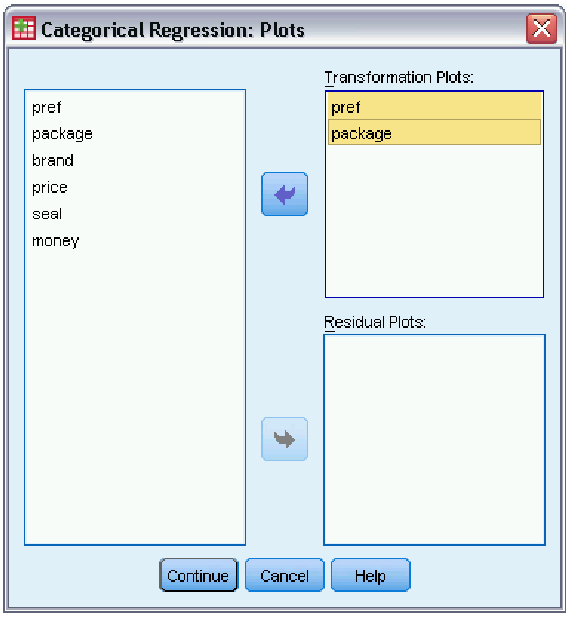
Regularized models and coefficients are saved whenever a regularization method is selected on the Regularization dialog. By default, the procedure creates a new dataset with a unique name, but you can of course specify a name of your own choosing or write to an external file.

Signs of regression coefficients are saved whenever multiple systematic starts are used as the initial configuration on the Options dialog. By default, the procedure creates a new dataset with a unique name, but you can of course specify a name of your own choosing or write to an external file.

Categorical Regression Transformation Plots

The Plots dialog box allows you to specify the variables that will produce transformation and residual plots.

Figure 2-9
Plots dialog box



Transformation Plots. For each of these variables, the category quantifications are plotted against the original category values. Empty categories appear on the horizontal axis but do not affect the computations. These categories are identified by breaks in the line connecting the quantifications.

Residual Plots. For each of these variables, residuals (computed for the dependent variable predicted from all predictor variables except the predictor variable in question) are plotted against category indicators and the optimal category quantifications multiplied with beta against category indicators.

CATREG Command Additional Features

You can customize your categorical regression if you paste your selections into a syntax window and edit the resulting `CATREG` command syntax. The command syntax language also allows you to:

- Specify rootnames for the transformed variables when saving them to the active dataset (with the `SAVE` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter

3

Categorical Principal Components Analysis (CATPCA)

This procedure simultaneously quantifies categorical variables while reducing the dimensionality of the data. Categorical principal components analysis is also known by the acronym CATPCA, for *cat*egorical principal components analysis.

The goal of principal components analysis is to reduce an original set of variables into a smaller set of uncorrelated components that represent most of the information found in the original variables. The technique is most useful when a large number of variables prohibits effective interpretation of the relationships between objects (subjects and units). By reducing the dimensionality, you interpret a few components rather than a large number of variables.

Standard principal components analysis assumes linear relationships between numeric variables. On the other hand, the optimal-scaling approach allows variables to be scaled at different levels. Categorical variables are optimally quantified in the specified dimensionality. As a result, nonlinear relationships between variables can be modeled.

Example. Categorical principal components analysis could be used to graphically display the relationship between job category, job division, region, amount of travel (high, medium, and low), and job satisfaction. You might find that two dimensions account for a large amount of variance. The first dimension might separate job category from region, whereas the second dimension might separate job division from amount of travel. You also might find that high job satisfaction is related to a medium amount of travel.

Statistics and plots. Frequencies, missing values, optimal scaling level, mode, variance accounted for by centroid coordinates, vector coordinates, total per variable and per dimension, component loadings for vector-quantified variables, category quantifications and coordinates, iteration history, correlations of the transformed variables and eigenvalues of the correlation matrix, correlations of the original variables and eigenvalues of the correlation matrix, object scores, category plots, joint category plots, transformation plots, residual plots, projected centroid plots, object plots, biplots, triplots, and component loadings plots.

Data. String variable values are always converted into positive integers by ascending alphanumeric order. User-defined missing values, system-missing values, and values less than 1 are considered missing; you can recode or add a constant to variables with values less than 1 to make them nonmissing.

Assumptions. The data must contain at least three valid cases. The analysis is based on positive integer data. The discretization option will automatically categorize a fractional-valued variable by grouping its values into categories with a close to "normal" distribution and will automatically convert values of string variables into positive integers. You can specify other discretization schemes.

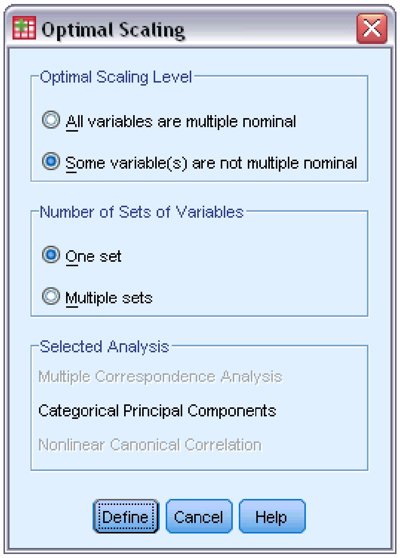
Related procedures. Scaling all variables at the numeric level corresponds to standard principal components analysis. Alternate plotting features are available by using the transformed variables in a standard linear principal components analysis. If all variables have multiple nominal scaling levels, categorical principal components analysis is identical to multiple correspondence analysis. If sets of variables are of interest, categorical (nonlinear) canonical correlation analysis should be used.

To Obtain a Categorical Principal Components Analysis

▶ From the menus choose:

Analyze > Dimension Reduction > Optimal Scaling...

Figure 3-1
Optimal Scaling dialog box



Optimal Scaling Level
○ All variables are multiple nominal
● Some variable(s) are not multiple nominal

Number of Sets of Variables
● One set
○ Multiple sets

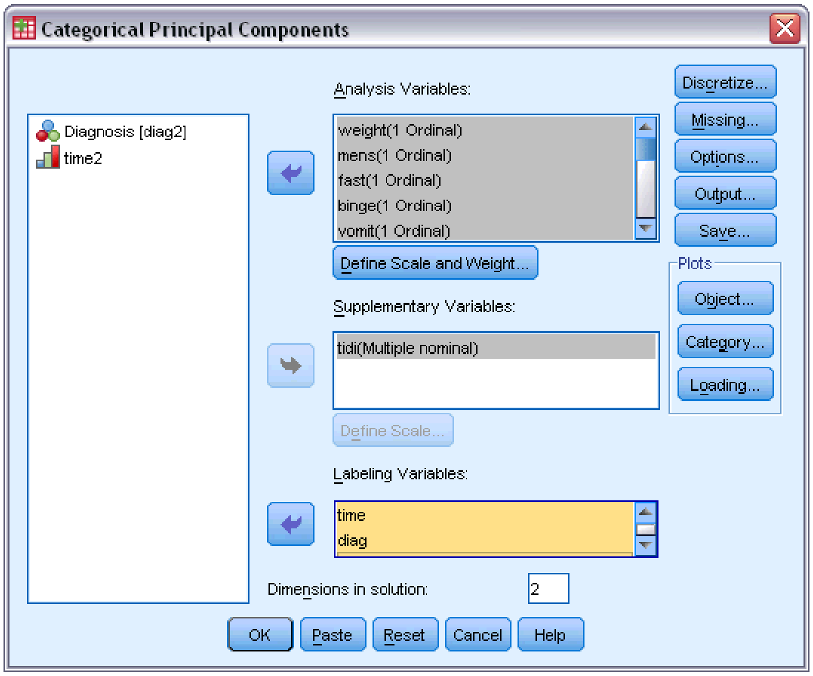
Selected Analysis
Multiple Correspondence Analysis
Categorical Principal Components
Nonlinear Canonical Correlation

▶ Select Some variable(s) not multiple nominal.

▶ Select One set.

▶ Click Define.

Figure 3-2
Categorical Principal Components dialog box



Diagnosis [diag2]
time2

Analysis Variables:
weight(1 Ordinal)
mens(1 Ordinal)
fast(1 Ordinal)
binge(1 Ordinal)
vomit(1 Ordinal)

Supplementary Variables:
tidi(Multiple nominal)

Labeling Variables:
time
diag

Dimensions in solution: 2

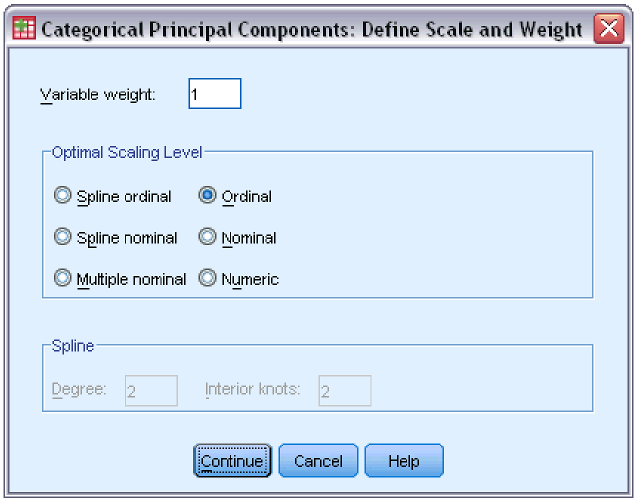
- ▶ Select at least two analysis variables and specify the number of dimensions in the solution.
- ▶ Click OK.

You may optionally specify supplementary variables, which are fitted into the solution found, or labeling variables for the plots.

Define Scale and Weight in CATPCA

You can set the optimal scaling level for analysis variables and supplementary variables. By default, they are scaled as second-degree monotonic splines (ordinal) with two interior knots. Additionally, you can set the weight for analysis variables.

Figure 3-3
Define Scale and Weight dialog box



Categorical Principal Components: Define Scale and Weight

Variable weight: 1

Optimal Scaling Level

○ Spline ordinal ● Ordinal

○ Spline nominal ○ Nominal

○ Multiple nominal ○ Numeric

Spline

Degree: 2 Interior knots: 2

Variable weight. You can choose to define a weight for each variable. The value specified must be a positive integer. The default value is 1.

Optimal Scaling Level. You can also select the scaling level to be used to quantify each variable.

- **Spline ordinal.** The order of the categories of the observed variable is preserved in the optimally scaled variable. Category points will be on a straight line (vector) through the origin. The resulting transformation is a smooth monotonic piecewise polynomial of the chosen degree. The pieces are specified by the user-specified number and procedure-determined placement of the interior knots.
- **Spline nominal.** The only information in the observed variable that is preserved in the optimally scaled variable is the grouping of objects in categories. The order of the categories of the observed variable is not preserved. Category points will be on a straight line (vector) through the origin. The resulting transformation is a smooth, possibly nonmonotonic, piecewise polynomial of the chosen degree. The pieces are specified by the user-specified number and procedure-determined placement of the interior knots.
- **Multiple nominal.** The only information in the observed variable that is preserved in the optimally scaled variable is the grouping of objects in categories. The order of the categories of the observed variable is not preserved. Category points will be in the centroid of the objects in the particular categories. *Multiple* indicates that different sets of quantifications are obtained for each dimension.
- **Ordinal.** The order of the categories of the observed variable is preserved in the optimally scaled variable. Category points will be on a straight line (vector) through the origin. The resulting transformation fits better than the spline ordinal transformation but is less smooth.
- **Nominal.** The only information in the observed variable that is preserved in the optimally scaled variable is the grouping of objects in categories. The order of the categories of the observed variable is not preserved. Category points will be on a straight line (vector) through

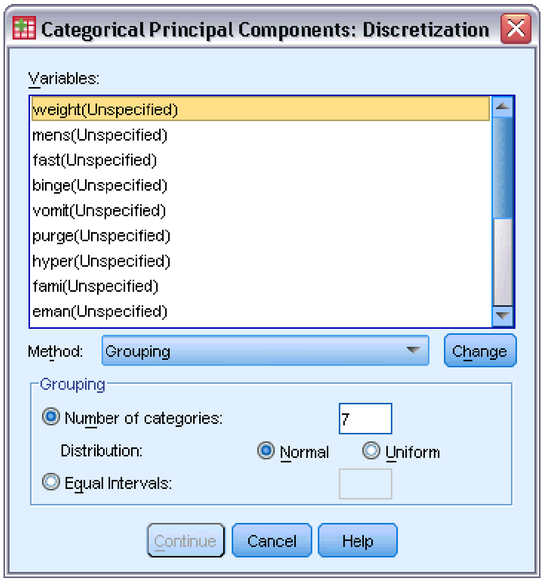
the origin. The resulting transformation fits better than the spline nominal transformation but is less smooth.

- **Numeric.** Categories are treated as ordered and equally spaced (interval level). The order of the categories and the equal distances between category numbers of the observed variable are preserved in the optimally scaled variable. Category points will be on a straight line (vector) through the origin. When all variables are at the numeric level, the analysis is analogous to standard principal components analysis.

Categorical Principal Components Analysis Discretization

The Discretization dialog box allows you to select a method of recoding your variables. Fractional-valued variables are grouped into seven categories (or into the number of distinct values of the variable if this number is less than seven) with an approximately normal distribution, unless specified otherwise. String variables are always converted into positive integers by assigning category indicators according to ascending alphanumeric order. Discretization for string variables applies to these integers. Other variables are left alone by default. The discretized variables are then used in the analysis.

Figure 3-4
Discretization dialog box



Categorical Principal Components: Discretization

Variables:
weight(Unspecified)
mens(Unspecified)
fast(Unspecified)
binge(Unspecified)
vomit(Unspecified)
purge(Unspecified)
hyper(Unspecified)
fami(Unspecified)
eman(Unspecified)

Method: Grouping

Grouping
● Number of categories: 7
Distribution: ● Normal ○ Uniform
○ Equal Intervals:

Method. Choose between grouping, ranking, and multiplying.

- **Grouping.** Recode into a specified number of categories or recode by interval.
- **Ranking.** The variable is discretized by ranking the cases.
- **Multiplying.** The current values of the variable are standardized, multiplied by 10, rounded, and have a constant added such that the lowest discretized value is 1.

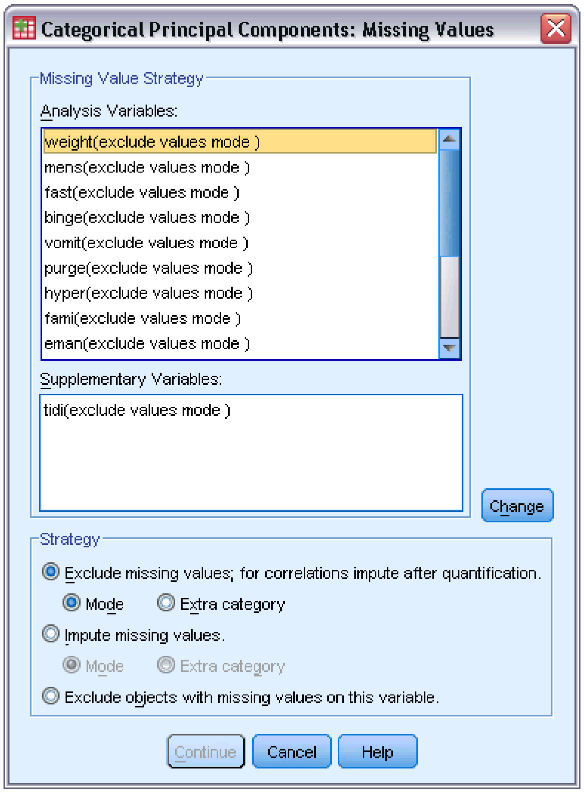
Grouping. The following options are available when you are discretizing variables by grouping:

- **Number of categories.** Specify a number of categories and whether the values of the variable should follow an approximately normal or uniform distribution across those categories.
- **Equal intervals.** Variables are recoded into categories defined by these equally sized intervals. You must specify the length of the intervals.

Categorical Principal Components Analysis Missing Values

The Missing Values dialog box allows you to choose the strategy for handling missing values in analysis variables and supplementary variables.

Figure 3-5
Missing Values dialog box



Categorical Principal Components: Missing Values

Missing Value Strategy

Analysis Variables:

weight(exclude values mode)
mens(exclude values mode)
fast(exclude values mode)
binge(exclude values mode)
vomit(exclude values mode)
purge(exclude values mode)
hyper(exclude values mode)
fami(exclude values mode)
eman(exclude values mode)

Supplementary Variables:

tidi(exclude values mode)

Strategy

● Exclude missing values; for correlations impute after quantification.
● Mode ○ Extra category
○ Impute missing values.
● Mode ○ Extra category
○ Exclude objects with missing values on this variable.

Strategy. Choose to exclude missing values (passive treatment), impute missing values (active treatment), or exclude objects with missing values (listwise deletion).

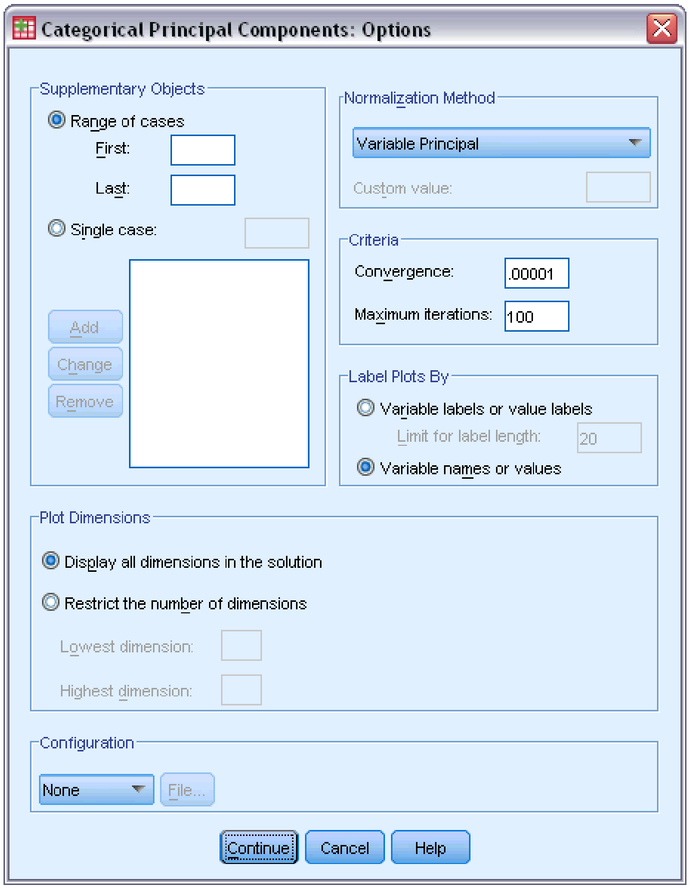
- **Exclude missing values; for correlations impute after quantification.** Objects with missing values on the selected variable do not contribute to the analysis for this variable. If all variables are given passive treatment, then objects with missing values on all variables are treated as supplementary. If correlations are specified in the Output dialog box, then (after analysis) missing values are imputed with the most frequent category, or mode, of the variable for the correlations of the original variables. For the correlations of the optimally scaled variables, you can choose the method of imputation. Select Mode to replace missing values with the mode of the optimally scaled variable. Select Extra category to replace missing values with the quantification of an extra category. This implies that objects with a missing value on this variable are considered to belong to the same (extra) category.

- **Impute missing values.** Objects with missing values on the selected variable have those values imputed. You can choose the method of imputation. Select Mode to replace missing values with the most frequent category. When there are multiple modes, the one with the smallest category indicator is used. Select Extra category to replace missing values with the same quantification of an extra category. This implies that objects with a missing value on this variable are considered to belong to the same (extra) category.
- **Exclude objects with missing values on this variable.** Objects with missing values on the selected variable are excluded from the analysis. This strategy is not available for supplementary variables.

Categorical Principal Components Analysis Options

The Options dialog box allows you to select the initial configuration, specify iteration and convergence criteria, select a normalization method, choose the method for labeling plots, and specify supplementary objects.

Figure 3-6
Options dialog box



Categorical Principal Components: Options

Supplementary Objects
● Range of cases
First:
Last:
○ Single case:

Normalization Method: Variable Principal
Custom value:

Criteria
Convergence: .00001
Maximum iterations: 100

Label Plots By
○ Variable labels or value labels
Limit for label length: 20
● Variable names or values

Plot Dimensions
● Display all dimensions in the solution
○ Restrict the number of dimensions
Lowest dimension:
Highest dimension:

Configuration: None

Supplementary Objects. Specify the case number of the object, or the first and last case numbers of a range of objects, that you want to make supplementary and then click Add. Continue until you have specified all of your supplementary objects. If an object is specified as supplementary, then case weights are ignored for that object.

Normalization Method. You can specify one of five options for normalizing the object scores and the variables. Only one normalization method can be used in a given analysis.

- **Variable Principal.** This option optimizes the association between variables. The coordinates of the variables in the object space are the component loadings (correlations with principal components, such as dimensions and object scores). This is useful when you are primarily interested in the correlation between the variables.
- **Object Principal.** This option optimizes distances between objects. This is useful when you are primarily interested in differences or similarities between the objects.
- **Symmetrical.** Use this normalization option if you are primarily interested in the relation between objects and variables.
- **Independent.** Use this normalization option if you want to examine distances between objects and correlations between variables separately.
- **Custom.** You can specify any real value in the closed interval [–1, 1]. A value of 1 is equal to the Object Principal method, a value of 0 is equal to the Symmetrical method, and a value of –1 is equal to the Variable Principal method. By specifying a value greater than –1 and less than 1, you can spread the eigenvalue over both objects and variables. This method is useful for making a tailor-made biplot or triplot.

Criteria. You can specify the maximum number of iterations the procedure can go through in its computations. You can also select a convergence criterion value. The algorithm stops iterating if the difference in total fit between the last two iterations is less than the convergence value or if the maximum number of iterations is reached.

Label Plots By. Allows you to specify whether variables and value labels or variable names and values will be used in the plots. You can also specify a maximum length for labels.

Plot Dimensions. Allows you to control the dimensions displayed in the output.

- **Display all dimensions in the solution.** All dimensions in the solution are displayed in a scatterplot matrix.
- **Restrict the number of dimensions.** The displayed dimensions are restricted to plotted pairs. If you restrict the dimensions, you must select the lowest and highest dimensions to be plotted. The lowest dimension can range from 1 to the number of dimensions in the solution minus 1 and is plotted against higher dimensions. The highest dimension value can range from 2 to the number of dimensions in the solution and indicates the highest dimension to be used in plotting the dimension pairs. This specification applies to all requested multidimensional plots.

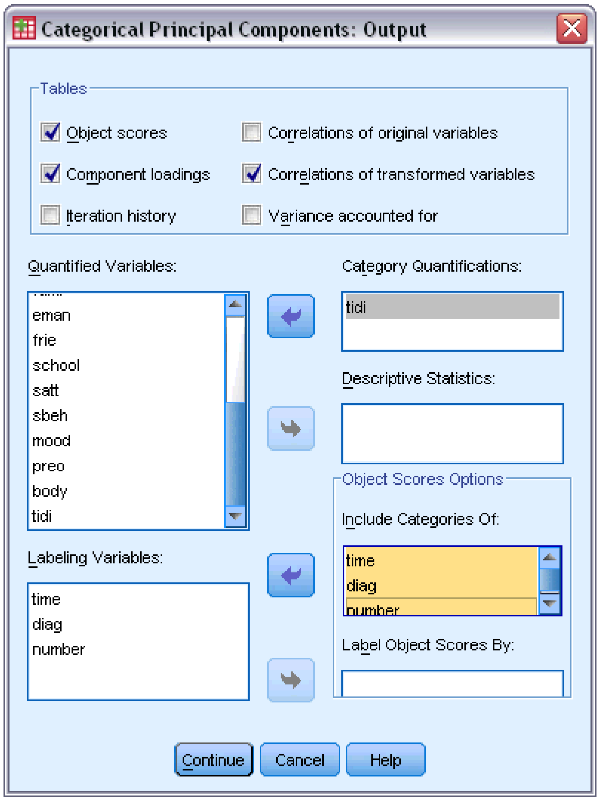
Configuration. You can read data from a file containing the coordinates of a configuration. The first variable in the file should contain the coordinates for the first dimension, the second variable should contain the coordinates for the second dimension, and so on.

- **Initial.** The configuration in the file specified will be used as the starting point of the analysis.
- **Fixed.** The configuration in the file specified will be used to fit in the variables. The variables that are fitted in must be selected as analysis variables, but because the configuration is fixed, they are treated as supplementary variables (so they do not need to be selected as supplementary variables).

Categorical Principal Components Analysis Output

The Output dialog box allows you to produce tables for object scores, component loadings, iteration history, correlations of original and transformed variables, the variance accounted for per variable and per dimension, category quantifications for selected variables, and descriptive statistics for selected variables.

Figure 3-7
Output dialog box



Categorical Principal Components: Output

Tables

☑ Object scores
☐ Correlations of original variables
☑ Component loadings
☑ Correlations of transformed variables
☐ Iteration history
☐ Variance accounted for

Quantified Variables:
eman
frie
school
satt
sbeh
mood
preo
body
tidi

Category Quantifications:
tidi

Descriptive Statistics:

Object Scores Options

Include Categories Of:
time
diag
number

Label Object Scores By:

Labeling Variables:
time
diag
number

Object scores. Displays the object scores and has the following options:

- **Include Categories Of.** Displays the category indicators of the analysis variables selected.
- **Label Object Scores By.** From the list of variables specified as labeling variables, you can select one to label the objects.

Component loadings. Displays the component loadings for all variables that were not given multiple nominal scaling levels.

Iteration history. For each iteration, the variance accounted for, loss, and increase in variance accounted for are shown.

Correlations of original variables. Shows the correlation matrix of the original variables and the eigenvalues of that matrix.

Correlations of transformed variables. Shows the correlation matrix of the transformed (optimally scaled) variables and the eigenvalues of that matrix.

Variance accounted for. Displays the amount of variance accounted for by centroid coordinates, vector coordinates, and total (centroid and vector coordinates combined) per variable and per dimension.

Category Quantifications. Gives the category quantifications and coordinates for each dimension of the variable(s) selected.

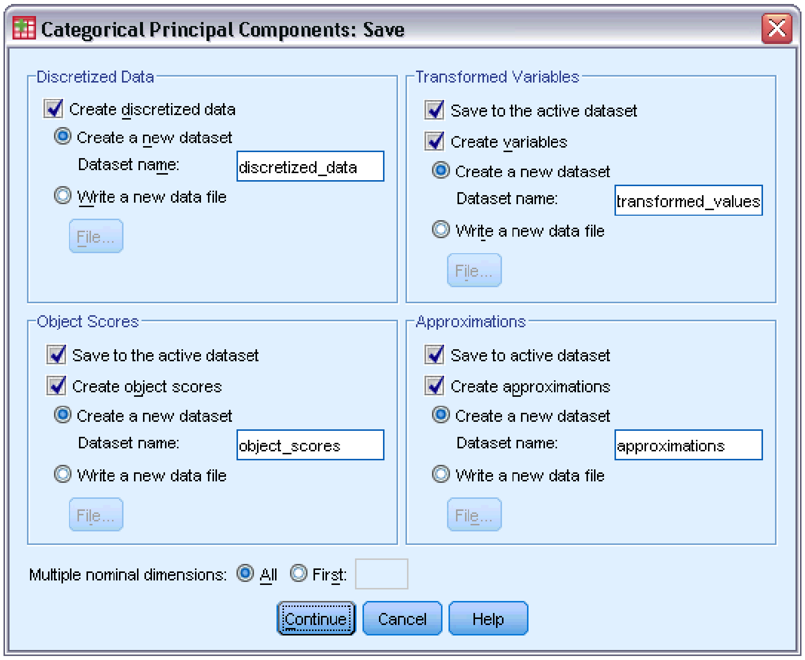
Descriptive Statistics. Displays frequencies, number of missing values, and mode of the variable(s) selected.

Categorical Principal Components Analysis Save

The Save dialog box allows you to save discretized data, object scores, transformed values, and approximations to an external IBM® SPSS® Statistics data file or dataset in the current session. You can also save transformed values, object scores, and approximations to the active dataset.

- Datasets are available during the current session but are not available in subsequent sessions unless you explicitly save them as data files. Dataset names must adhere to variable naming rules.
- Filenames or dataset names must be different for each type of data saved.
- If you save object scores or transformed values to the active dataset, you can specify the number of multiple nominal dimensions.

Figure 3-8
Save dialog box



Categorical Principal Components: Save

Discretized Data
☑ Create discretized data
● Create a new dataset
Dataset name: discretized_data
○ Write a new data file

Transformed Variables
☑ Save to the active dataset
☑ Create variables
● Create a new dataset
Dataset name: transformed_values
○ Write a new data file

Object Scores
☑ Save to the active dataset
☑ Create object scores
● Create a new dataset
Dataset name: object_scores
○ Write a new data file

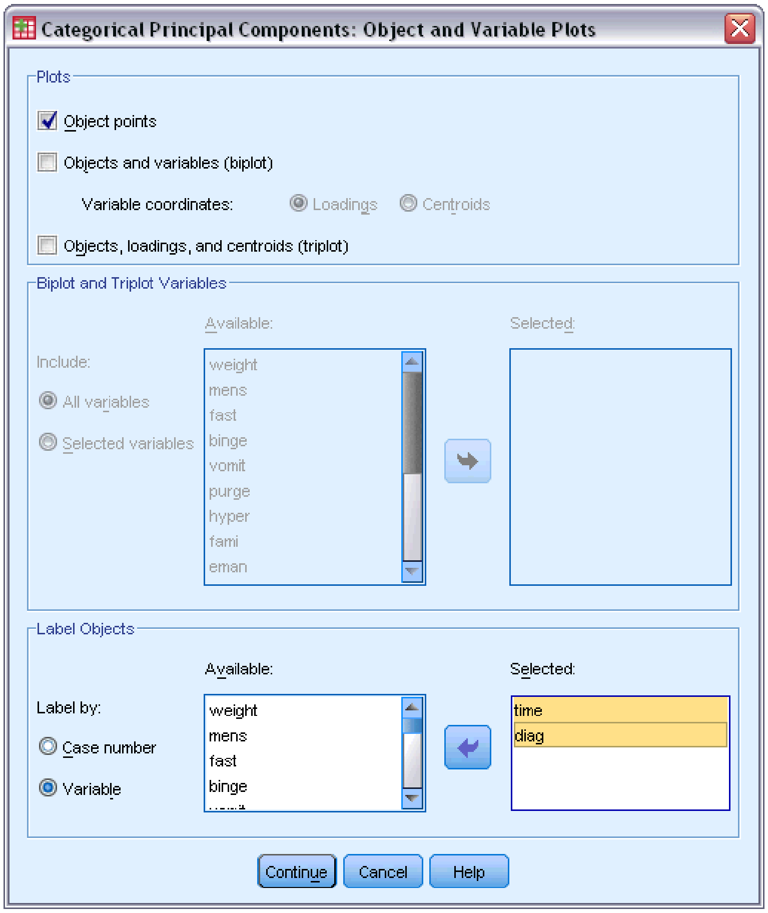
Approximations
☑ Save to active dataset
☑ Create approximations
● Create a new dataset
Dataset name: approximations
○ Write a new data file

Multiple nominal dimensions: ● All ○ First:

Categorical Principal Components Analysis Object Plots

The Object and Variable Plots dialog box allows you to specify the types of plots desired and the variables for which plots will be produced.

Figure 3-9
Object and Variable Plots dialog box



Categorical Principal Components: Object and Variable Plots

Plots

☑ Object points

□ Objects and variables (biplot)

Variable coordinates: ● Loadings ○ Centroids

□ Objects, loadings, and centroids (triplot)

Biplot and Triplot Variables

Include: ● All variables ○ Selected variables

Available: weight, mens, fast, binge, vomit, purge, hyper, fami, eman

Selected:

Label Objects

Label by: ○ Case number ● Variable

Available: weight, mens, fast, binge

Selected: time, diag

Object points. A plot of the object points is displayed.

Objects and variables (biplot). The object points are plotted with your choice of the variable coordinates—component loadings or variable centroids.

Objects, loadings, and centroids (triplot). The object points are plotted with the centroids of multiple nominal-scaling-level variables and the component loadings of other variables.

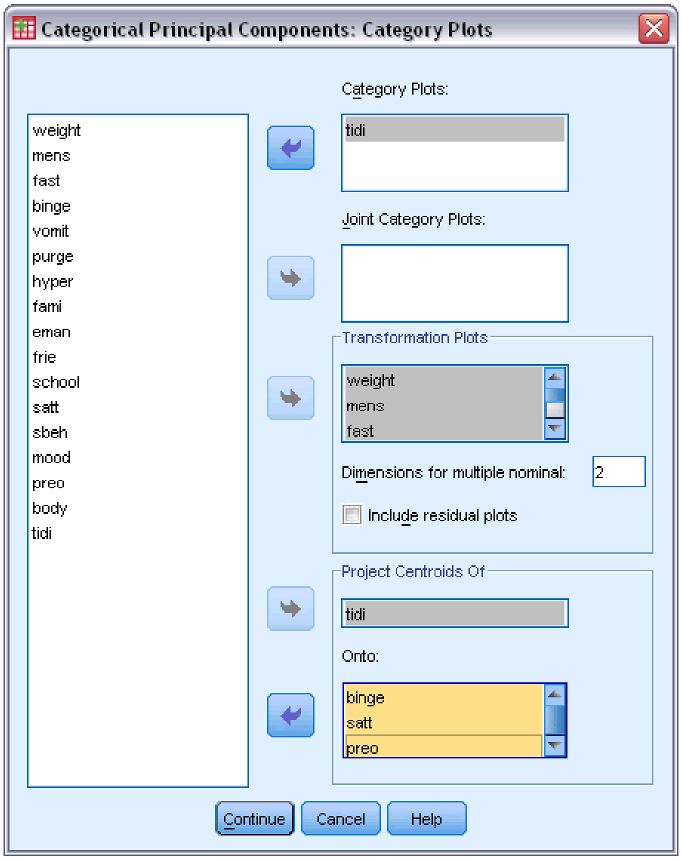
Biplot and Triplot Variables. You can choose to use all variables for the biplots and triplots or select a subset.

Label Objects. You can choose to have objects labeled with the categories of selected variables (you may choose category indicator values or value labels in the Options dialog box) or with their case numbers. One plot is produced per variable if Variable is selected.

Categorical Principal Components Analysis Category Plots

The Category Plots dialog box allows you to specify the types of plots desired and the variables for which plots will be produced.

Figure 3-10
Category Plots dialog box



Category Plots: tidi

Joint Category Plots:

Transformation Plots: weight, mens, fast

Dimensions for multiple nominal: 2

☐ Include residual plots

Project Centroids Of: tidi

Onto: binge, satt, preo

Category Plots. For each variable selected, a plot of the centroid and vector coordinates is plotted. For variables with multiple nominal scaling levels, categories are in the centroids of the objects in the particular categories. For all other scaling levels, categories are on a vector through the origin.

Joint Category Plots. This is a single plot of the centroid and vector coordinates of each selected variable.

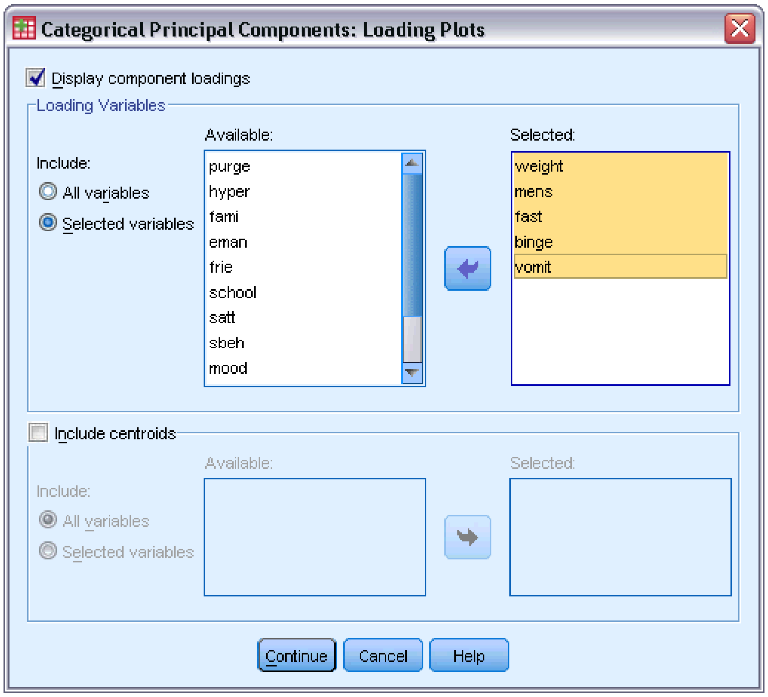
Transformation Plots. Displays a plot of the optimal category quantifications versus the category indicators. You can specify the number of dimensions desired for variables with multiple nominal scaling levels; one plot will be generated for each dimension. You can also choose to display residual plots for each variable selected.

Project Centroids Of. You may choose a variable and project its centroids onto selected variables. Variables with multiple nominal scaling levels cannot be selected to project on. When this plot is requested, a table with the coordinates of the projected centroids is also displayed.

Categorical Principal Components Analysis Loading Plots

The Loading Plots dialog box allows you to specify the variables that will be included in the plot, and whether or not to include centroids in the plot.

Figure 3-11
Loading Plots dialog box



Categorical Principal Components: Loading Plots

☑ Display component loadings

Loading Variables

Include:
○ All variables
● Selected variables

Available: purge, hyper, fami, eman, frie, school, satt, sbeh, mood

Selected: weight, mens, fast, binge, vomit

□ Include centroids

Include:
● All variables
○ Selected variables

Display component loadings. If selected, a plot of the component loadings is displayed.

Loading Variables. You can choose to use all variables for the component loadings plot or select a subset.

Include centroids. Variables with multiple nominal scaling levels do not have component loadings, but you may choose to include the centroids of those variables in the plot. You can choose to use all multiple nominal variables or select a subset.

CATPCA Command Additional Features

You can customize your categorical principal components analysis if you paste your selections into a syntax window and edit the resulting `CATPCA` command syntax. The command syntax language also allows you to:

- Specify rootnames for the transformed variables, object scores, and approximations when saving them to the active dataset (with the `SAVE` subcommand).
- Specify a maximum length for labels for each plot separately (with the `PLOT` subcommand).
- Specify a separate variable list for residual plots (with the `PLOT` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter

4

Nonlinear Canonical Correlation Analysis (OVERALS)

Nonlinear canonical correlation analysis corresponds to categorical canonical correlation analysis with optimal scaling. The purpose of this procedure is to determine how similar sets of categorical variables are to one another. Nonlinear canonical correlation analysis is also known by the acronym OVERALS.

Standard canonical correlation analysis is an extension of multiple regression, where the second set does not contain a single response variable but instead contain multiple response variables. The goal is to explain as much as possible of the variance in the relationships among two sets of numerical variables in a low dimensional space. Initially, the variables in each set are linearly combined such that the linear combinations have a maximal correlation. Given these combinations, subsequent linear combinations are determined that are uncorrelated with the previous combinations and that have the largest correlation possible.

The optimal scaling approach expands the standard analysis in three crucial ways. First, OVERALS allows more than two sets of variables. Second, variables can be scaled as either nominal, ordinal, or numerical. As a result, nonlinear relationships between variables can be analyzed. Finally, instead of maximizing correlations between the variable sets, the sets are compared to an unknown compromise set that is defined by the object scores.

Example. Categorical canonical correlation analysis with optimal scaling could be used to graphically display the relationship between one set of variables containing job category and years of education and another set of variables containing region of residence and gender. You might find that years of education and region of residence discriminate better than the remaining variables. You might also find that years of education discriminates best on the first dimension.

Statistics and plots. Frequencies, centroids, iteration history, object scores, category quantifications, weights, component loadings, single and multiple fit, object scores plots, category coordinates plots, component loadings plots, category centroids plots, transformation plots.

Data. Use integers to code categorical variables (nominal or ordinal scaling level). To minimize output, use consecutive integers beginning with 1 to code each variable. Variables that are scaled at the numerical level should not be recoded to consecutive integers. To minimize output, for each variable that is scaled at the numerical level, subtract the smallest observed value from every value and add 1. Fractional values are truncated after the decimal.

Assumptions. Variables can be classified into two or more sets. Variables in the analysis are scaled as multiple nominal, single nominal, ordinal, or numerical. The maximum number of dimensions that are used in the procedure depends on the optimal scaling level of the variables. If all variables are specified as ordinal, single nominal, or numerical, the maximum number of dimensions is the lesser of the following two values: the number of observations minus 1 or the total number of variables. However, if only two sets of variables are defined, the maximum number of dimensions is the number of variables in the smaller set. If some variables are multiple nominal, the maximum number of dimensions is the total number of multiple nominal categories plus the number of nonmultiple nominal variables minus the number of multiple nominal variables. For example, if the analysis involves five variables, one of which is multiple nominal with four categories, the maximum number of dimensions is (4 + 4 – 1), or 7. If you specify a number that is greater than the maximum, the maximum value is used.

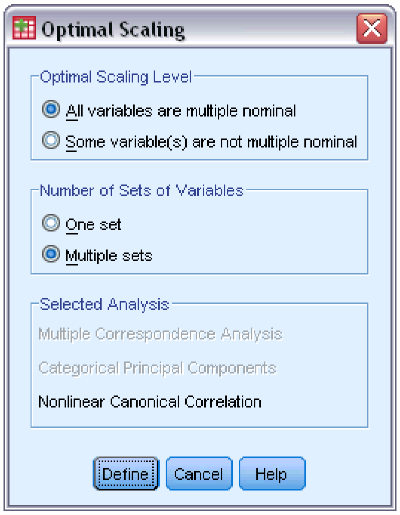
Related procedures. If each set contains one variable, nonlinear canonical correlation analysis is equivalent to principal components analysis with optimal scaling. If each of these variables is multiple nominal, the analysis corresponds to multiple correspondence analysis. If two sets of variables are involved, and one of the sets contains only one variable, the analysis is identical to categorical regression with optimal scaling.

To Obtain a Nonlinear Canonical Correlation Analysis

- ▶ From the menus choose:

  Analyze > Dimension Reduction > Optimal Scaling...

Figure 4-1
Optimal Scaling dialog box



Optimal Scaling

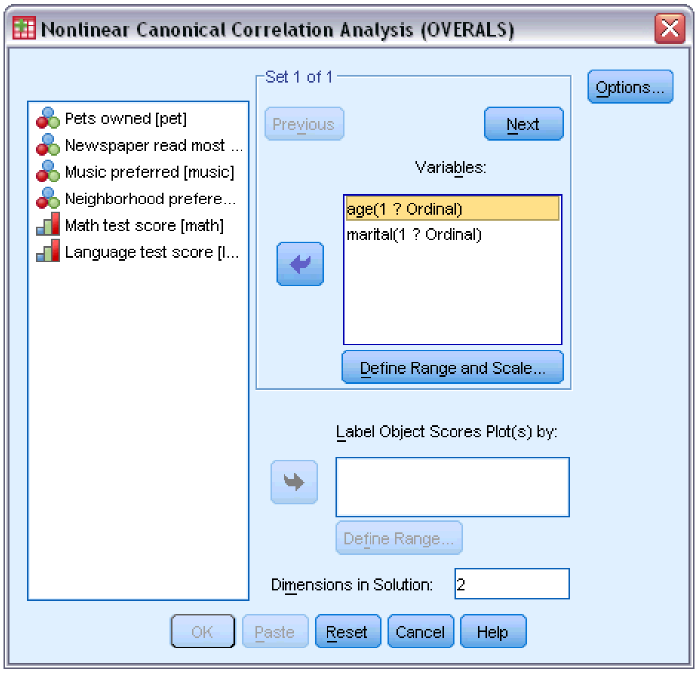
Optimal Scaling Level
● All variables are multiple nominal
○ Some variable(s) are not multiple nominal

Number of Sets of Variables
○ One set
● Multiple sets

Selected Analysis
Multiple Correspondence Analysis
Categorical Principal Components
Nonlinear Canonical Correlation

- ▶ Select either All variables multiple nominal or Some variable(s) not multiple nominal.
- ▶ Select Multiple sets.
- ▶ Click Define.

Figure 4-2
Nonlinear Canonical Correlation Analysis (OVERALS) dialog box



Variables:
age(1 ? Ordinal)
marital(1 ? Ordinal)

Dimensions in Solution: 2

- ▶ Define at least two sets of variables. Select the variable(s) that you want to include in the first set. To move to the next set, click Next, and select the variables that you want to include in the second set. You can add additional sets as desired. Click Previous to return to the previously defined variable set.

- ▶ Define the value range and measurement scale (optimal scaling level) for each selected variable.

- ▶ Click OK.

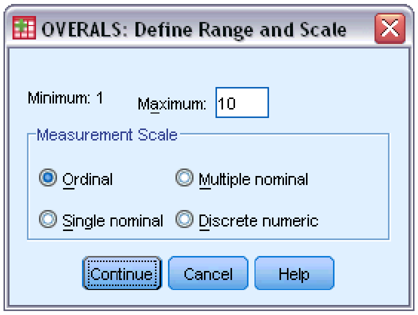
- ▶ Optionally:

  - ■ Select one or more variables to provide point labels for object scores plots. Each variable produces a separate plot, with the points labeled by the values of that variable. You must define a range for each of these plot label variables. When you are using the dialog box, a single variable cannot be used both in the analysis and as a labeling variable. If labeling the object scores plot with a variable that is used in the analysis is desired, use the Compute facility (available from the Transform menu) to create a copy of that variable. Use the new variable to label the plot. Alternatively, command syntax can be used.

  - ■ Specify the number of dimensions that you want in the solution. In general, choose as few dimensions as needed to explain most of the variation. If the analysis involves more than two dimensions, three-dimensional plots of the first three dimensions are produced. Other dimensions can be displayed by editing the chart.

Define Range and Scale

Figure 4-3
Define Range and Scale dialog box



OVERALS: Define Range and Scale

Minimum: 1 Maximum: 10

Measurement Scale

● Ordinal ○ Multiple nominal

○ Single nominal ○ Discrete numeric

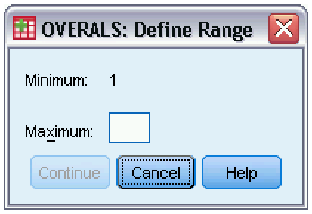
You must define a range for each variable. The maximum value that is specified must be an integer. Fractional data values are truncated in the analysis. A category value that is outside of the specified range is ignored in the analysis. To minimize output, use the Automatic Recode facility (available from the Transform menu) to create consecutive categories beginning with 1 for variables that are treated as nominal or ordinal. Recoding to consecutive integers is not recommended for variables that are scaled at the numerical level. To minimize output for variables that are treated as numerical, for each variable, subtract the minimum value from every value and add 1.

You must also select the scaling to be used to quantify each variable.

- **Ordinal.** The order of the categories of the observed variable is preserved in the quantified variable.
- **Single nominal.** In the quantified variable, objects in the same category receive the same score.
- **Multiple nominal.** The quantifications can be different for each dimension.
- **Discrete numeric.** Categories are treated as ordered and equally spaced. The differences between category numbers and the order of the categories of the observed variable are preserved in the quantified variable.

Define Range

Figure 4-4
Define Range dialog box



OVERALS: Define Range

Minimum: 1

Maximum:

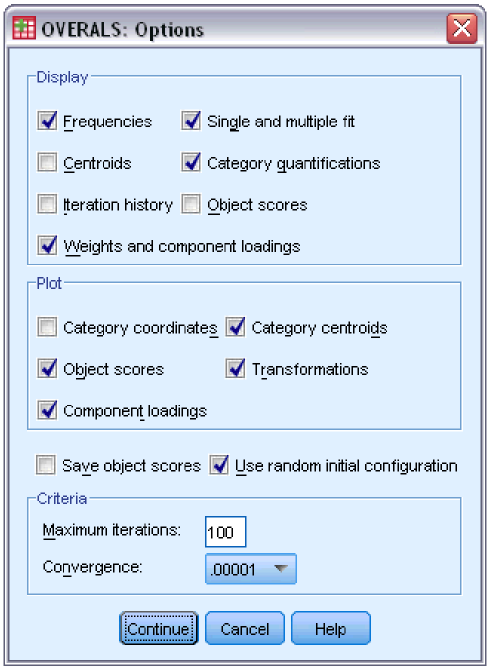
You must define a range for each variable. The maximum value that is specified must be an integer. Fractional data values are truncated in the analysis. A category value that is outside of the specified range is ignored in the analysis. To minimize output, use the Automatic Recode facility (available from the Transform menu) to create consecutive categories beginning with 1.

You must also define a range for each variable that is used to label the object scores plots. However, labels for categories with data values that are outside of the defined range for the variable do appear on the plots.

Nonlinear Canonical Correlation Analysis Options

The Options dialog box allows you to select optional statistics and plots, save object scores as new variables in the active dataset, specify iteration and convergence criteria, and specify an initial configuration for the analysis.

Figure 4-5
Options dialog box



☑ Frequencies ☑ Single and multiple fit

□ Centroids ☑ Category quantifications

□ Iteration history □ Object scores

☑ Weights and component loadings

□ Category coordinates ☑ Category centroids

☑ Object scores ☑ Transformations

☑ Component loadings

□ Save object scores ☑ Use random initial configuration

Maximum iterations: 100

Convergence: .00001

Display. Available statistics include marginal frequencies (counts), centroids, iteration history, weights and component loadings, category quantifications, object scores, and single and multiple fit statistics.

- **Centroids.** Category quantifications, and the projected and the actual averages of the object scores for the objects (cases) included in each set for those belonging to the same category of the variable.
- **Weights and component loadings.** The regression coefficients in each dimension for every quantified variable in a set, where the object scores are regressed on the quantified variables, and the projection of the quantified variable in the object space. Provides an indication of the contribution each variable makes to the dimension within each set.
- **Single and multiple fit.** Measures of goodness of fit of the single- and multiple-category coordinates/category quantifications with respect to the objects.
- **Category quantifications.** Optimal scale values assigned to the categories of a variable.
- **Object scores.** Optimal score assigned to an object (case) in a particular dimension.

Plot. You can produce plots of category coordinates, object scores, component loadings, category centroids, and transformations.

Save object scores. You can save the object scores as new variables in the active dataset. Object scores are saved for the number of dimensions that are specified in the main dialog box.

Use random initial configuration. A random initial configuration should be used if some or all of the variables are single nominal. If this option is not selected, a nested initial configuration is used.

Criteria. You can specify the maximum number of iterations that the nonlinear canonical correlation analysis can go through in its computations. You can also select a convergence criterion value. The analysis stops iterating if the difference in total fit between the last two iterations is less than the convergence value or if the maximum number of iterations is reached.

OVERALS Command Additional Features

You can customize your nonlinear canonical correlation analysis if you paste your selections into a syntax window and edit the resulting `OVERALS` command syntax. The command syntax language also allows you to:

- Specify the dimension pairs to be plotted, rather than plotting all extracted dimensions (using the`NDIM` keyword on the `PLOT` subcommand).
- Specify the number of value label characters that are used to label points on the plots (with the`PLOT` subcommand).
- Designate more than five variables as labeling variables for object scores plots (with the`PLOT` subcommand).
- Select variables that are used in the analysis as labeling variables for the object scores plots (with the `PLOT` subcommand).
- Select variables to provide point labels for the quantification score plot (with the `PLOT` subcommand).
- Specify the number of cases to be included in the analysis if you do not want to use all cases in the active dataset (with the `NOBSERVATIONS` subcommand).
- Specify rootnames for variables created by saving object scores (with the `SAVE` subcommand).
- Specify the number of dimensions to be saved, rather than saving all extracted dimensions (with the `SAVE` subcommand).
- Write category quantifications to a matrix file (using the `MATRIX` subcommand).
- Produce low-resolution plots that may be easier to read than the usual high-resolution plots (using the `SET` command).
- Produce centroid and transformation plots for specified variables only (with the `PLOT` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter

5

Correspondence Analysis

One of the goals of correspondence analysis is to describe the relationships between two nominal variables in a correspondence table in a low-dimensional space, while simultaneously describing the relationships between the categories for each variable. For each variable, the distances between category points in a plot reflect the relationships between the categories with similar categories plotted close to each other. Projecting points for one variable on the vector from the origin to a category point for the other variable describe the relationship between the variables.

An analysis of contingency tables often includes examining row and column profiles and testing for independence via the chi-square statistic. However, the number of profiles can be quite large, and the chi-square test does not reveal the dependence structure. The Crosstabs procedure offers several measures of association and tests of association but cannot graphically represent any relationships between the variables.

Factor analysis is a standard technique for describing relationships between variables in a low-dimensional space. However, factor analysis requires interval data, and the number of observations should be five times the number of variables. Correspondence analysis, on the other hand, assumes nominal variables and can describe the relationships between categories of each variable, as well as the relationship between the variables. In addition, correspondence analysis can be used to analyze any table of positive correspondence measures.

Example. Correspondence analysis could be used to graphically display the relationship between staff category and smoking habits. You might find that with regard to smoking, junior managers differ from secretaries, but secretaries do not differ from senior managers. You might also find that heavy smoking is associated with junior managers, whereas light smoking is associated with secretaries.

Statistics and plots. Correspondence measures, row and column profiles, singular values, row and column scores, inertia, mass, row and column score confidence statistics, singular value confidence statistics, transformation plots, row point plots, column point plots, and biplots.

Data. Categorical variables to be analyzed are scaled nominally. For aggregated data or for a correspondence measure other than frequencies, use a weighting variable with positive similarity values. Alternatively, for table data, use syntax to read the table.

Assumptions. The maximum number of dimensions used in the procedure depends on the number of active rows and column categories and the number of equality constraints. If no equality constraints are used and all categories are active, the maximum dimensionality is one fewer than the number of categories for the variable with the fewest categories. For example, if one variable has five categories and the other has four, the maximum number of dimensions is three. Supplementary categories are not active. For example, if one variable has five categories, two of which are supplementary, and the other variable has four categories, the maximum number of dimensions is two. Treat all sets of categories that are constrained to be equal as one category. For example, if a variable has five categories, three of which are constrained to be equal, that variable should be treated as having three categories when determining the maximum dimensionality. Two of the categories are unconstrained, and the third category corresponds to the three constrained categories. If you specify a number of dimensions greater than the maximum, the maximum value is used.

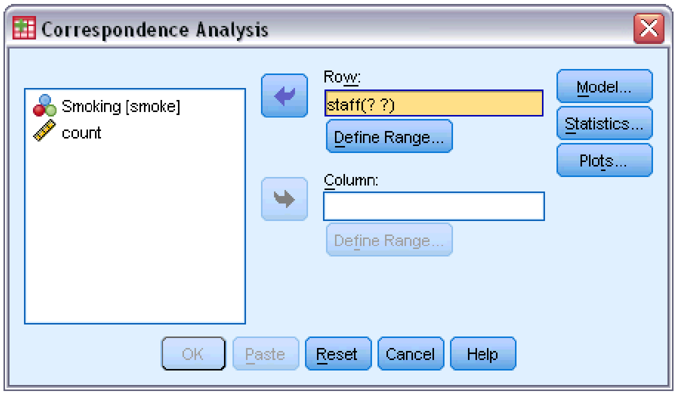
Related procedures. If more than two variables are involved, use multiple correspondence analysis. If the variables should be scaled ordinally, use categorical principal components analysis.

To Obtain a Correspondence Analysis

▶ From the menus choose:

Analyze > Dimension Reduction > Correspondence Analysis...

Figure 5-1
Correspondence Analysis dialog box



Correspondence Analysis

Row: staff(? ?)

Column:

▶ Select a row variable.

▶ Select a column variable.

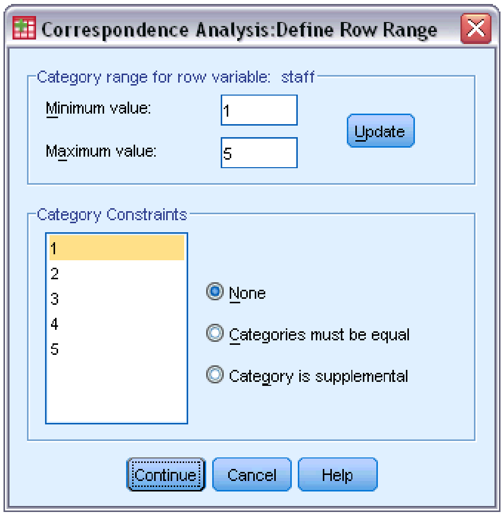
▶ Define the ranges for the variables.

▶ Click OK.

Define Row Range in Correspondence Analysis

You must define a range for the row variable. The minimum and maximum values specified must be integers. Fractional data values are truncated in the analysis. A category value that is outside of the specified range is ignored in the analysis.

Figure 5-2
Define Row Range dialog box



Correspondence Analysis:Define Row Range

Category range for row variable: staff

Minimum value: 1

Maximum value: 5

Category Constraints

1
2
3
4
5

● None

○ Categories must be equal

○ Category is supplemental

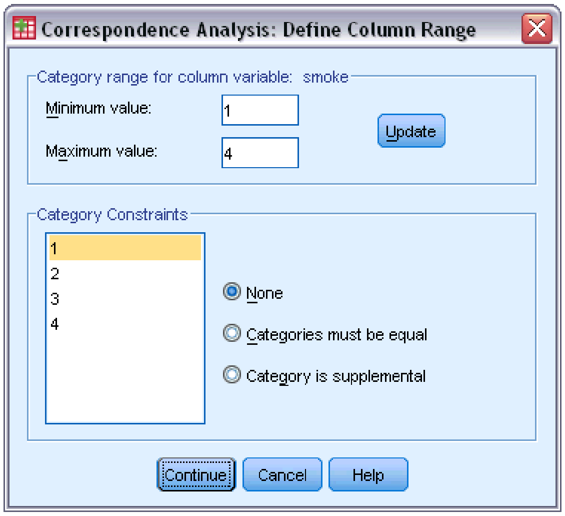
All categories are initially unconstrained and active. You can constrain row categories to equal other row categories, or you can define a row category as supplementary.

- **Categories must be equal.** Categories must have equal scores. Use equality constraints if the obtained order for the categories is undesirable or counterintuitive. The maximum number of row categories that can be constrained to be equal is the total number of active row categories minus 1. To impose different equality constraints on sets of categories, use syntax. For example, use syntax to constrain categories 1 and 2 to be equal and categories 3 and 4 to be equal.
- **Category is supplemental.** Supplementary categories do not influence the analysis but are represented in the space defined by the active categories. Supplementary categories play no role in defining the dimensions. The maximum number of supplementary row categories is the total number of row categories minus 2.

Define Column Range in Correspondence Analysis

You must define a range for the column variable. The minimum and maximum values specified must be integers. Fractional data values are truncated in the analysis. A category value that is outside of the specified range is ignored in the analysis.

Figure 5-3
Define Column Range dialog box



Correspondence Analysis: Define Column Range

Category range for column variable: smoke

Minimum value: 1

Maximum value: 4

Category Constraints

1
2
3
4

● None

○ Categories must be equal

○ Category is supplemental

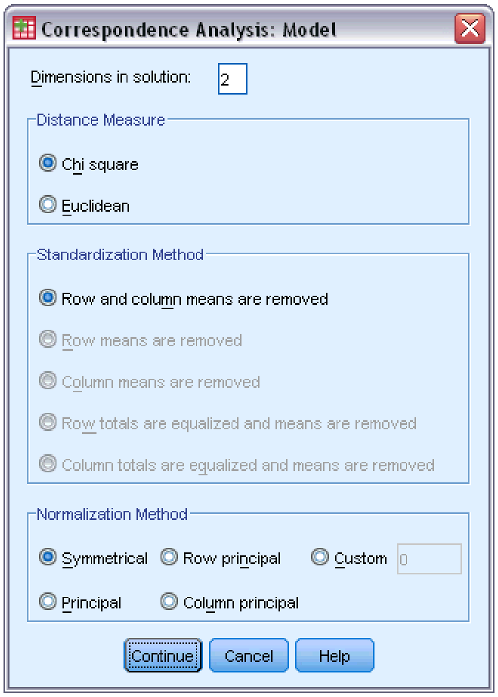
All categories are initially unconstrained and active. You can constrain column categories to equal other column categories, or you can define a column category as supplementary.

- **Categories must be equal.** Categories must have equal scores. Use equality constraints if the obtained order for the categories is undesirable or counterintuitive. The maximum number of column categories that can be constrained to be equal is the total number of active column categories minus 1. To impose different equality constraints on sets of categories, use syntax. For example, use syntax to constrain categories 1 and 2 to be equal and categories 3 and 4 to be equal.
- **Category is supplemental.** Supplementary categories do not influence the analysis but are represented in the space defined by the active categories. Supplementary categories play no role in defining the dimensions. The maximum number of supplementary column categories is the total number of column categories minus 2.

Correspondence Analysis Model

The Model dialog box allows you to specify the number of dimensions, the distance measure, the standardization method, and the normalization method.

Figure 5-4
Model dialog box



Correspondence Analysis: Model

Dimensions in solution: 2

Distance Measure

● Chi square

○ Euclidean

Standardization Method

● Row and column means are removed

○ Row means are removed

○ Column means are removed

○ Row totals are equalized and means are removed

○ Column totals are equalized and means are removed

Normalization Method

● Symmetrical ○ Row principal ○ Custom 0

○ Principal ○ Column principal

Dimensions in solution. Specify the number of dimensions. In general, choose as few dimensions as needed to explain most of the variation. The maximum number of dimensions depends on the number of active categories used in the analysis and on the equality constraints. The maximum number of dimensions is the smaller of:

- The number of active row categories minus the number of row categories constrained to be equal, plus the number of constrained row category sets.
- The number of active column categories minus the number of column categories constrained to be equal, plus the number of constrained column category sets.

Distance Measure. You can select the measure of distance among the rows and columns of the correspondence table. Choose one of the following alternatives:

- **Chi-square.** Use a weighted profile distance, where the weight is the mass of the rows or columns. This measure is required for standard correspondence analysis.
- **Euclidean.** Use the square root of the sum of squared differences between pairs of rows and pairs of columns.

Standardization Method. Choose one of the following alternatives:

- **Row and column means are removed.** Both the rows and columns are centered. This method is required for standard correspondence analysis.
- **Row means are removed.** Only the rows are centered.
- **Column means are removed.** Only the columns are centered.

- **Row totals are equalized and means are removed.** Before centering the rows, the row margins are equalized.
- **Column totals are equalized and means are removed.** Before centering the columns, the column margins are equalized.

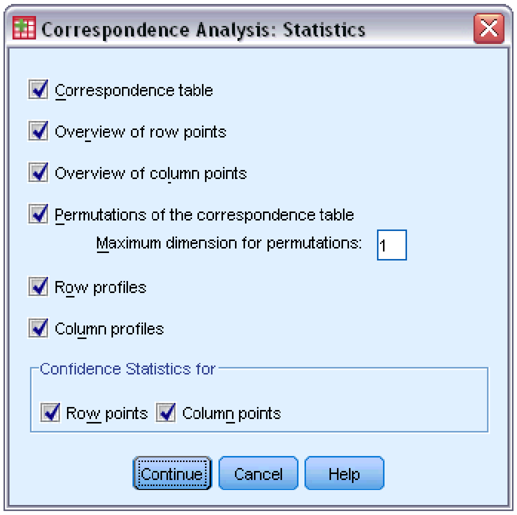
Normalization Method. Choose one of the following alternatives:

- **Symmetrical.** For each dimension, the row scores are the weighted average of the column scores divided by the matching singular value, and the column scores are the weighted average of row scores divided by the matching singular value. Use this method if you want to examine the differences or similarities between the categories of the two variables.
- **Principal**. The distances between row points and column points are approximations of the distances in the correspondence table according to the selected distance measure. Use this method if you want to examine differences between categories of either or both variables instead of differences between the two variables.
- **Row principal.** The distances between row points are approximations of the distances in the correspondence table according to the selected distance measure. The row scores are the weighted average of the column scores. Use this method if you want to examine differences or similarities between categories of the row variable.
- **Column principal.** The distances between column points are approximations of the distances in the correspondence table according to the selected distance measure. The column scores are the weighted average of the row scores. Use this method if you want to examine differences or similarities between categories of the column variable.
- **Custom.** You must specify a value between –1 and 1. A value of –1 corresponds to column principal. A value of 1 corresponds to row principal. A value of 0 corresponds to symmetrical. All other values spread the inertia over both the row and column scores to varying degrees. This method is useful for making tailor-made biplots.

Correspondence Analysis Statistics

The Statistics dialog box allows you to specify the numerical output produced.

Figure 5-5
Statistics dialog box



Correspondence Analysis: Statistics

☑ Correspondence table

☑ Overview of row points

☑ Overview of column points

☑ Permutations of the correspondence table

Maximum dimension for permutations: 1

☑ Row profiles

☑ Column profiles

Confidence Statistics for

☑ Row points ☑ Column points

Correspondence table. A crosstabulation of the input variables with row and column marginal totals.

Overview of row points. For each row category, the scores, mass, inertia, contribution to the inertia of the dimension, and the contribution of the dimension to the inertia of the point.

Overview of column points. For each column category, the scores, mass, inertia, contribution to the inertia of the dimension, and the contribution of the dimension to the inertia of the point.

Row profiles. For each row category, the distribution across the categories of the column variable.

Column profiles. For each column category, the distribution across the categories of the row variable.

Permutations of the correspondence table. The correspondence table reorganized such that the rows and columns are in increasing order according to the scores on the first dimension. Optionally, you can specify the maximum dimension number for which permuted tables will be produced. A permuted table for each dimension from 1 to the number specified is produced.

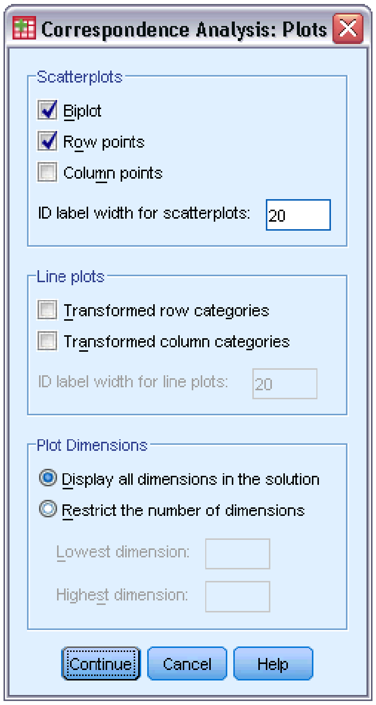
Confidence Statistics for Row points. Includes standard deviation and correlations for all nonsupplementary row points.

Confidence Statistics for Column points. Includes standard deviation and correlations for all nonsupplementary column points.

Correspondence Analysis Plots

The Plots dialog box allows you to specify which plots are produced.

Figure 5-6
Plots dialog box



Scatterplots

☑ Biplot

☑ Row points

☐ Column points

ID label width for scatterplots: 20

Line plots

☐ Transformed row categories

☐ Transformed column categories

ID label width for line plots: 20

Plot Dimensions

● Display all dimensions in the solution

○ Restrict the number of dimensions

Lowest dimension:

Highest dimension:

Scatterplots. Produces a matrix of all pairwise plots of the dimensions. Available scatterplots include:

- **Biplot.** Produces a matrix of joint plots of the row and column points. If principal normalization is selected, the biplot is not available.
- **Row points.** Produces a matrix of plots of the row points.
- **Column points.** Produces a matrix of plots of the column points.

Optionally, you can specify how many value label characters to use when labeling the points. This value must be a non-negative integer less than or equal to 20.

Line Plots. Produces a plot for every dimension of the selected variable. Available line plots include:

- **Transformed row categories.** Produces a plot of the original row category values against their corresponding row scores.
- **Transformed column categories.** Produces a plot of the original column category values against their corresponding column scores.

Optionally, you can specify how many value label characters to use when labeling the category axis. This value must be a non-negative integer less than or equal to 20.

Plot Dimensions. Allows you to control the dimensions displayed in the output.

- **Display all dimensions in the solution.** All dimensions in the solution are displayed in a scatterplot matrix.
- **Restrict the number of dimensions.** The displayed dimensions are restricted to plotted pairs. If you restrict the dimensions, you must select the lowest and highest dimensions to be plotted. The lowest dimension can range from 1 to the number of dimensions in the solution minus 1, and is plotted against higher dimensions. The highest dimension value can range from 2 to the number of dimensions in the solution, and indicates the highest dimension to be used in plotting the dimension pairs. This specification applies to all requested multidimensional plots.

CORRESPONDENCE Command Additional Features

You can customize your correspondence analysis if you paste your selections into a syntax window and edit the resulting `CORRESPONDENCE` command syntax. The command syntax language also allows you to:

- Specify table data as input instead of using casewise data (using the `TABLE = ALL` subcommand).
- Specify the number of value-label characters used to label points for each type of scatterplot matrix or biplot matrix (with the `PLOT` subcommand).
- Specify the number of value-label characters used to label points for each type of line plot (with the `PLOT` subcommand).
- Write a matrix of row and column scores to a matrix data file (with the `OUTFILE` subcommand).
- Write a matrix of confidence statistics (variances and covariances) for the singular values and the scores to a matrix data file (with the `OUTFILE` subcommand).
- Specify multiple sets of categories to be equal (with the `EQUAL` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter

6

Multiple Correspondence Analysis

Multiple Correspondence Analysis quantifies nominal (categorical) data by assigning numerical values to the cases (objects) and categories so that objects within the same category are close together and objects in different categories are far apart. Each object is as close as possible to the category points of categories that apply to the object. In this way, the categories divide the objects into homogeneous subgroups. Variables are considered homogeneous when they classify objects in the same categories into the same subgroups.

Example. Multiple Correspondence Analysis could be used to graphically display the relationship between job category, minority classification, and gender. You might find that minority classification and gender discriminate between people but that job category does not. You might also find that the Latino and African-American categories are similar to each other.

Statistics and plots. Object scores, discrimination measures, iteration history, correlations of original and transformed variables, category quantifications, descriptive statistics, object points plots, biplots, category plots, joint category plots, transformation plots, and discrimination measures plots.

Data. String variable values are always converted into positive integers by ascending alphanumeric order. User-defined missing values, system-missing values, and values less than 1 are considered missing; you can recode or add a constant to variables with values less than 1 to make them nonmissing.

Assumptions. All variables have the multiple nominal scaling level. The data must contain at least three valid cases. The analysis is based on positive integer data. The discretization option will automatically categorize a fractional-valued variable by grouping its values into categories with a close-to-normal distribution and will automatically convert values of string variables into positive integers. You can specify other discretization schemes.

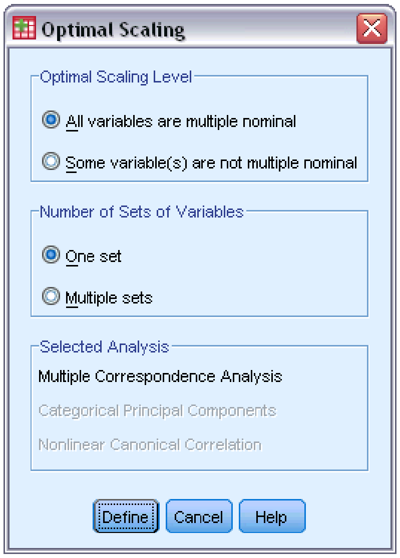
Related procedures. For two variables, Multiple Correspondence Analysis is analogous to Correspondence Analysis. If you believe that variables possess ordinal or numerical properties, Categorical Principal Components Analysis should be used. If sets of variables are of interest, Nonlinear Canonical Correlation Analysis should be used.

To Obtain a Multiple Correspondence Analysis

▶ From the menus choose:

Analyze > Dimension Reduction > Optimal Scaling...

Figure 6-1
Optimal Scaling dialog box


Optimal Scaling

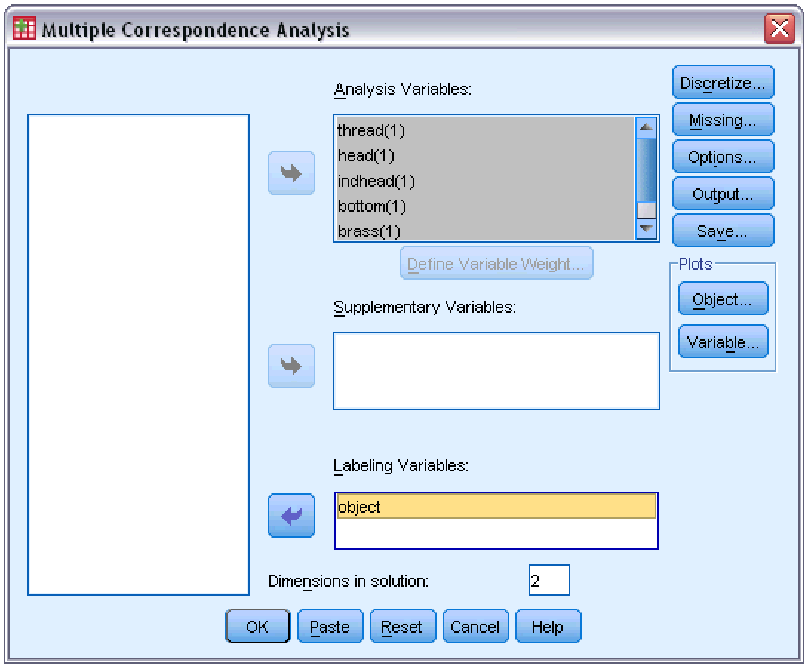
Optimal Scaling Level
- ● All variables are multiple nominal
- ○ Some variable(s) are not multiple nominal

Number of Sets of Variables
- ● One set
- ○ Multiple sets

Selected Analysis
Multiple Correspondence Analysis
Categorical Principal Components
Nonlinear Canonical Correlation

- ▶ Select All variables multiple nominal.
- ▶ Select One set.
- ▶ Click Define.

Figure 6-2
Multiple Correspondence Analysis dialog box


Multiple Correspondence Analysis

Analysis Variables:
thread(1)
head(1)
indhead(1)
bottom(1)
brass(1)

Supplementary Variables:

Labeling Variables:
object

Dimensions in solution: 2

- ▶ Select at least two analysis variables and specify the number of dimensions in the solution.

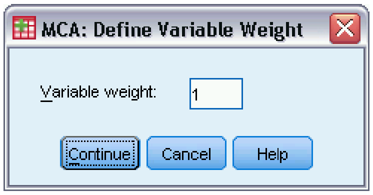
▶ Click OK.

You may optionally specify supplementary variables, which are fitted into the solution found, or labeling variables for the plots.

Define Variable Weight in Multiple Correspondence Analysis

You can set the weight for analysis variables.

Figure 6-3
Define Variable Weight dialog box



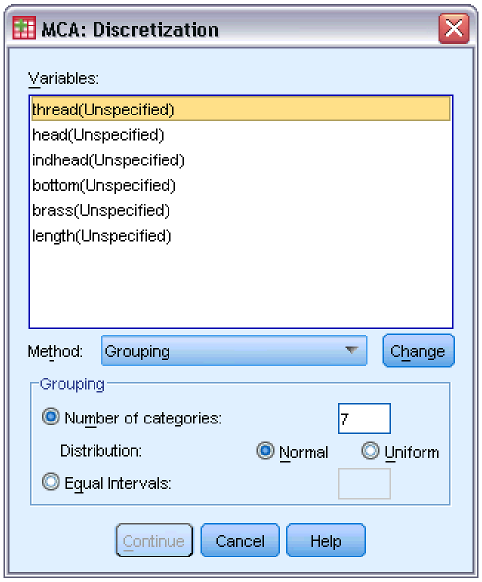
Variable weight: 1

Variable weight. You can choose to define a weight for each variable. The value specified must be a positive integer. The default value is 1.

Multiple Correspondence Analysis Discretization

The Discretization dialog box allows you to select a method of recoding your variables. Fractional-valued variables are grouped into seven categories (or into the number of distinct values of the variable if this number is less than seven) with an approximately normal distribution unless otherwise specified. String variables are always converted into positive integers by assigning category indicators according to ascending alphanumeric order. Discretization for string variables applies to these integers. Other variables are left alone by default. The discretized variables are then used in the analysis.

Figure 6-4
Discretization dialog box



MCA: Discretization

Variables:

thread(Unspecified)
head(Unspecified)
indhead(Unspecified)
bottom(Unspecified)
brass(Unspecified)
length(Unspecified)

Method: Grouping

Grouping

● Number of categories: 7

Distribution: ● Normal ○ Uniform

○ Equal Intervals:

Method. Choose between grouping, ranking, and multiplying.

- **Grouping.** Recode into a specified number of categories or recode by interval.
- **Ranking.** The variable is discretized by ranking the cases.
- **Multiplying.** The current values of the variable are standardized, multiplied by 10, rounded, and have a constant added so that the lowest discretized value is 1.

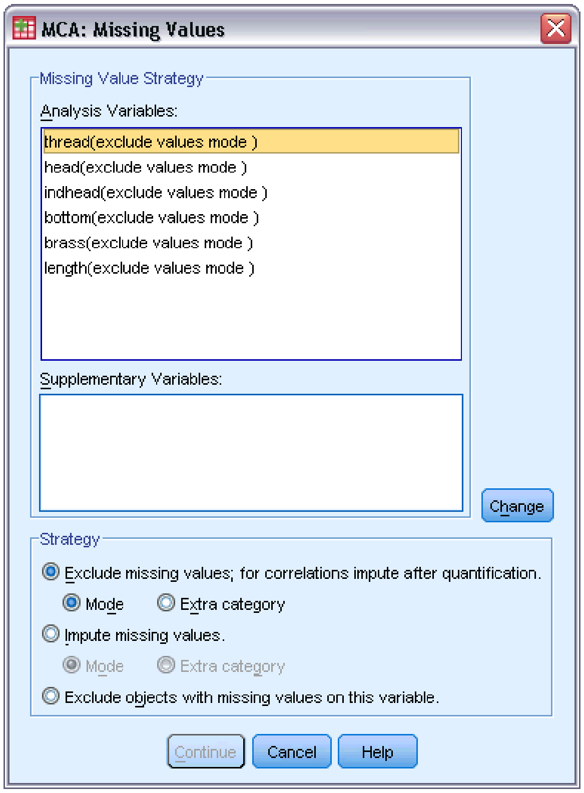
Grouping. The following options are available when discretizing variables by grouping:

- **Number of categories.** Specify a number of categories and whether the values of the variable should follow an approximately normal or uniform distribution across those categories.
- **Equal intervals.** Variables are recoded into categories defined by these equally sized intervals. You must specify the length of the intervals.

Multiple Correspondence Analysis Missing Values

The Missing Values dialog box allows you to choose the strategy for handling missing values in analysis variables and supplementary variables.

Figure 6-5
Missing Values dialog box



MCA: Missing Values

Missing Value Strategy

Analysis Variables:

thread(exclude values mode)
head(exclude values mode)
indhead(exclude values mode)
bottom(exclude values mode)
brass(exclude values mode)
length(exclude values mode)

Supplementary Variables:

Strategy

● Exclude missing values; for correlations impute after quantification.

● Mode ○ Extra category

○ Impute missing values.

● Mode ○ Extra category

○ Exclude objects with missing values on this variable.

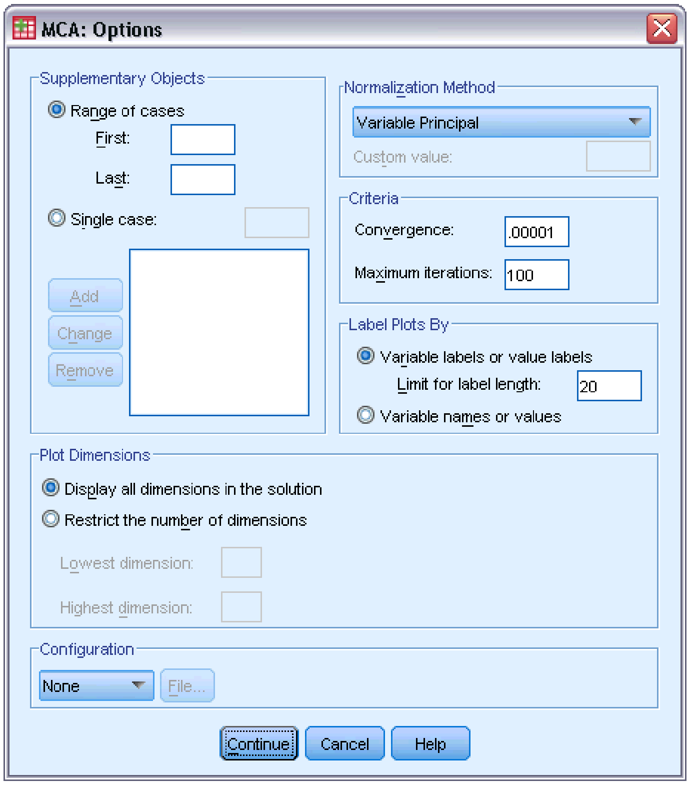
Missing Value Strategy. Choose to exclude missing values (passive treatment), impute missing values (active treatment), or exclude objects with missing values (listwise deletion).

- **Exclude missing values; for correlations impute after quantification.** Objects with missing values on the selected variable do not contribute to the analysis for this variable. If all variables are given passive treatment, then objects with missing values on all variables are treated as supplementary. If correlations are specified in the Output dialog box, then (after analysis) missing values are imputed with the most frequent category, or mode, of the variable for the correlations of the original variables. For the correlations of the optimally scaled variables, you can choose the method of imputation. Select Mode to replace missing values with the mode of the optimally scaled variable. Select Extra category to replace missing values with the quantification of an extra category. This implies that objects with a missing value on this variable are considered to belong to the same (extra) category.
- **Impute missing values.** Objects with missing values on the selected variable have those values imputed. You can choose the method of imputation. Select Mode to replace missing values with the most frequent category. When there are multiple modes, the one with the smallest category indicator is used. Select Extra category to replace missing values with the same quantification of an extra category. This implies that objects with a missing value on this variable are considered to belong to the same (extra) category.
- **Exclude objects with missing values on this variable.** Objects with missing values on the selected variable are excluded from the analysis. This strategy is not available for supplementary variables.

Multiple Correspondence Analysis Options

The Options dialog box allows you to select the initial configuration, specify iteration and convergence criteria, select a normalization method, choose the method for labeling plots, and specify supplementary objects.

Figure 6-6
Options dialog box



MCA: Options

Supplementary Objects
● Range of cases
First:
Last:
○ Single case:

Normalization Method
Variable Principal
Custom value:

Criteria
Convergence: .00001
Maximum iterations: 100

Label Plots By
● Variable labels or value labels
Limit for label length: 20
○ Variable names or values

Plot Dimensions
● Display all dimensions in the solution
○ Restrict the number of dimensions
Lowest dimension:
Highest dimension:

Configuration
None

Supplementary Objects. Specify the case number of the object (or the first and last case numbers of a range of objects) that you want to make supplementary, and then click Add. Continue until you have specified all of your supplementary objects. If an object is specified as supplementary, then case weights are ignored for that object.

Normalization Method. You can specify one of five options for normalizing the object scores and the variables. Only one normalization method can be used in a given analysis.

- **Variable Principal.** This option optimizes the association between variables. The coordinates of the variables in the object space are the component loadings (correlations with principal components, such as dimensions and object scores). This is useful when you are interested primarily in the correlation between the variables.
- **Object Principal.** This option optimizes distances between objects. This is useful when you are interested primarily in differences or similarities between the objects.
- **Symmetrical.** Use this normalization option if you are interested primarily in the relation between objects and variables.

- **Independent.** Use this normalization option if you want to examine distances between objects and correlations between variables separately.
- **Custom.** You can specify any real value in the closed interval [–1, 1]. A value of 1 is equal to the Object Principal method, a value of 0 is equal to the Symmetrical method, and a value of –1 is equal to the Variable Principal method. By specifying a value greater than –1 and less than 1, you can spread the eigenvalue over both objects and variables. This method is useful for making a tailor-made biplot or triplot.

Criteria. You can specify the maximum number of iterations the procedure can go through in its computations. You can also select a convergence criterion value. The algorithm stops iterating if the difference in total fit between the last two iterations is less than the convergence value or if the maximum number of iterations is reached.

Label Plots By. Allows you to specify whether variables and value labels or variable names and values will be used in the plots. You can also specify a maximum length for labels.

Plot Dimensions. Allows you to control the dimensions displayed in the output.

- **Display all dimensions in the solution.** All dimensions in the solution are displayed in a scatterplot matrix.
- **Restrict the number of dimensions.** The displayed dimensions are restricted to plotted pairs. If you restrict the dimensions, you must select the lowest and highest dimensions to be plotted. The lowest dimension can range from 1 to the number of dimensions in the solution minus 1 and is plotted against higher dimensions. The highest dimension value can range from 2 to the number of dimensions in the solution and indicates the highest dimension to be used in plotting the dimension pairs. This specification applies to all requested multidimensional plots.

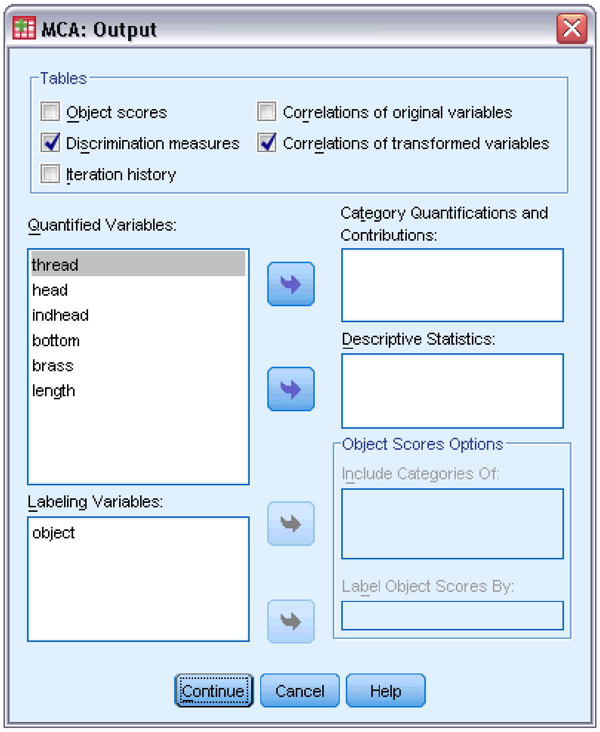
Configuration. You can read data from a file containing the coordinates of a configuration. The first variable in the file should contain the coordinates for the first dimension, the second variable should contain the coordinates for the second dimension, and so on.

- **Initial.** The configuration in the file specified will be used as the starting point of the analysis.
- **Fixed.** The configuration in the file specified will be used to fit in the variables. The variables that are fitted in must be selected as analysis variables, but, because the configuration is fixed, they are treated as supplementary variables (so they do not need to be selected as supplementary variables).

Multiple Correspondence Analysis Output

The Output dialog box allows you to produce tables for object scores, discrimination measures, iteration history, correlations of original and transformed variables, category quantifications for selected variables, and descriptive statistics for selected variables.

Figure 6-7
Output dialog box



MCA: Output

Tables

☐ Object scores
☑ Discrimination measures
☐ Iteration history
☐ Correlations of original variables
☑ Correlations of transformed variables

Quantified Variables: thread, head, indhead, bottom, brass, length

Labeling Variables: object

Object scores. Displays the object scores, including mass, inertia, and contributions, and has the following options:

- **Include Categories Of.** Displays the category indicators of the analysis variables selected.
- **Label Object Scores By.** From the list of variables specified as labeling variables, you can select one to label the objects.

Discrimination measures. Displays the discrimination measures per variable and per dimension.

Iteration history. For each iteration, the variance accounted for, loss, and increase in variance accounted for are shown.

Correlations of original variables. Shows the correlation matrix of the original variables and the eigenvalues of that matrix.

Correlations of transformed variables. Shows the correlation matrix of the transformed (optimally scaled) variables and the eigenvalues of that matrix.

Category Quantifications and Contributions. Gives the category quantifications (coordinates), including mass, inertia, and contributions, for each dimension of the variable(s) selected.

Note: the coordinates and contributions (including the mass and inertia) are displayed in separate layers of the pivot table output, with the coordinates shown by default. To display the contributions, activate (double-click) on the table and select Contributions from the Layer dropdown list.

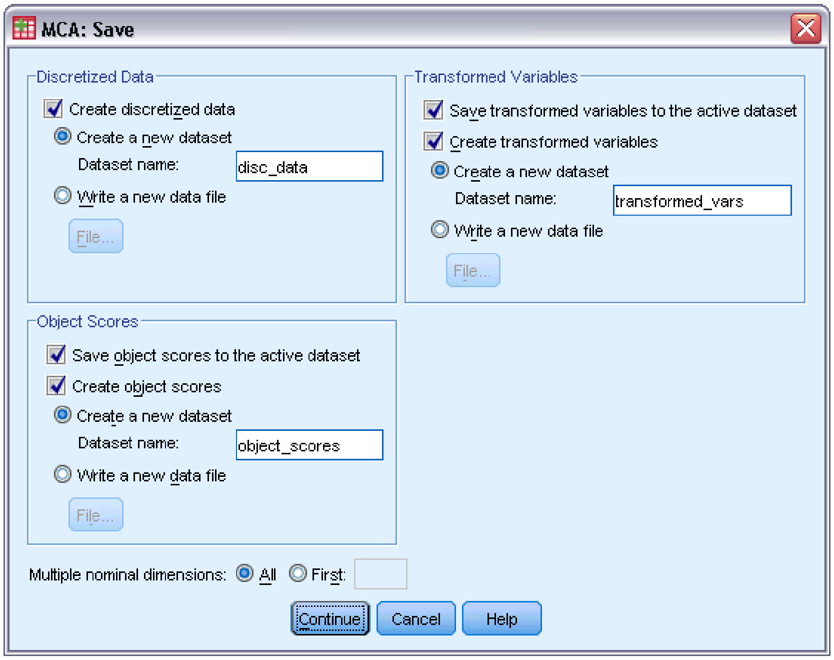
Descriptive Statistics. Displays frequencies, number of missing values, and mode of the variable(s) selected.

Multiple Correspondence Analysis Save

The Save dialog box allows you to save discretized data, object scores, and transformed values to an external IBM® SPSS® Statistics data file or dataset in the current session. You can also save transformed values and object scores to the active dataset.

- Datasets are available during the current session but are not available in subsequent sessions unless you explicitly save them as data files. Dataset names must adhere to variable naming rules.
- Filenames or dataset names must be different for each type of data saved.
- If you save object scores or transformed values to the active dataset, you can specify the number of multiple nominal dimensions.

Figure 6-8
Save dialog box



MCA: Save

Discretized Data
☑ Create discretized data
● Create a new dataset
Dataset name: disc_data
○ Write a new data file

Transformed Variables
☑ Save transformed variables to the active dataset
☑ Create transformed variables
● Create a new dataset
Dataset name: transformed_vars
○ Write a new data file

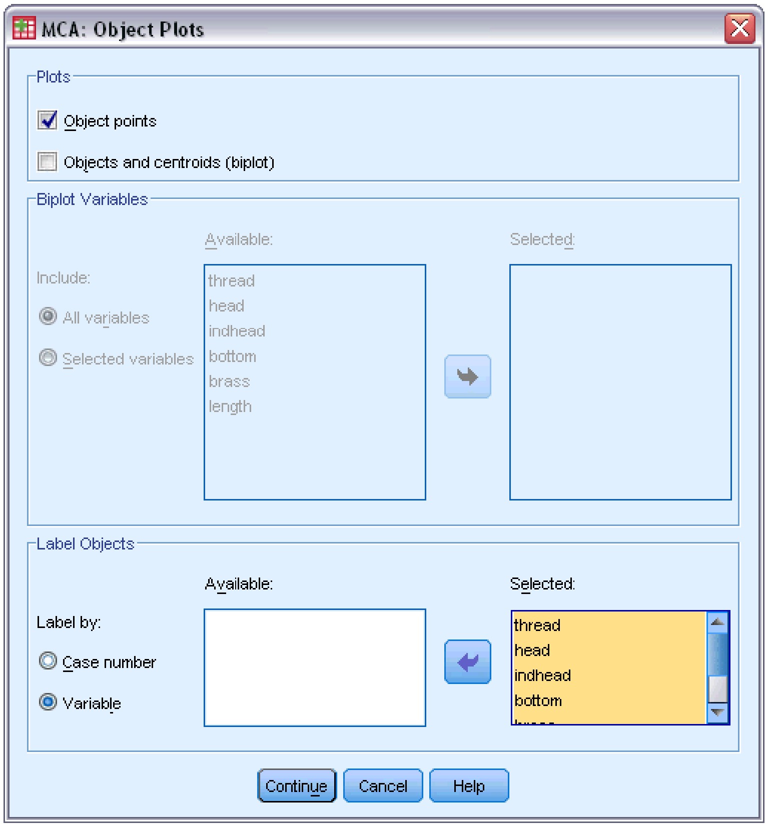
Object Scores
☑ Save object scores to the active dataset
☑ Create object scores
● Create a new dataset
Dataset name: object_scores
○ Write a new data file

Multiple nominal dimensions: ● All ○ First:

Multiple Correspondence Analysis Object Plots

The Object Plots dialog box allows you to specify the types of plots desired and the variables to be plotted

Figure 6-9
Object Plots dialog box



MCA: Object Plots

Plots

☑ Object points

□ Objects and centroids (biplot)

Biplot Variables

Include:

● All variables

○ Selected variables

Available:

thread
head
indhead
bottom
brass
length

Selected:

Label Objects

Label by:

○ Case number

● Variable

Available:

Selected:

thread
head
indhead
bottom

Object points. A plot of the object points is displayed.

Objects and centroids (biplot). The object points are plotted with the variable centroids.

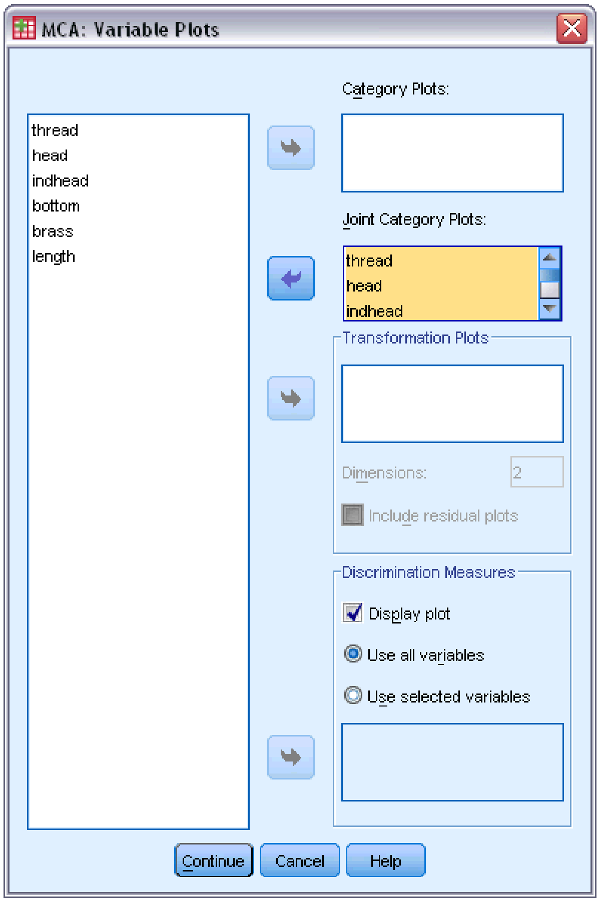
Biplot Variables. You can choose to use all variables for the biplots or select a subset.

Label Objects. You can choose to have objects labeled with the categories of selected variables (you may choose category indicator values or value labels in the Options dialog box) or with their case numbers. One plot is produced per variable if Variable is selected.

Multiple Correspondence Analysis Variable Plots

The Variable Plots dialog box allows you to specify the types of plots desired and the variables to be plotted.

Figure 6-10
Variable Plots dialog box



thread
head
indhead
bottom
brass
length

Category Plots:

Joint Category Plots:
thread
head
indhead

Transformation Plots

Dimensions: 2

□ Include residual plots

Discrimination Measures

☑ Display plot

● Use all variables

○ Use selected variables

Category Plots. For each variable selected, a plot of the centroid coordinates is plotted. Categories are in the centroids of the objects in the particular categories.

Joint Category Plots. This is a single plot of the centroid coordinates of each selected variable.

Transformation Plots. Displays a plot of the optimal category quantifications versus the category indicators. You can specify the number of dimensions desired; one plot will be generated for each dimension. You can also choose to display residual plots for each variable selected.

Discrimination Measures. Produces a single plot of the discrimination measures for the selected variables.

MULTIPLE CORRESPONDENCE Command Additional Features

You can customize your Multiple Correspondence Analysis if you paste your selections into a syntax window and edit the resulting `MULTIPLE CORRESPONDENCE` command syntax. The command syntax language also allows you to:

- Specify rootnames for the transformed variables, object scores, and approximations when saving them to the active dataset (with the `SAVE` subcommand).

- Specify a maximum length for labels for each plot separately (with the PLOT subcommand).
- Specify a separate variable list for residual plots (with the PLOT subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter

7

Multidimensional Scaling (PROXSCAL)

Multidimensional scaling attempts to find the structure in a set of proximity measures between objects. This process is accomplished by assigning observations to specific locations in a conceptual low-dimensional space such that the distances between points in the space match the given (dis)similarities as closely as possible. The result is a least-squares representation of the objects in that low-dimensional space, which, in many cases, will help you to further understand your data.

Example. Multidimensional scaling can be very useful in determining perceptual relationships. For example, when considering your product image, you can conduct a survey to obtain a dataset that describes the perceived similarity (or proximity) of your product to those of your competitors. Using these proximities and independent variables (such as price), you can try to determine which variables are important to how people view these products, and you can adjust your image accordingly.

Statistics and plots. Iteration history, stress measures, stress decomposition, coordinates of the common space, object distances within the final configuration, individual space weights, individual spaces, transformed proximities, transformed independent variables, stress plots, common space scatterplots, individual space weight scatterplots, individual spaces scatterplots, transformation plots, Shepard residual plots, and independent variables transformation plots.

Data. Data can be supplied in the form of proximity matrices or variables that are converted into proximity matrices. The matrices can be formatted in columns or across columns. The proximities can be treated on the ratio, interval, ordinal, or spline scaling levels.

Assumptions. At least three variables must be specified. The number of dimensions cannot exceed the number of objects minus one. Dimensionality reduction is omitted if combined with multiple random starts. If only one source is specified, all models are equivalent to the identity model; therefore, the analysis defaults to the identity model.

Related procedures. Scaling all variables at the numerical level corresponds to standard multidimensional scaling analysis.

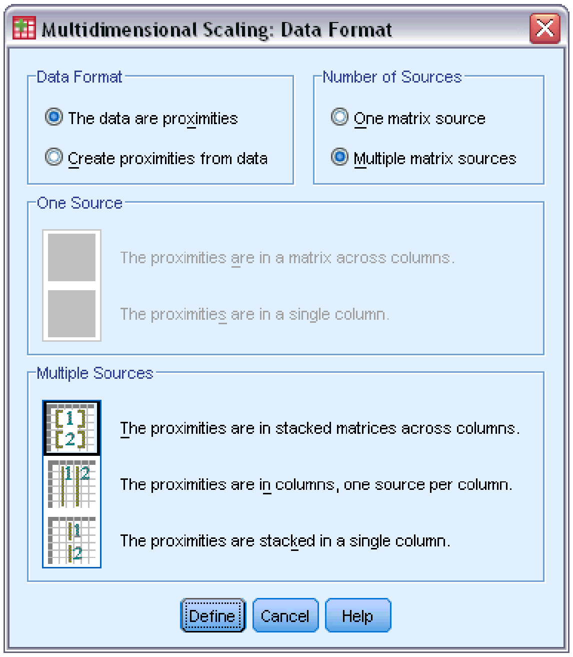
To Obtain a Multidimensional Scaling

- ▶ From the menus choose:

  Analyze > Scale > Multidimensional Scaling (PROXSCAL)...

  This opens the Data Format dialog box.

Figure 7-1
Data Format dialog box



Multidimensional Scaling: Data Format

Data Format
● The data are proximities
○ Create proximities from data

Number of Sources
○ One matrix source
● Multiple matrix sources

One Source
The proximities are in a matrix across columns.
The proximities are in a single column.

Multiple Sources
The proximities are in stacked matrices across columns.
The proximities are in columns, one source per column.
The proximities are stacked in a single column.

- ▶ Specify the format of your data:

Data Format. Specify whether your data consist of proximity measures or you want to create proximities from the data.

Number of Sources. If your data are proximities, specify whether you have a single source or multiple sources of proximity measures.

One Source. If there is one source of proximities, specify whether your dataset is formatted with the proximities in a matrix across the columns or in a single column with two separate variables to identify the row and column of each proximity.

- **The proximities are in a matrix across columns.** The proximity matrix is spread across a number of columns equal to the number of objects. This leads to the Proximities in Matrices across Columns dialog box.
- **The proximities are in a single column.** The proximity matrix is collapsed into a single column, or variable. Two additional variables, identifying the row and column for each cell, are necessary. This leads to the Proximities in One Column dialog box.

Multiple Sources. If there are multiple sources of proximities, specify whether the dataset is formatted with the proximities in stacked matrices across columns, in multiple columns with one source per column, or in a single column.

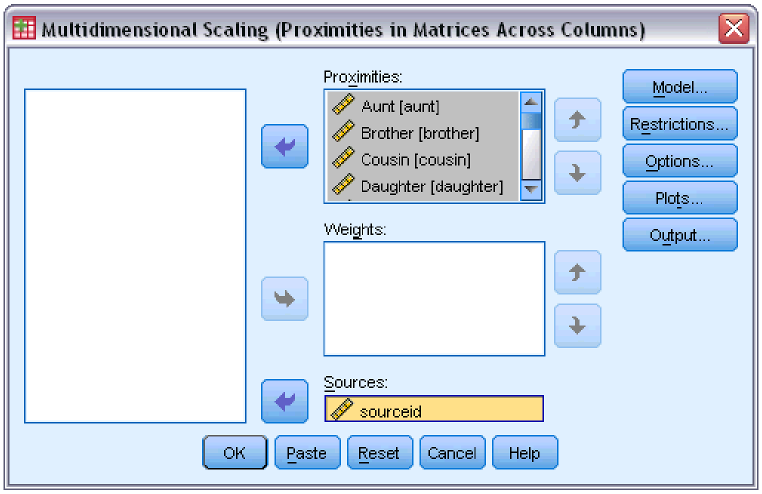
- **The proximities are in stacked matrices across columns.** The proximity matrices are spread across a number of columns equal to the number of objects and are stacked above one another across a number of rows equal to the number of objects times the number of sources. This leads to the Proximities in Matrices across Columns dialog box.
- **The proximities are in columns, one source per column.** The proximity matrices are collapsed into multiple columns, or variables. Two additional variables, identifying the row and column for each cell, are necessary. This leads to the Proximities in Columns dialog box.
- **The proximites are stacked in a single column.** The proximity matrices are collapsed into a single column, or variable. Three additional variables, identifying the row, column, and source for each cell, are necessary. This leads to the Proximities in One Column dialog box.

▶ Click Define.

Proximities in Matrices across Columns

If you select the proximities in matrices data model for either one source or multiple sources in the Data Format dialog box, the main dialog box will appear as follows:

Figure 7-2
Proximities in Matrices across Columns dialog box



Multidimensional Scaling (Proximities in Matrices Across Columns)

Proximities:
Aunt [aunt]
Brother [brother]
Cousin [cousin]
Daughter [daughter]

Weights:

Sources:
sourceid

▶ Select three or more proximities variables. (Be sure that the order of the variables in the list matches the order of the columns of the proximities.)

▶ Optionally, select a number of weights variables equal to the number of proximities variables. (Be sure that the order of the weights matches the order of the proximities that they weight.)

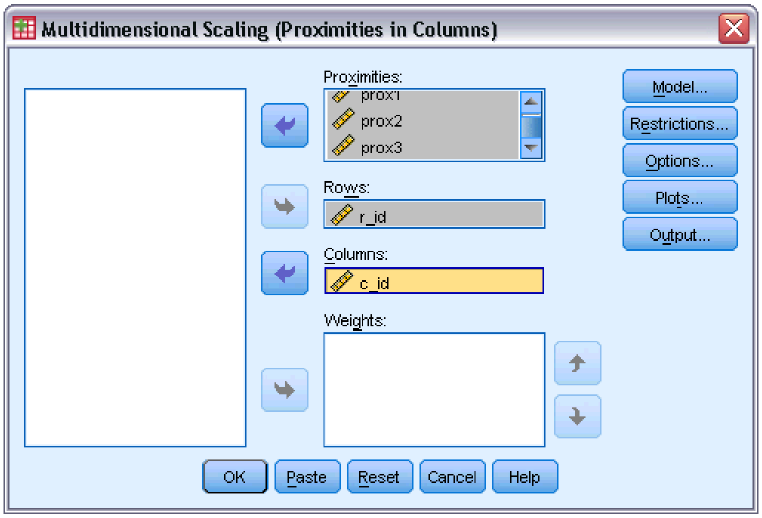
▶ Optionally, if there are multiple sources, select a sources variable. (The number of cases in each proximities variable should equal the number of proximities variables times the number of sources.)

Additionally, you can define a model for the multidimensional scaling, place restrictions on the common space, set convergence criteria, specify the initial configuration to be used, and choose plots and output.

Proximities in Columns

If you select the multiple columns model for multiple sources in the Data Format dialog box, the main dialog box will appear as follows:

Figure 7-3
Proximities in Columns dialog box



Proximities:
prox1
prox2
prox3

Rows:
r_id

Columns:
c_id

- ▶ Select two or more proximities variables. (Each variable is assumed to be a matrix of proximities from a separate source.)

- ▶ Select a rows variable to define the row locations for the proximities in each proximities variable.

- ▶ Select a columns variable to define the column locations for the proximities in each proximities variable. (Cells of the proximity matrix that are not given a row/column designation are treated as missing.)

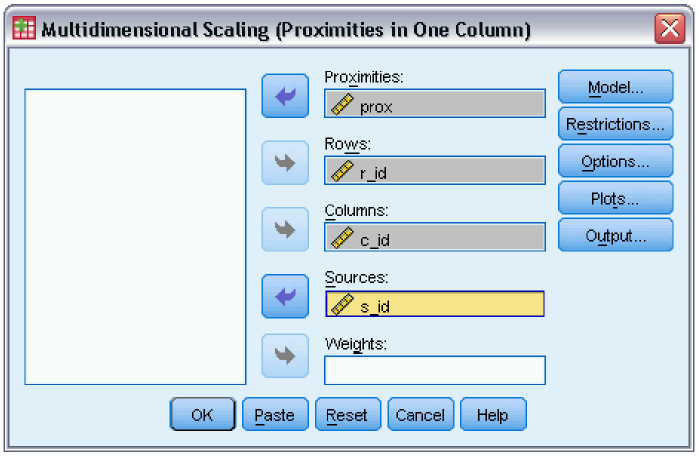
- ▶ Optionally, select a number of weights variables equal to the number of proximities variables.

Additionally, you can define a model for the multidimensional scaling, place restrictions on the common space, set convergence criteria, specify the initial configuration to be used, and choose plots and output.

Proximities in One Column

If you select the one column model for either one source or multiple sources in the Data Format dialog box, the main dialog box will appear as follows:

Figure 7-4
Proximities in One Column dialog box



Multidimensional Scaling (Proximities in One Column)

Proximities: prox

Rows: r_id

Columns: c_id

Sources: s_id

Weights:

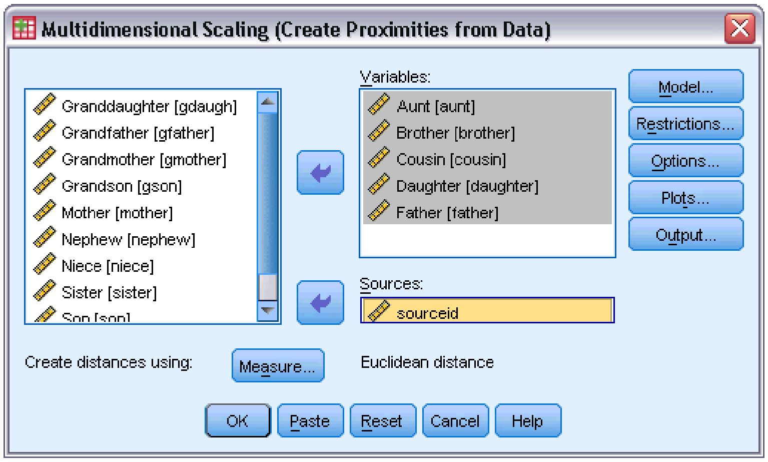
- ▶ Select a proximities variable. (t is assumed to be one or more matrices of proximities.)
- ▶ Select a rows variable to define the row locations for the proximities in the proximities variable.
- ▶ Select a columns variable to define the column locations for the proximities in the proximities variable.
- ▶ If there are multiple sources, select a sources variable. (For each source, cells of the proximity matrix that are not given a row/column designation are treated as missing.)
- ▶ Optionally, select a weights variable.

Additionally, you can define a model for the multidimensional scaling, place restrictions on the common space, set convergence criteria, specify the initial configuration to be used, and choose plots and output.

Create Proximities from Data

If you choose to create proximities from the data in the Data Format dialog box, the main dialog box will appear as follows:

Figure 7-5
Create Proximities from Data dialog box



Multidimensional Scaling (Create Proximities from Data)

Granddaughter [gdaugh]
Grandfather [gfather]
Grandmother [gmother]
Grandson [gson]
Mother [mother]
Nephew [nephew]
Niece [niece]
Sister [sister]
Son [son]

Variables:
Aunt [aunt]
Brother [brother]
Cousin [cousin]
Daughter [daughter]
Father [father]

Sources:
sourceid

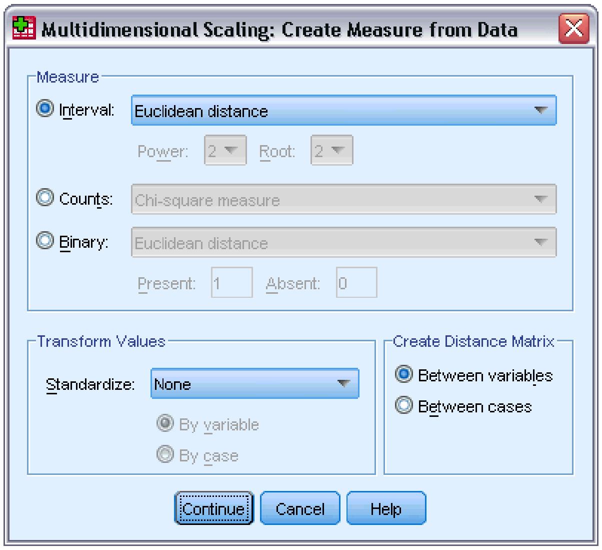
Create distances using: Euclidean distance

- ▶ If you create distances between variables (see the Create Measure from Data dialog box), select at least three variables. These variables will be used to create the proximity matrix (or matrices, if there are multiple sources). If you create distances between cases, only one variable is needed.
- ▶ If there are multiple sources, select a sources variable.
- ▶ Optionally, choose a measure for creating proximities.

Additionally, you can define a model for the multidimensional scaling, place restrictions on the common space, set convergence criteria, specify the initial configuration to be used, and choose plots and output.

Create Measure from Data

Figure 7-6
Create Measure from Data dialog box



Multidimensional Scaling: Create Measure from Data

Measure

● Interval: Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-square measure

○ Binary: Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: None

● By variable

○ By case

Create Distance Matrix

● Between variables

○ Between cases

Multidimensional scaling uses dissimilarity data to create a scaling solution. If your data are multivariate data (values of measured variables), you must create dissimilarity data in order to compute a multidimensional scaling solution. You can specify the details of creating dissimilarity measures from your data.

Measure. Allows you to specify the dissimilarity measure for your analysis. Select one alternative from the Measure group corresponding to your type of data, and then select one of the measures from the drop-down list corresponding to that type of measure. Available alternatives are:

- **Interval**. Euclidean distance, Squared Euclidean distance, Chebychev, Block, Minkowski, or Customized.
- **Counts**. Chi-square measure or Phi-square measure.
- **Binary**. Euclidean distance, Squared Euclidean distance, Size difference, Pattern difference, Variance, or Lance and Williams.

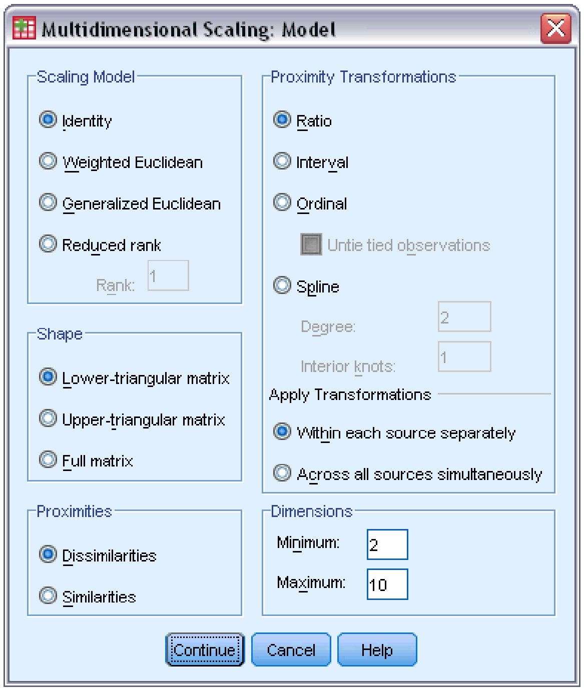
Create Distance Matrix. Allows you to choose the unit of analysis. Alternatives are Between variables or Between cases.

Transform Values. In certain cases, such as when variables are measured on very different scales, you want to standardize values before computing proximities (not applicable to binary data). Select a standardization method from the Standardize drop-down list (if no standardization is required, select None).

Define a Multidimensional Scaling Model

The Model dialog box allows you to specify a scaling model, its minimum and maximum number of dimensions, the structure of the proximity matrix, the transformation to use on the proximities, and whether proximities are transformed within each source separately or unconditionally on the source.

Figure 7-7
Model dialog box



Multidimensional Scaling: Model

Scaling Model
● Identity
○ Weighted Euclidean
○ Generalized Euclidean
○ Reduced rank
Rank: 1

Shape
● Lower-triangular matrix
○ Upper-triangular matrix
○ Full matrix

Proximities
● Dissimilarities
○ Similarities

Proximity Transformations
● Ratio
○ Interval
○ Ordinal
□ Untie tied observations
○ Spline
Degree: 2
Interior knots: 1

Apply Transformations
● Within each source separately
○ Across all sources simultaneously

Dimensions
Minimum: 2
Maximum: 10

Scaling Model. Choose from the following alternatives:

- **Identity.** All sources have the same configuration.
- **Weighted Euclidean.** This model is an individual differences model. Each source has an individual space in which every dimension of the common space is weighted differentially.
- **Generalized Euclidean.** This model is an individual differences model. Each source has an individual space that is equal to a rotation of the common space, followed by a differential weighting of the dimensions.
- **Reduced rank.** This model is a generalized Euclidean model for which you can specify the rank of the individual space. You must specify a rank that is greater than or equal to 1 and less than the maximum number of dimensions.

Shape. Specify whether the proximities should be taken from the lower-triangular part or the upper-triangular part of the proximity matrix. You can specify that the full matrix be used, in which case the weighted sum of the upper-triangular part and the lower-triangular part will be analyzed. In any case, the complete matrix should be specified, including the diagonal, though only the specified parts will be used.

Proximities. Specify whether your proximity matrix contains measures of similarity or dissimilarity.

Proximity Transformations. Choose from the following alternatives:

- **Ratio.** The transformed proximities are proportional to the original proximities. This is allowed only for positively valued proximities.
- **Interval.** The transformed proximities are proportional to the original proximities, plus an intercept term. The intercept assures all transformed proximities to be positive.
- **Ordinal.** The transformed proximities have the same order as the original proximities. You specify whether tied proximities should be kept tied or allowed to become untied.
- **Spline.** The transformed proximities are a smooth nondecreasing piecewise polynomial transformation of the original proximities. You specify the degree of the polynomial and the number of interior knots.

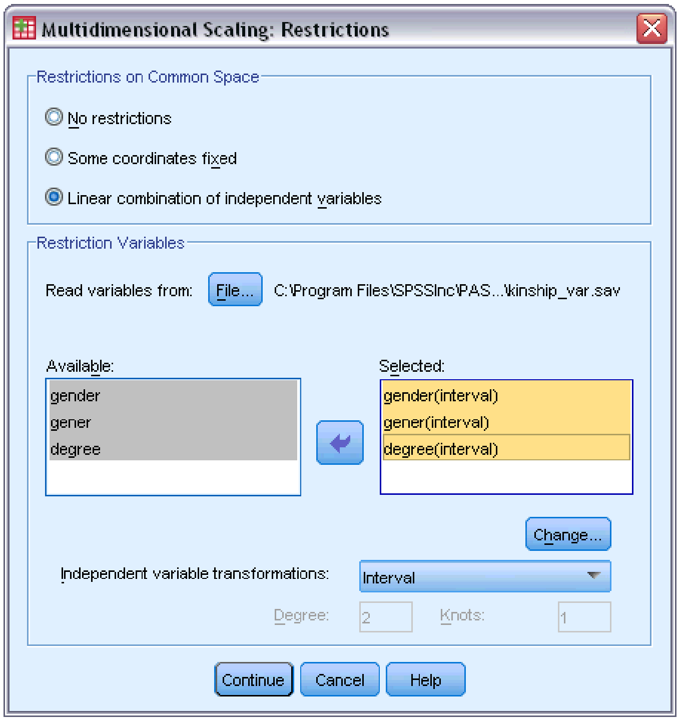
Apply Transformations. Specify whether only proximities within each source are compared with each other or whether the comparisons are unconditional on the source.

Dimensions. By default, a solution is computed in two dimensions (Minimum = 2, Maximum = 2). You choose an integer minimum and maximum from 1 to the number of objects minus 1 (as long as the minimum is less than or equal to the maximum). The procedure computes a solution in the maximum dimensions and then reduces the dimensionality in steps until the lowest is reached.

Multidimensional Scaling Restrictions

The Restrictions dialog box allows you to place restrictions on the common space.

Figure 7-8
Restrictions dialog box



Multidimensional Scaling: Restrictions

Restrictions on Common Space

○ No restrictions
○ Some coordinates fixed
● Linear combination of independent variables

Restriction Variables

Read variables from: File... C:\Program Files\SPSSInc\PAS...\kinship_var.sav

Available:
gender
gener
degree

Selected:
gender(interval)
gener(interval)
degree(interval)

Independent variable transformations: Interval

Degree: 2 Knots: 1

Restrictions on Common Space. Specify the type of restriction desired.

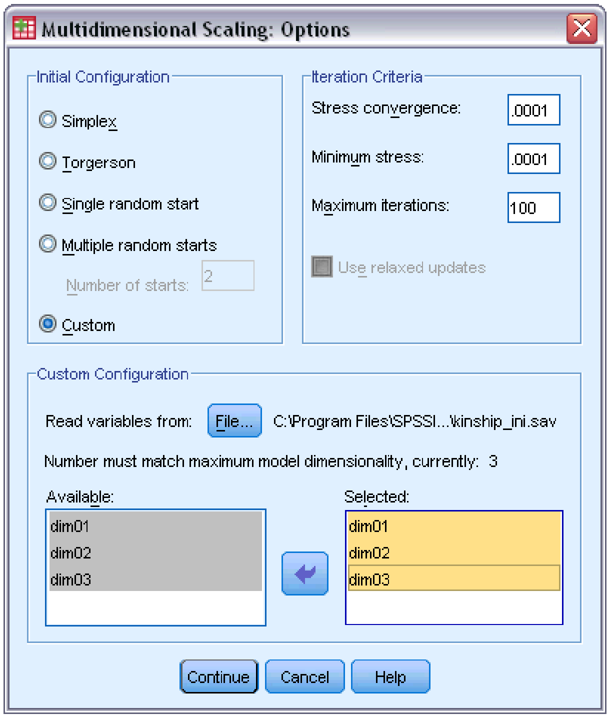
- **No restrictions.** No restrictions are placed on the common space.
- **Some coordinates fixed.** The first variable selected contains the coordinates of the objects on the first dimension, the second variable corresponds to coordinates on the second dimension, and so on. A missing value indicates that a coordinate on a dimension is free. The number of variables selected must equal the maximum number of dimensions requested.
- **Linear combination of independent variables.** The common space is restricted to be a linear combination of the variables selected.

Restriction Variables. Select the variables that define the restrictions on the common space. If you specified a linear combination, you specify an interval, nominal, ordinal, or spline transformation for the restriction variables. In either case, the number of cases for each variable must equal the number of objects.

Multidimensional Scaling Options

The Options dialog box allows you to select the initial configuration style, specify iteration and convergence criteria, and select standard or relaxed updates.

Figure 7-9
Options dialog box



Initial Configuration

○ Simplex
○ Torgerson
○ Single random start
○ Multiple random starts
Number of starts: 2
● Custom

Iteration Criteria

Stress convergence: .0001
Minimum stress: .0001
Maximum iterations: 100
□ Use relaxed updates

Custom Configuration

Read variables from: File... C:\Program Files\SPSSI...\kinship_ini.sav

Number must match maximum model dimensionality, currently: 3

Available: dim01, dim02, dim03

Selected: dim01, dim02, dim03

Initial Configuration. Choose one of the following alternatives:

- **Simplex.** Objects are placed at the same distance from each other in the maximum dimension. One iteration is taken to improve this high-dimensional configuration, followed by a dimension reduction operation to obtain an initial configuration that has the maximum number of dimensions that you specified in the Model dialog box.
- **Torgerson.** A classical scaling solution is used as the initial configuration.
- **Single random start.** A configuration is chosen at random.
- **Multiple random starts.** Several configurations are chosen at random, and the configuration with the lowest normalized raw stress is used as the initial configuration.
- **Custom.** You select variables that contain the coordinates of your own initial configuration. The number of variables selected should equal the maximum number of dimensions specified, with the first variable corresponding to coordinates on dimension 1, the second variable corresponding to coordinates on dimension 2, and so on. The number of cases in each variable should equal the number of objects.

Iteration Criteria. Specify the iteration criteria values.

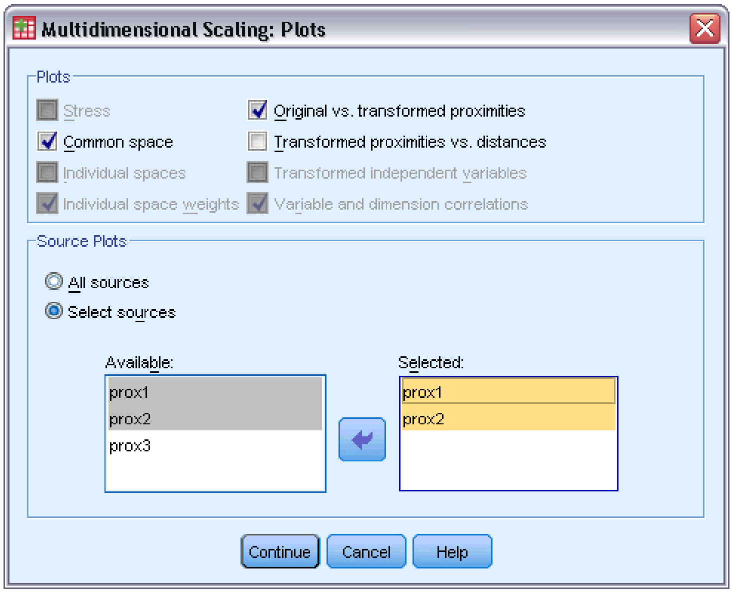
- **Stress convergence.** The algorithm will stop iterating when the difference in consecutive normalized raw stress values is less than the number that is specified here, which must lie between 0.0 and 1.0.
- **Minimum stress.** The algorithm will stop when the normalized raw stress falls below the number that is specified here, which must lie between 0.0 and 1.0.

- **Maximum iterations.** The algorithm will perform the number of specified iterations, unless one of the above criteria is satisfied first.
- **Use relaxed updates.** Relaxed updates will speed up the algorithm; these updates cannot be used with models other than the identity model or used with restrictions.

Multidimensional Scaling Plots, Version 1

The Plots dialog box allows you to specify which plots will be produced. If you have the Proximities in Columns data format, the following Plots dialog box is displayed. For Individual space weights, Original vs. transformed proximities, and Transformed proximities vs. distances plots, you specify the sources for which the plots should be produced. The list of available sources is the list of proximities variables in the main dialog box.

Figure 7-10
Plots dialog box, version 1



Multidimensional Scaling: Plots

Plots

☐ Stress
☑ Common space
☐ Individual spaces
☑ Individual space weights
☑ Original vs. transformed proximities
☐ Transformed proximities vs. distances
☐ Transformed independent variables
☑ Variable and dimension correlations

Source Plots

○ All sources
● Select sources

Available: prox1, prox2, prox3

Selected: prox1, prox2

Stress. A plot is produced of normalized raw stress versus dimensions. This plot is produced only if the maximum number of dimensions is larger than the minimum number of dimensions.

Common space. A scatterplot matrix of coordinates of the common space is displayed.

Individual spaces. For each source, the coordinates of the individual spaces are displayed in scatterplot matrices. This is possible only if one of the individual differences models is specified in the Model dialog box.

Individual space weights. A scatterplot is produced of the individual space weights. This is possible only if one of the individual differences models is specified in the Model dialog box. For the weighted Euclidean model, the weights are printed in plots, with one dimension on each axis. For the generalized Euclidean model, one plot is produced per dimension, indicating both rotation and weighting of that dimension. The reduced rank model produces the same plot as the generalized Euclidean model but reduces the number of dimensions for the individual spaces.

Original vs. transformed proximities. Plots are produced of the original proximities versus the transformed proximities.

Transformed proximities vs. distances. The transformed proximities versus the distances are plotted.

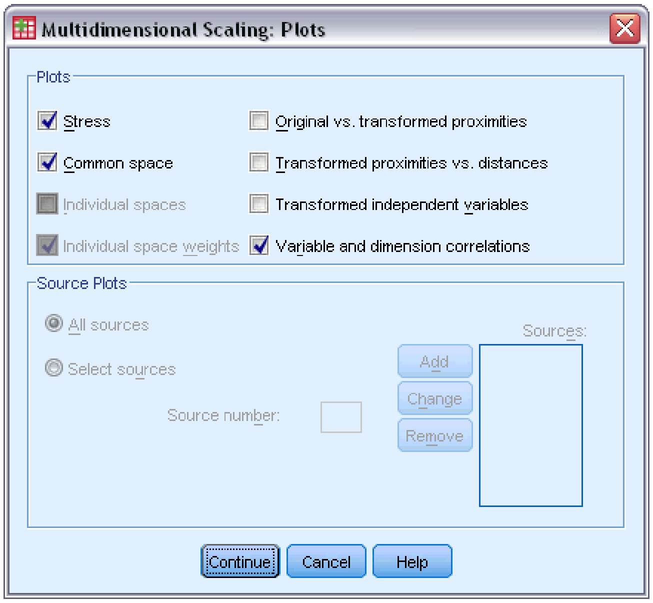
Transformed independent variables. Transformation plots are produced for the independent variables.

Variable and dimension correlations. A plot of correlations between the independent variables and the dimensions of the common space is displayed.

Multidimensional Scaling Plots, Version 2

The Plots dialog box allows you to specify which plots will be produced. If your data format is anything other than Proximities in Columns, the following Plots dialog box is displayed. For Individual space weights, Original vs. transformed proximities, and Transformed proximities vs. distances plots, you specify the sources for which the plots should be produced. The source numbers entered must be values of the sources variable that is specified in the main dialog box and must range from 1 to the number of sources.

Figure 7-11
Plots dialog box, version 2



Multidimensional Scaling: Plots

Plots

☑ Stress
☐ Original vs. transformed proximities
☑ Common space
☐ Transformed proximities vs. distances
☐ Individual spaces
☐ Transformed independent variables
☑ Individual space weights
☑ Variable and dimension correlations

Source Plots

● All sources
○ Select sources

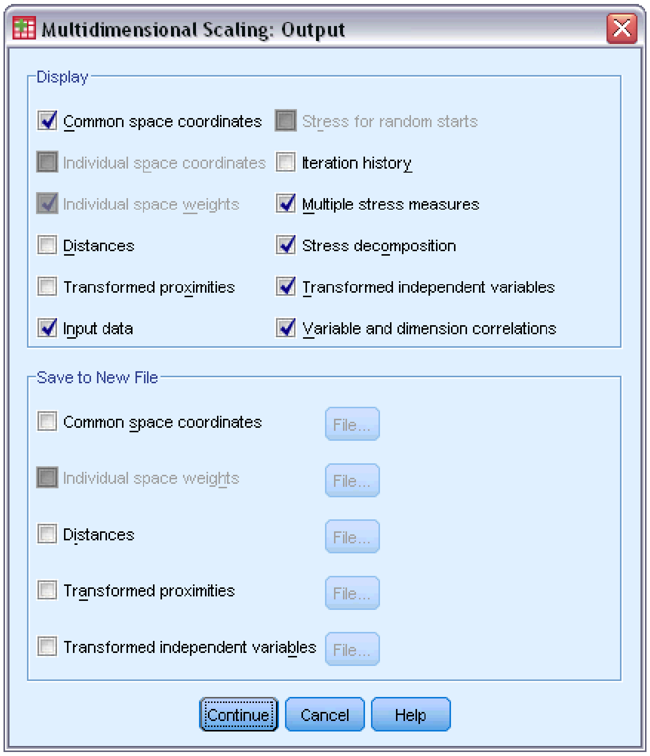
Source number:

Sources:

Multidimensional Scaling Output

The Output dialog box allows you to control the amount of displayed output and save some of it to separate files.

Figure 7-12
Output dialog box



Display

☑ Common space coordinates
□ Individual space coordinates
☑ Individual space weights
□ Distances
□ Transformed proximities
☑ Input data
□ Stress for random starts
□ Iteration history
☑ Multiple stress measures
☑ Stress decomposition
☑ Transformed independent variables
☑ Variable and dimension correlations

Save to New File

□ Common space coordinates
□ Individual space weights
□ Distances
□ Transformed proximities
□ Transformed independent variables

Display. Select one or more of the following items for display:

- **Common space coordinates.** Displays the coordinates of the common space.
- **Individual space coordinates.** The coordinates of the individual spaces are displayed only if the model is not the identity model.
- **Individual space weights.** Displays the individual space weights only if one of the individual differences models is specified. Depending on the model, the space weights are decomposed in rotation weights and dimension weights, which are also displayed.
- **Distances.** Displays the distances between the objects in the configuration.
- **Transformed proximities.** Displays the transformed proximities between the objects in the configuration.
- **Input data.** Includes the original proximities and, if present, the data weights, the initial configuration, and the fixed coordinates of the independent variables.
- **Stress for random starts.** Displays the random number seed and normalized raw stress value of each random start.
- **Iteration history.** Displays the history of iterations of the main algorithm.
- **Multiple stress measures.** Displays different stress values. The table contains values for normalized raw stress, Stress-I, Stress-II, S-Stress, Dispersion Accounted For (DAF), and Tucker's Coefficient of Congruence.
- **Stress decomposition.** Displays an objects and sources decomposition of final normalized raw stress, including the average per object and the average per source.

- **Transformed independent variables.** If a linear combination restriction was selected, the transformed independent variables and the corresponding regression weights are displayed.
- **Variable and dimension correlations.** If a linear combination restriction was selected, the correlations between the independent variables and the dimensions of the common space are displayed.

Save to New File. You can save the common space coordinates, individual space weights, distances, transformed proximities, and transformed independent variables to separate IBM® SPSS® Statistics data files.

PROXSCAL Command Additional Features

You can customize your multidimensional scaling of proximities analysis if you paste your selections into a syntax window and edit the resulting `PROXSCAL` command syntax. The command syntax language also allows you to:

- Specify separate variable lists for transformations and residuals plots (with the `PLOT` subcommand).
- Specify separate source lists for individual space weights, transformations, and residuals plots (with the `PLOT` subcommand).
- Specify a subset of the independent variables transformation plots to be displayed (with the `PLOT` subcommand).

See the *Command Syntax Reference* for complete syntax information.

Chapter

8

Multidimensional Unfolding (PREFSCAL)

The Multidimensional Unfolding procedure attempts to find a common quantitative scale that allows you to visually examine the relationships between two sets of objects.

Examples. You have asked 21 individuals to rank 15 breakfast items in order of preference, 1 to 15. Using Multidimensional Unfolding, you can determine that the individuals discriminate between breakfast items in two primary ways: between soft and hard breads, and between fattening and non-fattening items.

Alternatively, you have asked a group of drivers to rate 26 models of cars on 10 attributes on a 6-point scale ranging from 1="not true at all" to 6="very true." Averaged over individuals, the values are taken as similarities. Using Multidimensional Unfolding, you find clusterings of similar models and the attributes with which they are most closely associated.

Statistics and plots. The Multidimensional Unfolding procedure can produce an iteration history, stress measures, stress decomposition, coordinates of the common space, object distances within the final configuration, individual space weights, individual spaces, transformed proximities, stress plots, common space scatterplots, individual space weight scatterplots, individual spaces scatterplots, transformation plots, and Shepard residual plots.

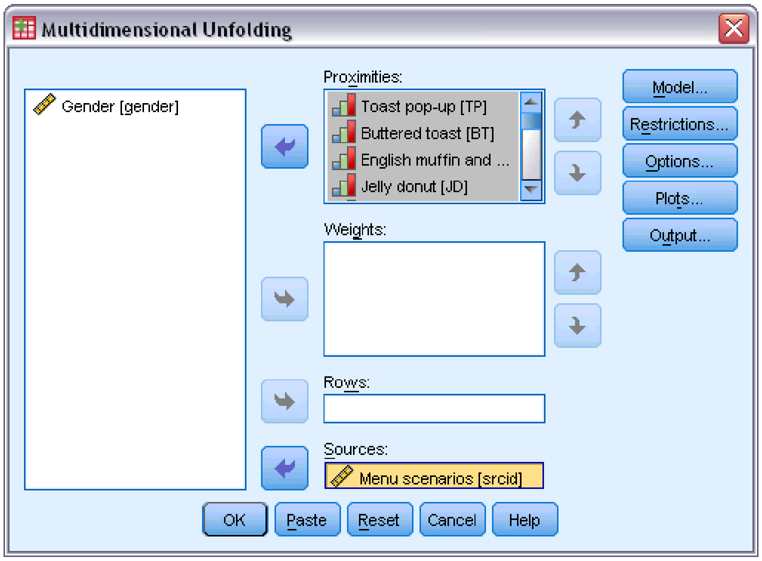
Data. Data are supplied in the form of rectangular proximity matrices. Each column is considered a separate column object. Each row of a proximity matrix is considered a separate row object. When there are multiple sources of proximities, the matrices are stacked.

Assumptions. At least two variables must be specified. The number of dimensions in the solution may not exceed the number of objects minus one. If only one source is specified, all models are equivalent to the identity model; therefore, the analysis defaults to the identity model.

To Obtain a Multidimensional Unfolding

▶ From the menus choose:
Analyze > Scale > Multidimensional Unfolding (PREFSCAL)...

Figure 8-1
Multidimensional Unfolding main dialog box



Multidimensional Unfolding

Gender [gender]

Proximities:
Toast pop-up [TP]
Buttered toast [BT]
English muffin and ...
Jelly donut [JD]

Weights:

Rows:

Sources:
Menu scenarios [srcid]

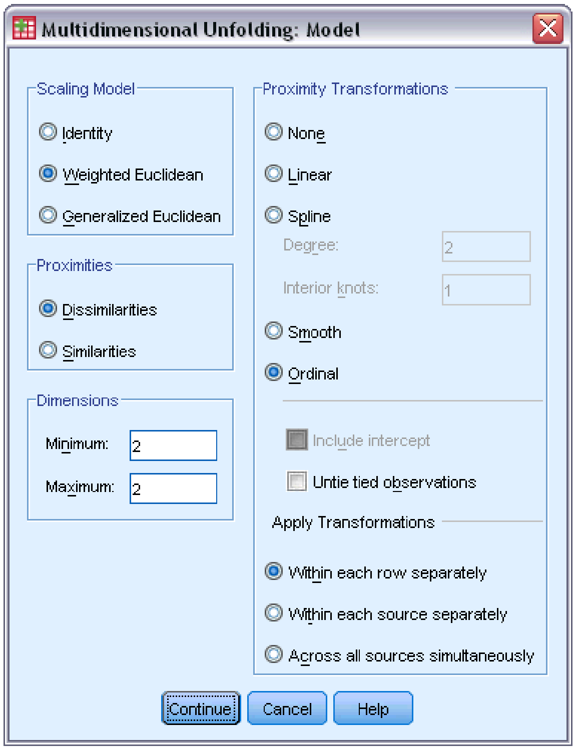
- ▶ Select two or more variables that identify the columns in the rectangular proximity matrix. Each variable represents a separate column object.
- ▶ Optionally, select a number of weights variables equal to the number of column object variables. The order of the weights variables should match the order of the column objects they weight.
- ▶ Optionally, select a rows variable. The values (or value labels) of this variable are used to label row objects in the output.
- ▶ If there are multiple sources, optionally select a sources variable. The number of cases in the data file should equal the number of row objects times the number of sources.

Additionally, you can define a model for the multidimensional unfolding, place restrictions on the common space, set convergence criteria, specify the initial configuration to be used, and choose plots and output.

Define a Multidimensional Unfolding Model

The Model dialog box allows you to specify a scaling model, its minimum and maximum number of dimensions, the structure of the proximity matrix, the transformation to use on the proximities, and whether proximities are transformed conditonal upon the row, conditional upon the source, or unconditionally on the source.

Figure 8-2
Model dialog box



Multidimensional Unfolding: Model

Scaling Model

○ Identity
● Weighted Euclidean
○ Generalized Euclidean

Proximities

● Dissimilarities
○ Similarities

Dimensions

Minimum: 2
Maximum: 2

Proximity Transformations

○ None
○ Linear
○ Spline
Degree: 2
Interior knots: 1
○ Smooth
● Ordinal

■ Include intercept
☐ Untie tied observations

Apply Transformations

● Within each row separately
○ Within each source separately
○ Across all sources simultaneously

Scaling Model. Choose from the following alternatives:

- **Identity.** All sources have the same configuration.
- **Weighted Euclidean.** This model is an individual differences model. Each source has an individual space in which every dimension of the common space is weighted differentially.
- **Generalized Euclidean.** This model is an individual differences model. Each source has an individual space that is equal to a rotation of the common space, followed by a differential weighting of the dimensions.

Proximities. Specify whether your proximity matrix contains measures of similarity or dissimilarity.

Dimensions. By default, a solution is computed in two dimensions (Minimum = 2, Maximum = 2). You can choose an integer minimum and maximum from 1 to the number of objects minus 1 as long as the minimum is less than or equal to the maximum. The procedure computes a solution in the maximum dimensionality and then reduces the dimensionality in steps until the lowest is reached.

Proximity Transformations. Choose from the following alternatives:

- **None.** The proximities are not transformed. You can optionally select Include intercept, in which case the proximities can be shifted by a constant term.

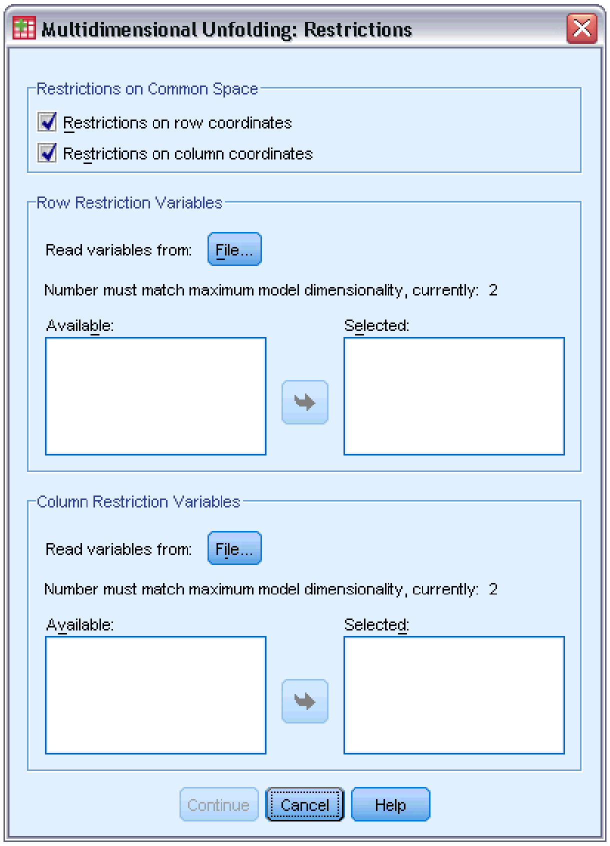
- **Linear.** The transformed proximities are proportional to the original proximities; that is, the transformation function estimates a slope and the intercept is fixed at 0. This is also called a ratio transformation. You can optionally select Include intercept, in which case the proximities can also be shifted by a constant term. This is also called an interval transformation.
- **Spline.** The transformed proximities are a smooth nondecreasing piecewise polynomial transformation of the original proximities. You can specify the degree of the polynomial and the number of interior knots. You can optionally select Include intercept, in which case the proximities can also be shifted by a constant term.
- **Smooth.** The transformed proximities have the same order as the original proximities, including a restriction that takes the differences between subsequent values into account. The result is a "smooth ordinal" transformation. You can specify whether tied proximities should be kept tied or allowed to become untied.
- **Ordinal.** The transformed proximities have the same order as the original proximities. You can specify whether tied proximities should be kept tied or allowed to become untied.

Apply Transformations. Specify whether only proximities within each row are compared with each other, or only proximities within each source are compared with each other, or the comparisons are unconditional on the row or source; that is, whether the transformations are performed per row, per source, or over all proximities at once.

Multidimensional Unfolding Restrictions

The Restrictions dialog box allows you to place restrictions on the common space.

Figure 8-3
Restrictions dialog box



Multidimensional Unfolding: Restrictions

Restrictions on Common Space

☑ Restrictions on row coordinates

☑ Restrictions on column coordinates

Row Restriction Variables

Read variables from: File...

Number must match maximum model dimensionality, currently: 2

Column Restriction Variables

Read variables from: File...

Number must match maximum model dimensionality, currently: 2

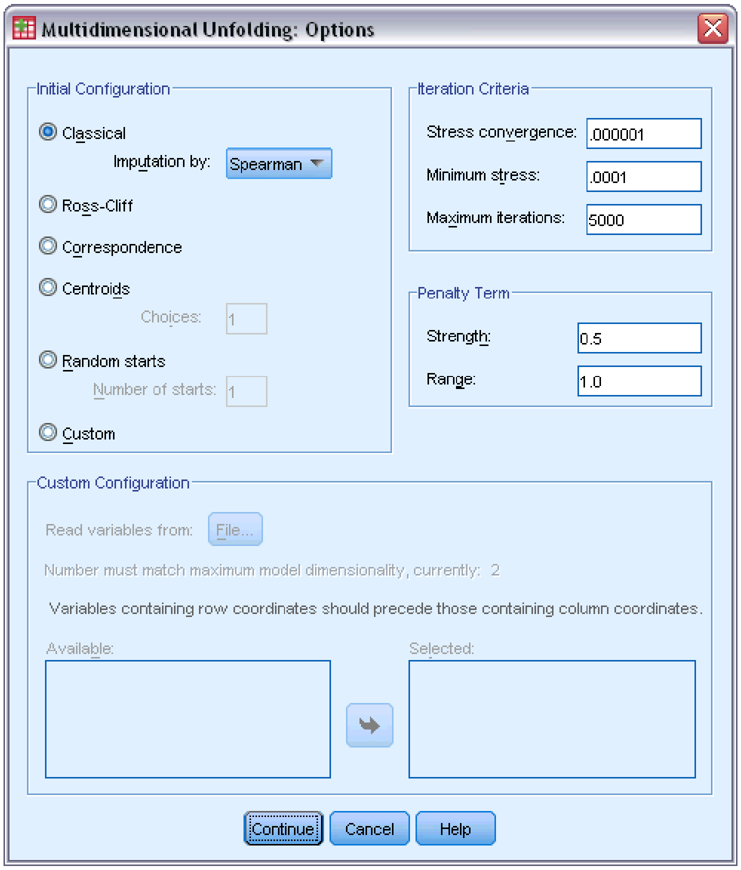
Restrictions on Common Space. You can choose to fix the coordinates of row and/or column objects in the common space.

Row/Column Restriction Variables. Choose the file containing the restrictions and select the variables that define the restrictions on the common space. The first variable selected contains the coordinates of the objects on the first dimension, the second variable corresponds to coordinates on the second dimension, and so on. A missing value indicates that a coordinate on a dimension is free. The number of variables selected must equal the maximum number of dimensions requested. The number of cases for each variable must equal the number of objects.

Multidimensional Unfolding Options

The Options dialog box allows you to select the initial configuration style, specify iteration and convergence criteria, and set the penalty term for stress.

Figure 8-4
Options dialog box



Multidimensional Unfolding: Options

Initial Configuration

- ● Classical
  - Imputation by: Spearman
- ○ Ross-Cliff
- ○ Correspondence
- ○ Centroids
  - Choices: 1
- ○ Random starts
  - Number of starts: 1
- ○ Custom

Iteration Criteria

- Stress convergence: .000001
- Minimum stress: .0001
- Maximum iterations: 5000

Penalty Term

- Strength: 0.5
- Range: 1.0

Custom Configuration

Read variables from: File...

Number must match maximum model dimensionality, currently: 2

Variables containing row coordinates should precede those containing column coordinates.

Initial Configuration. Choose one of the following alternatives:

- **Classical.** The rectangular proximity matrix is used to supplement the intra-blocks (values between rows and between columns) of the complete symmetrical MDS matrix. Once the complete matrix is formed, a classical scaling solution is used as the initial configuration. The intra-blocks can be filled via imputation using the triangle inequality or Spearman distances.
- **Ross-Cliff.** The Ross-Cliff start uses the results of a singular value decomposition on the double centered and squared proximity matrix as the initial values for the row and column objects.
- **Correspondence.** The correspondence start uses the results of a correspondence analysis on the reversed data (similarities instead of dissimilarities), with symmetric normalization of row and column scores.
- **Centroids.** The procedure starts by positioning the row objects in the configuration using an eigenvalue decomposition. Then the column objects are positioned at the centroid of the specified choices. For the number of choices, specify a positive integer between 1 and the number of proximities variables.

- **Multiple random starts.** Solutions are computed for several initial configurations chosen at random, and the one with the lowest penalized stress is shown as the best solution.
- **Custom.** You can select variables that contain the coordinates of your own initial configuration. The number of variables selected should equal the maximum number of dimensions specified, with the first variable corresponding to coordinates on dimension 1, the second variable corresponding to coordinates on dimension 2, and so on. The number of cases in each variable should equal the combined number of row and column objects. The row and column coordinates should be stacked, with the column coordinates following the row coordinates.

Iteration Criteria. Specify the iteration criteria values.

- **Stress convergence.** The algorithm will stop iterating when the relative difference in consecutive penalized stress values is less than the number specified here, which must be non-negative.
- **Minimum stress.** The algorithm will stop when the penalized stress falls below the number specified here, which must be non-negative.
- **Maximum iterations.** The algorithm will perform the number of iterations specified here unless one of the above criteria is satisfied first.

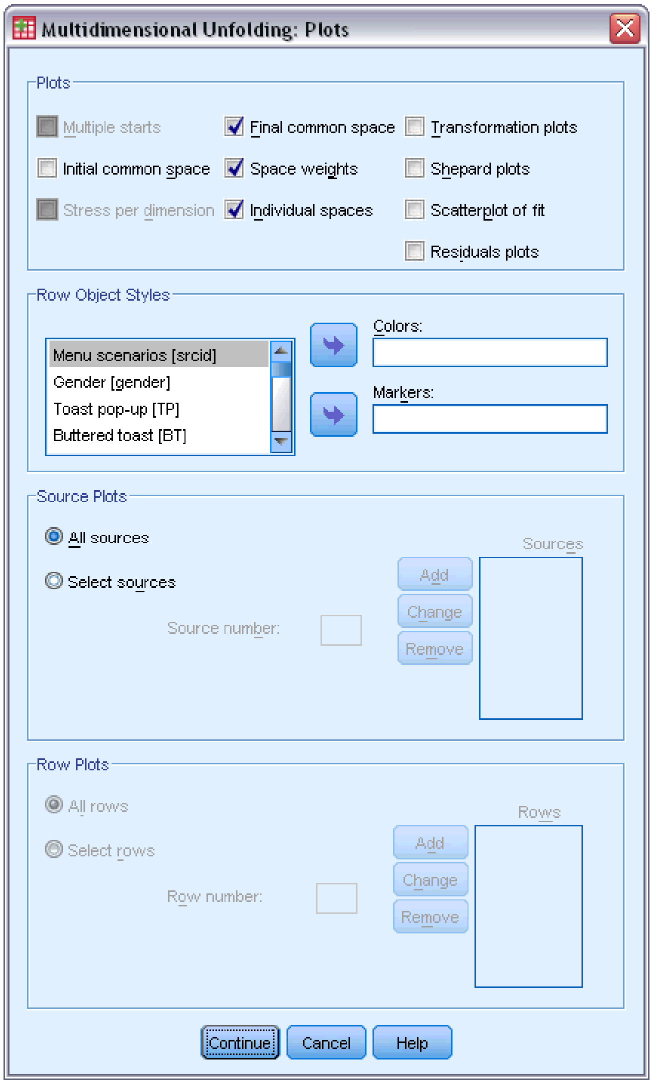
Penalty Term. The algorithm attempts to minimize penalized stress, a goodness-of-fit measure equal to the product of Kruskal's Stress-I and a penalty term based on the coefficient of variation of the transformed proximities. These controls allow you to set the strength and range of the penalty term.

- **Strength.** The smaller the value of the strength parameter, the stronger the penalty. Specify a value between 0.0 and 1.0.
- **Range.** This parameter sets the moment at which the penalty becomes active. If set to 0.0, the penalty is inactive. Increasing the value causes the algorithm to search for a solution with greater variation among the transformed proximities. Specify a non-negative value.

Multidimensional Unfolding Plots

The Plots dialog box allows you to specify which plots will be produced.

Figure 8-5
Plots dialog box



Multidimensional Unfolding: Plots

Plots

☐ Multiple starts ☑ Final common space ☐ Transformation plots

☐ Initial common space ☑ Space weights ☐ Shepard plots

☐ Stress per dimension ☑ Individual spaces ☐ Scatterplot of fit

☐ Residuals plots

Row Object Styles

Menu scenarios [srcid]
Gender [gender]
Toast pop-up [TP]
Buttered toast [BT]

Colors:

Markers:

Source Plots

● All sources

○ Select sources

Source number:

Sources

Row Plots

● All rows

○ Select rows

Row number:

Rows

Plots. The following plots are available:

- **Multiple starts.** Displays a stacked histogram of penalized stress displaying both stress and penalty.
- **Initial common space.** Displays a scatterplot matrix of the coordinates of the initial common space.
- **Stress per dimension.** Produces a lineplot of penalized stress versus dimensionality. This plot is produced only if the maximum number of dimensions is larger than the minimum number of dimensions.
- **Final common space.** A scatterplot matrix of coordinates of the common space is displayed.

- **Space weights.** A scatterplot is produced of the individual space weights. This is possible only if one of the individual differences models is specified in the Model dialog box. For the weighted Euclidean model, the weights for all sources are displayed in a plot, with one dimension on each axis. For the generalized Euclidean model, one plot is produced per dimension, indicating both rotation and weighting of that dimension for each source.
- **Individual spaces.** A scatterplot matrix of coordinates of the individual space of each source is displayed. This is possible only if one of the individual differences models is specified in the Model dialog box.
- **Transformation plots.** A scatterplot is produced of the original proximities versus the transformed proximities. Depending on how transformations are applied, a separate color is assigned to each row or source. An unconditional transformation produces a single color.
- **Shepard plots.** The original proximities versus both transformed proximities and distances. The distances are indicated by points, and the transformed proximities are indicated by a line. Depending on how transformations are applied, a separate line is produced for each row or source. An unconditional transformation produces one line.
- **Scatterplot of fit.** A scatterplot of the transformed proximities versus the distances is displayed. A separate color is assigned to each source if multiple sources are specified.
- **Residuals plots.** A scatterplot of the transformed proximities versus the residuals (transformed proximities minus distances) is displayed. A separate color is assigned to each source if multiple sources are specified.

Row Object Styles. These give you further control of the display of row objects in plots. The values of the optional colors variable are used to cycle through all colors. The values of the optional markers variable are used to cycle through all possible markers.

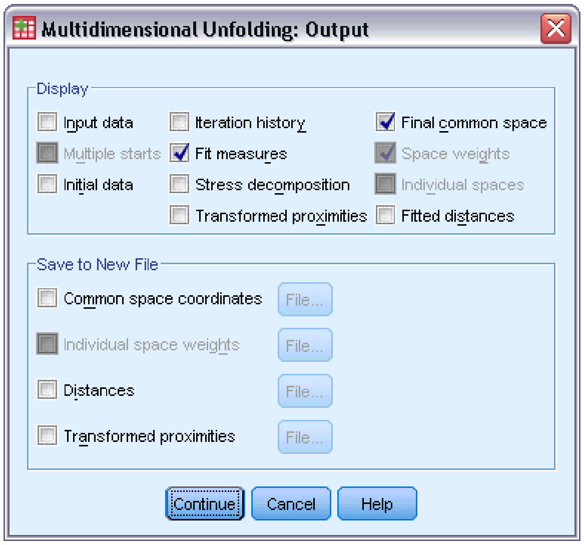
Source Plots. For Individual spaces, Scatterplot of fit, and Residuals plots—and if transformations are applied by source, for Transformation plots and Shepard plots—you can specify the sources for which the plots should be produced. The source numbers entered must be values of the sources variable specified in the main dialog box and range from 1 to the number of sources.

Row Plots. If transformations are applied by row, for Transformation plots and Shepard plots, you can specify the row for which the plots should be produced. The row numbers entered must range from 1 to the number of rows.

Multidimensional Unfolding Output

The Output dialog box allows you to control the amount of displayed output and save some of it to separate files.

Figure 8-6
Output dialog box



Multidimensional Unfolding: Output

Display

□ Input data
□ Multiple starts
□ Initial data
□ Iteration history
☑ Fit measures
□ Stress decomposition
□ Transformed proximities
☑ Final common space
☑ Space weights
□ Individual spaces
□ Fitted distances

Save to New File

□ Common space coordinates
□ Individual space weights
□ Distances
□ Transformed proximities

Display. Select one or more of the following for display:

- **Input data.** Includes the original proximities and, if present, the data weights, the initial configuration, and the fixed coordinates.
- **Multiple starts.** Displays the random number seed and penalized stress value of each random start.
- **Initial data.** Displays the coordinates of the initial common space.
- **Iteration history.** Displays the history of iterations of the main algorithm.
- **Fit measures.** Displays different measures. The table contains several goodness-of-fit, badness-of-fit, correlation, variation, and nondegeneracy measures.
- **Stress decomposition.** Displays an objects, rows, and sources decomposition of penalized stress, including row, column, and source means and standard deviations.
- **Transformed proximities.** Displays the transformed proximities.
- **Final common space.** Displays the coordinates of the common space.
- **Space weights.** Displays the individual space weights. This option is available only if one of the individual differences models is specified. Depending on the model, the space weights are decomposed in rotation weights and dimension weights, which are also displayed.
- **Individual spaces.** The coordinates of the individual spaces are displayed. This option is available only if one of the individual differences models is specified.
- **Fitted distances.** Displays the distances between the objects in the configuration.

Save to New File. You can save the common space coordinates, individual space weights, distances, and transformed proximities to separate IBM® SPSS® Statistics data files.

PREFSCAL Command Additional Features

You can customize your Multidimensional Unfolding of proximities analysis if you paste your selections into a syntax window and edit the resulting PREFSCAL command syntax. The command syntax language also allows you to:

- Specify multiple source lists for Individual spaces, Scatterplots of fit, and Residuals plots—and in the case of matrix conditional transformations, for Transformation plots and Shepard plots—when multiple sources are available (with the PLOT subcommand).
- Specify multiple row lists for Transformation plots and Shepard plots in the case of row conditional transformations (with the PLOT subcommand).
- Specify a number of rows instead of a row ID variable (with the INPUT subcommand).
- Specify a number of sources instead of a source ID variable (with the INPUT subcommand).

See the *Command Syntax Reference* for complete syntax information.

Part II: Examples

Chapter

9

Categorical Regression

The goal of categorical regression with optimal scaling is to describe the relationship between a response variable and a set of predictors. By quantifying this relationship, values of the response can be predicted for any combination of predictors.

In this chapter, two examples serve to illustrate the analyses involved in optimal scaling regression. The first example uses a small data set to illustrate the basic concepts. The second example uses a much larger set of variables and observations in a practical example.

Example: Carpet Cleaner Data

In a popular example (Green and Wind, 1973), a company interested in marketing a new carpet cleaner wants to examine the influence of five factors on consumer preference—package design, brand name, price, a *Good Housekeeping* seal, and a money-back guarantee. There are three factor levels for package design, each one differing in the location of the applicator brush; three brand names (*K2R*, *Glory*, and *Bissell*); three price levels; and two levels (either no or yes) for each of the last two factors. The following table displays the variables used in the carpet-cleaner study, with their variable labels and values.

Table 9-1
Explanatory variables in the carpet-cleaner study

Variable name	Variable label	Value label
package	Package design	A*, B*, C*
brand	Brand name	K2R, Glory, Bissell
price	Price	$1.19, $1.39, $1.59
seal	*Good Housekeeping* seal	No, yes
money	Money-back guarantee	No, yes

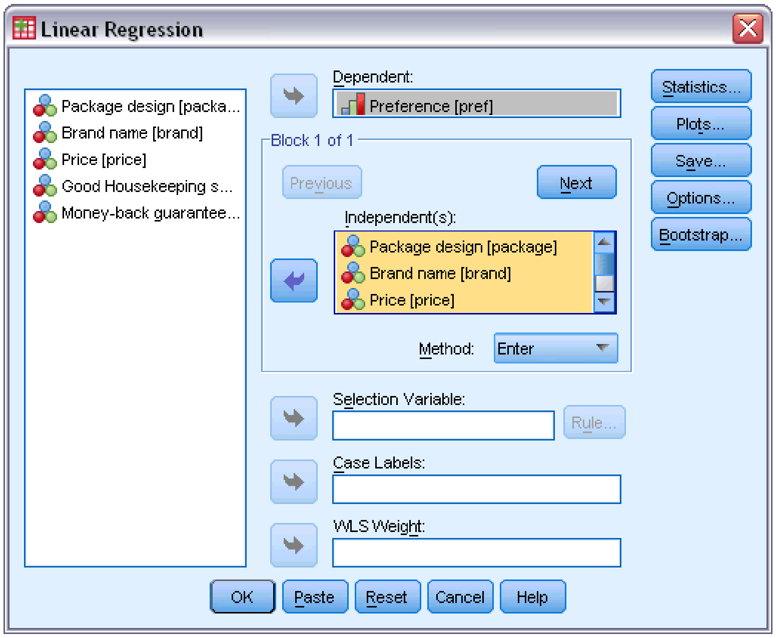
Ten consumers rank 22 profiles defined by these factors. The variable *Preference* contains the rank of the average rankings for each profile. Low rankings correspond to high preference. This variable reflects an overall measure of preference for each profile. Using categorical regression, you will explore how the five factors are related to preference. This data set can be found in *carpet.sav*. For more information, see the topic Sample Files in Appendix A on p. 292.

A Standard Linear Regression Analysis

▶ To produce standard linear regression output, from the menus choose:
Analyze > Regression > Linear...

Note: This feature requires the Statistics Base option.

Figure 9-1
Linear Regression dialog



Dependent: Preference [pref]

Independent(s): Package design [package], Brand name [brand], Price [price]

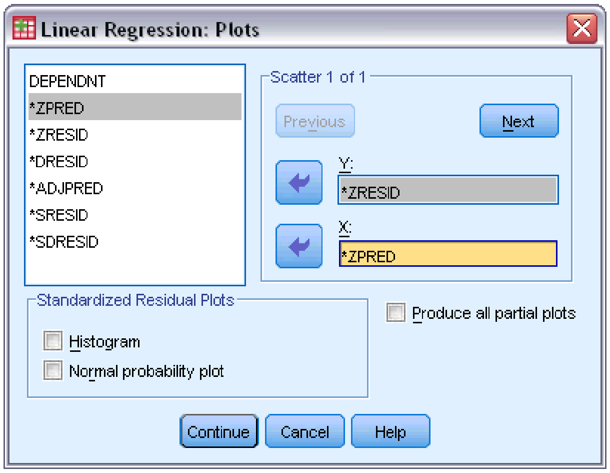
Method: Enter

- ▶ Select *Preference* as the dependent variable.

- ▶ Select *Package design* through *Money-back guarantee* as independent variables.

- ▶ Click Plots.

Figure 9-2
Plots dialog



Y: *ZRESID

X: *ZPRED

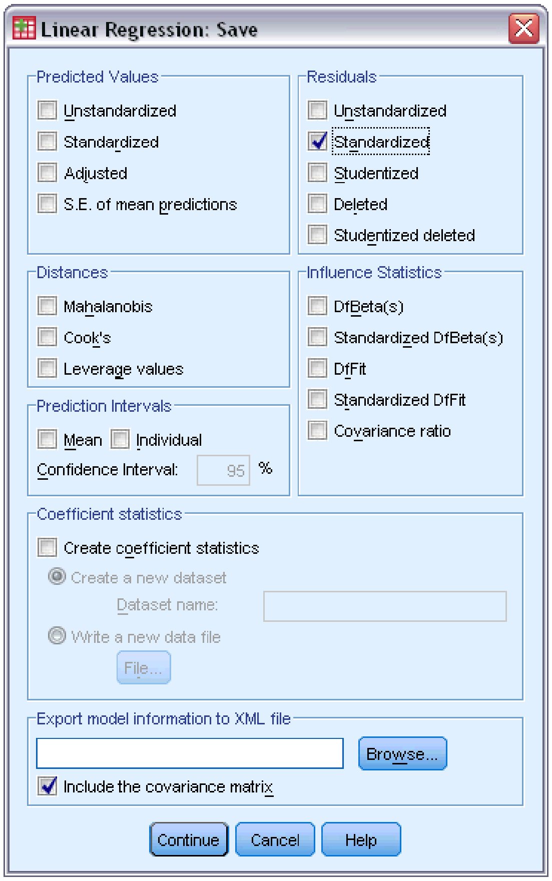
- ▶ Select **ZRESID* as the *y*-axis variable.

- ▶ Select **ZPRED* as the *x*-axis variable.

- ▶ Click Continue.

- ▶ Click Save in the Linear Regression dialog.

Figure 9-3
Save dialog



Linear Regression: Save

Predicted Values
☐ Unstandardized
☐ Standardized
☐ Adjusted
☐ S.E. of mean predictions

Residuals
☐ Unstandardized
☑ Standardized
☐ Studentized
☐ Deleted
☐ Studentized deleted

Distances
☐ Mahalanobis
☐ Cook's
☐ Leverage values

Influence Statistics
☐ DfBeta(s)
☐ Standardized DfBeta(s)
☐ DfFit
☐ Standardized DfFit
☐ Covariance ratio

Prediction Intervals
☐ Mean ☐ Individual
Confidence Interval: 95 %

Coefficient statistics
☐ Create coefficient statistics
● Create a new dataset
Dataset name:
○ Write a new data file

Export model information to XML file
☑ Include the covariance matrix

- ▶ Select Standardized in the Residuals group.
- ▶ Click Continue.
- ▶ Click OK in the Linear Regression dialog.

Model Summary

Figure 9-4
Model summary for standard linear regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.841[a]	.707	.615	3.99810

a. Predictors: (Constant), Money-back guarantee, Price, Good Housekeeping seal, Brand name, Package design

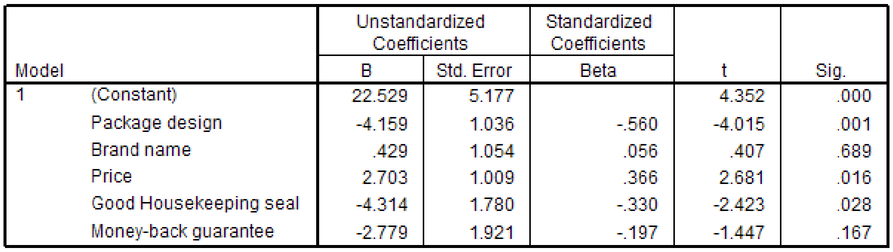
The standard approach for describing the relationships in this problem is linear regression. The most common measure of how well a regression model fits the data is R^2. This statistic represents how much of the variance in the response is explained by the weighted combination of predictors.

The closer R^2 is to 1, the better the model fits. Regressing *Preference* on the five predictors results in an R^2 of 0.707, indicating that approximately 71% of the variance in the preference rankings is explained by the predictor variables in the linear regression.

Coefficients

The standardized coefficients are shown in the table. The sign of the coefficient indicates whether the predicted response increases or decreases when the predictor increases, all other predictors being constant. For categorical data, the category coding determines the meaning of an increase in a predictor. For instance, an increase in *Money-back guarantee*, *Package design*, or *Good Housekeeping seal* will result in a decrease in predicted preference ranking. *Money-back guarantee* is coded 1 for *no money-back guarantee* and 2 for *money-back guarantee*. An increase in *Money-back guarantee* corresponds to the addition of a money-back guarantee. Thus, adding a money-back guarantee reduces the predicted preference ranking, which corresponds to an increased predicted preference.

Figure 9-5
Regression coefficients

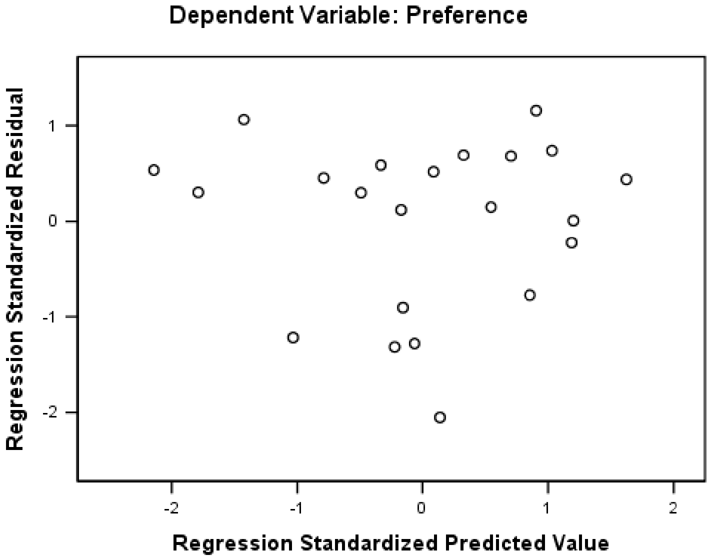


		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	22.529	5.177		4.352	.000
	Package design	-4.159	1.036	-.560	-4.015	.001
	Brand name	.429	1.054	.056	.407	.689
	Price	2.703	1.009	.366	2.681	.016
	Good Housekeeping seal	-4.314	1.780	-.330	-2.423	.028
	Money-back guarantee	-2.779	1.921	-.197	-1.447	.167

The value of the coefficient reflects the amount of change in the predicted preference ranking. Using standardized coefficients, interpretations are based on the standard deviations of the variables. Each coefficient indicates the number of standard deviations that the predicted response changes for a one standard deviation change in a predictor, all other predictors remaining constant. For example, a one standard deviation change in *Brand name* yields an increase in predicted preference of 0.056 standard deviations. The standard deviation of *Preference* is 6.44, so *Preference* increases by $0.056 \times 6.44 = 0.361$. Changes in *Package design* yield the greatest changes in predicted preference.

Residual Scatterplots

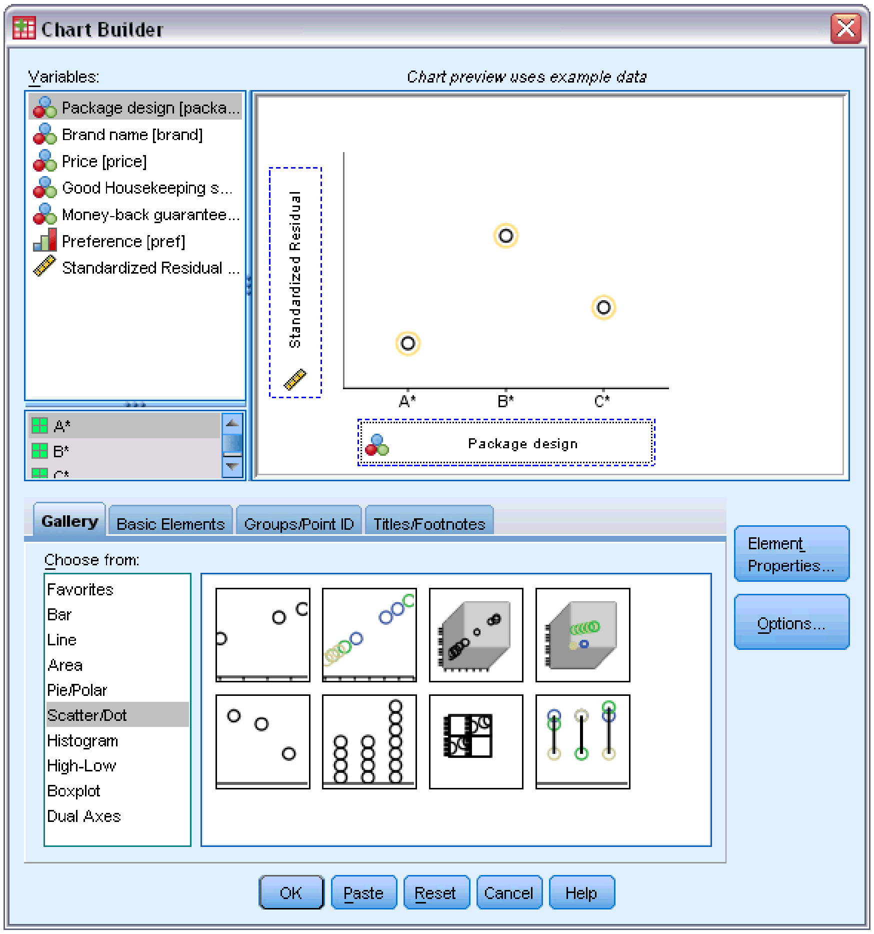
Figure 9-6
Residuals versus predicted values



The standardized residuals are plotted against the standardized predicted values. No patterns should be present if the model fits well. Here you see a U-shape in which both low and high standardized predicted values have positive residuals. Standardized predicted values near 0 tend to have negative residuals.

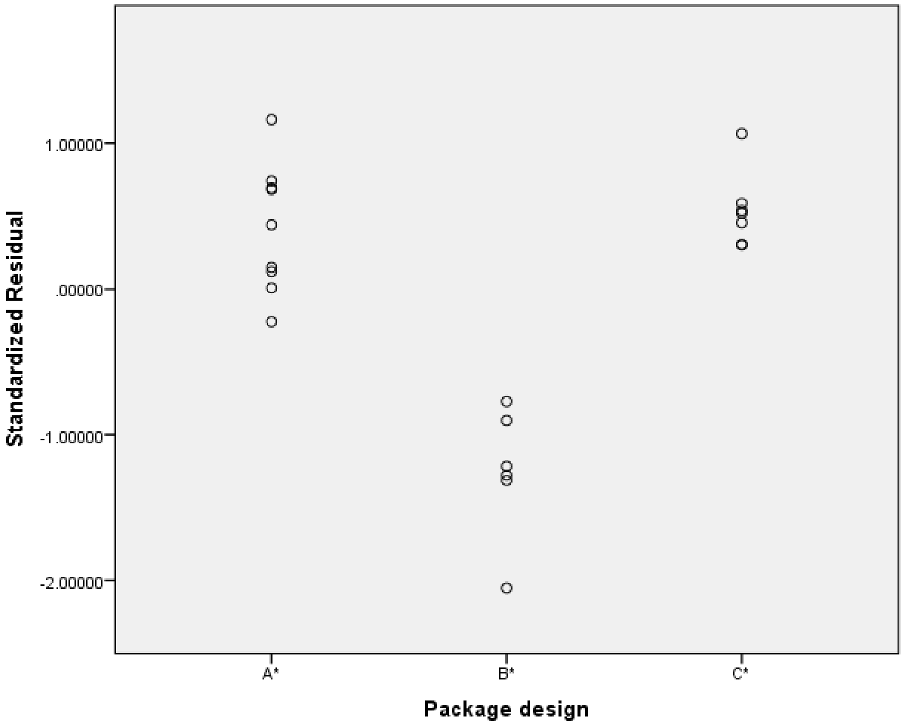
▶ To produce a scatterplot of the residuals by the predictor *Package design*, from the menus choose:
Graphs > Chart Builder...

Figure 9-7
Chart Builder



- ▶ Select the Scatter/Dot gallery and choose Simple Scatter.
- ▶ Select *Standardized Residual* as the *y*-axis variable and *Package design* as the *x*-axis variable.
- ▶ Click OK.

Figure 9-8
Residuals versus package design



The U-shape is more pronounced in the plot of the standardized residuals against package. Every residual for Design B* is negative, whereas all but one of the residuals is positive for the other two designs. Because the linear regression model fits one parameter for each variable, the relationship cannot be captured by the standard approach.

A Categorical Regression Analysis

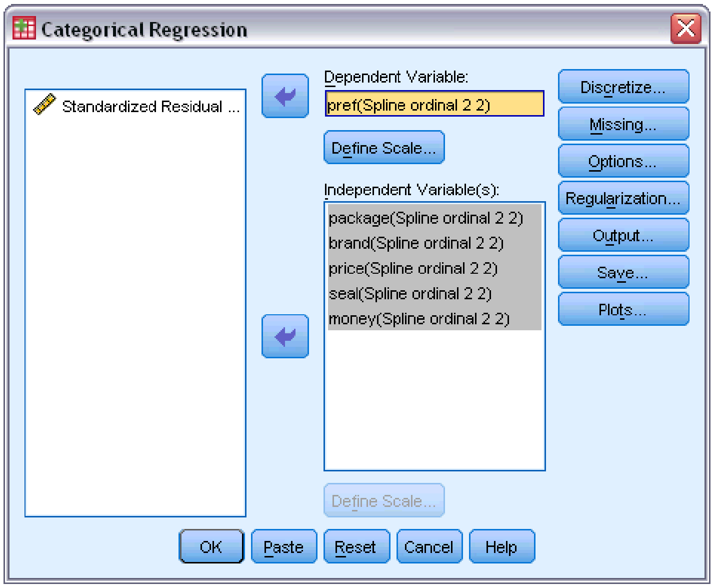
The categorical nature of the variables and the nonlinear relationship between *Preference* and *Package design* suggest that regression on optimal scores may perform better than standard regression. The U-shape of the residual plots indicates that a nominal treatment of *Package design* should be used. All other predictors will be treated at the numerical scaling level.

The response variable warrants special consideration. You want to predict the values of *Preference*. Thus, recovering as many properties of its categories as possible in the quantifications is desirable. Using an ordinal or nominal scaling level ignores the differences between the response categories. However, linearly transforming the response categories preserves category differences. Consequently, scaling the response numerically is generally preferred and will be employed here.

Running the Analysis

▶ To run a Categorical Regression analysis, from the menus choose:
Analyze > Regression > Optimal Scaling (CATREG)...

Figure 9-9
Categorical Regression dialog


Categorical Regression

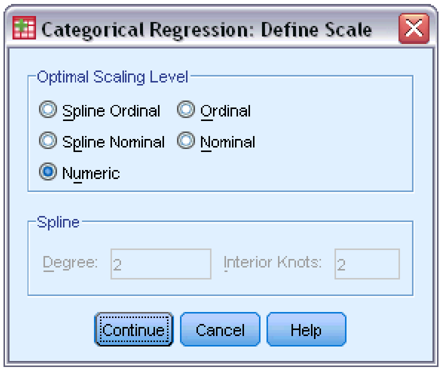
Standardized Residual ...

Dependent Variable: pref(Spline ordinal 2 2)

Independent Variable(s):
- package(Spline ordinal 2 2)
- brand(Spline ordinal 2 2)
- price(Spline ordinal 2 2)
- seal(Spline ordinal 2 2)
- money(Spline ordinal 2 2)

- ▶ Select *Preference* as the dependent variable.
- ▶ Select *Package design* through *Money-back guarantee* as independent variables.
- ▶ Select *Preference* and click Define Scale.

Figure 9-10
Define Scale dialog


Categorical Regression: Define Scale

Optimal Scaling Level

○ Spline Ordinal ○ Ordinal

○ Spline Nominal ○ Nominal

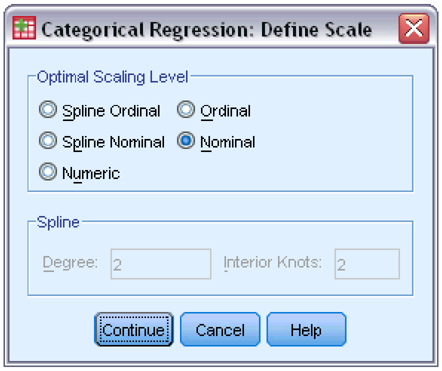
● Numeric

Spline

Degree: 2 Interior Knots: 2

- ▶ Select Numeric as the optimal scaling level.
- ▶ Click Continue.
- ▶ Select *Package design* and click Define Scale in the Categorical Regression dialog.

Figure 9-11
Define Scale dialog



Categorical Regression: Define Scale

Optimal Scaling Level

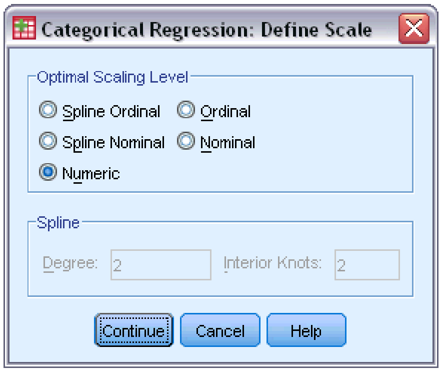
○ Spline Ordinal ○ Ordinal
○ Spline Nominal ● Nominal
○ Numeric

Spline

Degree: 2 Interior Knots: 2

- ▶ Select Nominal as the optimal scaling level.
- ▶ Click Continue.
- ▶ Select *Brand name* through *Money-back guarantee* and click Define Scale in the Categorical Regression dialog.

Figure 9-12
Define Scale dialog



Categorical Regression: Define Scale

Optimal Scaling Level

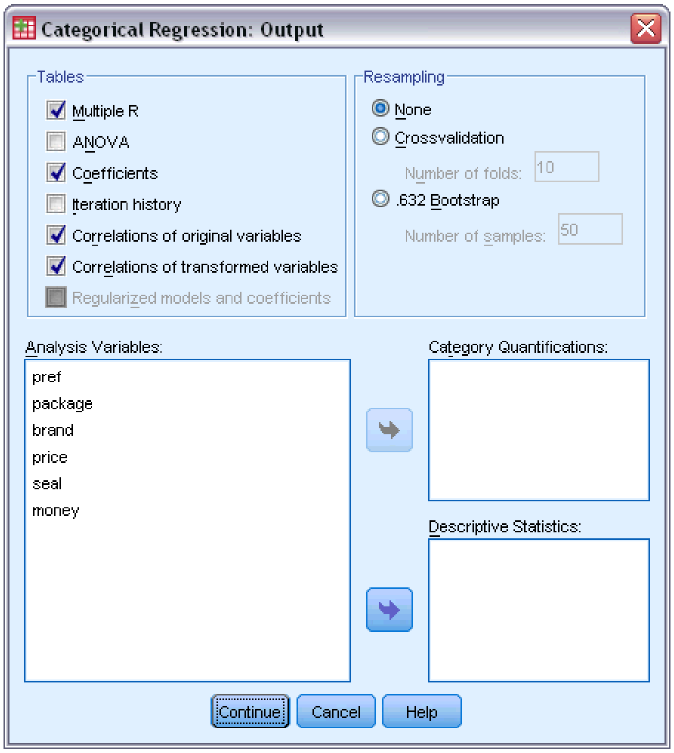
○ Spline Ordinal ○ Ordinal
○ Spline Nominal ○ Nominal
● Numeric

Spline

Degree: 2 Interior Knots: 2

- ▶ Select Numeric as the optimal scaling level.
- ▶ Click Continue.
- ▶ Click Output in the Categorical Regression dialog.

Figure 9-13
Output dialog



Categorical Regression: Output

Tables

☑ Multiple R
□ ANOVA
☑ Coefficients
□ Iteration history
☑ Correlations of original variables
☑ Correlations of transformed variables
□ Regularized models and coefficients

Resampling

● None
○ Crossvalidation
Number of folds: 10
○ .632 Bootstrap
Number of samples: 50

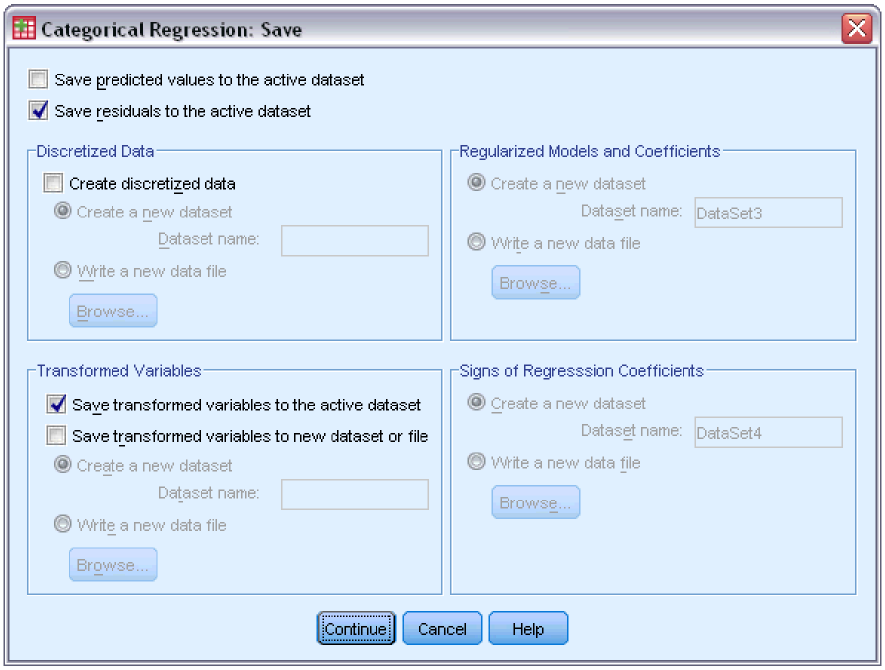
Analysis Variables:
pref
package
brand
price
seal
money

Category Quantifications:

Descriptive Statistics:

- ▶ Select Correlations of original variables and Correlations of transformed variables.
- ▶ Deselect ANOVA.
- ▶ Click Continue.
- ▶ Click Save in the Categorical Regression dialog.

Figure 9-14
Save dialog



Categorical Regression: Save

☐ Save predicted values to the active dataset
☑ Save residuals to the active dataset

Discretized Data
☐ Create discretized data
● Create a new dataset
Dataset name:
○ Write a new data file

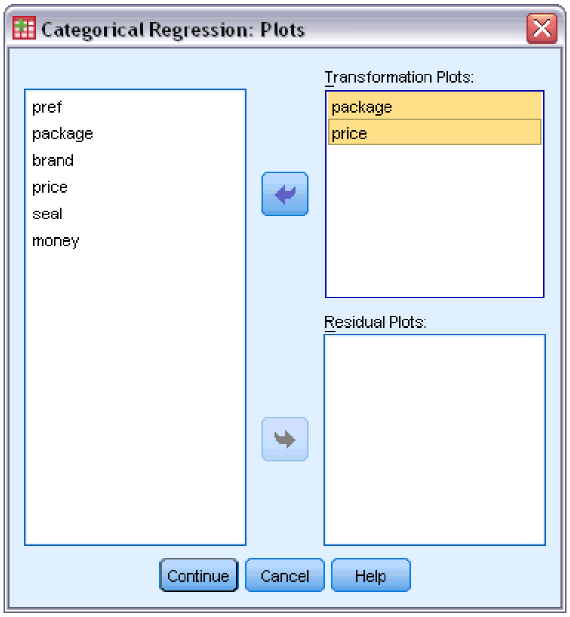
Regularized Models and Coefficients
● Create a new dataset
Dataset name: DataSet3
○ Write a new data file

Transformed Variables
☑ Save transformed variables to the active dataset
☐ Save transformed variables to new dataset or file
● Create a new dataset
Dataset name:
○ Write a new data file

Signs of Regresssion Coefficients
● Create a new dataset
Dataset name: DataSet4
○ Write a new data file

- ▶ Select Save residuals to the active dataset.
- ▶ Select Save transformed variables to the active dataset in the Transformed Variables group.
- ▶ Click Continue.
- ▶ Click Plots in the Categorical Regression dialog.

Figure 9-15
Plots dialog



Transformation Plots: package, price

- ▶ Choose to create transformation plots for *Package design* and *Price*.
- ▶ Click Continue.
- ▶ Click OK in the Categorical Regression dialog.

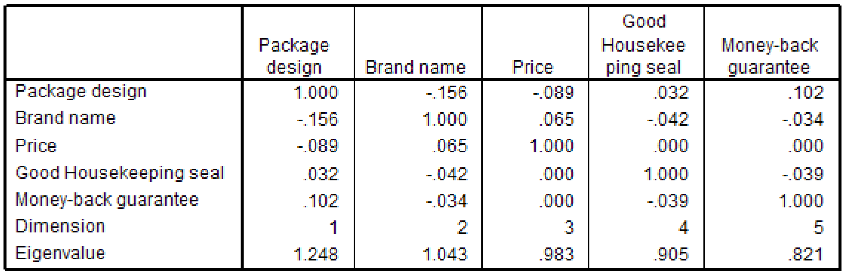
Intercorrelations

The intercorrelations among the predictors are useful for identifying multicollinearity in the regression. Variables that are highly correlated will lead to unstable regression estimates. However, due to their high correlation, omitting one of them from the model only minimally affects prediction. The variance in the response that can be explained by the omitted variable is still explained by the remaining correlated variable. However, zero-order correlations are sensitive to outliers and also cannot identify multicollinearity due to a high correlation between a predictor and a combination of other predictors.

Figure 9-16
Original predictor correlations

	Package design	Brand name	Price	Good Housekeeping seal	Money-back guarantee
Package design	1.000	-.189	-.126	.081	.066
Brand name	-.189	1.000	.065	-.042	-.034
Price	-.126	.065	1.000	.000	.000
Good Housekeeping seal	.081	-.042	.000	1.000	-.039
Money-back guarantee	.066	-.034	.000	-.039	1.000
Dimension	1	2	3	4	5
Eigenvalue	1.291	1.038	.980	.905	.785

Figure 9-17
Transformed predictor correlations



	Package design	Brand name	Price	Good Housekeeping seal	Money-back guarantee
Package design	1.000	-.156	-.089	.032	.102
Brand name	-.156	1.000	.065	-.042	-.034
Price	-.089	.065	1.000	.000	.000
Good Housekeeping seal	.032	-.042	.000	1.000	-.039
Money-back guarantee	.102	-.034	.000	-.039	1.000
Dimension	1	2	3	4	5
Eigenvalue	1.248	1.043	.983	.905	.821

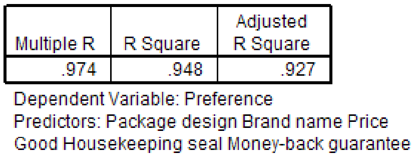
The intercorrelations of the predictors for both the untransformed and transformed predictors are displayed. All values are near 0, indicating that multicollinearity between individual variables is not a concern.

Notice that the only correlations that change involve *Package design*. Because all other predictors are treated numerically, the differences between the categories and the order of the categories are preserved for these variables. Consequently, the correlations cannot change.

Model Fit and Coefficients

The Categorical Regression procedure yields an R^2 of 0.948, indicating that almost 95% of the variance in the transformed preference rankings is explained by the regression on the optimally transformed predictors. Transforming the predictors improves the fit over the standard approach.

Figure 9-18
Model summary for categorical regression

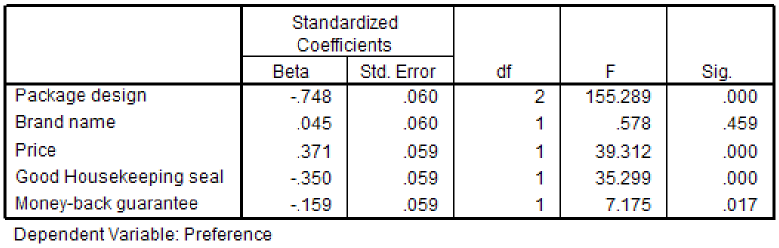


Multiple R	R Square	Adjusted R Square
.974	.948	.927

Dependent Variable: Preference
Predictors: Package design Brand name Price
Good Housekeeping seal Money-back guarantee

The following table shows the standardized regression coefficients. Categorical regression standardizes the variables, so only standardized coefficients are reported. These values are divided by their corresponding standard errors, yielding an *F* test for each variable. However, the test for each variable is contingent upon the other predictors being in the model. In other words, the test determines if omission of a predictor variable from the model with all other predictors present significantly worsens the predictive capabilities of the model. These values should not be used to omit several variables at one time for a subsequent model. Moreover, alternating least squares optimizes the quantifications, implying that these tests must be interpreted conservatively.

Figure 9-19
Standardized coefficients for transformed predictors



	Standardized Coefficients				
	Beta	Std. Error	df	F	Sig.
Package design	-.748	.060	2	155.289	.000
Brand name	.045	.060	1	.578	.459
Price	.371	.059	1	39.312	.000
Good Housekeeping seal	-.350	.059	1	35.299	.000
Money-back guarantee	-.159	.059	1	7.175	.017

Dependent Variable: Preference

The largest coefficient occurs for *Package design*. A one standard deviation increase in *Package design* yields a 0.748 standard deviation decrease in predicted preference ranking. However, *Package design* is treated nominally, so an increase in the quantifications need not correspond to an increase in the original category codes.

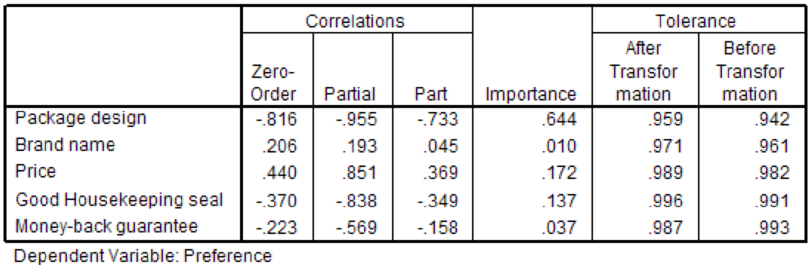
Standardized coefficients are often interpreted as reflecting the importance of each predictor. However, regression coefficients cannot fully describe the impact of a predictor or the relationships between the predictors. Alternative statistics must be used in conjunction with the standardized coefficients to fully explore predictor effects.

Correlations and Importance

To interpret the contributions of the predictors to the regression, it is not sufficient to only inspect the regression coefficients. In addition, the correlations, partial correlations, and part correlations should be inspected. The following table contains these correlational measures for each variable.

The zero-order correlation is the correlation between the transformed predictor and the transformed response. For this data, the largest correlation occurs for *Package design*. However, if you can explain some of the variation in either the predictor or the response, you will get a better representation of how well the predictor is doing.

Figure 9-20
Zero-order, part, and partial correlations (transformed variables)



	Correlations				Tolerance	
	Zero-Order	Partial	Part	Importance	After Transformation	Before Transformation
Package design	-.816	-.955	-.733	.644	.959	.942
Brand name	.206	.193	.045	.010	.971	.961
Price	.440	.851	.369	.172	.989	.982
Good Housekeeping seal	-.370	-.838	-.349	.137	.996	.991
Money-back guarantee	-.223	-.569	-.158	.037	.987	.993

Dependent Variable: Preference

Other variables in the model can confound the performance of a given predictor in predicting the response. The partial correlation coefficient removes the linear effects of other predictors from both the predictor and the response. This measure equals the correlation between the residuals from regressing the predictor on the other predictors and the residuals from regressing the response on the other predictors. The squared partial correlation corresponds to the proportion of the variance explained relative to the residual variance of the response remaining after removing the effects of the other variables. For example, *Package design* has a partial correlation of –0.955. Removing the effects of the other variables, *Package design* explains $(–0.955)^2$ = 0.91 = 91% of the variation in the preference rankings. Both *Price* and *Good Housekeeping seal* also explain a large portion of variance if the effects of the other variables are removed.

As an alternative to removing the effects of variables from both the response and a predictor, you can remove the effects from just the predictor. The correlation between the response and the residuals from regressing a predictor on the other predictors is the part correlation. Squaring this value yields a measure of the proportion of variance explained relative to the total variance of response. If you remove the effects of *Brand name*, *Good Housekeeping seal*, *Money back guarantee*, and *Price* from *Package design*, the remaining part of *Package design* explains $(–0.733)^2$ = 0.54 = 54% of the variation in preference rankings.

Importance

In addition to the regression coefficients and the correlations, Pratt's measure of relative importance (Pratt, 1987) aids in interpreting predictor contributions to the regression. Large individual importances relative to the other importances correspond to predictors that are crucial to the regression. Also, the presence of suppressor variables is signaled by a low importance for a variable that has a coefficient of similar size to the important predictors.

In contrast to the regression coefficients, this measure defines the importance of the predictors additively—that is, the importance of a set of predictors is the sum of the individual importances of the predictors. Pratt's measure equals the product of the regression coefficient and the zero-order correlation for a predictor. These products add to R^2, so they are divided by R^2, yielding a sum of 1. The set of predictors *Package design* and *Brand name*, for example, have an importance of 0.654. The largest importance corresponds to *Package design*, with *Package design*, *Price*, and *Good Housekeeping seal* accounting for 95% of the importance for this combination of predictors.

Multicollinearity

Large correlations between predictors will dramatically reduce a regression model's stability. Correlated predictors result in unstable parameter estimates. Tolerance reflects how much the independent variables are linearly related to one another. This measure is the proportion of a variable's variance not accounted for by other independent variables in the equation. If the other predictors can explain a large amount of a predictor's variance, that predictor is not needed in the model. A tolerance value near 1 indicates that the variable cannot be predicted very well from the other predictors. In contrast, a variable with a very low tolerance contributes little information to a model, and can cause computational problems. Moreover, large negative values of Pratt's importance measure indicate multicollinearity.

All of the tolerance measures are very high. None of the predictors are predicted very well by the other predictors and multicollinearity is not present.

Transformation Plots

Plotting the original category values against their corresponding quantifications can reveal trends that might not be noticed in a list of the quantifications. Such plots are commonly referred to as transformation plots. Attention should be given to categories that receive similar quantifications. These categories affect the predicted response in the same manner. However, the transformation type dictates the basic appearance of the plot.

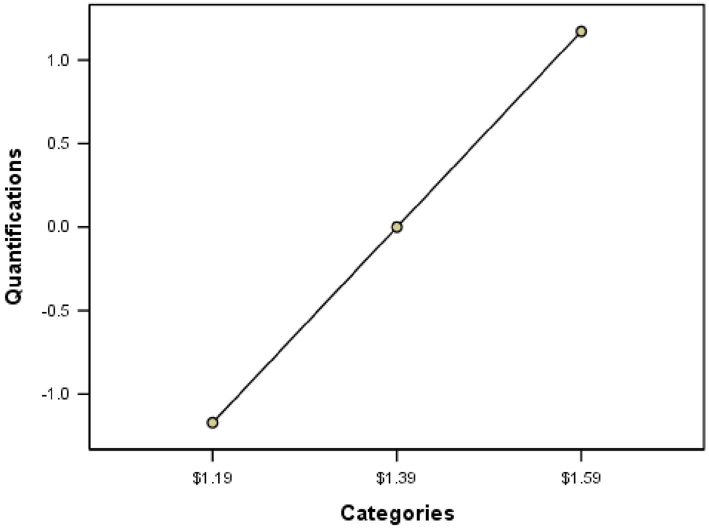
Variables treated as numerical result in a linear relationship between the quantifications and the original categories, corresponding to a straight line in the transformation plot. The order and the difference between the original categories is preserved in the quantifications.

The order of the quantifications for variables treated as ordinal correspond to the order of the original categories. However, the differences between the categories are not preserved. As a result, the transformation plot is nondecreasing but need not be a straight line. If consecutive categories correspond to similar quantifications, the category distinction may be unnecessary and the categories could be combined. Such categories result in a plateau on the transformation plot. However, this pattern can also result from imposing an ordinal structure on a variable that should be treated as nominal. If a subsequent nominal treatment of the variable reveals the same pattern, combining categories is warranted. Moreover, if the quantifications for a variable treated as ordinal fall along a straight line, a numerical transformation may be more appropriate.

For variables treated as nominal, the order of the categories along the horizontal axis corresponds to the order of the codes used to represent the categories. Interpretations of category order or of the distance between the categories is unfounded. The plot can assume any nonlinear or linear form. If an increasing trend is present, an ordinal treatment should be attempted. If the nominal transformation plot displays a linear trend, a numerical transformation may be more appropriate.

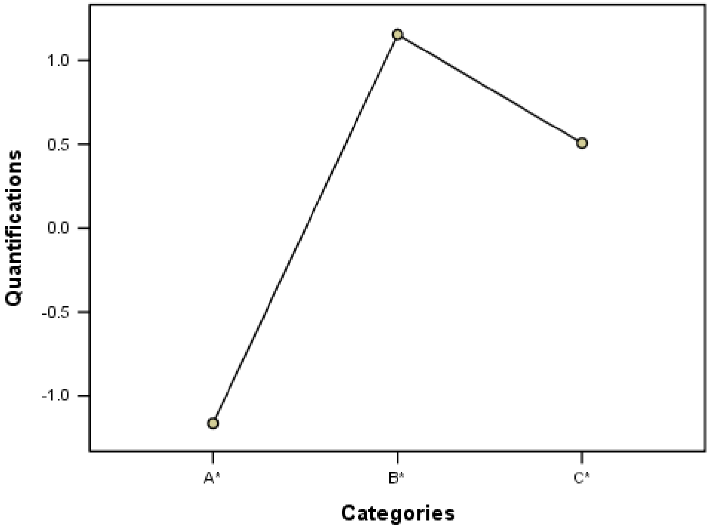
The following figure displays the transformation plot for *Price*, which was treated as numerical. Notice that the order of the categories along the straight line correspond to the order of the original categories. Also, the difference between the quantifications for *$1.19* and *$1.39* (–1.173 and 0) is the same as the difference between the quantifications for *$1.39* and *$1.59* (0 and 1.173). The fact that categories 1 and 3 are the same distance from category 2 is preserved in the quantifications.

Figure 9-21
Transformation plot of Price (numerical)



The nominal transformation of *Package design* yields the following transformation plot. Notice the distinct nonlinear shape in which the second category has the largest quantification. In terms of the regression, the second category decreases predicted preference ranking, whereas the first and third categories have the opposite effect.

Figure 9-22
Transformation plot of Package design (nominal)

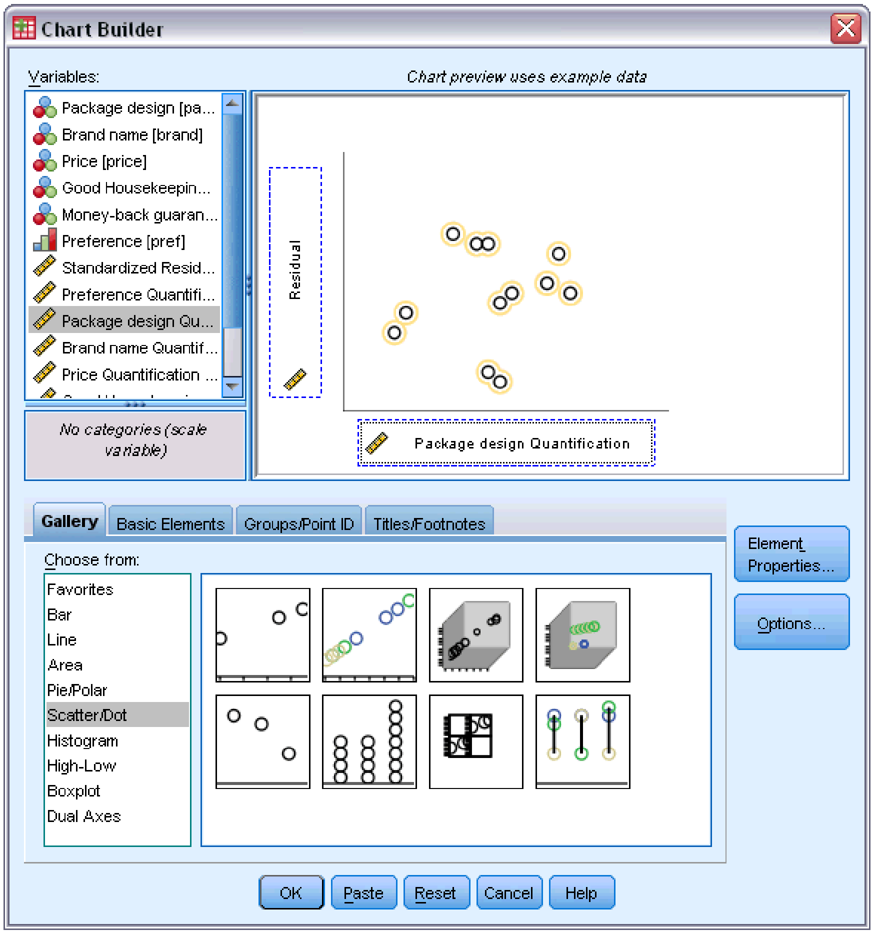


Residual Analysis

Using the transformed data and residuals that you saved to the active dataset allows you to create a scatterplot of the predicted values by the transformed values of *Package design*.

To obtain such a scatterplot, recall the Chart Builder and click Reset to clear your previous selections and restore the default options.

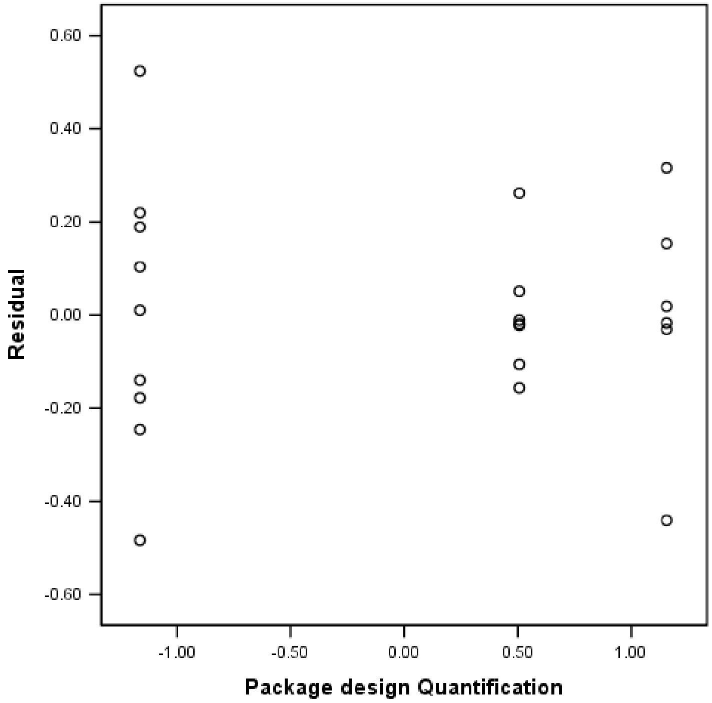
Figure 9-23
Chart Builder



- ▶ Select the Scatter/Dot gallery and choose Simple Scatter.
- ▶ Select *Residual* as the *y*-axis variable.
- ▶ Select *Package design Quantification* as the *x*-axis variable.
- ▶ Click OK.

The scatterplot shows the standardized residuals plotted against the optimal scores for *Package design*. All of the residuals are within two standard deviations of 0. A random scatter of points replaces the U-shape present in the scatterplot from the standard linear regression. Predictive abilities are improved by optimally quantifying the categories.

Figure 9-24
Residuals for Categorical Regression



Example: Ozone Data

In this example, you will use a larger set of data to illustrate the selection and effects of optimal scaling transformations. The data include 330 observations on six meteorological variables previously analyzed by Breiman and Friedman (Breiman and Friedman, 1985), and Hastie and Tibshirani (Hastie and Tibshirani, 1990), among others. The following table describes the original variables. Your categorical regression attempts to predict the ozone concentration from the remaining variables. Previous researchers found nonlinearities among these variables, which hinder standard regression approaches.

Table 9-2
Original variables

Variable	**Description**
ozon	daily ozone level; categorized into one of 38 categories
ibh	inversion base height
dpg	pressure gradient (mm Hg)
vis	visibility (miles)
temp	temperature (degrees F)
doy	day of the year

This dataset can be found in *ozone.sav*.For more information, see the topic Sample Files in Appendix A on p. 292.

Discretizing Variables

If a variable has more categories than is practically interpretable, you should modify the categories using the Discretization dialog to reduce the category range to a more manageable number.

The variable *Day of the year* has a minimum value of 3 and a maximum value of 365. Using this variable in a categorical regression corresponds to using a variable with 365 categories. Similarly, *Visibility (miles)* ranges from 0 to 350. To simplify interpretation of analyses, discretize these variables into equal intervals of length 10.

The variable *Inversion base height* ranges from 111 to 5000. A variable with this many categories results in very complex relationships. However, discretizing this variable into equal intervals of length 100 yields roughly 50 categories. Using a 50-category variable rather than a 5000-category variable simplifies interpretations significantly.

Pressure gradient (mm Hg) ranges from –69 to 107. The procedure omits any categories coded with negative numbers from the analysis, but discretizing this variable into equal intervals of length 10 yields roughly 19 categories.

Temperature (degrees F) ranges from 25 to 93 on the Fahrenheit scale. In order to analyze the data as if it were on the Celsius scale, discretize this variable into equal intervals of length 1.8.

Different discretizations for variables may be desired. The choices used here are purely subjective. If you desire fewer categories, choose larger intervals. For example, *Day of the year* could have been divided into months of the year or seasons.

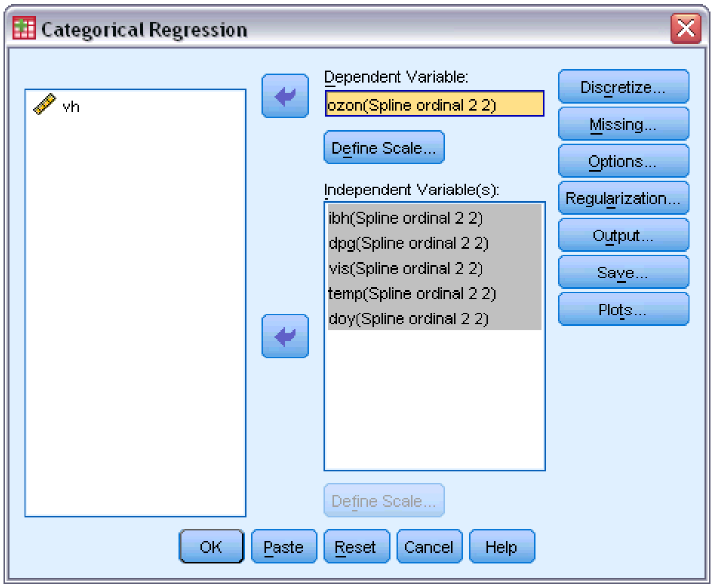
Selection of Transformation Type

Each variable can be analyzed at one of several different levels. However, because prediction of the response is the goal, you should scale the response “as is” by employing the numerical optimal scaling level. Consequently, the order and the differences between categories will be preserved in the transformed variable.

▶ To run a Categorical Regression analysis, from the menus choose:

Analyze > Regression > Optimal Scaling (CATREG)...

Figure 9-25
Categorical Regression dialog



Categorical Regression

vh

Dependent Variable: ozon(Spline ordinal 2 2)

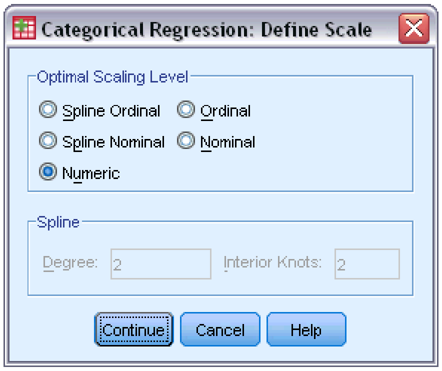
Independent Variable(s):
ibh(Spline ordinal 2 2)
dpg(Spline ordinal 2 2)
vis(Spline ordinal 2 2)
temp(Spline ordinal 2 2)
doy(Spline ordinal 2 2)

▶ Select *Daily ozone level* as the dependent variable.

▶ Select *Inversion base height* through *Day of the year* as independent variables.

▶ Select *Daily ozone level* and click Define Scale.

Figure 9-26
Define Scale dialog



Categorical Regression: Define Scale

Optimal Scaling Level
○ Spline Ordinal ○ Ordinal
○ Spline Nominal ○ Nominal
● Numeric

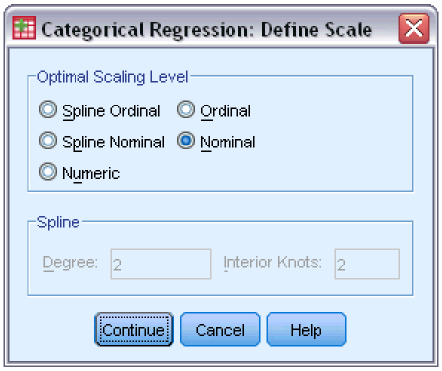
Spline
Degree: 2 Interior Knots: 2

▶ Select Numeric as the optimal scaling level.

▶ Click Continue.

▶ Select *Inversion base height* through *Day of the year*, and click Define Scale in the Categorical Regression dialog.

Figure 9-27
Define Scale dialog



Categorical Regression: Define Scale

Optimal Scaling Level

○ Spline Ordinal ○ Ordinal

○ Spline Nominal ● Nominal

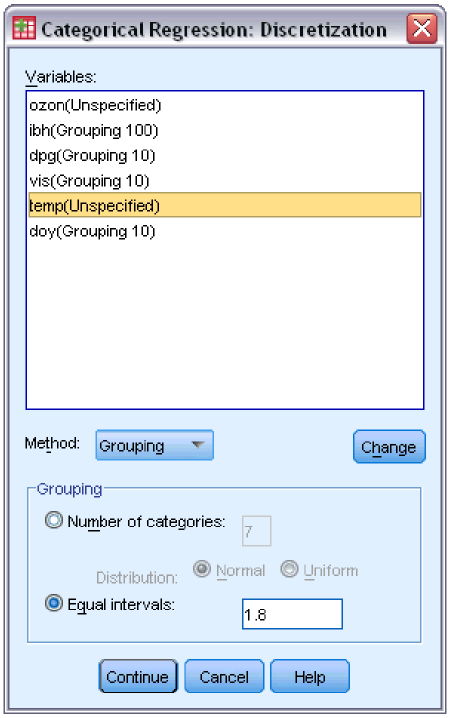
○ Numeric

Spline

Degree: 2 Interior Knots: 2

- Select Nominal as the optimal scaling level.
- Click Continue.
- Click Discretize in the Categorical Regression dialog.

Figure 9-28
Discretization dialog



Categorical Regression: Discretization

Variables:

ozon(Unspecified)
ibh(Grouping 100)
dpg(Grouping 10)
vis(Grouping 10)
temp(Unspecified)
doy(Grouping 10)

Method: Grouping

Grouping

○ Number of categories: 7

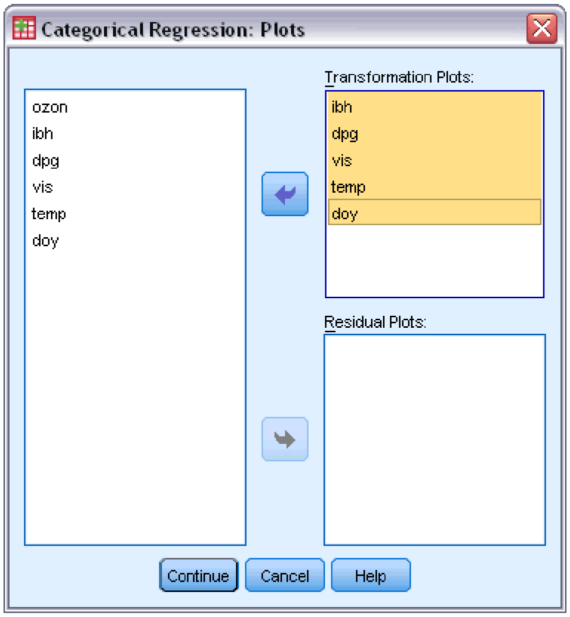
Distribution: ● Normal ○ Uniform

● Equal intervals: 1.8

- Select *ibh*.
- Select Equal intervals and type 100 as the interval length.
- Click Change.
- Select *dpg*, *vis*, and *doy*.

- ▶ Type 10 as the interval length.
- ▶ Click Change.
- ▶ Select *temp*.
- ▶ Type 1.8 as the interval length.
- ▶ Click Change.
- ▶ Click Continue.
- ▶ Click Plots in the Categorical Regression dialog.

Figure 9-29
Plots dialog

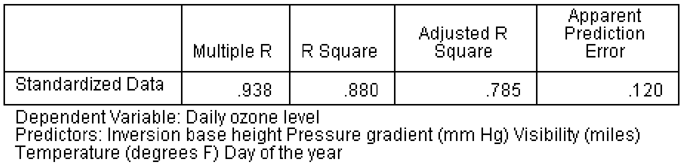


Transformation Plots:
ibh
dpg
vis
temp
doy

Residual Plots:

- ▶ Select transformation plots for *Inversion base height* through *Day of the year*.
- ▶ Click Continue.
- ▶ Click OK in the Categorical Regression dialog.

Figure 9-30
Model summary

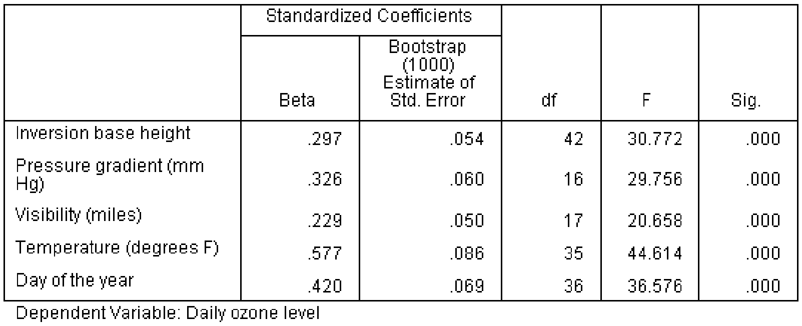


	Multiple R	R Square	Adjusted R Square	Apparent Prediction Error
Standardized Data	.938	.880	.785	.120

Dependent Variable: Daily ozone level
Predictors: Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

Treating all predictors as nominal yields an R^2 of 0.880. This large amount of variance accounted for is not surprising because nominal treatment imposes no restrictions on the quantifications. However, interpreting the results can be quite difficult.

Figure 9-31
Regression coefficients (all predictors nominal)

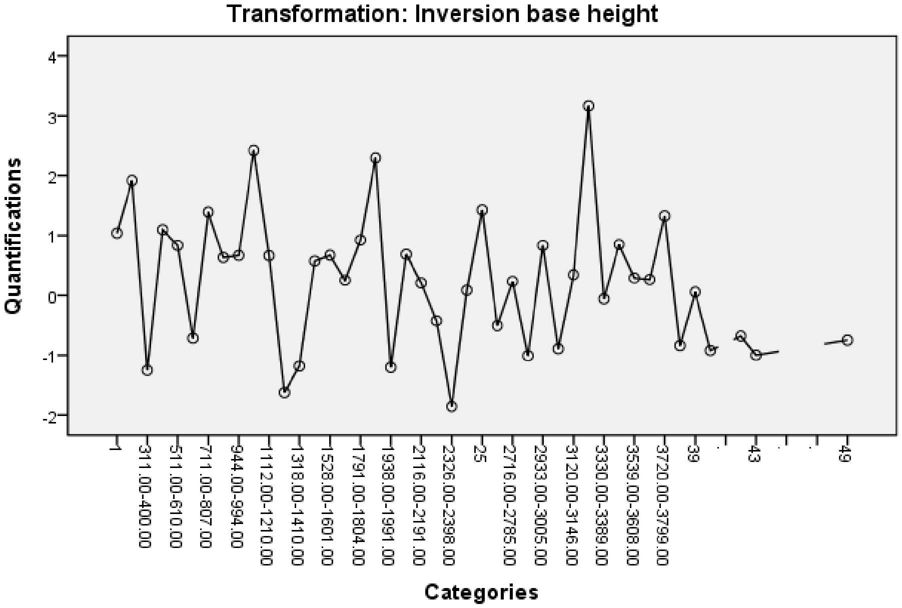


	Standardized Coefficients				
	Beta	Bootstrap (1000) Estimate of Std. Error	df	F	Sig.
Inversion base height	.297	.054	42	30.772	.000
Pressure gradient (mm Hg)	.326	.060	16	29.756	.000
Visibility (miles)	.229	.050	17	20.658	.000
Temperature (degrees F)	.577	.086	35	44.614	.000
Day of the year	.420	.069	36	36.576	.000

Dependent Variable: Daily ozone level

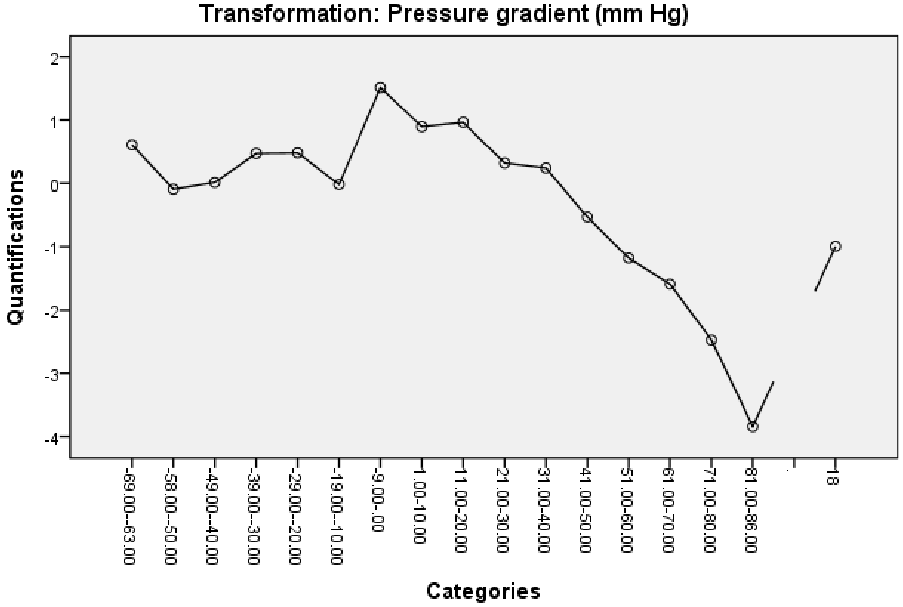
This table shows the standardized regression coefficients of the predictors. A common mistake made when interpreting these values involves focusing on the coefficients while neglecting the quantifications. You cannot simply assert that a positive value of *Inversion base height*, for example, implies that as the predictor increases, predicted *Ozone* increases. All interpretations must be relative to the transformed variables, so that as the quantifications for *Inversion base height* increase, predicted *Ozone* increases. To examine the effects of the original variables, you must relate the categories to the quantifications.

Figure 9-32
Transformation plot of Inversion base height (nominal)



The transformation plot of *Inversion base height* shows no apparent pattern. As evidenced by the jagged nature of the plot, moving from low categories to high categories yields fluctuations in the quantifications in both directions. Thus, describing the effects of this variable requires focusing on the individual categories. Imposing ordinal or linear restrictions on the quantifications for this variable might significantly reduce the fit.

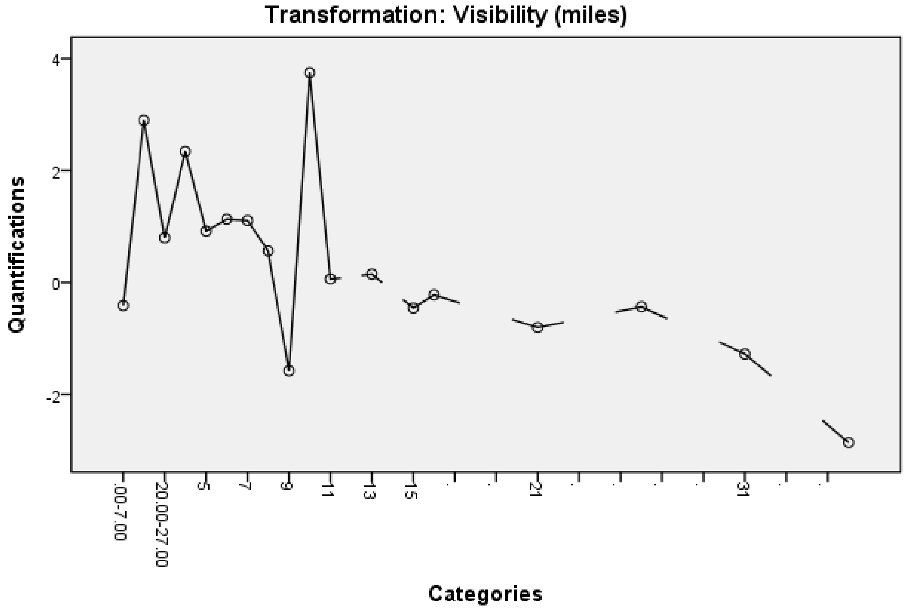
Figure 9-33
Transformation plot of Pressure gradient (nominal)



This figure displays the transformation plot of *Pressure gradient*. The initial discretized categories (*1* through *6*) receive small quantifications and thus have minimal contributions to the predicted response. The next three categories receive somewhat higher, positive values, resulting in a moderate increase in predicted ozone.

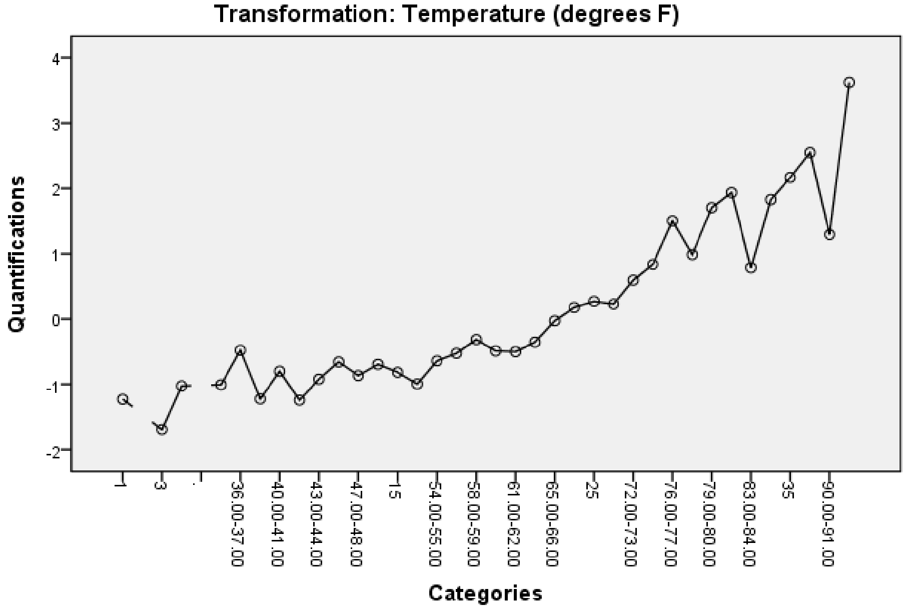
The quantifications decrease up to category *16*, where *Pressure gradient* has its greatest decreasing effect on predicted ozone. Although the line increases after this category, using an ordinal scaling level for *Pressure gradient* may not significantly reduce the fit, while simplifying the interpretations of the effects. However, the importance measure of 0.04 and the regression coefficient for *Pressure gradient* indicates that this variable is not very useful in the regression.

Figure 9-34
Transformation plot of Visibility (nominal)



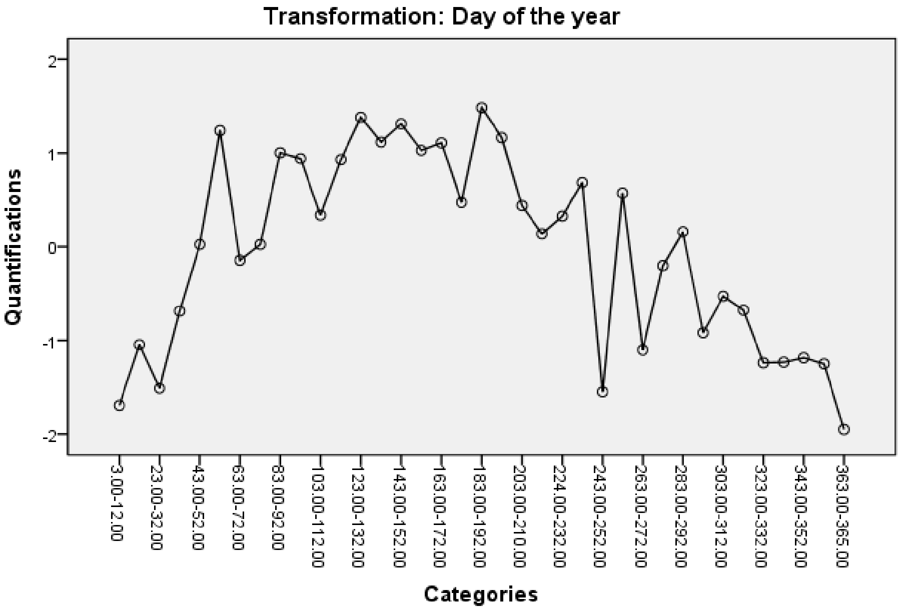
The transformation plot of *Visibility*, like that for *Inversion base height*, shows no apparent pattern. Imposing ordinal or linear restrictions on the quantifications for this variable might significantly reduce the fit.

Figure 9-35
Transformation plot of Temperature (nominal)



The transformation plot of *Temperature* displays an alternative pattern. As the categories increase, the quantifications tend to increase. As a result, as *Temperature* increases, predicted ozone tends to increase. This pattern suggests scaling *Temperature* at the ordinal level.

Figure 9-36
Transformation plot of Day of the year (nominal)

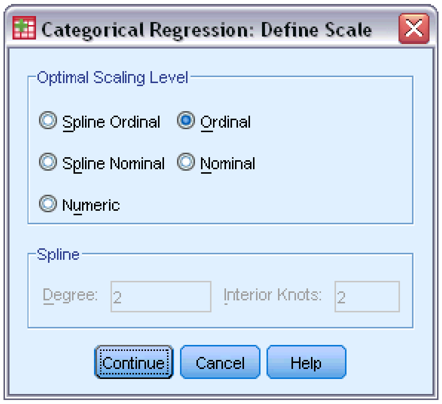


This figure shows the transformation plot of *Day of the year*. The quantifications tend to increase up to the midpoint of the graph, at which point they tend to decrease, yielding an inverted U-shape. Considering the sign of the regression coefficient for *Day of the year*, the initial categories receive quantifications that have a decreasing effect on predicted ozone. For the middle categories, the effect of the quantifications on predicted ozone increases, reaching a maximum around the midpoint of the graph.

Beyond that point, the quantifications tend to decrease the predicted ozone. Although the line is quite jagged, the general shape is still identifiable. Thus, the transformation plots suggest scaling *Temperature* at the ordinal level while keeping all other predictors nominally scaled.

To recompute the regression, scaling *Temperature* at the ordinal level, recall the Categorical Regression dialog.

Figure 9-37
Define Scale dialog


Categorical Regression: Define Scale

Optimal Scaling Level

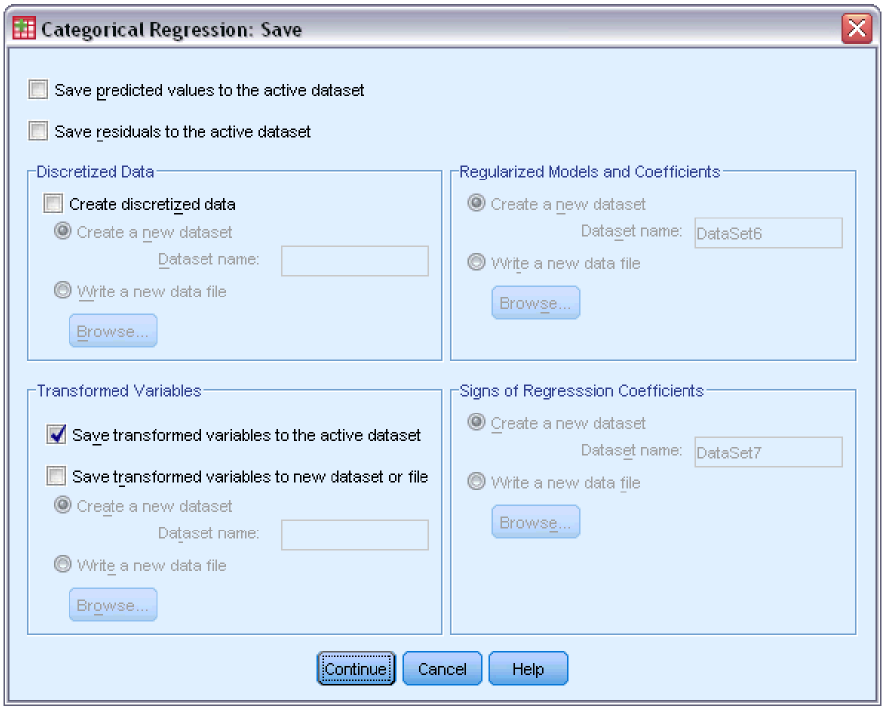
○ Spline Ordinal ● Ordinal
○ Spline Nominal ○ Nominal
○ Numeric

Spline

Degree: 2 Interior Knots: 2

- ▶ Select *Temperature* and click Define Scale.
- ▶ Select Ordinal as the optimal scaling level.
- ▶ Click Continue.
- ▶ Click Save in the Categorical Regression dialog.

Figure 9-38
Save dialog



Categorical Regression: Save

□ Save predicted values to the active dataset

□ Save residuals to the active dataset

Discretized Data
□ Create discretized data
● Create a new dataset
Dataset name:
○ Write a new data file

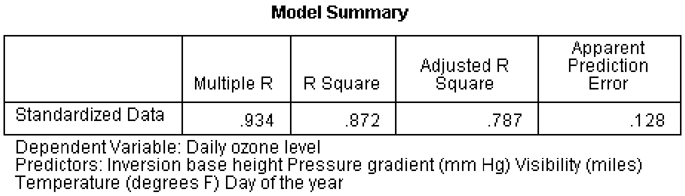
Regularized Models and Coefficients
● Create a new dataset
Dataset name: DataSet6
○ Write a new data file

Transformed Variables
☑ Save transformed variables to the active dataset
□ Save transformed variables to new dataset or file
● Create a new dataset
Dataset name:
○ Write a new data file

Signs of Regresssion Coefficients
● Create a new dataset
Dataset name: DataSet7
○ Write a new data file

- ▶ Select Save transformed variables to the active dataset in the Transformed Variables group.
- ▶ Click Continue.
- ▶ Click OK in the Categorical Regression dialog.

Figure 9-39
Model summary for regression with Temperature (ordinal)



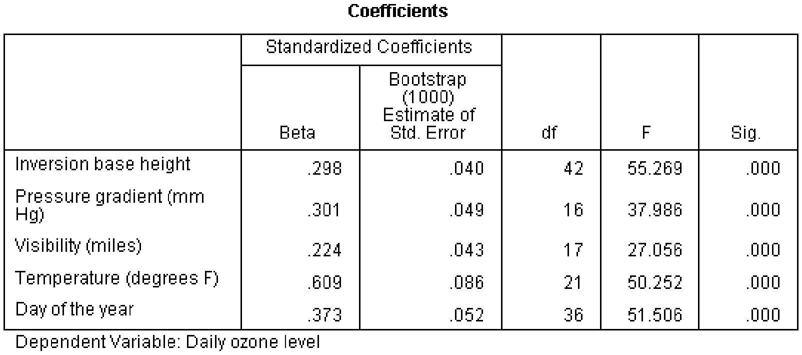
Model Summary

	Multiple R	R Square	Adjusted R Square	Apparent Prediction Error
Standardized Data	.934	.872	.787	.128

Dependent Variable: Daily ozone level
Predictors: Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

This model results in an R^2 of 0.872, so the variance accounted for decreases negligibly when the quantifications for *Temperature* are restricted to be ordered.

Figure 9-40
Regression coefficients with Temperature (ordinal)



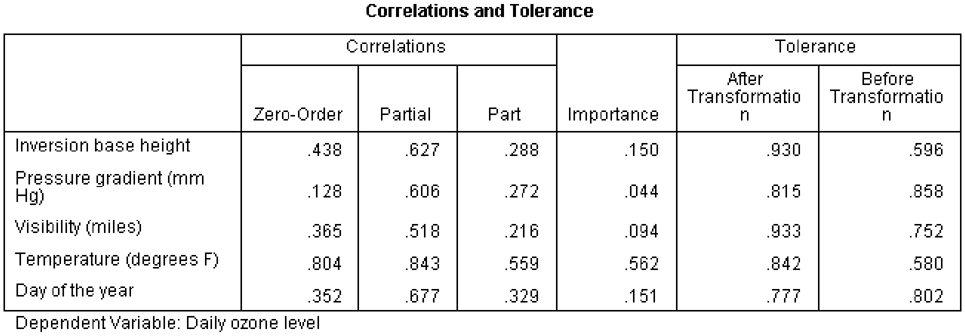
Coefficients

	Standardized Coefficients				
	Beta	Bootstrap (1000) Estimate of Std. Error	df	F	Sig.
Inversion base height	.298	.040	42	55.269	.000
Pressure gradient (mm Hg)	.301	.049	16	37.986	.000
Visibility (miles)	.224	.043	17	27.056	.000
Temperature (degrees F)	.609	.086	21	50.252	.000
Day of the year	.373	.052	36	51.506	.000

Dependent Variable: Daily ozone level

This table displays the coefficients for the model in which *Temperature* is scaled as ordinal. Comparing the coefficients to those for the model in which *Temperature* is scaled as nominal, no large changes occur.

Figure 9-41
Correlations, importance, and tolerance



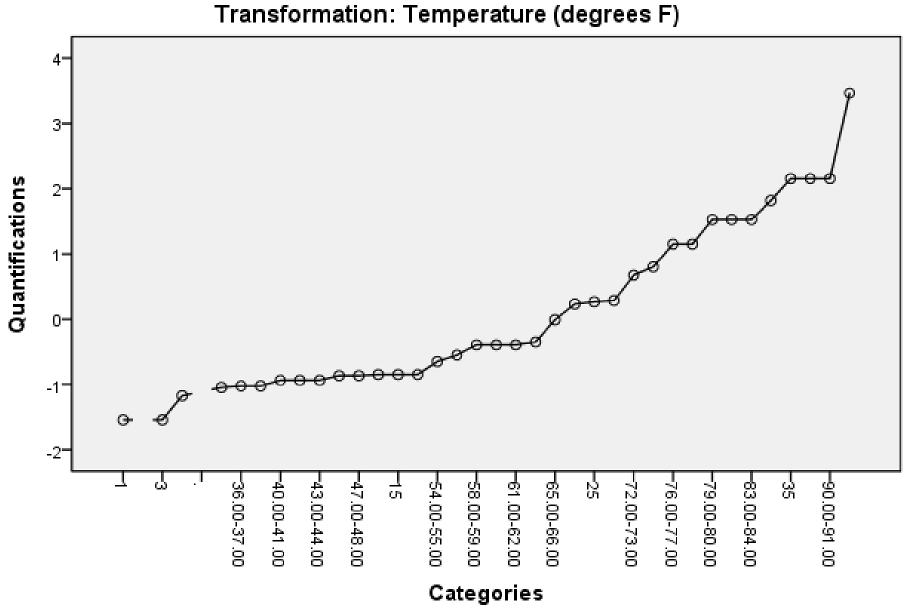
Correlations and Tolerance

	Correlations				Tolerance	
	Zero-Order	Partial	Part	Importance	After Transformatio n	Before Transformatio n
Inversion base height	.438	.627	.288	.150	.930	.596
Pressure gradient (mm Hg)	.128	.606	.272	.044	.815	.858
Visibility (miles)	.365	.518	.216	.094	.933	.752
Temperature (degrees F)	.804	.843	.559	.562	.842	.580
Day of the year	.352	.677	.329	.151	.777	.802

Dependent Variable: Daily ozone level

Moreover, the importance measures suggest that *Temperature* is still much more important to the regression than the other variables. Now, however, as a result of the ordinal scaling level of *Temperature* and the positive regression coefficient, you can assert that as *Temperature* increases, predicted ozone increases.

Figure 9-42
Transformation plot of Temperature (ordinal)



The transformation plot illustrates the ordinal restriction on the quantifications for *Temperature*. The jagged line from the nominal transformation is replaced here by a smooth ascending line. Moreover, no long plateaus are present, indicating that collapsing categories is not needed.

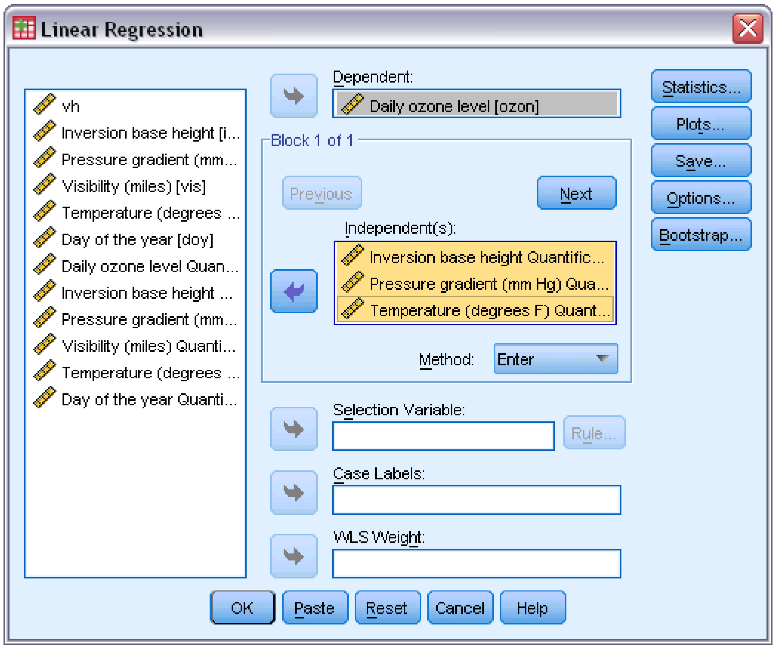
Optimality of the Quantifications

The transformed variables from a categorical regression can be used in a standard linear regression, yielding identical results. However, the quantifications are optimal only for the model that produced them. Using a subset of the predictors in linear regression does not correspond to an optimal scaling regression on the same subset.

For example, the categorical regression that you have computed has an R^2 of 0.875. You have saved the transformed variables, so in order to fit a linear regression using only *Temperature*, *Pressure gradient*, and *Inversion base height* as predictors, from the menus choose:

Analyze > Regression > Linear...

Figure 9-43
Linear Regression dialog



Dependent: Daily ozone level [ozon]

Independent(s):
Inversion base height Quantific...
Pressure gradient (mm Hg) Qua...
Temperature (degrees F) Quant...

Method: Enter

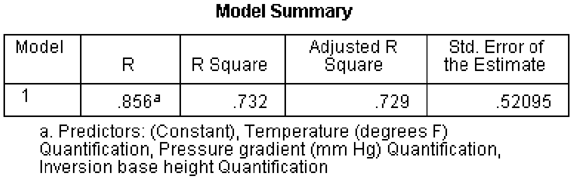
Selection Variable:

Case Labels:

WLS Weight:

- ▶ Select *Daily ozone level Quantification* as the dependent variable.
- ▶ Select *Inversion base height Quantification*, *Pressure gradient (mm Hg) Quantification*, and *Temperature (degrees F) Quantification* as independent variables.
- ▶ Click OK.

Figure 9-44
Model summary for regression with subset of optimally scaled predictors



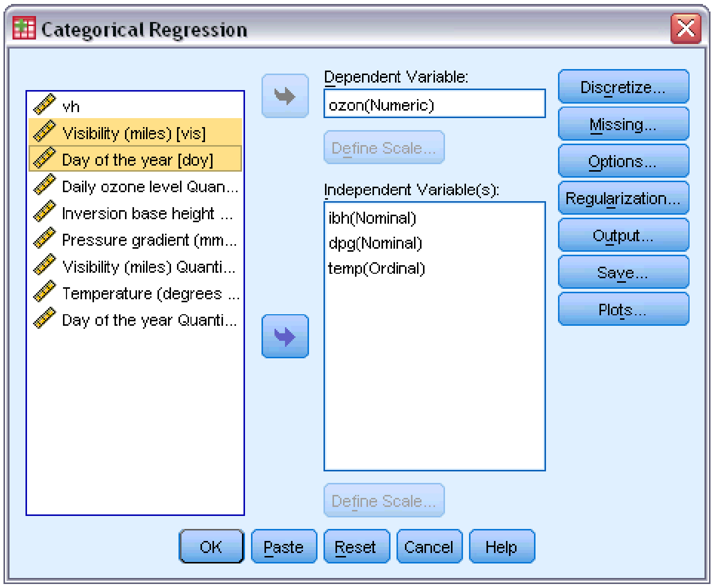
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.856[a]	.732	.729	.52095

a. Predictors: (Constant), Temperature (degrees F) Quantification, Pressure gradient (mm Hg) Quantification, Inversion base height Quantification

Using the quantifications for the response, *Temperature*, *Pressure gradient*, and *Inversion base height* in a standard linear regression results in a fit of 0.732. To compare this to the fit of a categorical regression using just those three predictors, recall the Categorical Regression dialog.

Figure 9-45
Categorical Regression dialog


Categorical Regression

vh; Visibility (miles) [vis]; Day of the year [doy]; Daily ozone level Quan...; Inversion base height ...; Pressure gradient (mm...; Visibility (miles) Quanti...; Temperature (degrees ...; Day of the year Quanti...

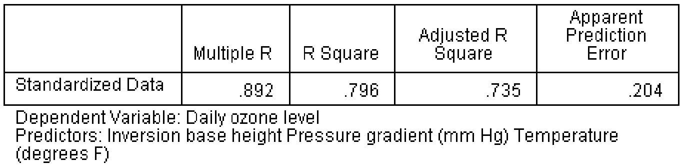
Dependent Variable: ozon(Numeric)

Independent Variable(s): ibh(Nominal), dpg(Nominal), temp(Ordinal)

- ▶ Deselect *Visibility (miles)* and *Day of the year* as independent variables.

- ▶ Click OK.

Figure 9-46
Model summary for categorical regression on three predictors



	Multiple R	R Square	Adjusted R Square	Apparent Prediction Error
Standardized Data	.892	.796	.735	.204

Dependent Variable: Daily ozone level
Predictors: Inversion base height Pressure gradient (mm Hg) Temperature (degrees F)

The categorical regression analysis has a fit of 0.796, which is better than the fit of 0.732. This demonstrates the property of the scalings that the quantifications obtained in the original regression are only optimal when all five variables are included in the model.

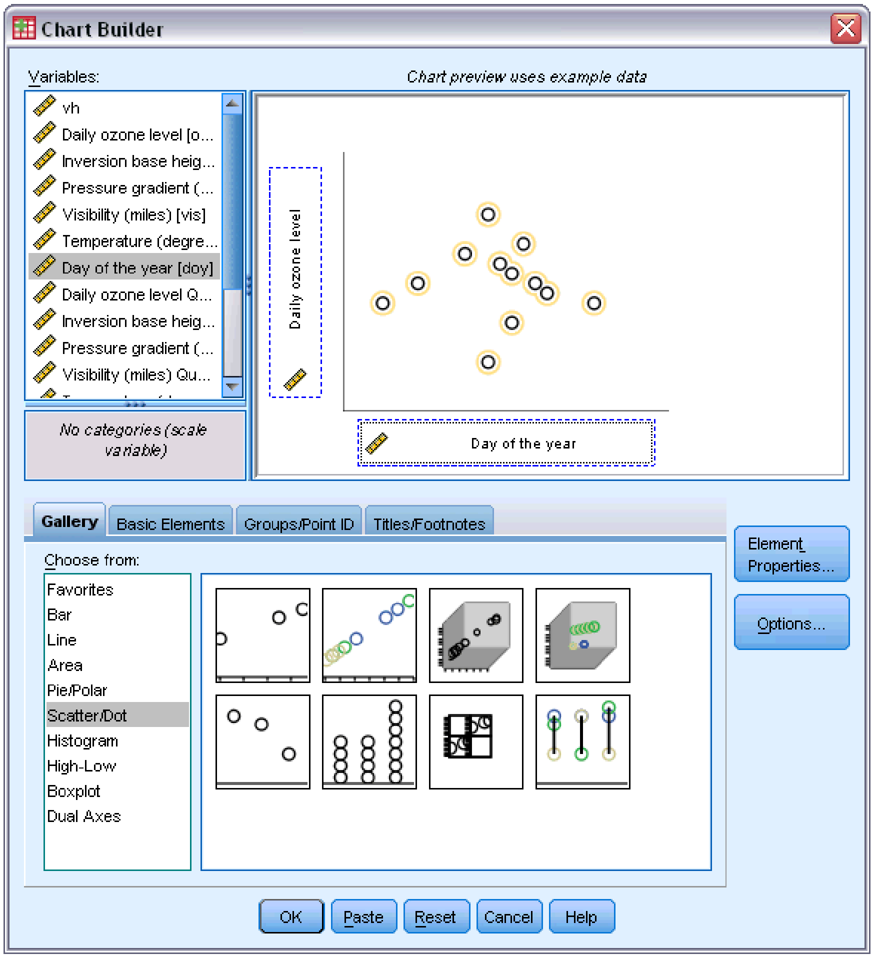
Effects of Transformations

Transforming the variables makes a nonlinear relationship between the original response and the original set of predictors linear for the transformed variables. However, when there are multiple predictors, pairwise relationships are confounded by the other variables in the model.

To focus your analysis on the relationship between *Daily ozone level* and *Day of the year*, begin by looking at a scatterplot. From the menus choose:

Graphs > Chart Builder...

Figure 9-47
Chart Builder dialog

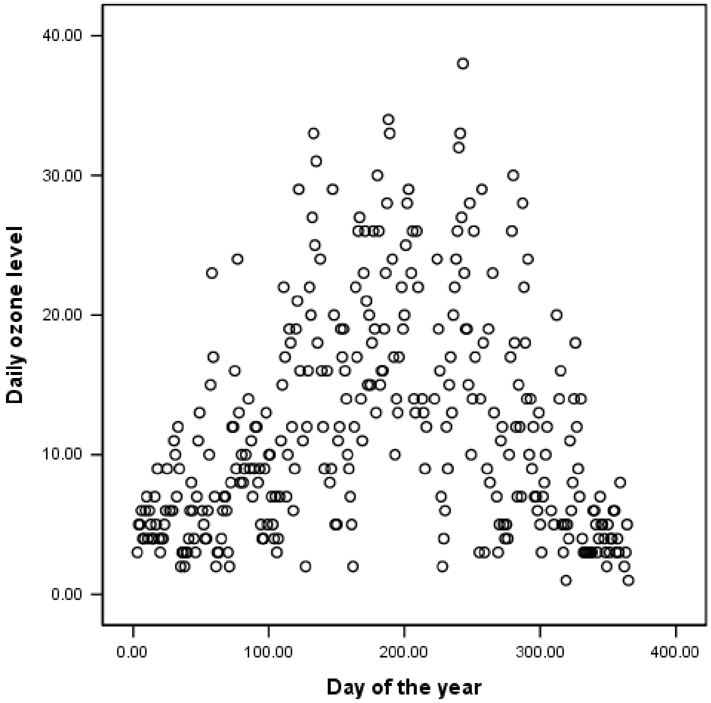


- ▶ Select the Scatter/Dot gallery and choose Simple Scatter.

- ▶ Select *Daily ozone level* as the *y*-axis variable and *Day of the year* as the *x*-axis variable.

- ▶ Click OK.

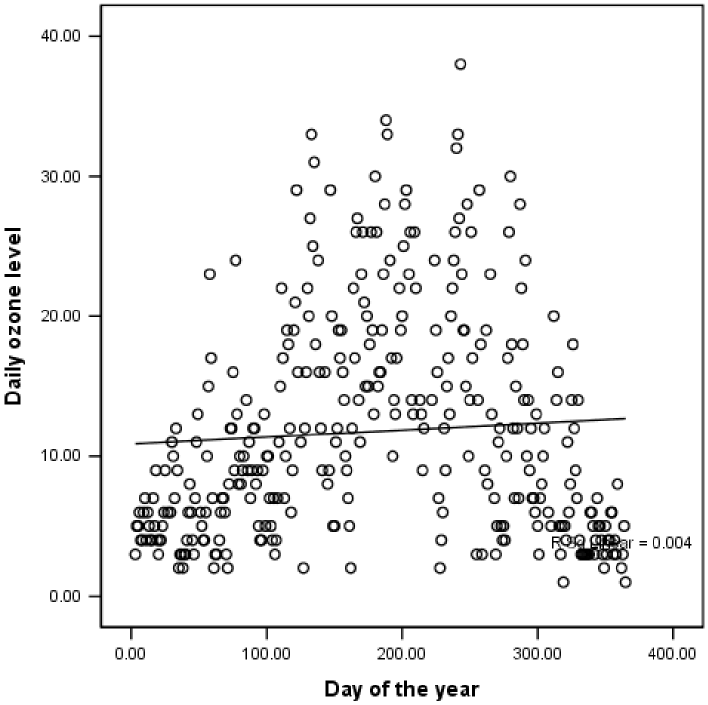
Figure 9-48
Scatterplot of Daily ozone level and Day of the year



This figure illustrates the relationship between *Daily ozone level* and *Day of the year*. As *Day of the year* increases to approximately 200, *Daily ozone level* increases. However, for *Day of the year* values greater than 200, *Daily ozone level* decreases. This inverted U pattern suggests a quadratic relationship between the two variables. A linear regression cannot capture this relationship.

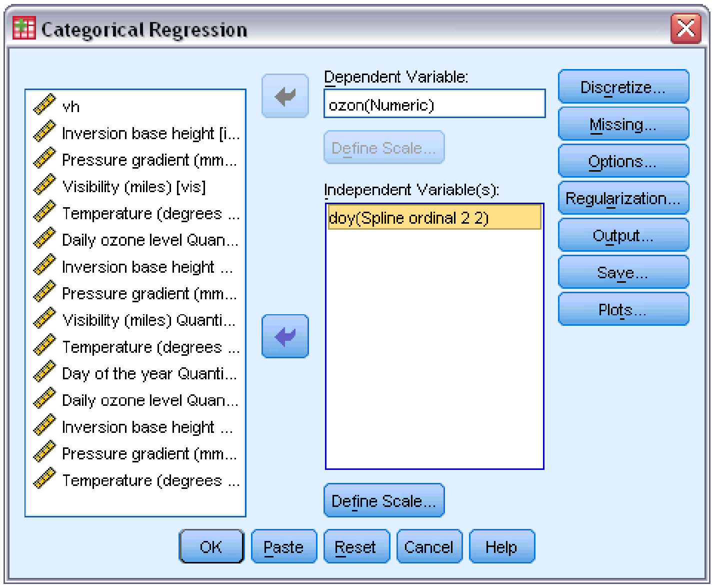
- ▶ To see a best-fit line overlaid on the points in the scatterplot, activate the graph by double-clicking on it.
- ▶ Select a point in the Chart Editor.
- ▶ Click the Add Fit Line at Total tool, and close the Chart Editor.

Figure 9-49
Scatterplot showing best-fit line



A linear regression of *Daily ozone level* on *Day of the year* yields an R^2 of 0.004. This fit suggests that *Day of the year* has no predictive value for *Daily ozone level*. This is not surprising, given the pattern in the figure. By using optimal scaling, however, you can linearize the quadratic relationship and use the transformed *Day of the year* to predict the response.

Figure 9-50
Categorical Regression dialog



Categorical Regression

vh
Inversion base height [i...
Pressure gradient (mm...
Visibility (miles) [vis]
Temperature (degrees ...
Daily ozone level Quan...
Inversion base height ...
Pressure gradient (mm...
Visibility (miles) Quanti...
Temperature (degrees ...
Day of the year Quanti...
Daily ozone level Quan...
Inversion base height ...
Pressure gradient (mm...
Temperature (degrees ...

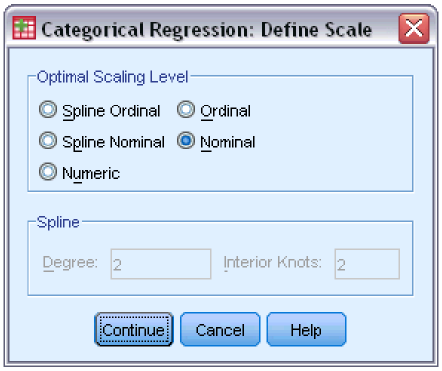
Dependent Variable: ozon(Numeric)

Independent Variable(s): doy(Spline ordinal 2 2)

To obtain a categorical regression of *Daily ozone level* on *Day of the year*, recall the Categorical Regression dialog.

- Deselect *Inversion base height* through *Temperature (degrees F)* as independent variables.
- Select *Day of the year* as an independent variable.
- Click Define Scale.

Figure 9-51
Define Scale dialog



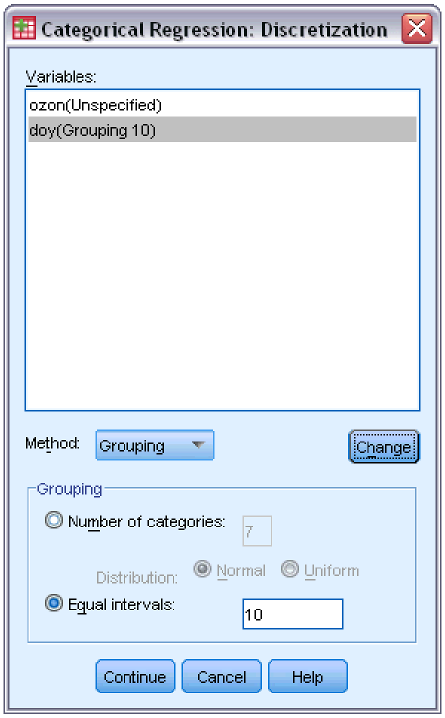
Categorical Regression: Define Scale

Optimal Scaling Level
○ Spline Ordinal ○ Ordinal
○ Spline Nominal ● Nominal
○ Numeric

Spline
Degree: 2 Interior Knots: 2

- Select Nominal as the optimal scaling level.
- Click Continue.
- Click Discretize in the Categorical Regression dialog.

Figure 9-52
Discretization dialog



Variables:
ozon(Unspecified)
doy(Grouping 10)

Method: Grouping

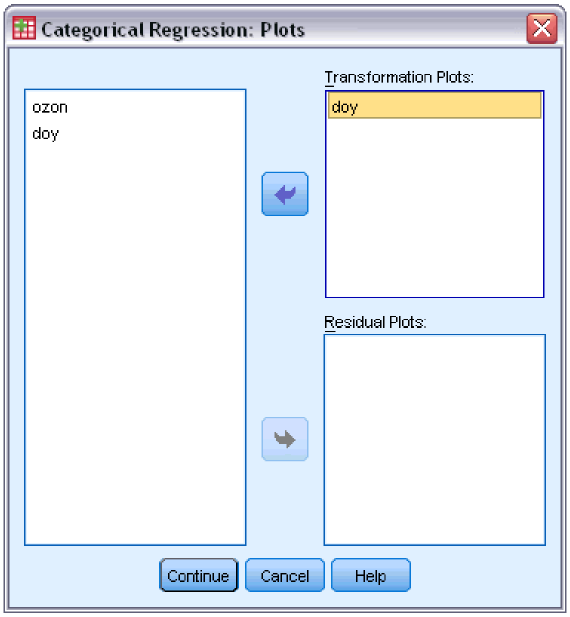
○ Number of categories: 7

Distribution: ● Normal ○ Uniform

● Equal intervals: 10

- ▶ Select *doy*.
- ▶ Select Equal intervals.
- ▶ Type 10 as the interval length.
- ▶ Click Change.
- ▶ Click Continue.
- ▶ Click Plots in the Categorical Regression dialog.

Figure 9-53
Plots dialog

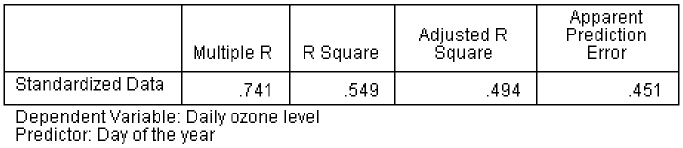


Transformation Plots: doy

Residual Plots:

- ▶ Select *doy* for transformation plots.
- ▶ Click Continue.
- ▶ Click OK in the Categorical Regression dialog.

Figure 9-54
Model summary for categorical regression of Daily ozone level on Day of the year

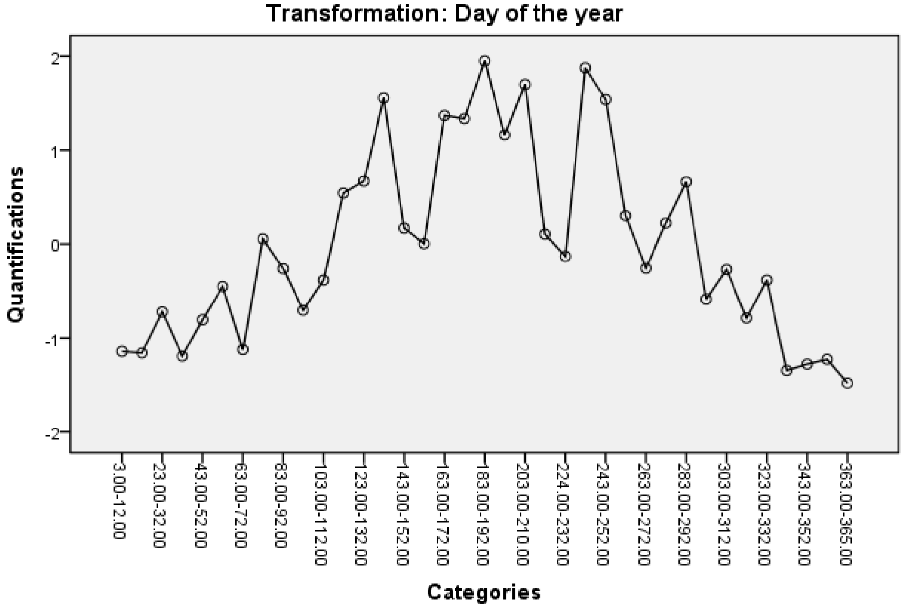


	Multiple R	R Square	Adjusted R Square	Apparent Prediction Error
Standardized Data	.741	.549	.494	.451

Dependent Variable: Daily ozone level
Predictor: Day of the year

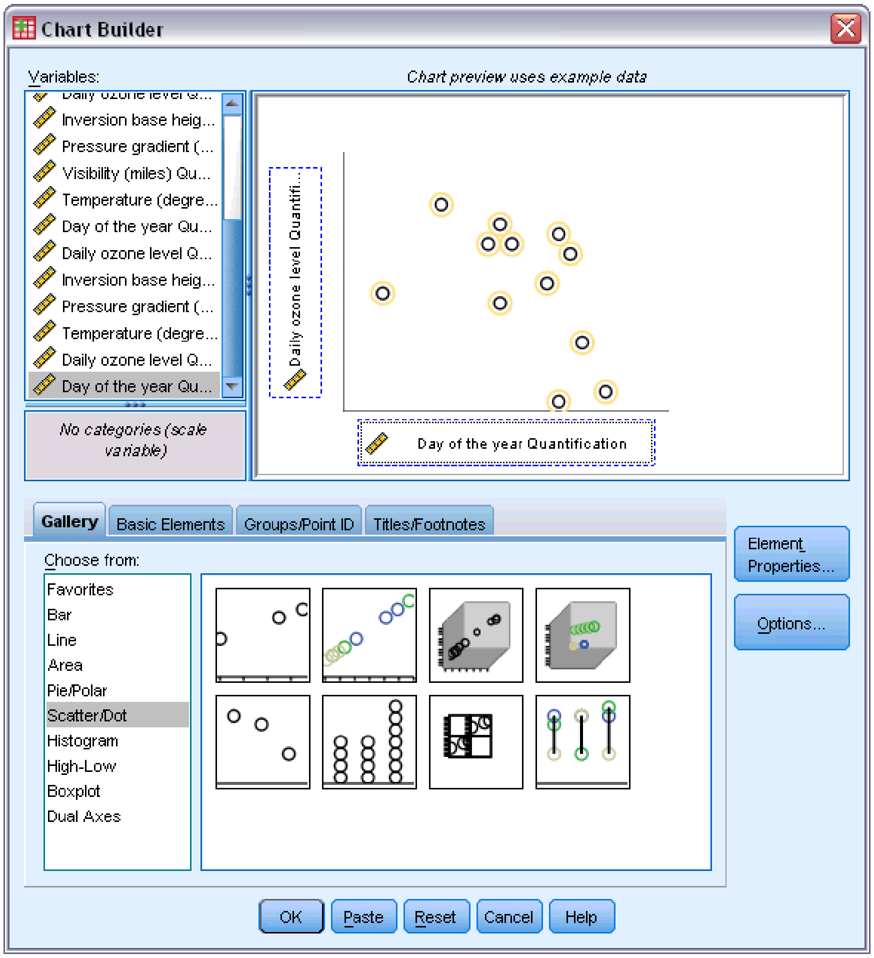
The optimal scaling regression treats *Daily ozone level* as numerical and *Day of the year* as nominal. This results in an R^2 of 0.549. Although only 55% of the variation in *Daily ozone level* is accounted for by the categorical regression, this is a substantial improvement over the original regression. Transforming *Day of the year* allows for the prediction of *Daily ozone level*.

Figure 9-55
Transformation plot of Day of the year (nominal)



This figure displays the transformation plot of *Day of the year*. The extremes of *Day of the year* both receive negative quantifications, whereas the central values have positive quantifications. By applying this transformation, the low and high *Day of the year* values have similar effects on predicted *Daily ozone level*.

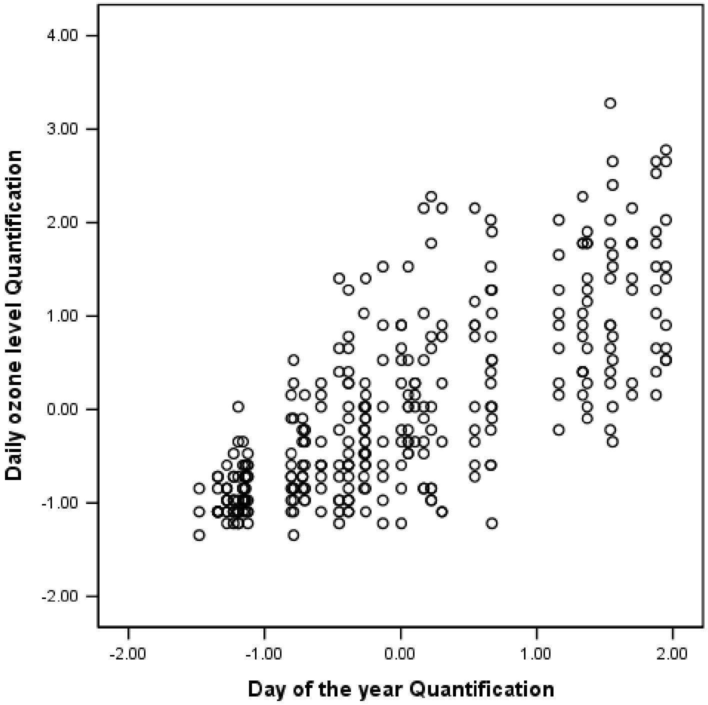
Figure 9-56
Chart Builder



To see a scatterplot of the transformed variables, recall the Chart Builder, and click Reset to clear your previous selections.

- Select the Scatter/Dot gallery and choose Simple Scatter.
- Select *Daily ozone level Quantification [TRA1_3]* as the *y*-axis variable and *Day of the year Quantification [TRA2_3]* as the *x*-axis variable.
- Click OK.

Figure 9-57
Scatterplot of the transformed variables



This figure depicts the relationship between the transformed variables. An increasing trend replaces the inverted U. The regression line has a positive slope, indicating that as transformed *Day of the year* increases, predicted *Daily ozone level* increases. Using optimal scaling linearizes the relationship and allows interpretations that would otherwise go unnoticed.

Recommended Readings

See the following texts for more information on categorical regression:

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, 1032–1069.

Hastie, T., R. Tibshirani, and A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, 1255–1270.

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, 93–96.

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, 251–263.

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, 493–517.

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, 425–441.

Van der Kooij, A. J., and J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. In: *Softstat '97,* F. Faulbaum, and W. Bandilla, eds. Stuttgart: Gustav Fisher, 99–106.

Winsberg, S., and J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, 669–674.

Winsberg, S., and J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, 575–595.

Young, F. W., J. De Leeuw, and Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 505–528.

Chapter

10

Categorical Principal Components Analysis

Categorical principal components analysis can be thought of as a method of dimension reduction. A set of variables is analyzed to reveal major dimensions of variation. The original data set can then be replaced by a new, smaller data set with minimal loss of information. The method reveals relationships among variables, among cases, and among variables and cases.

The criterion used by categorical principal components analysis for quantifying the observed data is that the object scores (component scores) should have large correlations with each of the quantified variables. A solution is good to the extent that this criterion is satisfied.

Two examples of categorical principal components analysis will be presented. The first employs a rather small data set useful for illustrating the basic concepts and interpretations associated with the procedure. The second example examines a practical application.

Example: Examining Interrelations of Social Systems

This example examines Guttman's (Guttman, 1968) adaptation of a table by Bell (Bell, 1961). The data are also discussed by Lingoes (Lingoes, 1968).

Bell presented a table to illustrate possible social groups. Guttman used a portion of this table, in which five variables describing such things as social interaction, feelings of belonging to a group, physical proximity of members, and formality of the relationship were crossed with seven theoretical social groups, including crowds (for example, people at a football game), audiences (for example, people at a theater or classroom lecture), public (for example, newspaper or television audiences), mobs (like a crowd but with much more intense interaction), primary groups (intimate), secondary groups (voluntary), and the modern community (loose confederation resulting from close physical proximity and a need for specialized services).

The following table shows the variables in the dataset resulting from the classification into seven social groups used in the Guttman-Bell data, with their variable labels and the value labels (categories) associated with the levels of each variable. This dataset can be found in *guttman.sav*. For more information, see the topic Sample Files in Appendix A on p. 292. In addition to selecting variables to be included in the computation of the categorical principal components analysis, you can select variables that are used to label objects in plots. In this example, the first five variables in the data are included in the analysis, while cluster is used exclusively as a labeling variable. When you specify a categorical principal components analysis, you must specify the optimal scaling level for each analysis variable. In this example, an ordinal level is specified for all analysis variables.

Table 10-1
Variables in the Guttman-Bell dataset

Variable name	**Variable label**	**Value label**
intnsity	Intensity of interaction	Slight, low, moderate, high
frquency	Frequency of interaction	Slight, nonrecurring, infrequent, frequent

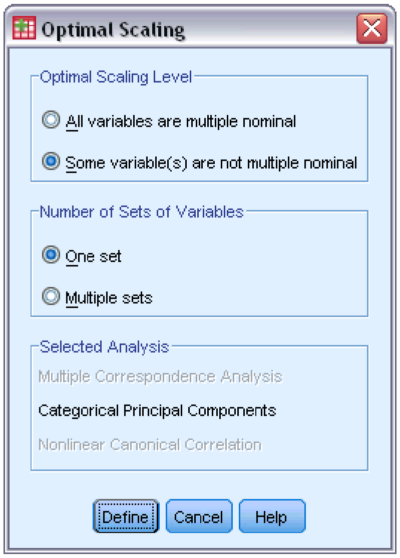
Variable name	Variable label	Value label
blonging	Feeling of belonging	None, slight, variable, high
proxmity	Physical proximity	Distant, close
formlity	Formality of relationship	No relationship, formal, informal
cluster		Crowds, audiences, public, mobs, primary groups, secondary groups, modern community

Running the Analysis

- ▶ To produce categorical principal components output for this dataset, from the menus choose:

  Analyze > Dimension Reduction > Optimal Scaling...

Figure 10-1
Optimal Scaling dialog



Optimal Scaling Level

○ All variables are multiple nominal

● Some variable(s) are not multiple nominal

Number of Sets of Variables

● One set

○ Multiple sets

Selected Analysis

Multiple Correspondence Analysis

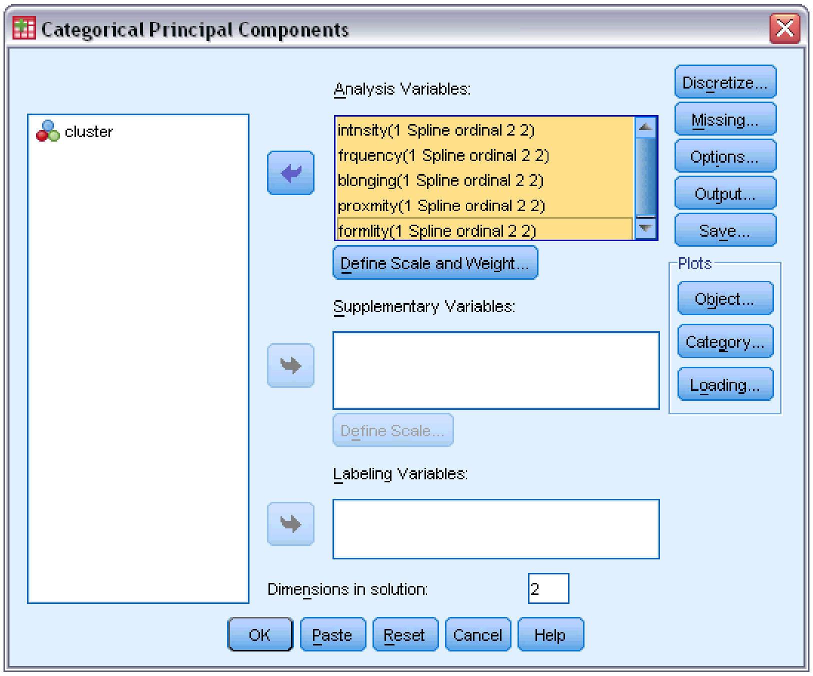
Categorical Principal Components

Nonlinear Canonical Correlation

- ▶ Select Some variable(s) not multiple nominal in the Optimal Scaling Level group.

- ▶ Click Define.

Figure 10-2
Categorical Principal Components dialog



cluster

Analysis Variables:

intnsity(1 Spline ordinal 2 2)
frquency(1 Spline ordinal 2 2)
blonging(1 Spline ordinal 2 2)
proxmity(1 Spline ordinal 2 2)
formlity(1 Spline ordinal 2 2)

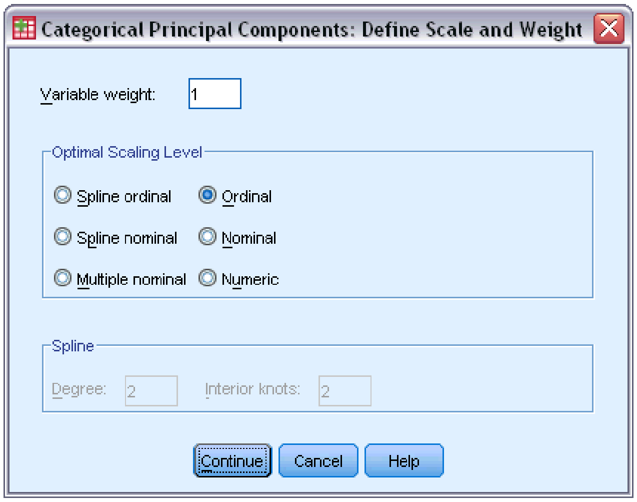
Supplementary Variables:

Labeling Variables:

Dimensions in solution: 2

- ▶ Select *Intensity of interaction* through *Formality of relationship* as analysis variables.
- ▶ Click Define Scale and Weight.

Figure 10-3
Define Scale and Weight dialog



Variable weight: 1

Optimal Scaling Level

○ Spline ordinal ● Ordinal
○ Spline nominal ○ Nominal
○ Multiple nominal ○ Numeric

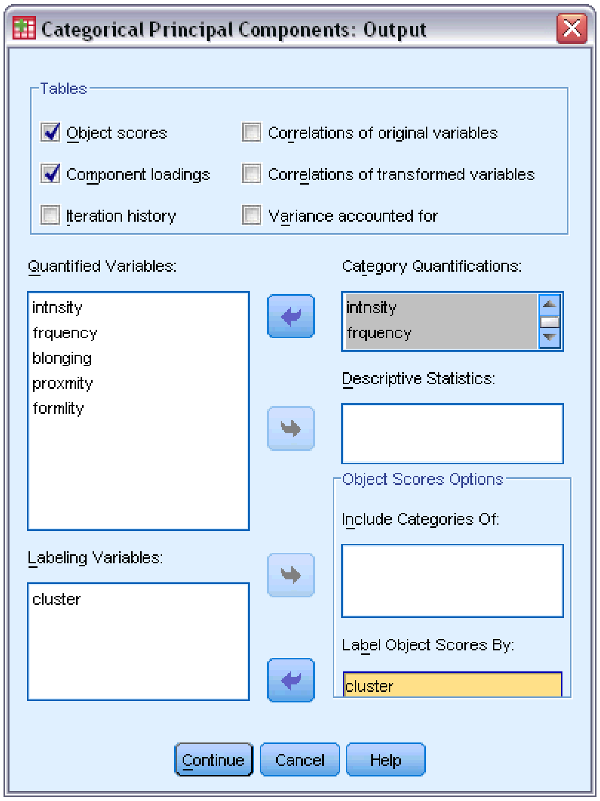
Spline

Degree: 2 Interior knots: 2

- ▶ Select Ordinal in the Optimal Scaling Level group.
- ▶ Click Continue.
- ▶ Select *cluster* as a labeling variable in the Categorical Principal Components dialog.

- ▶ Click Output.

Figure 10-4
Output dialog



Categorical Principal Components: Output

Tables

☑ Object scores
☑ Component loadings
☐ Iteration history
☐ Correlations of original variables
☐ Correlations of transformed variables
☐ Variance accounted for

Quantified Variables:
intnsity
frquency
blonging
proxmity
formlity

Category Quantifications:
intnsity
frquency

Descriptive Statistics:

Labeling Variables:
cluster

Object Scores Options

Include Categories Of:

Label Object Scores By:
cluster

- ▶ Select Object scores and deselect Correlations of transformed variables in the Tables group.

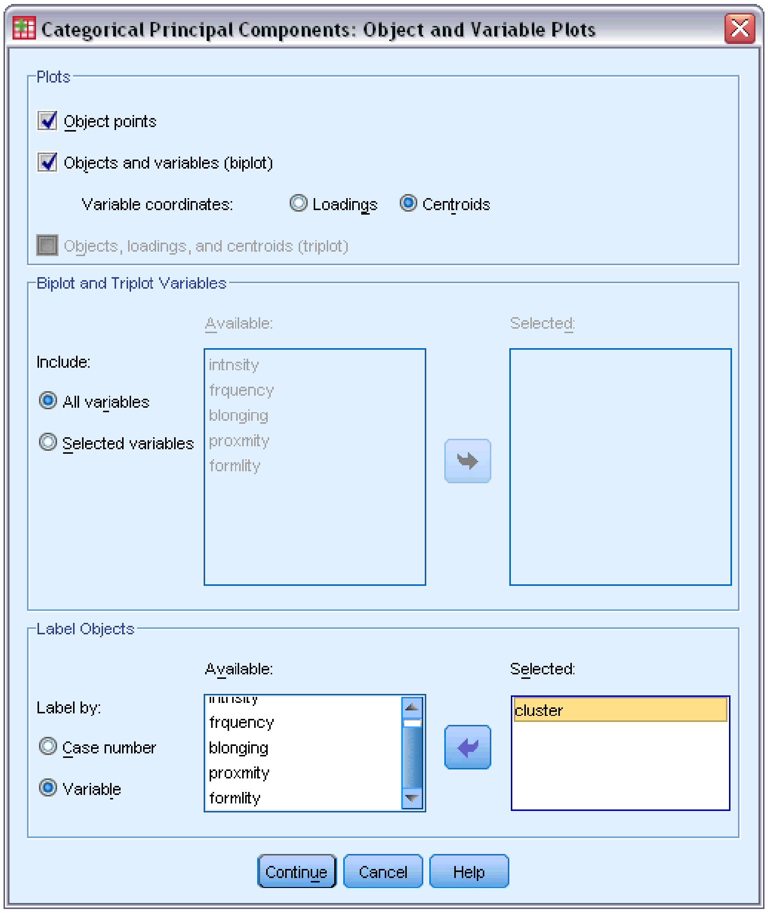
- ▶ Choose to produce category quantifications for *intnsity (Intensity of interaction)* through *formlity (Formality of relationship)*.

- ▶ Choose to label object scores by *cluster*.

- ▶ Click Continue.

- ▶ Click Object in the Plots group of the Categorical Principal Components dialog.

Figure 10-5
Object and Variable Plots dialog



Categorical Principal Components: Object and Variable Plots

Plots

☑ Object points

☑ Objects and variables (biplot)

Variable coordinates: ○ Loadings ● Centroids

□ Objects, loadings, and centroids (triplot)

Biplot and Triplot Variables

Include: ● All variables ○ Selected variables

Available: intnsity, frquency, blonging, proxmity, formlity

Selected:

Label Objects

Label by: ○ Case number ● Variable

Available: intnsity, frquency, blonging, proxmity, formlity

Selected: cluster

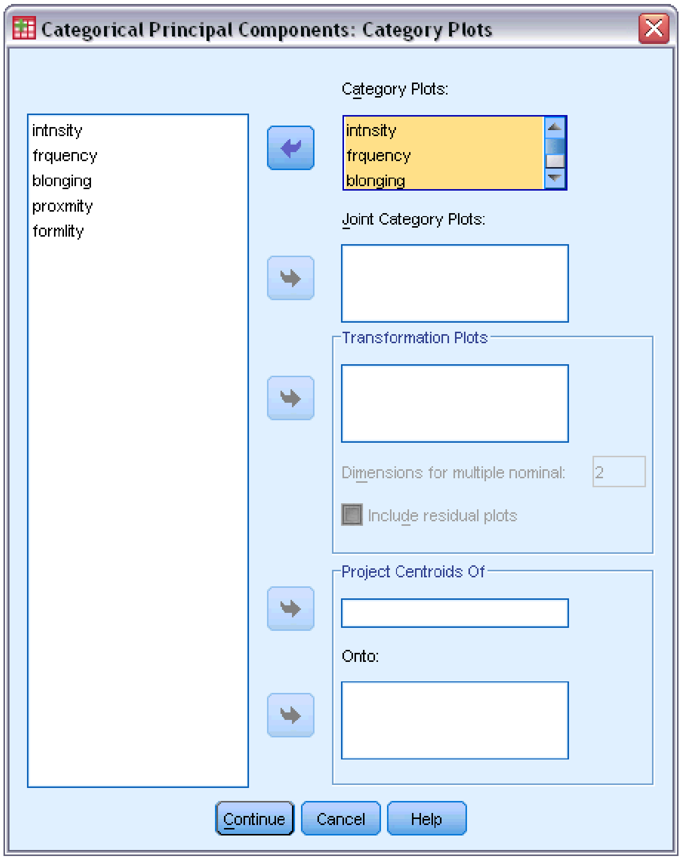
- ▶ Select Objects and variables (biplot) in the Plots group.

- ▶ Choose to label objects by Variable in the Label Objects group, and then select *cluster* as the variable to label objects by.

- ▶ Click Continue.

- ▶ Click Category in the Plots group of the Categorical Principal Components dialog.

Figure 10-6
Category Plots dialog



Categorical Principal Components: Category Plots

intnsity
frquency
blonging
proxmity
formlity

Category Plots:

intnsity
frquency
blonging

Joint Category Plots:

Dimensions for multiple nominal: 2

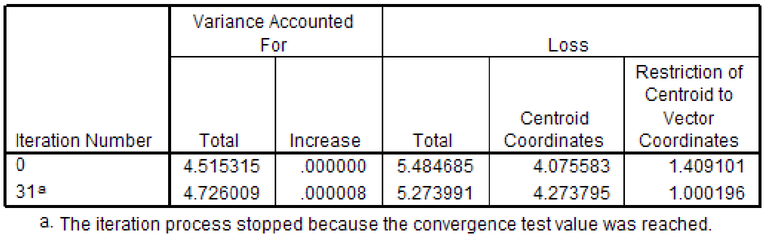
□ Include residual plots

- ▶ Choose to produce joint category plots for *intnsity (Intensity of interaction)* through *formlity (Formality of relationship)*.
- ▶ Click Continue.
- ▶ Click OK in the Categorical Principal Components dialog.

Number of Dimensions

These figures show some of the initial output for the categorical principal components analysis. After the iteration history of the algorithm, the model summary, including the eigenvalues of each dimension, is displayed. These eigenvalues are equivalent to those of classical principal components analysis. They are measures of how much variance is accounted for by each dimension.

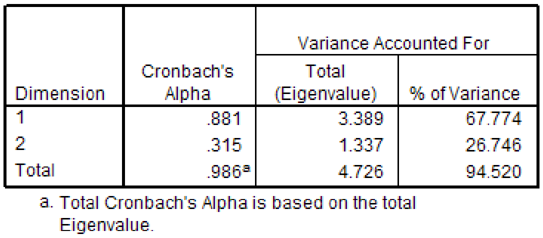
Figure 10-7
Iteration history



	Variance Accounted For		Loss		
Iteration Number	Total	Increase	Total	Centroid Coordinates	Restriction of Centroid to Vector Coordinates
0	4.515315	.000000	5.484685	4.075583	1.409101
31[a]	4.726009	.000008	5.273991	4.273795	1.000196

a. The iteration process stopped because the convergence test value was reached.

Figure 10-8
Model summary



		Variance Accounted For	
Dimension	Cronbach's Alpha	Total (Eigenvalue)	% of Variance
1	.881	3.389	67.774
2	.315	1.337	26.746
Total	.986[a]	4.726	94.520

a. Total Cronbach's Alpha is based on the total Eigenvalue.

The eigenvalues can be used as an indication of how many dimensions are needed. In this example, the default number of dimensions, 2, was used. Is this the right number? As a general rule, when all variables are either single nominal, ordinal, or numerical, the eigenvalue for a dimension should be larger than 1. Since the two-dimensional solution accounts for 94.52% of the variance, a third dimension probably would not add much more information.

For multiple nominal variables, there is no easy rule of thumb to determine the appropriate number of dimensions. If the number of variables is replaced by the total number of categories minus the number of variables, the above rule still holds. But this rule alone would probably allow more dimensions than are needed. When choosing the number of dimensions, the most useful guideline is to keep the number small enough so that meaningful interpretations are possible. The model summary table also shows Cronbach's alpha (a measure of reliability), which is maximized by the procedure.

Quantifications

For each variable, the quantifications, the vector coordinates, and the centroid coordinates for each dimension are presented. The quantifications are the values assigned to each category. The centroid coordinates are the average of the object scores of objects in the same category. The vector coordinates are the coordinates of the categories when they are required to be on a line, representing the variable in the object space. This is required for variables with the ordinal and numerical scaling level.

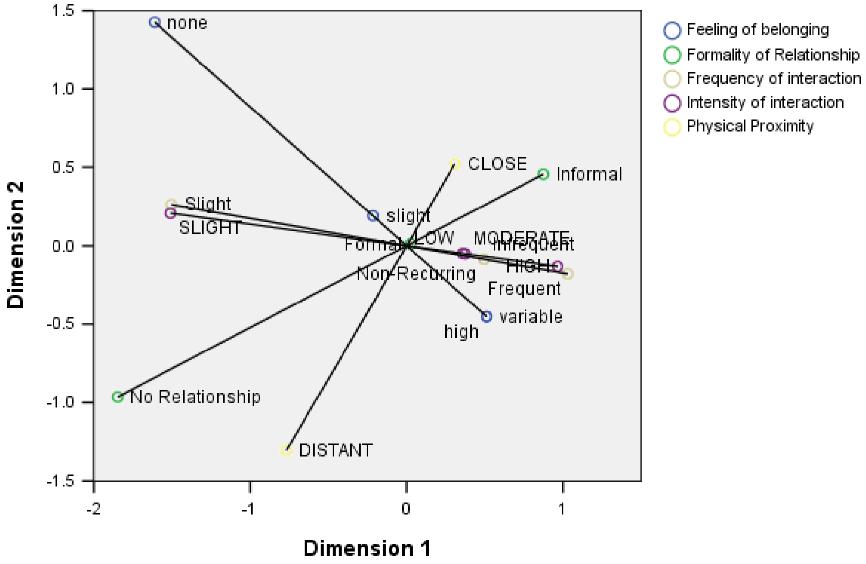
Figure 10-9
Quantifications for Intensity of interaction

Category	Frequency	Quantification	Centroid Coordinates		Vector Coordinates	
			Dimension		Dimension	
			1	2	1	2
SLIGHT	2	-1.530	-1.496	.308	-1.510	.208
LOW	2	.362	.392	.202	.358	-.049
MODERATE	1	.379	.188	-1.408	.374	-.051
HIGH	2	.978	1.010	.194	.965	-.133

Variable Principal Normalization.

Glancing at the quantifications in the joint plot of the category points, you can see that some of the categories of some variables were not clearly separated by the categorical principal components analysis as cleanly as would have been expected if the level had been truly ordinal. Variables *Intensity of interaction* and *Frequency of interaction*, for example, have equal or almost equal quantifications for their two middle categories. This kind of result might suggest trying alternative categorical principal components analyses, perhaps with some categories collapsed, or perhaps with a different level of analysis, such as (multiple) nominal.

Figure 10-10
Joint plot category points



The joint plot of category points resembles the plot for the component loadings, but it also shows where the endpoints are located that correspond to the lowest quantifications (for example, *slight* for *Intensity of interaction* and *none* for *Feeling of belonging*). The two variables measuring interaction, *Intensity of interaction* and *Frequency of interaction*, appear very close together and account for much of the variance in dimension 1. *Formality of Relationship* also appears close to *Physical Proximity*.

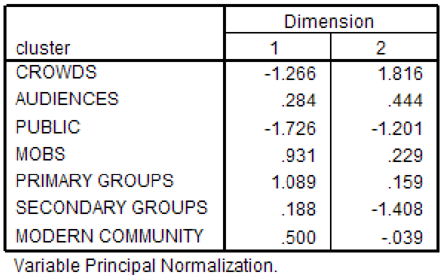
By focusing on the category points, you can see the relationships even more clearly. Not only are *Intensity of interaction* and *Frequency of interaction* close, but the directions of their scales are similar; that is, slight intensity is close to slight frequency, and frequent interaction is near high intensity of interaction. You also see that close physical proximity seems to go hand-in-hand with an informal type of relationship, and physical distance is related to no relationship.

Object Scores

You can also request a listing and plot of object scores. The plot of the object scores can be useful for detecting outliers, detecting typical groups of objects, or revealing some special patterns.

The object scores table shows the listing of object scores labeled by social group for the Guttman-Bell data. By examining the values for the object points, you can identify specific objects in the plot.

Figure 10-11
Object scores

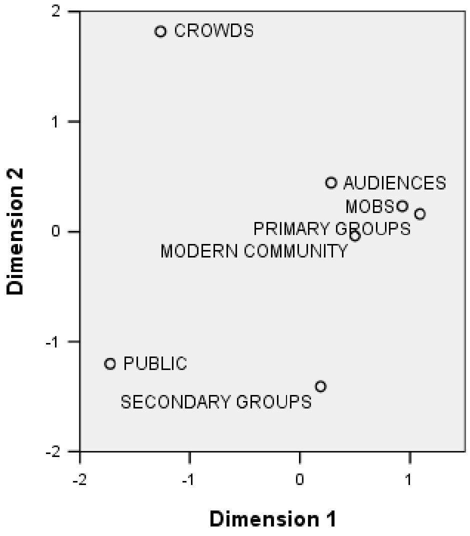


	Dimension	
cluster	1	2
CROWDS	-1.266	1.816
AUDIENCES	.284	.444
PUBLIC	-1.726	-1.201
MOBS	.931	.229
PRIMARY GROUPS	1.089	.159
SECONDARY GROUPS	.188	-1.408
MODERN COMMUNITY	.500	-.039

Variable Principal Normalization.

The first dimension appears to separate *CROWDS* and *PUBLIC*, which have relatively large negative scores, from *MOBS* and *PRIMARY GROUPS*, which have relatively large positive scores. The second dimension has three clumps: *PUBLIC* and *SECONDARY GROUPS* with large negative values, *CROWDS* with large positive values, and the other social groups in between. This is easier to see by inspecting the plot of the object scores.

Figure 10-12
Object scores plot

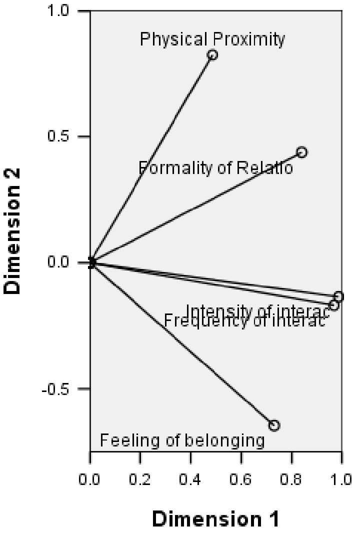


In the plot, you see *PUBLIC* and *SECONDARY GROUPS* at the bottom, *CROWDS* at the top, and the other social groups in the middle. Examining patterns among individual objects depends on the additional information available for the units of analysis. In this case, you know the classification of the objects. In other cases, you can use supplementary variables to label the objects. You can also see that the categorical principal components analysis does not separate *MOBS* from *PRIMARY GROUPS*. Although most people usually don't think of their families as mobs, on the variables used, these two groups received the same score on four of the five variables! Obviously, you might want to explore possible shortcomings of the variables and categories used. For example, high intensity of interaction and informal relationships probably mean different things to these two groups. Alternatively, you might consider a higher dimensional solution.

Component Loadings

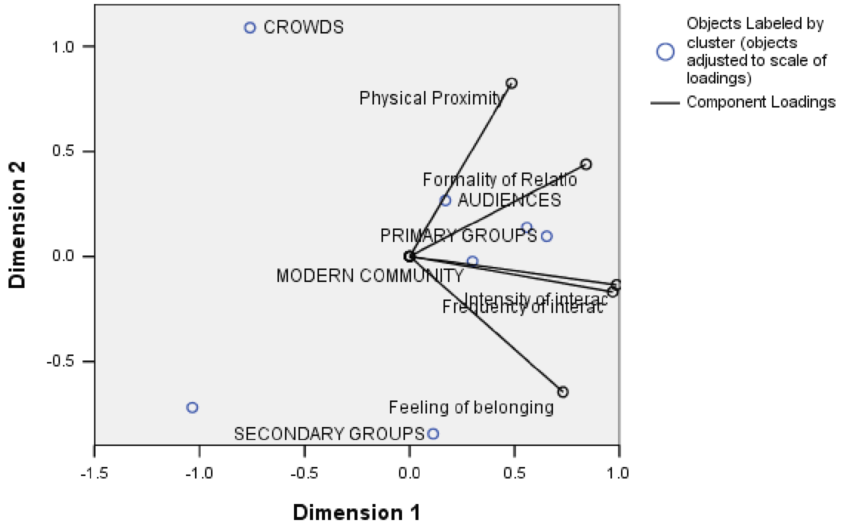
This figure shows the plot of component loadings. The vectors (lines) are relatively long, indicating again that the first two dimensions account for most of the variance of all of the quantified variables. On the first dimension, all variables have high (positive) component loadings. The second dimension is correlated mainly with the quantified variables *Feeling of belonging* and *Physical Proximity*, in opposite directions. This means that objects with a large negative score in dimension 2 will have a high score in feeling of belonging and a low score in physical proximity. The second dimension, therefore, reveals a contrast between these two variables while having little relation with the quantified variables *Intensity of interaction* and *Frequency of interaction*.

Figure 10-13
Component loadings



To examine the relation between the objects and the variables, look at the biplot of objects and component loadings. The vector of a variable points into the direction of the highest category of the variable. For example, for *Physical Proximity* and *Feeling of belonging* the highest categories are *close* and *high*, respectively. Therefore, *CROWDS* are characterized by close physical proximity and no feeling of belonging, and *SECONDARY GROUPS*, by distant physical proximity and a high feeling of belonging.

Figure 10-14
Biplot



Additional Dimensions

Increasing the number of dimensions will increase the amount of variation accounted for and may reveal differences concealed in lower dimensional solutions. As noted previously, in two dimensions *MOBS* and *PRIMARY GROUPS* cannot be separated. However, increasing the dimensionality may allow the two groups to be differentiated.

Running the Analysis

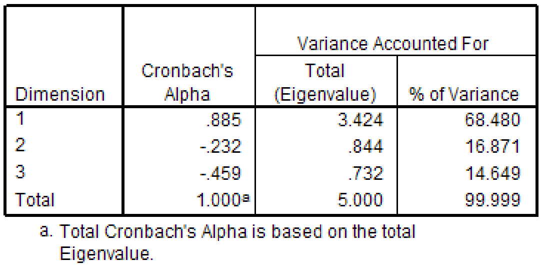
▶ To obtain a three-dimensional solution, recall the Categorical Principal Components dialog.

▶ Type 3 as the number of dimensions in the solution.

▶ Click OK in the Categorical Principal Components dialog.

Model Summary

Figure 10-15
Model summary



Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	.885	3.424	68.480
2	-.232	.844	16.871
3	-.459	.732	14.649
Total	1.000[a]	5.000	99.999

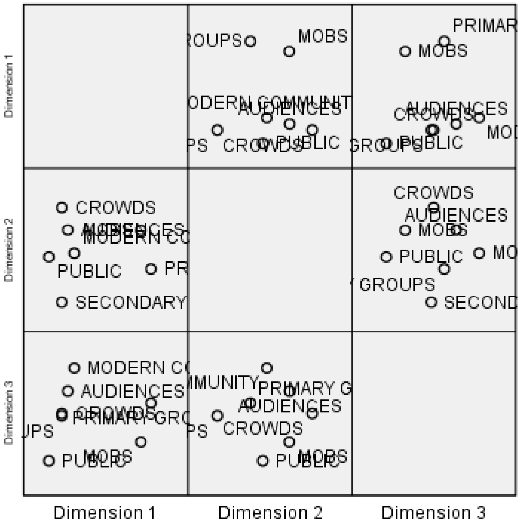
a. Total Cronbach's Alpha is based on the total Eigenvalue.

A three-dimensional solution has eigenvalues of 3.424, 0.844, and 0.732, accounting for nearly all of the variance.

Object Scores

The object scores for the three-dimensional solution are plotted in a scatterplot matrix. In a scatterplot matrix, every dimension is plotted against every other dimension in a series of two-dimensional scatterplots. Note that the first two eigenvalues in three dimensions are not equal to the eigenvalues in the two-dimensional solution; in other words, the solutions are not nested. Because the eigenvalues in dimensions 2 and 3 are now smaller than 1 (giving a Cronbach's alpha that is negative), you should prefer the two-dimensional solution. The three-dimensional solution is included for purposes of illustration.

Figure 10-16
Three-dimensional object scores scatterplot matrix



The top row of plots reveals that the first dimension separates *PRIMARY GROUPS* and *MOBS* from the other groups. Notice that the order of the objects along the vertical axis does not change in any of the plots in the top row; each of these plots employs dimension 1 as the *y* axis.

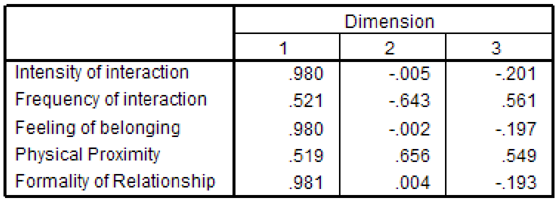
The middle row of plots allows for interpretation of dimension 2. The second dimension has changed slightly from the two-dimensional solution. Previously, the second dimension had three distinct clumps, but now the objects are more spread out along the axis.

The third dimension helps to separate *MOBS* from *PRIMARY GROUPS*, which did not occur in the two-dimensional solution.

Look more closely at the dimension 2 versus dimension 3 and dimension 1 versus dimension 2 plots. On the plane defined by dimensions 2 and 3, the objects form a rough rectangle, with *CROWDS*, *MODERN COMMUNITY*, *SECONDARY GROUPS*, and *PUBLIC* at the vertices. On this plane, *MOBS* and *PRIMARY GROUPS* appear to be convex combinations of *PUBLIC-CROWDS* and *SECONDARY GROUPS-MODERN COMMUNITY*, respectively. However, as previously mentioned, they are separated from the other groups along dimension 1. *AUDIENCES* is not separated from the other groups along dimension 1 and appears to be a combination of *CROWDS* and *MODERN COMMUNITY*.

Component Loadings

Figure 10-17
Three-dimensional component loadings



	Dimension		
	1	2	3
Intensity of interaction	.980	-.005	-.201
Frequency of interaction	.521	-.643	.561
Feeling of belonging	.980	-.002	-.197
Physical Proximity	.519	.656	.549
Formality of Relationship	.981	.004	-.193

Knowing how the objects are separated does not reveal which variables correspond to which dimensions. This is accomplished using the component loadings. The first dimension corresponds primarily to *Feeling of belonging*, *Intensity of interaction*, and *Formality of Relationship*; the second dimension separates *Frequency of interaction* and *Physical Proximity*; and the third dimension separates these from the others.

Example: Symptomatology of Eating Disorders

Eating disorders are debilitating illnesses associated with disturbances in eating behavior, severe body image distortion, and an obsession with weight that affects the mind and body simultaneously. Millions of people are affected each year, with adolescents particularly at risk. Treatments are available and most are helpful when the condition is identified early.

A health professional can attempt to diagnose an eating disorder through a psychological and medical evaluation. However, it can be difficult to assign a patient to one of several different classes of eating disorders because there is no standardized symptomatology of anorectic/bulimic behavior. Are there symptoms that clearly differentiate patients into the four groups? Which symptoms do they have in common?

In order to try to answer these questions, researchers (Van der Ham, Meulman, Van Strien, and Van Engeland, 1997) made a study of 55 adolescents with known eating disorders, as shown in the following table.

Table 10-2
Patient diagnoses

Diagnosis	Number of Patients
Anorexia nervosa	25
Anorexia with bulimia nervosa	9
Bulimia nervosa after anorexia	14
Atypical eating disorder	7
Total	55

Each patient was seen four times over four years, for a total of 220 observations. At each observation, the patients were scored for each of the 16 symptoms outlined in the following table. Symptom scores are missing for patient 71 at time 2, patient 76 at time 2, and patient 47 at time 3, leaving 217 valid observations. The data can be found in *anorectic.sav*.For more information, see the topic Sample Files in Appendix A on p. 292.

Table 10-3
Modified Morgan-Russell subscales measuring well-being

Variable name	Variable label	Lower end (score1)	Upper end (score 3 or 4)
weight	Body weight	Outside normal range	Normal
mens	Menstruation	Amenorrhea	Regular periods
fast	Restriction of food intake (fasting)	Less than 1200 calories	Normal/regular meals
binge	Binge eating	Greater than once a week	No bingeing
vomit	Vomiting	Greater than once a week	No vomiting
purge	Purging	Greater than once a week	No purging
hyper	Hyperactivity	Not able to be at rest	No hyperactivity
fami	Family relations	Poor	Good
eman	Emancipation from family	Very dependent	Adequate
frie	Friends	No good friends	Two or more good friends
school	School/employment record	Stopped school/work	Moderate to good record
satt	Sexual attitude	Inadequate	Adequate
sbeh	Sexual behavior	Inadequate	Can enjoy sex
mood	Mental state (mood)	Very depressed	Normal
preo	Preoccupation with food and weight	Complete	No preoccupation
body	Body perception	Disturbed	Normal

Principal components analysis is ideal for this situation, since the purpose of the study is to ascertain the relationships between symptoms and the different classes of eating disorders. Moreover, categorical principal components analysis is likely to be more useful than classical principal components analysis because the symptoms are scored on an ordinal scale.

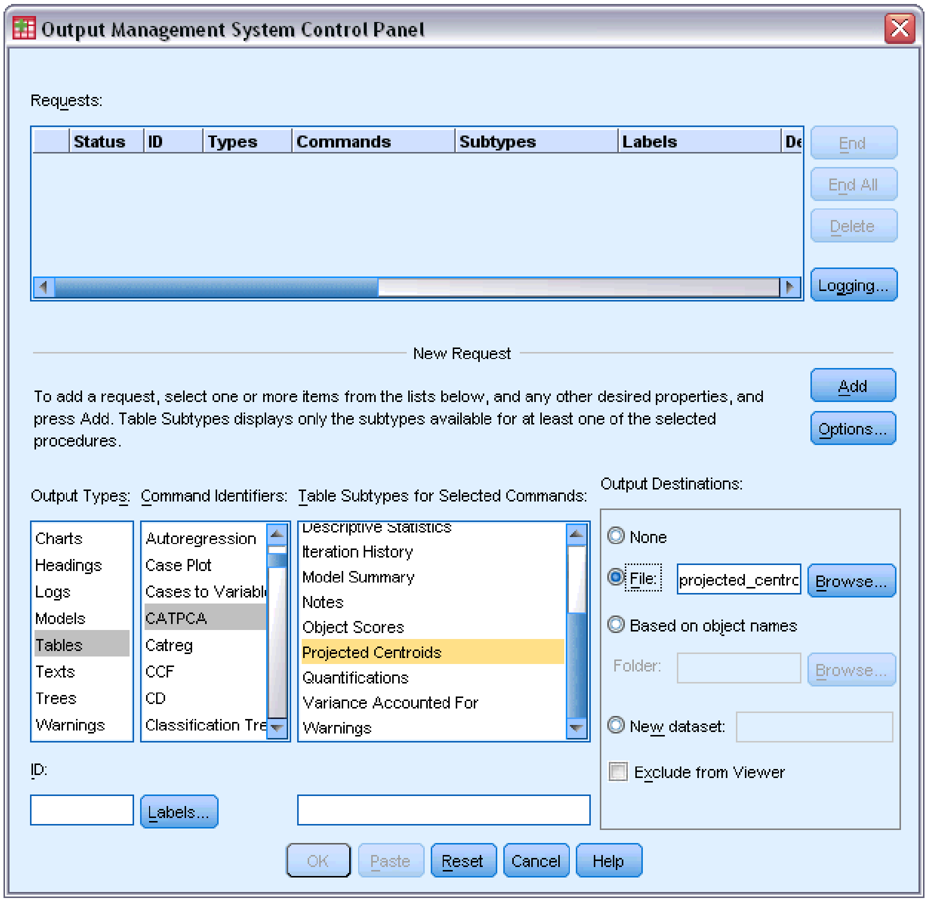
Running the Analysis

In order to properly examine the structure of the course of illness for each diagnosis, you will want to make the results of the projected centroids table available as data for scatterplots. You can accomplish this using the Output Management System.

- ▶ To begin an OMS request, from the menus choose:

  Utilities > OMS Control Panel...

Figure 10-18
Output Management System Control Panel



Output Management System Control Panel

Requests:

	Status	ID	Types	Commands	Subtypes	Labels	De

New Request

To add a request, select one or more items from the lists below, and any other desired properties, and press Add. Table Subtypes displays only the subtypes available for at least one of the selected procedures.

Output Types: Tables

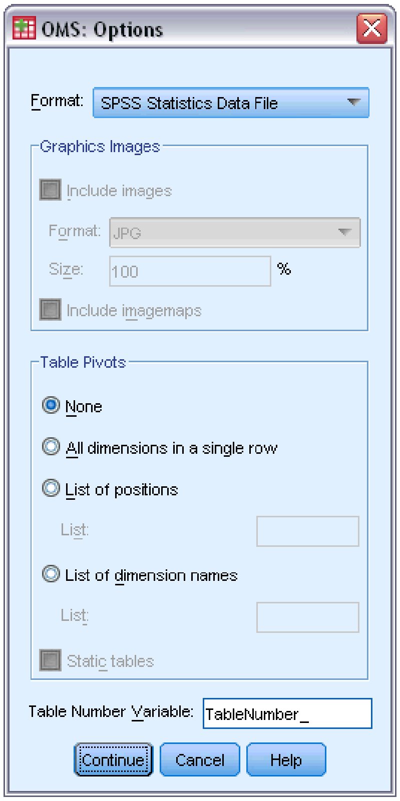
Command Identifiers: CATPCA

Table Subtypes for Selected Commands: Projected Centroids

Output Destinations: ● File: projected_centrc

- ▶ Select Tables as the output type.
- ▶ Select CATPCA as the command.
- ▶ Select Projected Centroids as the table type.
- ▶ Select File in the Output Destinations group and type projected_centroids.sav as the filename.
- ▶ Click Options.

Figure 10-19
Options dialog



OMS: Options

Format: SPSS Statistics Data File

Graphics Images

□ Include images

Format: JPG

Size: 100 %

□ Include imagemaps

Table Pivots

● None

○ All dimensions in a single row

○ List of positions

List:

○ List of dimension names

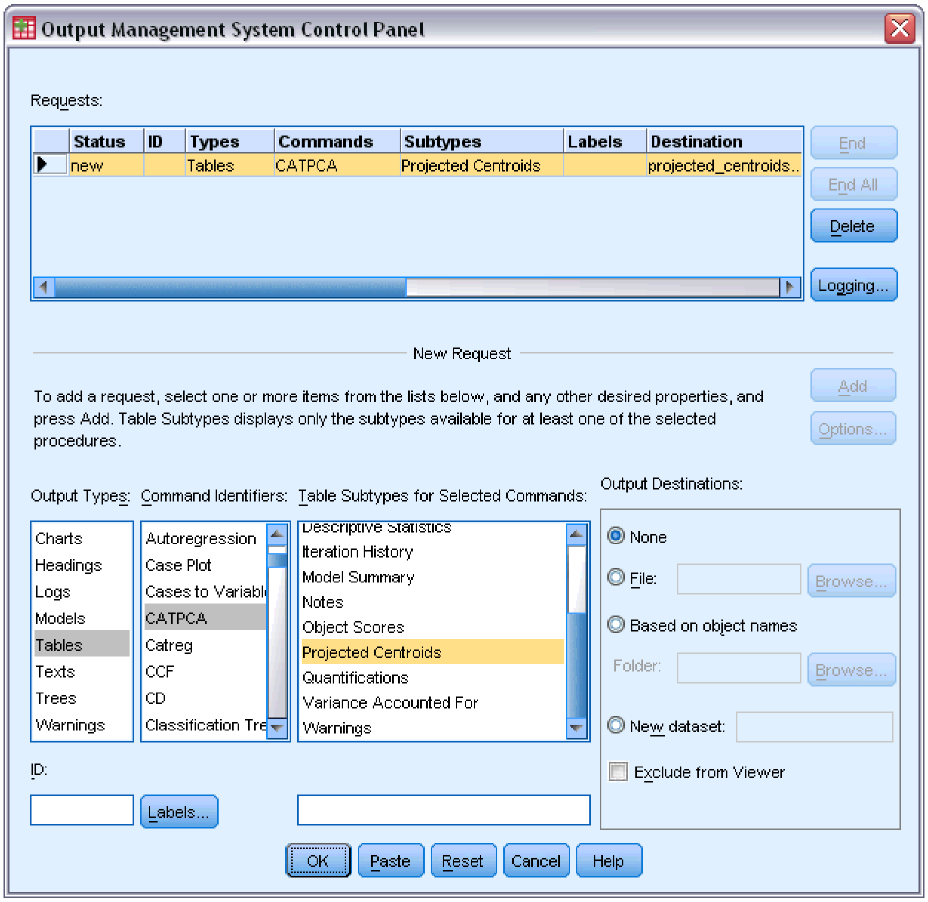
List:

□ Static tables

Table Number Variable: TableNumber_

- Select IBM® SPSS® Statistics Data File as the output format.
- Type TableNumber_1 as the table number variable.
- Click Continue.

Figure 10-20
Output Management System Control Panel



Requests:

	Status	ID	Types	Commands	Subtypes	Labels	Destination
▶	new		Tables	CATPCA	Projected Centroids		projected_centroids...

To add a request, select one or more items from the lists below, and any other desired properties, and press Add. Table Subtypes displays only the subtypes available for at least one of the selected procedures.

Output Types: Tables

Command Identifiers: CATPCA

Table Subtypes for Selected Commands: Projected Centroids

Output Destinations: ● None ○ File: ○ Based on object names ○ New dataset:

□ Exclude from Viewer

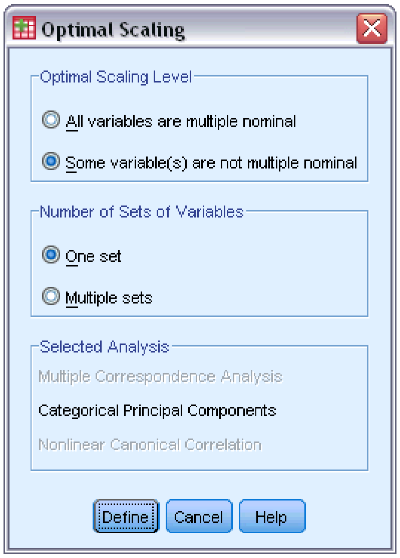
- ▶ Click Add.

- ▶ Click OK, and then click OK to confirm the OMS session.

  The Output Management System is now set to write the results of the projected centroids table to the file *projected_centroids.sav*.

- ▶ To produce categorical principal components output for this dataset, from the menus choose:
  Analyze > Dimension Reduction > Optimal Scaling...

Figure 10-21
Optimal Scaling dialog



Optimal Scaling

Optimal Scaling Level

○ All variables are multiple nominal

● Some variable(s) are not multiple nominal

Number of Sets of Variables

● One set

○ Multiple sets

Selected Analysis

Multiple Correspondence Analysis

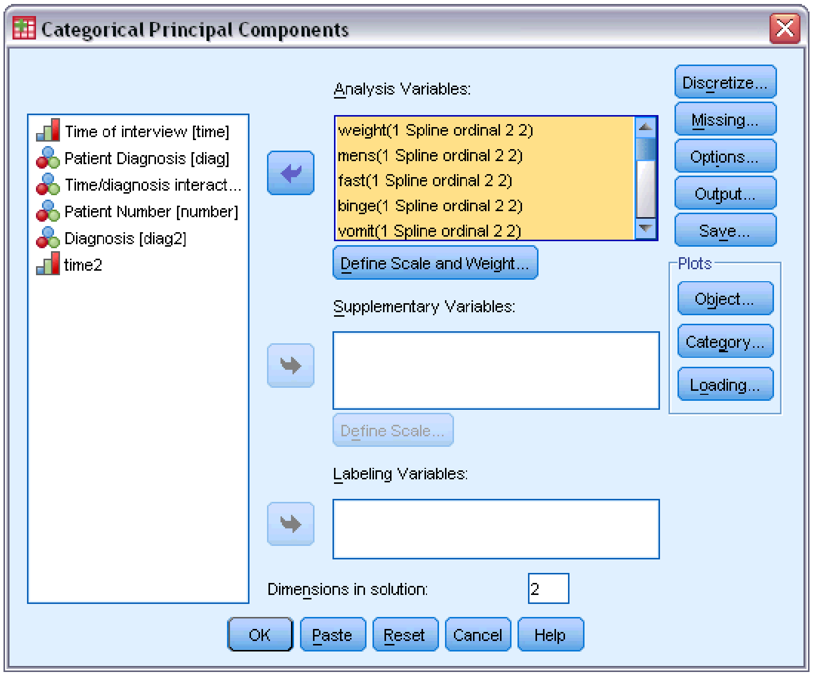
Categorical Principal Components

Nonlinear Canonical Correlation

▶ Select Some variable(s) not multiple nominal in the Optimal Scaling Level group.

▶ Click Define.

Figure 10-22
Categorical Principal Components dialog



Categorical Principal Components

Time of interview [time]
Patient Diagnosis [diag]
Time/diagnosis interact...
Patient Number [number]
Diagnosis [diag2]
time2

Analysis Variables:

weight(1 Spline ordinal 2 2)
mens(1 Spline ordinal 2 2)
fast(1 Spline ordinal 2 2)
binge(1 Spline ordinal 2 2)
vomit(1 Spline ordinal 2 2)

Supplementary Variables:

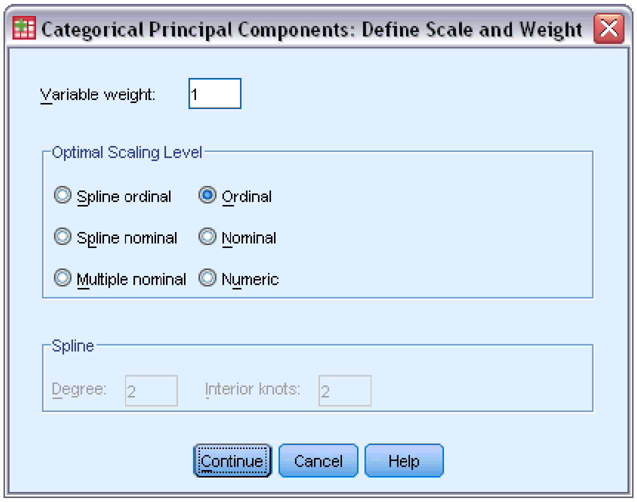
Labeling Variables:

Dimensions in solution: 2

▶ Select *Body weight* through *Body perception* as analysis variables.

▶ Click Define Scale and Weight.

Figure 10-23
Define Scale and Weight dialog



Variable weight: 1

Optimal Scaling Level

○ Spline ordinal ● Ordinal
○ Spline nominal ○ Nominal
○ Multiple nominal ○ Numeric

Spline

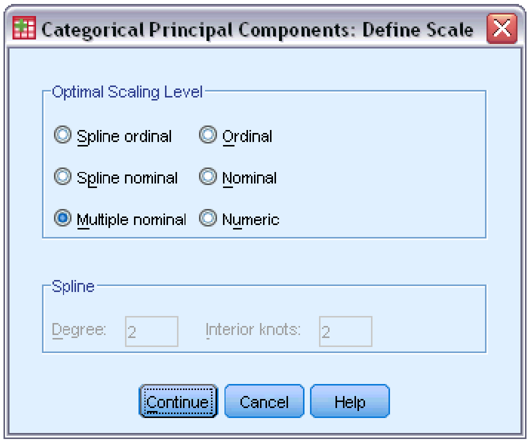
Degree: 2 Interior knots: 2

▶ Select Ordinal as the optimal scaling level.

▶ Click Continue.

▶ Select *Time/diagnosis interaction* as a supplementary variable and click Define Scale in the Categorical Principal Components dialog.

Figure 10-24
Define Scale dialog



Optimal Scaling Level

○ Spline ordinal ○ Ordinal
○ Spline nominal ○ Nominal
● Multiple nominal ○ Numeric

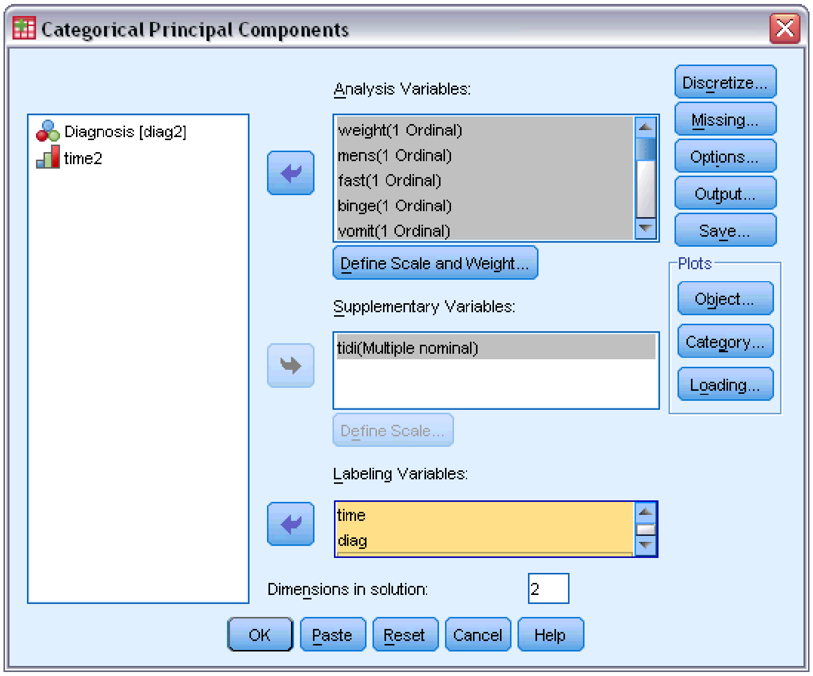
Spline

Degree: 2 Interior knots: 2

▶ Select Multiple nominal as the optimal scaling level.

▶ Click Continue.

Figure 10-25
Categorical Principal Components dialog



Categorical Principal Components

Diagnosis [diag2]
time2

Analysis Variables:
weight(1 Ordinal)
mens(1 Ordinal)
fast(1 Ordinal)
binge(1 Ordinal)
vomit(1 Ordinal)

Supplementary Variables:
tidi(Multiple nominal)

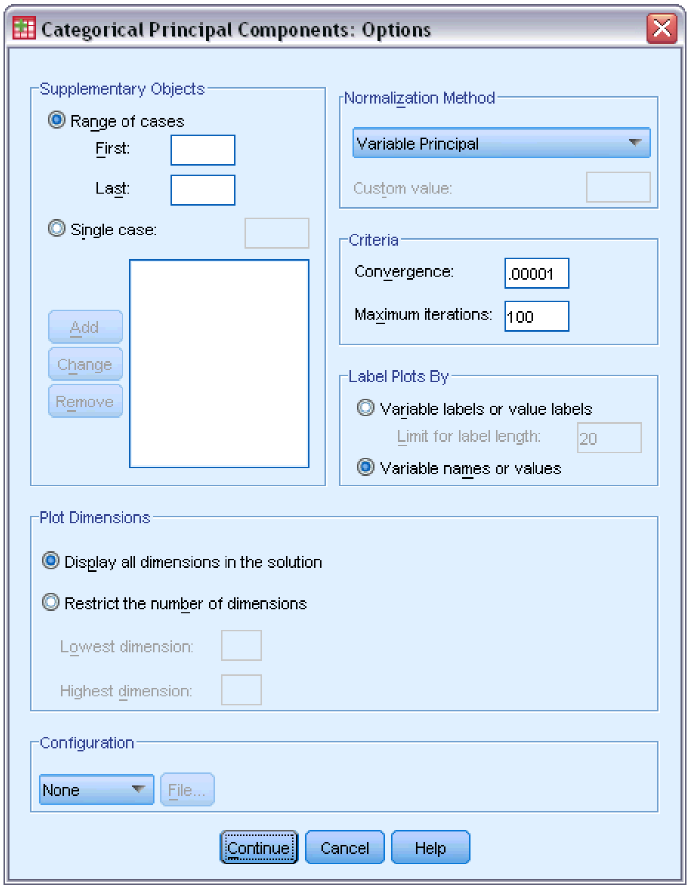
Labeling Variables:
time
diag

Dimensions in solution: 2

- ▶ Select *Time of interview* through *Patient number* as labeling variables.

- ▶ Click Options.

Figure 10-26
Options dialog



Categorical Principal Components: Options

Supplementary Objects
- ● Range of cases
  - First:
  - Last:
- ○ Single case:

Normalization Method: Variable Principal

Custom value:

Criteria
- Convergence: .00001
- Maximum iterations: 100

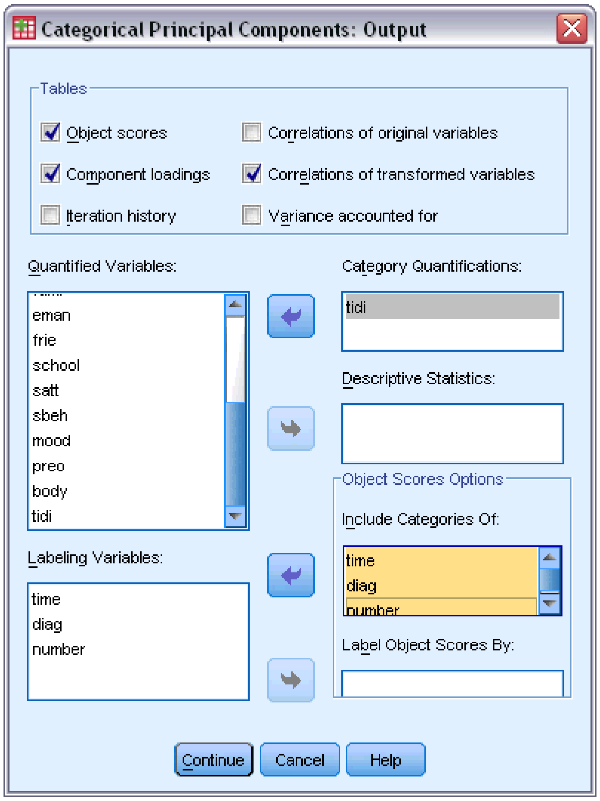
Label Plots By
- ○ Variable labels or value labels
  - Limit for label length: 20
- ● Variable names or values

Plot Dimensions
- ● Display all dimensions in the solution
- ○ Restrict the number of dimensions
  - Lowest dimension:
  - Highest dimension:

Configuration: None

- ▶ Choose to label plots by Variable names or values.
- ▶ Click Continue.
- ▶ Click Output in the Categorical Principal Components dialog.

Figure 10-27
Output dialog



Categorical Principal Components: Output

Tables

☑ Object scores
☑ Component loadings
□ Iteration history
□ Correlations of original variables
☑ Correlations of transformed variables
□ Variance accounted for

Quantified Variables:
eman
frie
school
satt
sbeh
mood
preo
body
tidi

Category Quantifications:
tidi

Descriptive Statistics:

Object Scores Options

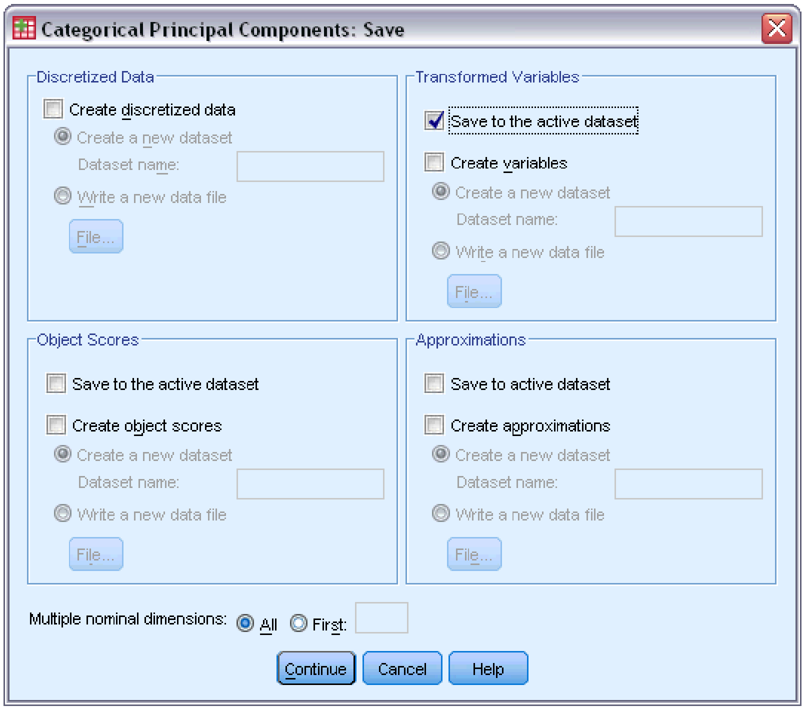
Include Categories Of:
time
diag
number

Label Object Scores By:

Labeling Variables:
time
diag
number

- Select Object scores in the Tables group.
- Request category quantifications for *tidi*.
- Choose to include categories of *time*, *diag*, and *number*.
- Click Continue.
- Click Save in the Categorical Principal Components dialog.

Figure 10-28
Save dialog



Categorical Principal Components: Save

Discretized Data
☐ Create discretized data
● Create a new dataset
Dataset name:
○ Write a new data file

Transformed Variables
☑ Save to the active dataset
☐ Create variables
● Create a new dataset
Dataset name:
○ Write a new data file

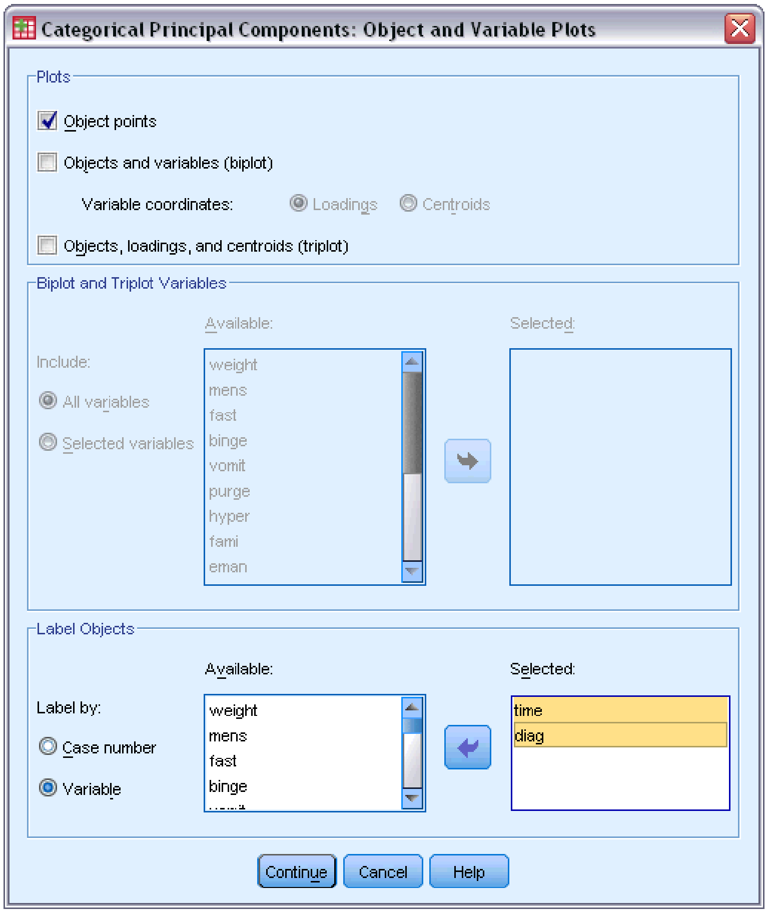
Object Scores
☐ Save to the active dataset
☐ Create object scores
● Create a new dataset
Dataset name:
○ Write a new data file

Approximations
☐ Save to active dataset
☐ Create approximations
● Create a new dataset
Dataset name:
○ Write a new data file

Multiple nominal dimensions: ● All ○ First:

- ▶ In the Transformed Variables group, select Save to the active dataset.
- ▶ Click Continue.
- ▶ Click Object in the Categorical Principal Components dialog.

Figure 10-29
Object and Variable Plots dialog



Categorical Principal Components: Object and Variable Plots

Plots

☑ Object points

□ Objects and variables (biplot)

Variable coordinates: ● Loadings ○ Centroids

□ Objects, loadings, and centroids (triplot)

Biplot and Triplot Variables

Include:

● All variables

○ Selected variables

Available: weight, mens, fast, binge, vomit, purge, hyper, fami, eman

Selected:

Label Objects

Label by:

○ Case number

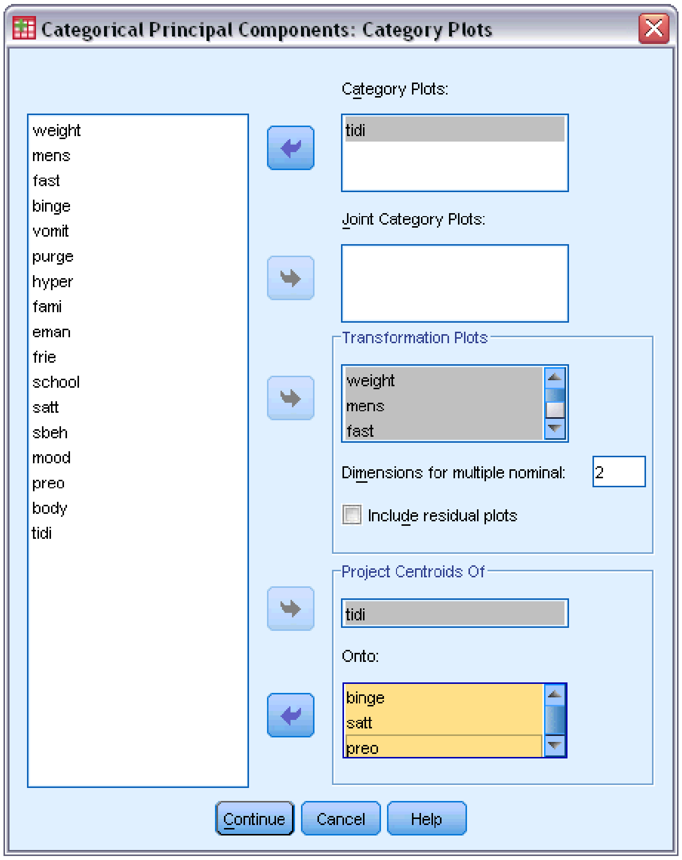
● Variable

Available: weight, mens, fast, binge

Selected: time, diag

- ▶ Choose to label objects by Variable.
- ▶ Select *time* and *diag* as the variables to label objects by.
- ▶ Click Continue.
- ▶ Click Category in the Categorical Principal Components dialog.

Figure 10-30
Category Plots dialog



Categorical Principal Components: Category Plots

weight
mens
fast
binge
vomit
purge
hyper
fami
eman
frie
school
satt
sbeh
mood
preo
body
tidi

Category Plots: tidi

Joint Category Plots:

Transformation Plots: weight, mens, fast

Dimensions for multiple nominal: 2

□ Include residual plots

Project Centroids Of: tidi

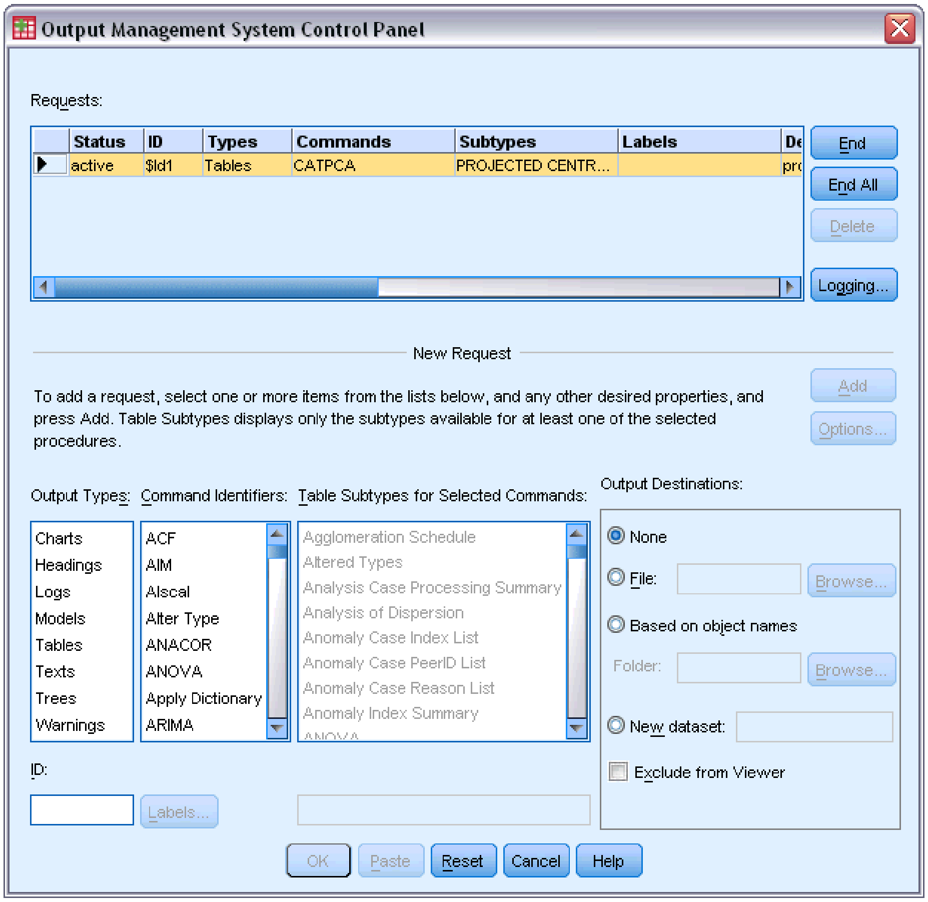
Onto: binge, satt, preo

- ▶ Request category plots for *tidi*.
- ▶ Request transformation plots for *weight* through *body*.
- ▶ Choose to project centroids of *tidi* onto *binge*, *satt*, and *preo*.
- ▶ Click Continue.
- ▶ Click OK in the Categorical Principal Components dialog.

The procedure results in scores for the subjects (with mean 0 and unit variance) and quantifications of the categories that maximize the mean squared correlation of the subject scores and the transformed variables. In the present analysis, the category quantifications were constrained to reflect the ordinal information.

Finally, to write the projected centroids table information to *projected_centroids.sav*, you need to end the OMS request. Recall the OMS Control Panel.

Figure 10-31
Output Management System Control Panel



Requests:

	Status	ID	Types	Commands	Subtypes	Labels	De
▶	active	$Id1	Tables	CATPCA	PROJECTED CENTR...		pro

To add a request, select one or more items from the lists below, and any other desired properties, and press Add. Table Subtypes displays only the subtypes available for at least one of the selected procedures.

Output Destinations: ● None

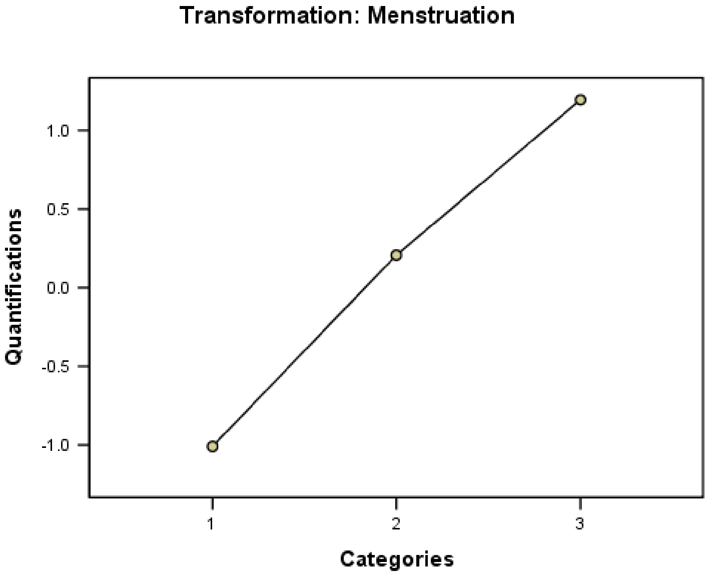
▶ Click End.

▶ Click OK, and then click OK to confirm.

Transformation Plots

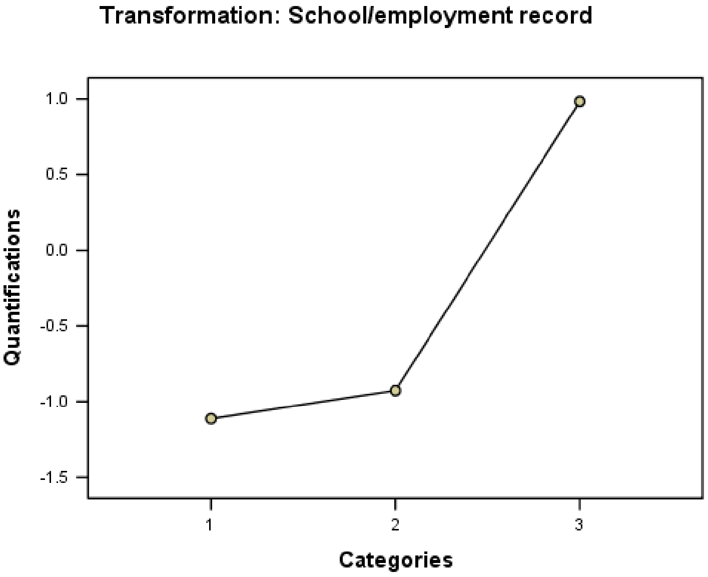
The transformation plots display the original category number on the horizontal axes; the vertical axes give the optimal quantifications.

Figure 10-32
Transformation plot for menstruation



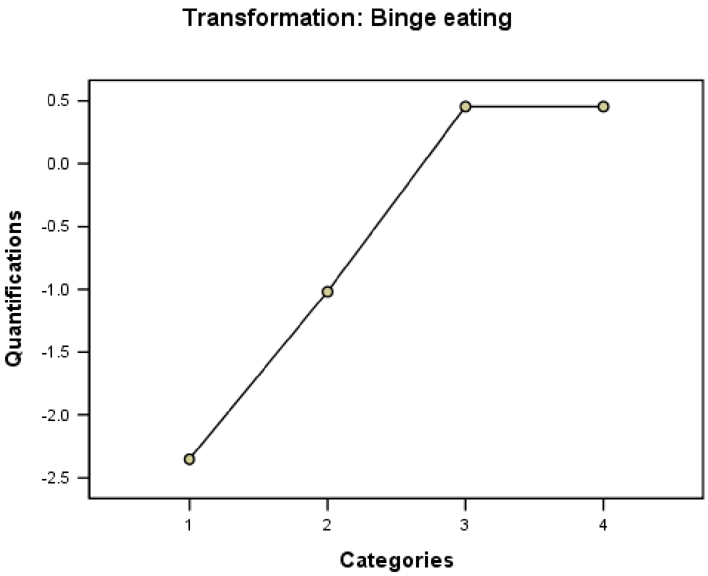
Some variables, like *Menstruation*, obtained nearly linear transformations, so in this analysis you may interpret them as numerical.

Figure 10-33
Transformation plot for School/employment record



The quantifications for other variables like *School/employment record* did not obtain linear transformations and should be interpreted at the ordinal scaling level. The difference between the second and third categories is much more important than that between the first and second categories.

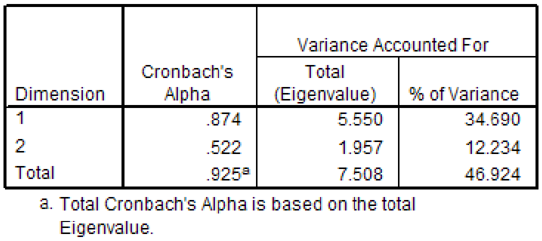
Figure 10-34
Transformation plot for Binge eating



An interesting case arises in the quantifications for *Binge eating*. The transformation obtained is linear for categories 1 through 3, but the quantified values for categories 3 and 4 are equal. This result shows that scores of 3 and 4 do not differentiate between patients and suggests that you could use the numerical scaling level in a two-component solution by recoding 4's as 3's.

Model Summary

Figure 10-35
Model summary



		Variance Accounted For	
Dimension	Cronbach's Alpha	Total (Eigenvalue)	% of Variance
1	.874	5.550	34.690
2	.522	1.957	12.234
Total	.925[a]	7.508	46.924

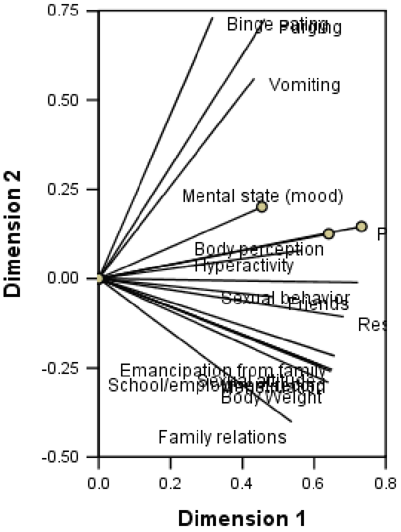
[a] Total Cronbach's Alpha is based on the total Eigenvalue.

To see how well your model fits the data, look at the model summary. About 47% of the total variance is explained by the two-component model, 35% by the first dimension and 12% by the second. So, almost half of the variability on the individual objects level is explained by the two-component model.

Component Loadings

To begin to interpret the two dimensions of your solution, look at the component loadings. All variables have a positive component loading in the first dimension, which means that there is a common factor that correlates positively with all of the variables.

Figure 10-36
Component loadings plot



The second dimension separates the variables. The variables *Binge eating*, *Vomiting*, and *Purging* form a bundle having large positive loadings in the second dimension. These symptoms are typically considered to be representative of bulimic behavior.

The variables *Emancipation from family*, *School/employment record*, *Sexual attitude*, *Body weight*, and *Menstruation* form another bundle, and you can include *Restriction of food intake (fasting)* and *Family relations* in this bundle, because their vectors are close to the main cluster, and these variables are considered to be anorectic symptoms (fasting, weight, menstruation) or are psychosocial in nature (emancipation, school/work record, sexual attitude, family relations). The vectors of this bundle are orthogonal (perpendicular) to the vectors of binge, vomit, and purge, which means that this set of variables is uncorrelated with the set of bulimic variables.

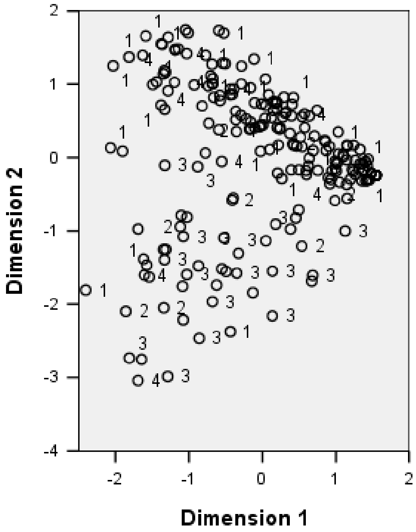
The variables *Friends*, *Mental state (mood)*, and *Hyperactivity* do not appear to fit very well into the solution. You can see this in the plot by observing the lengths of each vector. The length of a given variable's vector corresponds to its fit, and these variables have the shortest vectors. Based on a two-component solution, you would probably drop these variables from a proposed symptomatology for eating disorders. They may, however, fit better in a higher dimensional solution.

The variables *Sexual behavior*, *Preoccupation with food and weight*, and *Body perception* form another theoretic group of symptoms, pertaining to how the patient experiences his or her body. While correlated with the two orthogonal bundles of variables, these variables have fairly long vectors and are strongly associated with the first dimension and therefore may provide some useful information about the "common" factor.

Object Scores

The following figure shows a plot of the object scores, in which the subjects are labeled with their diagnosis category.

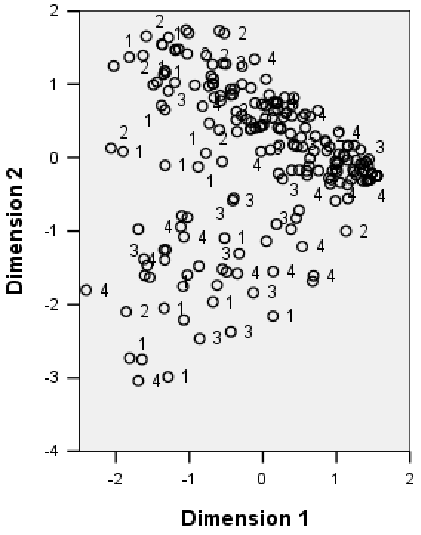
Figure 10-37
Object scores plot labeled by diagnosis



This plot does not help to interpret the first dimension, because patients are not separated by diagnosis along it. However, there is some information about the second dimension. Anorexia subjects (1) and patients with atypical eating disorder (4) form a group, located above subjects with some form of bulimia (2 and 3). Thus, the second dimension separates bulimic patients from others, as you have also seen in the previous section (the variables in the bulimic bundle have large positive component loadings in the second dimension). This makes sense, given that the component loadings of the symptoms that are traditionally associated with bulimia have large values in the second dimension.

This figure shows a plot of the object scores, in which the subjects are labeled with their time of diagnosis.

Figure 10-38
Object scores labeled by time of interview



Labeling the object scores by time reveals that the first dimension has a relation to time because there seems to be a progression of times of diagnosis from the 1's mostly to the left and others to the right. Note that you can connect the time points in this plot by saving the object scores and creating a scatterplot using the dimension 1 scores on the *x* axis, the dimension 2 scores on the *y* axis, and setting the markers using the patient numbers.

Comparing the object scores plot labeled by time with the one labeled by diagnosis can give you some insight into unusual objects. For example, in the plot labeled by time, there is a patient whose diagnosis at time 4 lies to the left of all other points in the plot. This is unusual because the general trend of the points is for the later times to lie further to the right. Interestingly, this point that seems out of place in time also has an unusual diagnosis, in that the patient is an anorectic whose scores place the patient in the bulimic cluster. By looking in the table of object scores, you find that this is patient 43, diagnosed with anorexia nervosa, whose object scores are shown in the following table.

Table 10-4
Object scores for patient 43

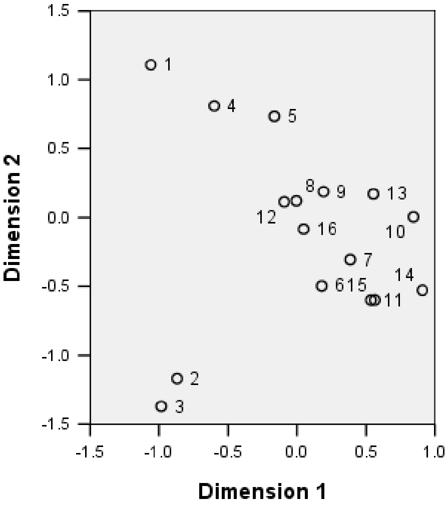
Time	Dimension 1	Dimension 2
1	–2.031	1.250
2	–2.067	0.131
3	–1.575	–1.467
4	–2.405	–1.807

The patient's scores at time 1 are prototypical for anorectics, with the large negative score in dimension 1 corresponding to poor body image and the positive score in dimension 2 corresponding to anorectic symptoms or poor psychosocial behavior. However, unlike the majority of patients, there is little or no progress in dimension 1. In dimension 2, there is apparently some progress toward “normal” (around 0, between anorectic and bulimic behavior), but then the patient shifts to exhibit bulimic symptoms.

Examining the Structure of the Course of Illness

To find out more about how the two dimensions were related to the four diagnosis categories and the four time points, a supplementary variable *Time/diagnosis interaction* was created by a cross-classification of the four categories of *Patient diagnosis* and the four categories of *Time of interview*. Thus, *Time/diagnosis interaction* has 16 categories, where the first category indicates the anorexia nervosa patients at their first visit. The fifth category indicates the anorexia nervosa patients at time point 2, and so on, with the sixteenth category indicating the atypical eating disorder patients at time point 4. The use of the supplementary variable *Time/diagnosis interaction* allows for the study of the courses of illness for the different groups over time. The variable was given a multiple nominal scaling level, and the category points are displayed in the following figure.

Figure 10-39
Category points for time/diagnosis interaction

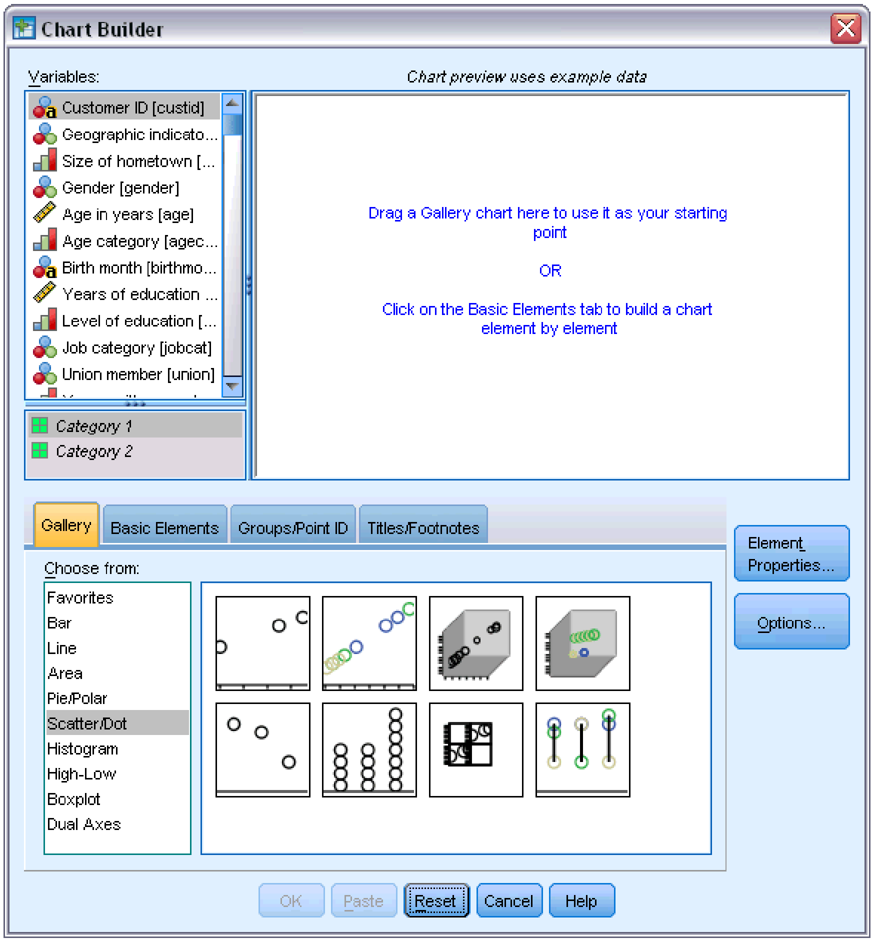


Some of the structure is apparent from this plot: the diagnosis categories at time point 1 clearly separate anorexia nervosa and atypical eating disorder from anorexia nervosa with bulimia nervosa and bulimia nervosa after anorexia nervosa in the second dimension. After that, it’s a little more difficult to see the patterns.

However, you can make the patterns more easily visible by creating a scatterplot based on the quantifications. To do this, from the menus choose:

Graphs > Chart Builder...

Figure 10-40
Scatter/Dot gallery



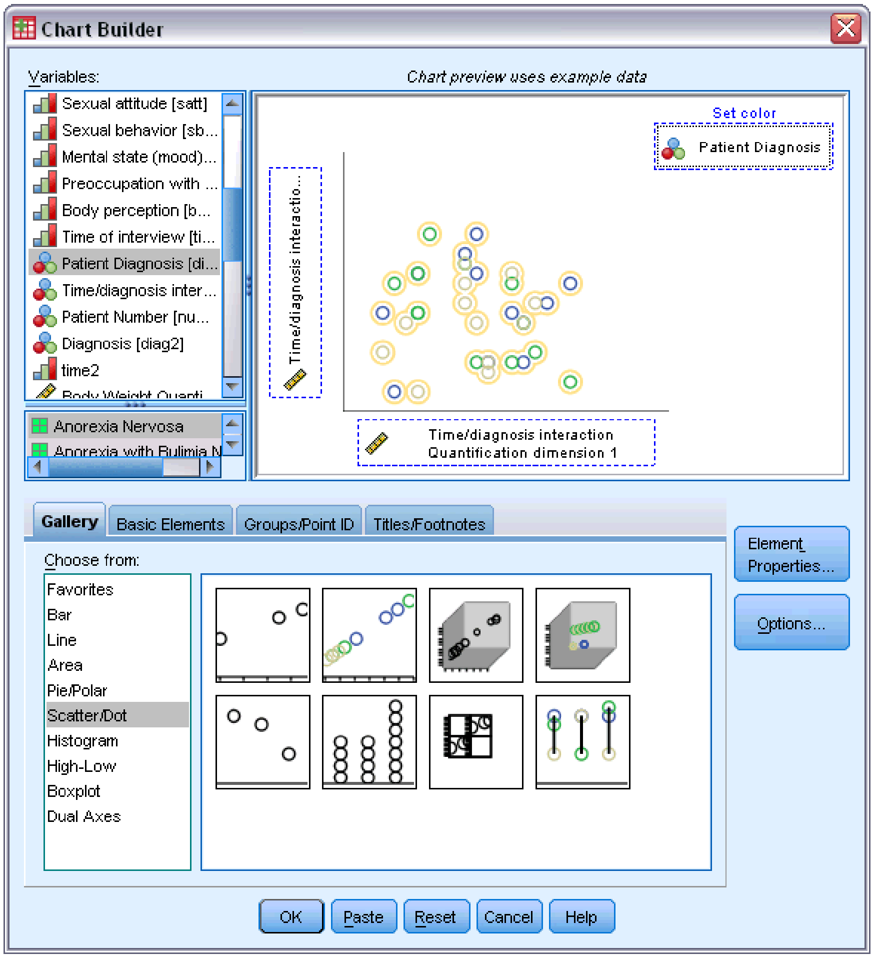
Drag a Gallery chart here to use it as your starting point

OR

Click on the Basic Elements tab to build a chart element by element

▶ Select the Scatter/Dot gallery and choose Grouped Scatter.

Figure 10-41
Chart Builder



- ▶ Select *Time/diagnosis interaction Quantification dimension 2* as the *y*-axis variable and *Time/diagnosis interaction Quantification dimension 1* as the *x*-axis variable.

- ▶ Choose to set color by *Patient Diagnosis*.

- ▶ Click OK.

Figure 10-42
Structures of the courses of illness

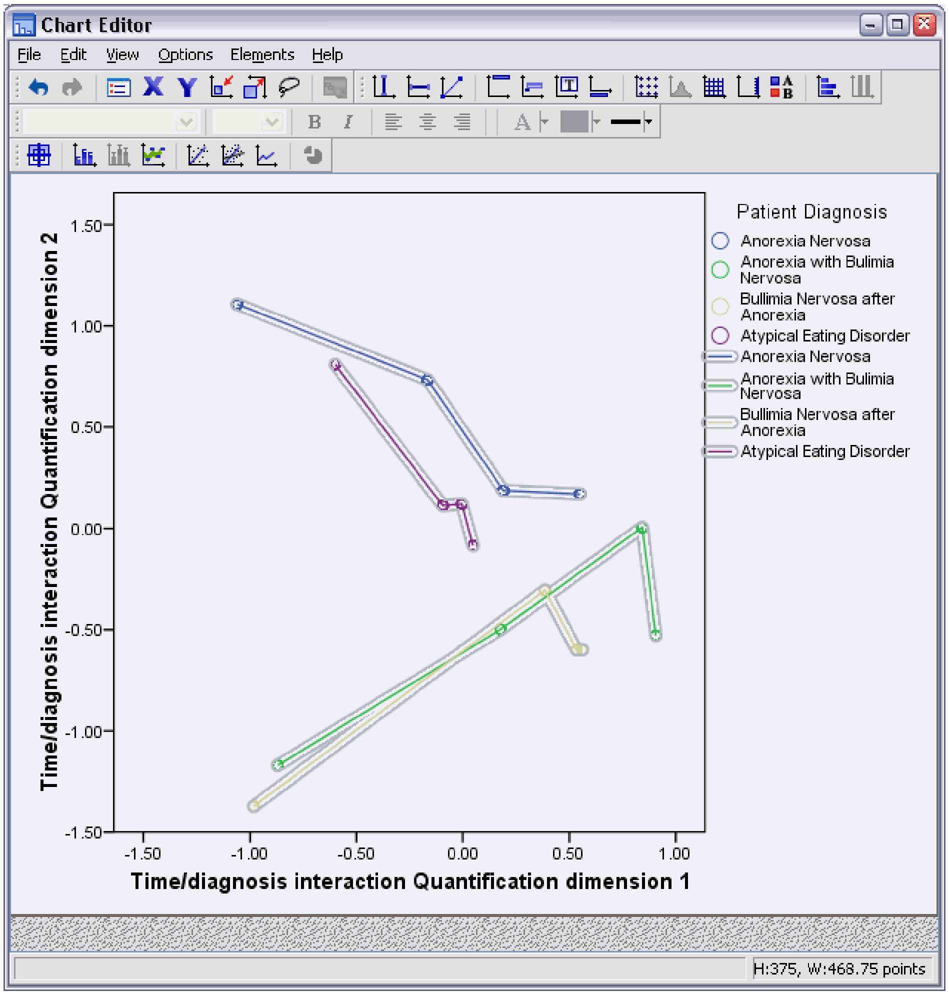
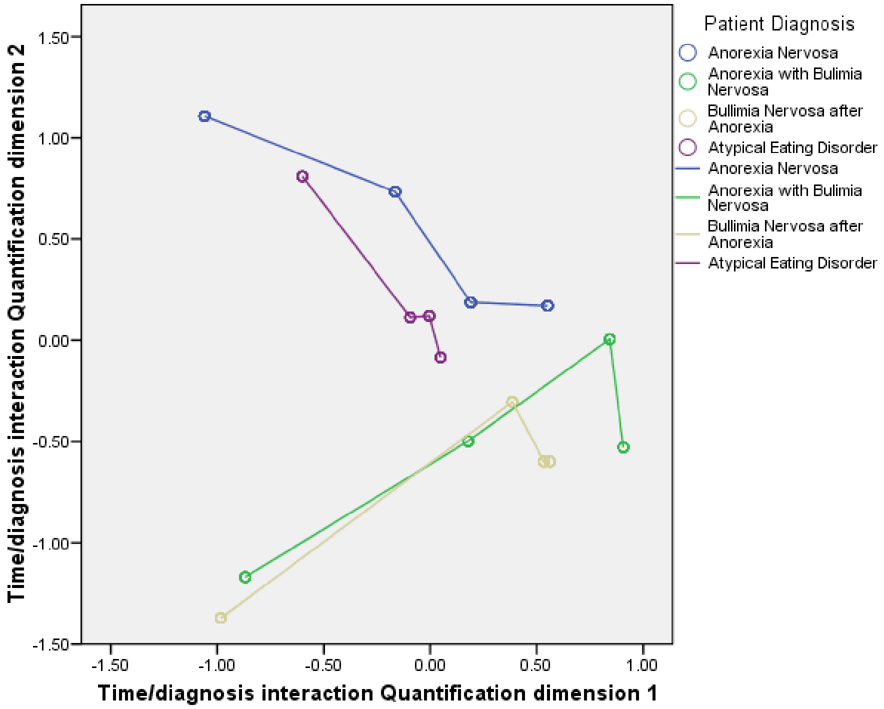


Chart Editor
File Edit View Options Elements Help
Patient Diagnosis
Anorexia Nervosa
Anorexia with Bulimia Nervosa
Bullimia Nervosa after Anorexia
Atypical Eating Disorder
Anorexia Nervosa
Anorexia with Bulimia Nervosa
Bullimia Nervosa after Anorexia
Atypical Eating Disorder
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
Time/diagnosis interaction Quantification dimension 2
-1.50
-1.00
-0.50
0.00
0.50
1.00
Time/diagnosis interaction Quantification dimension 1
H:375, W:468.75 points

- Then, to connect the points, double-click on the graph, and click the Add interpolation line tool in the Chart Editor.

- Close the Chart Editor.

Figure 10-43
Structures of the courses of illness



By connecting the category points for each diagnostic category across time, the patterns immediately suggest that the first dimension is related to time and the second, to diagnosis, as you previously determined from the object scores plots.

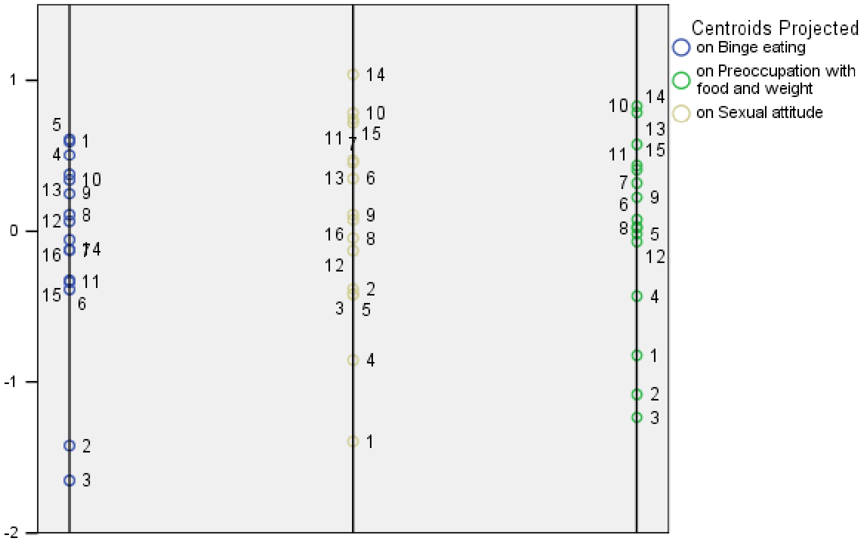
However, this plot further shows that, over time, the illnesses tend to become more alike. Moreover, for all groups, the progress is greatest between time points 1 and 2; the anorectic patients show some more progress from 2 to 3, but the other groups show little progress.

Differential Development for Selected Variables

One variable from each bundle of symptoms identified by the component loadings was selected as "representative" of the bundle. Binge eating was selected from the bulimic bundle; sexual attitude, from the anorectic/psychosocial bundle; and body preoccupation, from the third bundle.

In order to examine the possible differential courses of illness, the projections of *Time/diagnosis interaction* on *Binge eating*, *Sexual attitude*, and *Preoccupation with food and weight* were computed and plotted in the following figure.

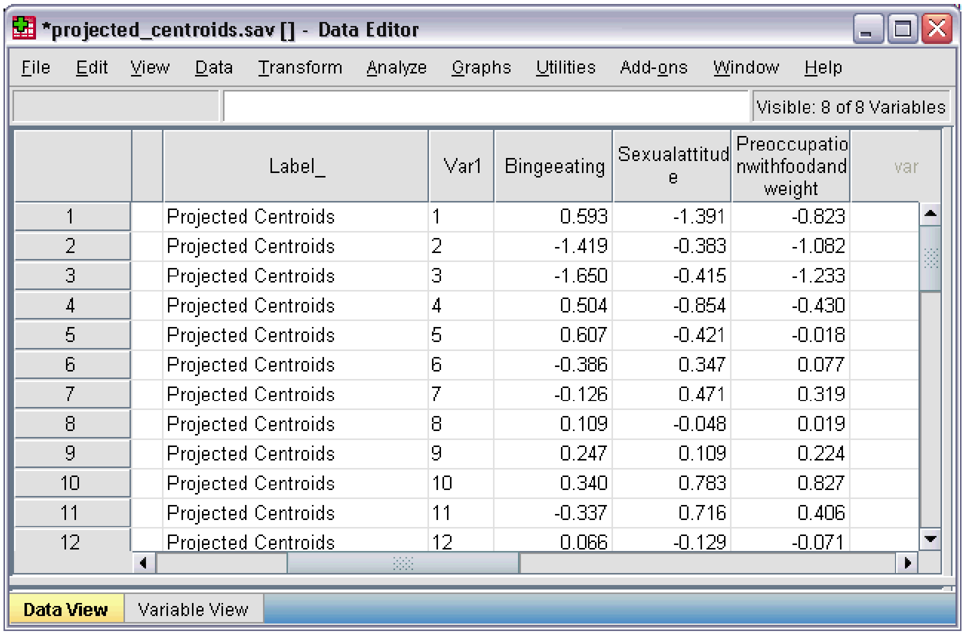
Figure 10-44
Projected centroids of Time/diagnosis interaction on Binge eating, Sexual attitude, and Preoccupation with food and weight



This plot shows that at the first time point, the symptom binge eating separates bulimic patients (2 and 3) from others (1 and 4); sexual attitude separates anorectic and atypical patients (1 and 4) from others (2 and 3); and body preoccupation does not really separate the patients. In many applications, this plot would be sufficient to describe the relationship between the symptoms and diagnosis, but because of the complication of multiple time points, the picture becomes muddled.

In order to view these projections over time, you need to be able to plot the contents of the projected centroids table. This is made possible by the OMS request that saved this information to *projected_centroids.sav*.

Figure 10-45
Projected_centroids.sav



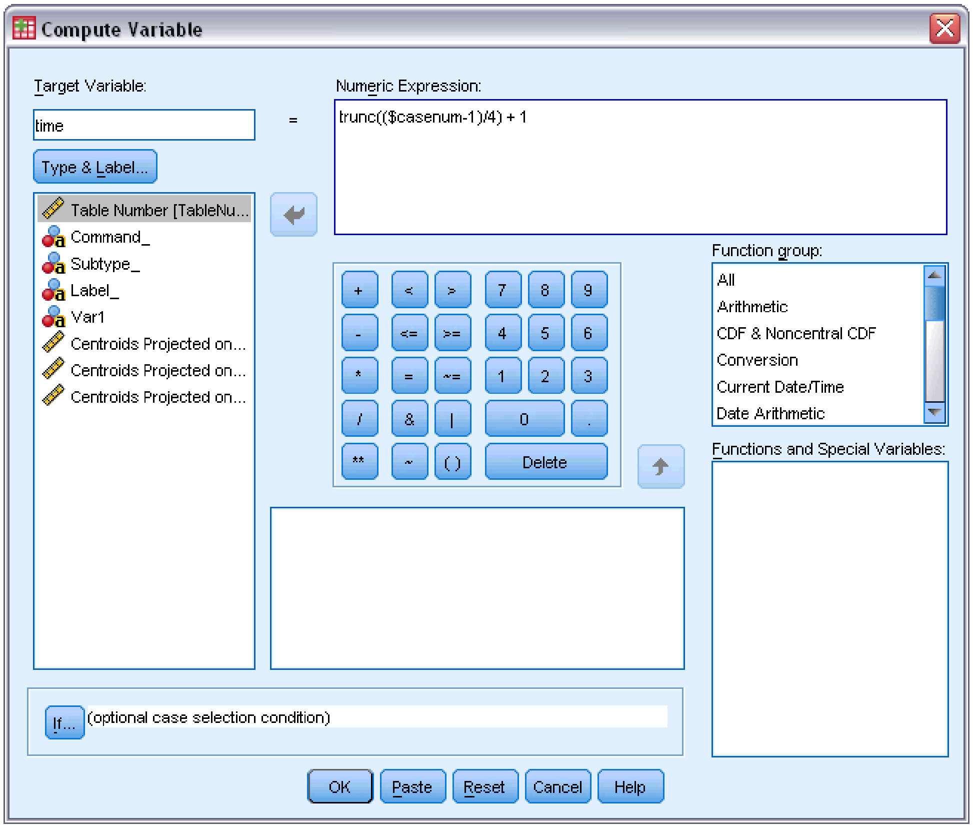
		Label_	Var1	Bingeeating	Sexualattitude	Preoccupationwithfoodandweight	var
1		Projected Centroids	1	0.593	-1.391	-0.823	
2		Projected Centroids	2	-1.419	-0.383	-1.082	
3		Projected Centroids	3	-1.650	-0.415	-1.233	
4		Projected Centroids	4	0.504	-0.854	-0.430	
5		Projected Centroids	5	0.607	-0.421	-0.018	
6		Projected Centroids	6	-0.386	0.347	0.077	
7		Projected Centroids	7	-0.126	0.471	0.319	
8		Projected Centroids	8	0.109	-0.048	0.019	
9		Projected Centroids	9	0.247	0.109	0.224	
10		Projected Centroids	10	0.340	0.783	0.827	
11		Projected Centroids	11	-0.337	0.716	0.406	
12		Projected Centroids	12	0.066	-0.129	-0.071	

The variables *Bingeeating*, *Sexualattitude*, and *Preoccupationwithfoodandweight* contain the values of the centroids projected on each of the symptoms of interest. The case number (1 through 16) corresponds to the time/diagnosis interaction. You will need to compute new variables that separate out the Time and Diagnosis values.

▶ From the menus choose:

Transform > Compute Variable...

Figure 10-46
Compute Variable dialog

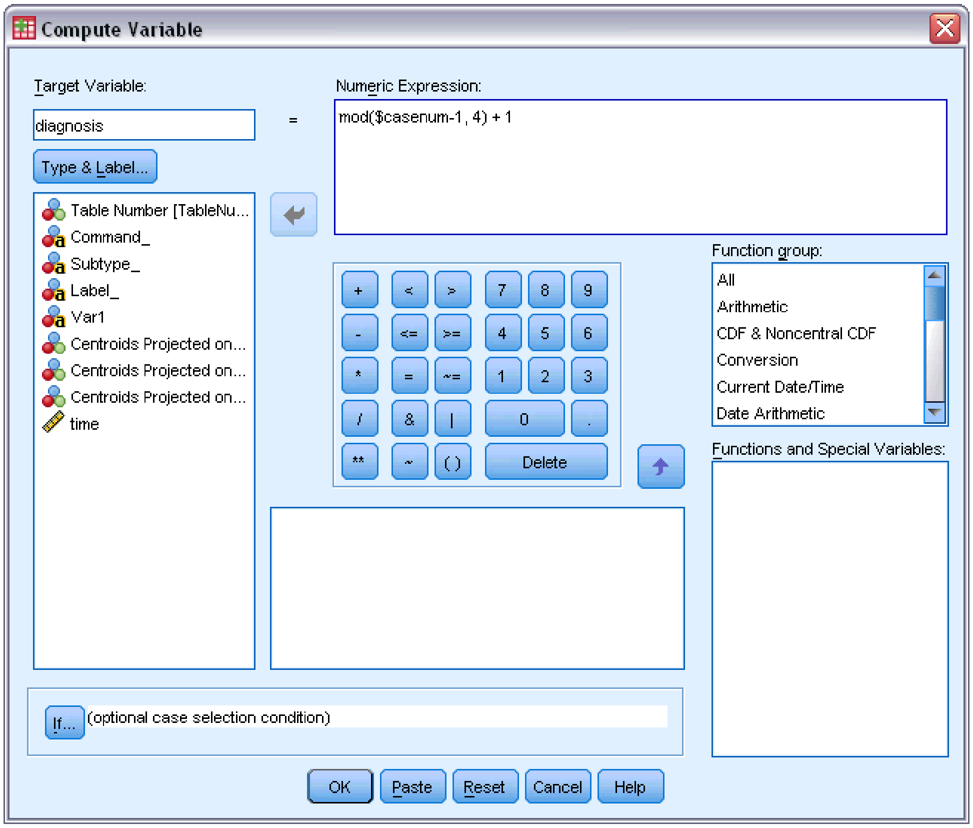


Target Variable: time

Numeric Expression: trunc(($casenum-1)/4) + 1

- ▶ Type *time* as the target variable.
- ▶ Type trunc(($casenum-1)/4) + 1 as the numeric expression.
- ▶ Click OK.

Figure 10-47
Compute Variable dialog

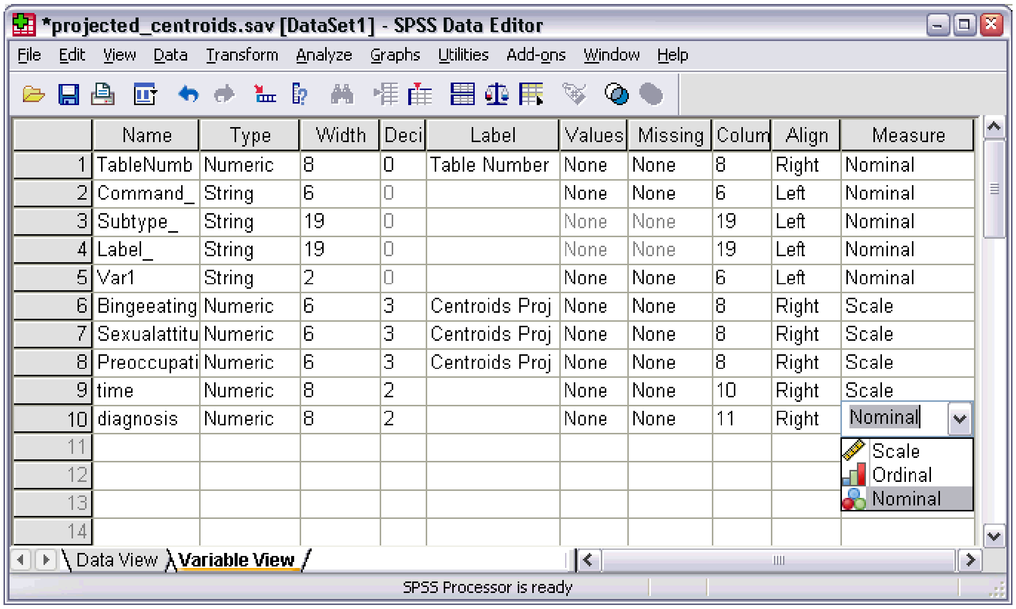


Target Variable: diagnosis

Numeric Expression: mod($casenum-1, 4) + 1

- ▶ Recall the Compute Variable dialog.
- ▶ Type *diagnosis* as the target variable.
- ▶ Type mod($casenum-1, 4) + 1 as the numeric expression.
- ▶ Click OK.

Figure 10-48
Projected_centroids.sav

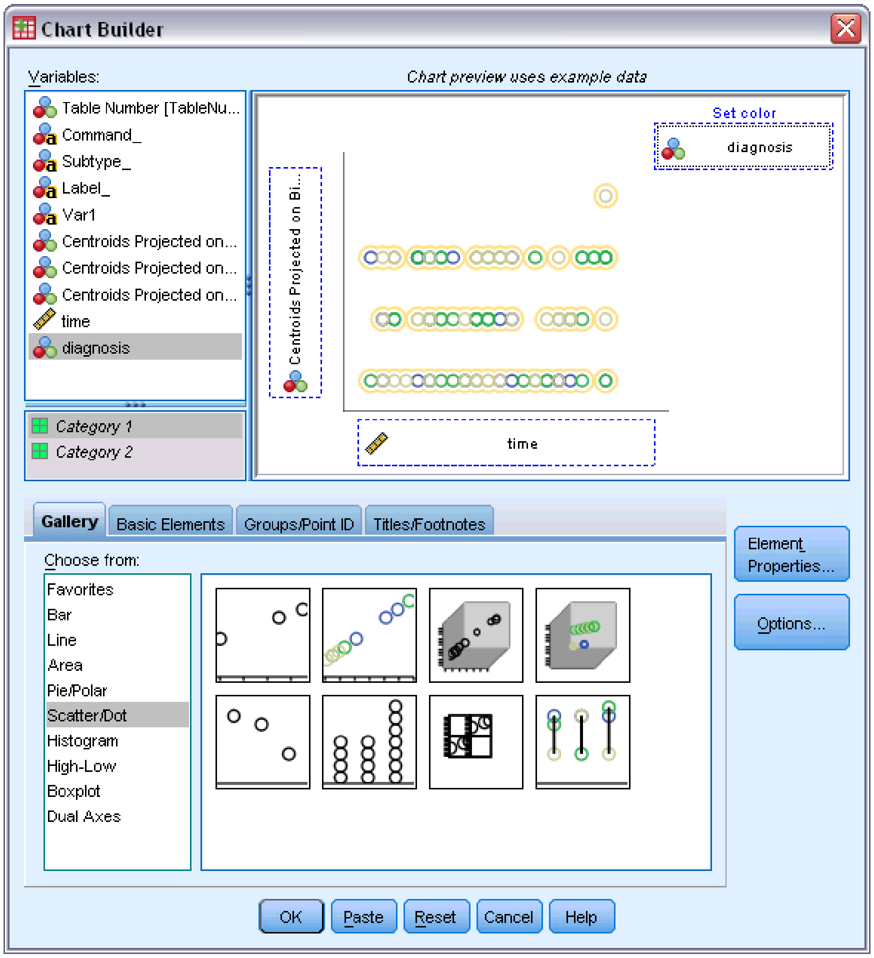


*projected_centroids.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Deci	Label	Values	Missing	Colum	Align	Measure
1	TableNumb	Numeric	8	0	Table Number	None	None	8	Right	Nominal
2	Command_	String	6	0		None	None	6	Left	Nominal
3	Subtype_	String	19	0		None	None	19	Left	Nominal
4	Label_	String	19	0		None	None	19	Left	Nominal
5	Var1	String	2	0		None	None	6	Left	Nominal
6	Bingeeating	Numeric	6	3	Centroids Proj	None	None	8	Right	Scale
7	Sexualattitu	Numeric	6	3	Centroids Proj	None	None	8	Right	Scale
8	Preoccupati	Numeric	6	3	Centroids Proj	None	None	8	Right	Scale
9	time	Numeric	8	2		None	None	10	Right	Scale
10	diagnosis	Numeric	8	2		None	None	11	Right	Nominal
11										Scale Ordinal Nominal
12										
13										
14										

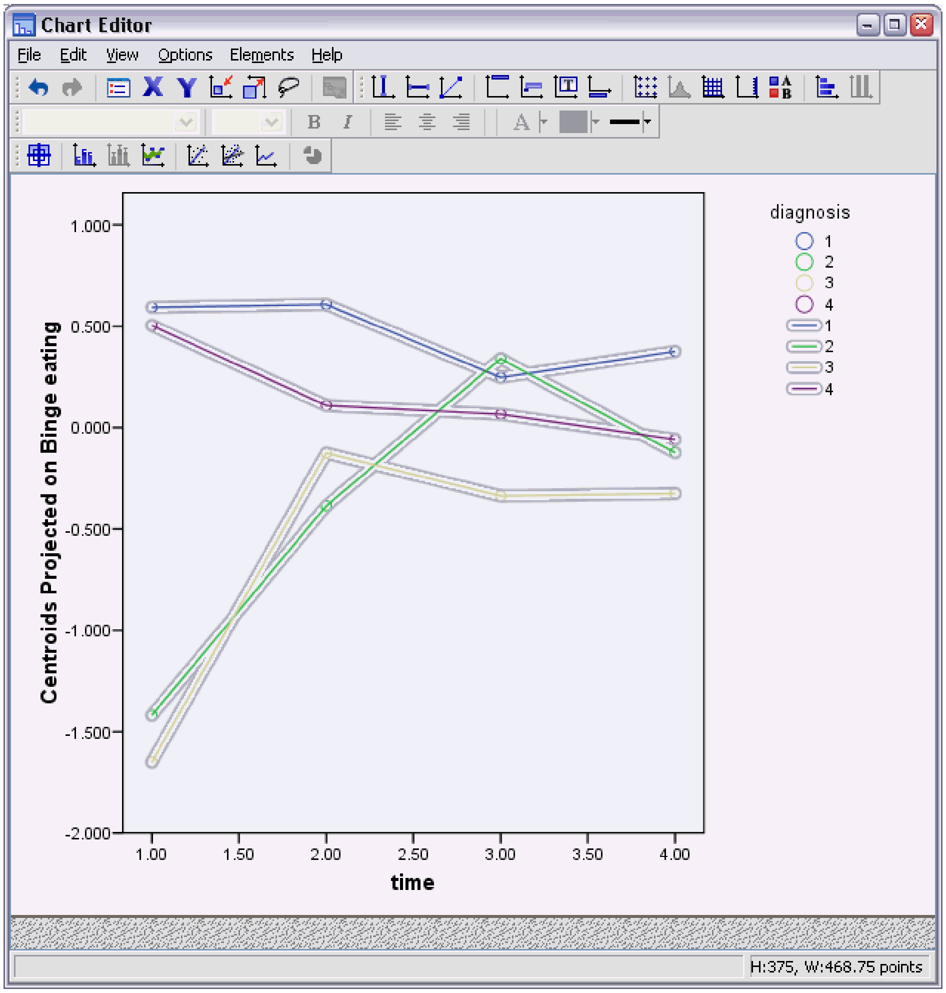
In the Variable View, change the measure for *diagnosis* from Scale to Nominal.

Figure 10-49
Chart Builder



- Finally, to view the projected centroids of time of diagnosis on binging over time, recall the Chart Builder and click Reset to clear your previous selections.
- Select the Scatter/Dot gallery and choose Grouped Scatter.
- Select *Centroids Projected on Binge eating* as the *y*-axis variable and *time* as the *x*-axis variable.
- Choose to set colors by *diagnosis*.
- Click OK.

Figure 10-50
Projected centroids of Time of diagnosis on Binge eating over time

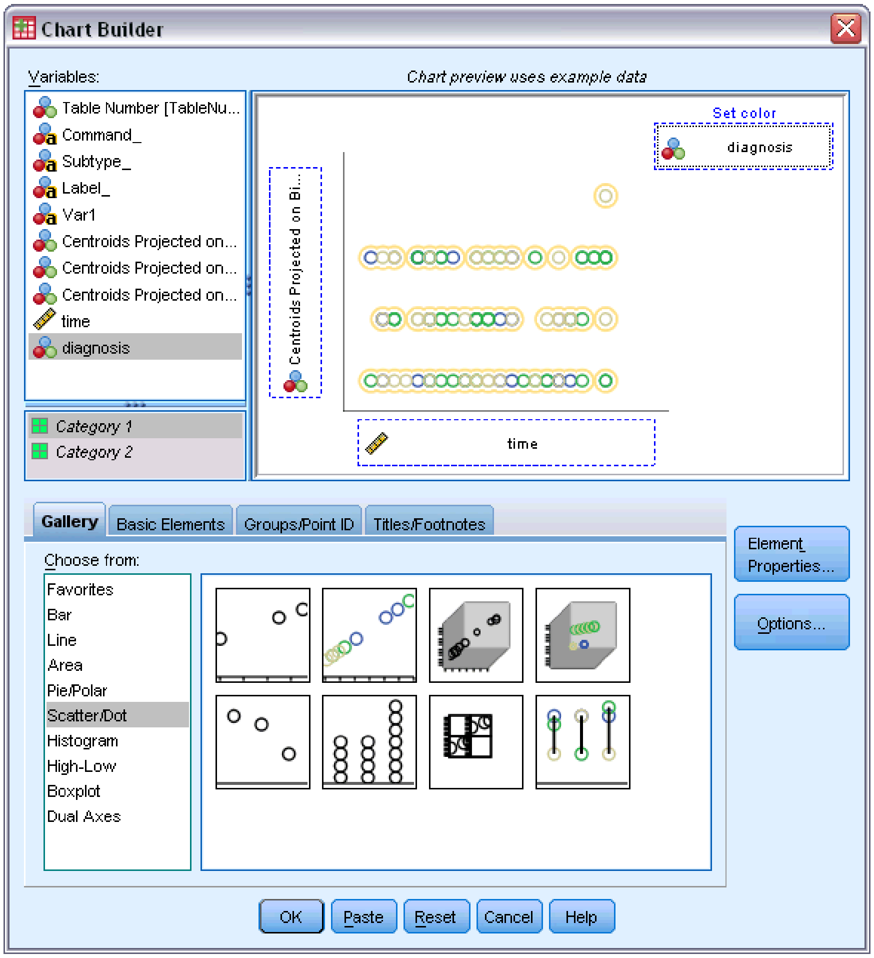


- ▶ Then, to connect the points, double-click on the graph, and click the Add interpolation line tool in the Chart Editor.

- ▶ Close the Chart Editor.

With respect to binge eating, it is clear that the anorectic groups have different starting values from the bulimic groups. This difference shrinks over time, as the anorectic groups hardly change, while the bulimic groups show progress.

Figure 10-51
Chart Builder

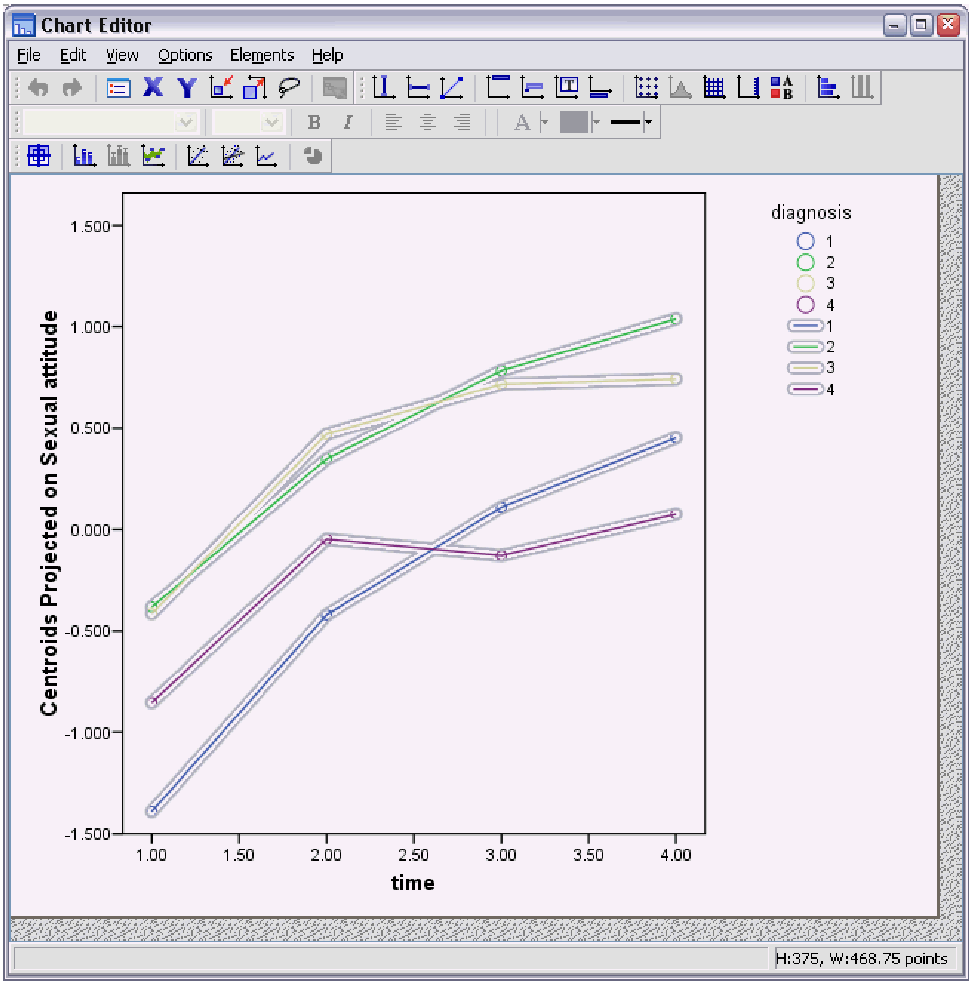


- ▶ Recall the Chart Builder.

- ▶ Deselect *Centroids Projected on Binge eating* as the *y*-axis variable and select *Centroids Projected on Sexual attitude* as the *y*-axis variable.

- ▶ Click OK.

Figure 10-52
Projected centroids of Time of diagnosis on Sexual attitude over time

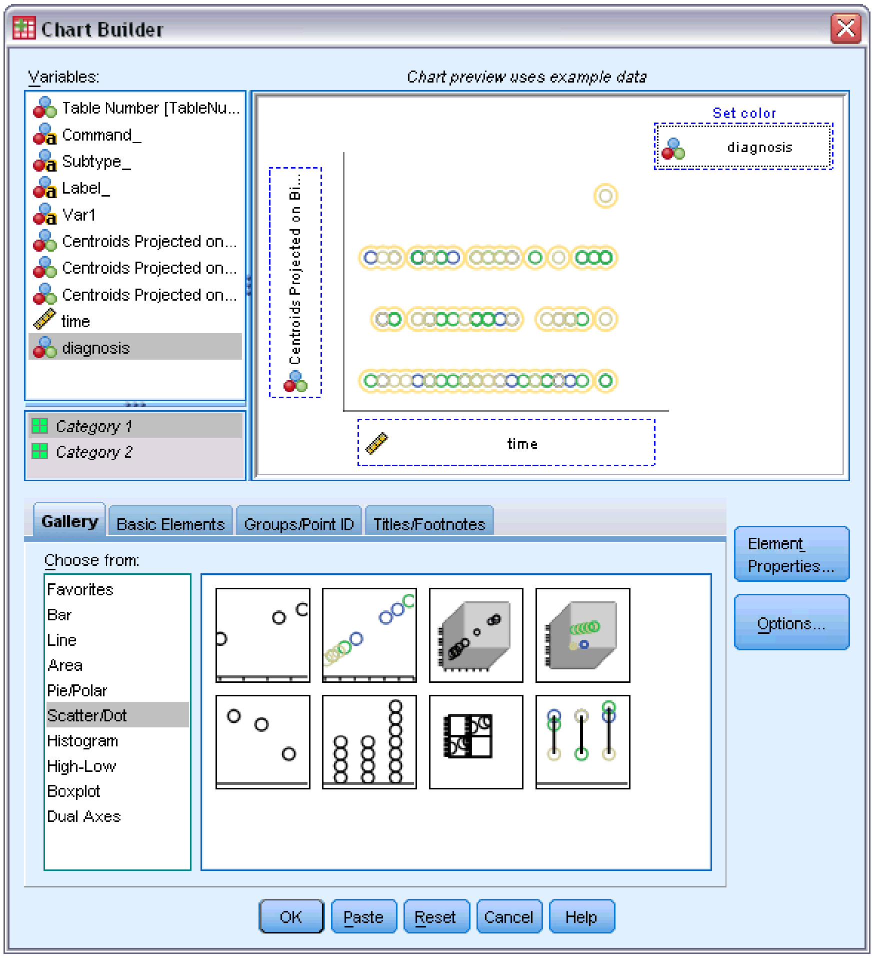


- ▶ Then, to connect the points, double-click on the graph, and click the Add interpolation line tool in the Chart Editor.

- ▶ Close the Chart Editor.

With respect to sexual attitude, the four trajectories are more or less parallel over time, and all groups show progress. The bulimic groups, however, have higher (better) scores than the anorectic group.

Figure 10-53
Chart Builder

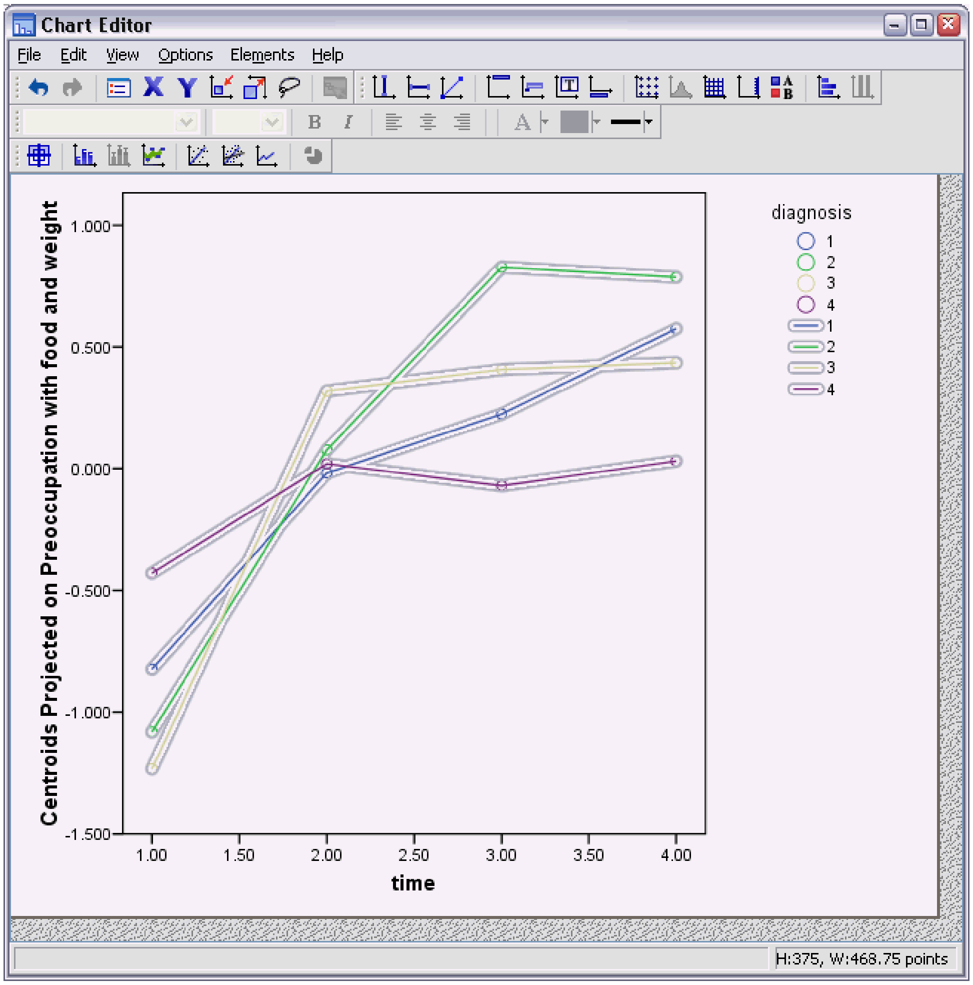


- ▶ Recall the Chart Builder.

- ▶ Deselect *Centroids Projected on Sexual attitude* as the *y*-axis variable and select *Centroids Projected on Preoccupation with food and weight* as the *y*-axis variable.

- ▶ Click OK.

Figure 10-54
Projected centroids of Time of diagnosis on Body preoccupation over time



- ▶ Then, to connect the points, double-click on the graph, and click the Add interpolation line tool in the Chart Editor.

- ▶ Close the Chart Editor.

Body preoccupation is a variable that represents the core symptoms, which are shared by the four different groups. Apart from the atypical eating disorder patients, the anorectic group and the two bulimic groups have very similar levels both at the beginning and at the end.

Recommended Readings

See the following texts for more information on categorical principal components analysis:

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, and J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, 579–595.

De Leeuw, J. 1982. Nonlinear principal components analysis. In: *COMPSTAT Proceedings in Computational Statistics,* Vienna: Physica Verlag, 77–89.

Eckart, C., and G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, 211–218.

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, 453–467.

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02.* Leiden: Department of Data Theory, University of Leiden.

Gower, J. C., and J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, 43–51.

Heiser, W. J., and J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, and J. Blasius, eds. New York: Academic Press, 179–209.

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. In: *International Encyclopedia of Statistics,* W. H. Kruskal, and J. M. Tanur, eds. New York: The Free Press, 307–330.

Kruskal, J. B., and R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, 123–157.

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, 287–300.

Meulman, J. J., and P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, 7–35.

Meulman, J. J., A. J. Van der Kooij, and A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. In: *Classification, Automation and New Media,* W. Gaul, and G. Ritter, eds. Berlin: Springer-Verlag, 207–217.

Meulman, J. J., A. J. Van der Kooij, and W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc., 49–70.

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, and J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, 109–126.

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. In: *Psychological Scaling: Theory & Applications,* H. Gulliksen, and S. Messick, eds. NewYork: John Wiley and Sons, 155–167.

Vlek, C., and P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, 235–271.

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. London: Lawrence Erlbaum Associates, Inc.

Young, F. W., Y. Takane, and J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, 279–281.

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, and J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, 281–302.

Chapter
11

Nonlinear Canonical Correlation Analysis

The purpose of nonlinear canonical correlation analysis is to determine how similar two or more sets of variables are to one another. As in linear canonical correlation analysis, the aim is to account for as much of the variance in the relationships among the sets as possible in a low-dimensional space. Unlike linear canonical correlation analysis, however, nonlinear canonical correlation analysis does not assume an interval level of measurement or assume that the relationships are linear. Another important difference is that nonlinear canonical correlation analysis establishes the similarity between the sets by simultaneously comparing linear combinations of the variables in each set to an unknown set—the object scores.

Example: An Analysis of Survey Results

The example in this chapter is from a survey (Verdegaal, 1985). The responses of 15 subjects to 8 variables were recorded. The variables, variable labels, and value labels (categories) in the dataset are shown in the following table.

Table 11-1
Survey data

Variable name	Variable label	Value label
age	*Age in years*	20–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70
marital	*Marital status*	Single, Married, Other
pet	*Pets owned*	No, Cat(s), Dog(s), Other than cat or dog, Various domestic animals
news	*Newspaper read most often*	None, Telegraaf, Volkskrant, NRC, Other
music	*Music preferred*	Classical, New wave, Popular, Variety, Don't like music
live	*Neighborhood preference*	Town, Village, Countryside
math	*Math test score*	0–5, 6–10, 11–15
language	*Language test score*	0–5, 6–10, 11–15, 16–20

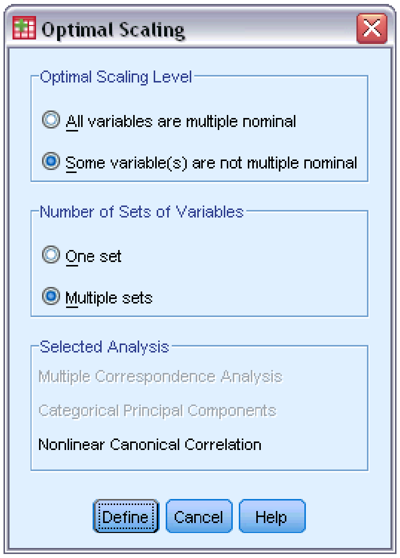
This dataset can be found in *verd1985.sav*. For more information, see the topic Sample Files in Appendix A on p. 292. The variables of interest are the first six variables, and they are divided into three sets. Set 1 includes *age* and *marital*, set 2 includes *pet* and *news*, and set 3 includes *music* and *live*. *Pet* is scaled as multiple nominal, and *age* is scaled as ordinal; all other variables are scaled as single nominal. This analysis requests a random initial configuration. By default, the initial configuration is numerical. However, when some of the variables are treated as single nominal with no possibility of ordering, it is best to choose a random initial configuration. This is the case with most of the variables in this study.

Examining the Data

▶ To obtain a nonlinear canonical correlation analysis for this dataset, from the menus choose:

Analyze > Dimension Reduction > Optimal Scaling...

Figure 11-1
Optimal Scaling dialog box



Optimal Scaling

Optimal Scaling Level

○ All variables are multiple nominal

● Some variable(s) are not multiple nominal

Number of Sets of Variables

○ One set

● Multiple sets

Selected Analysis

Multiple Correspondence Analysis

Categorical Principal Components

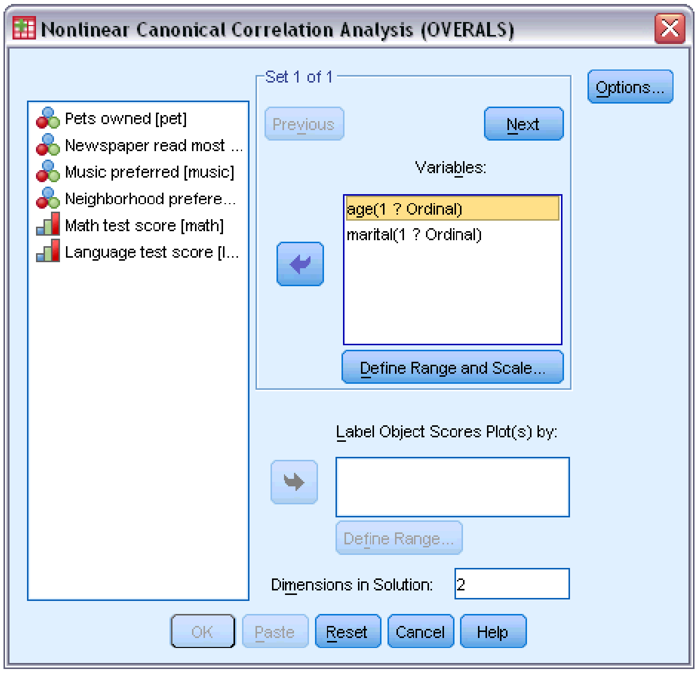
Nonlinear Canonical Correlation

▶ Select Some variable(s) not multiple nominal in the Optimal Scaling Level group.

▶ Select Multiple sets in the Number of Sets of Variables group.

▶ Click Define.

Figure 11-2
Nonlinear Canonical Correlation Analysis dialog box

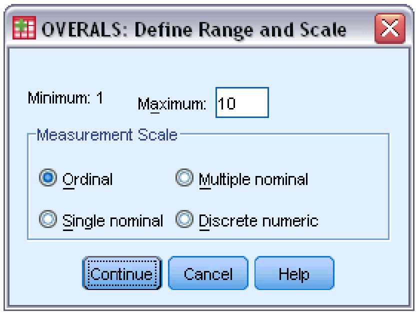


Variables:
age(1 ? Ordinal)
marital(1 ? Ordinal)

Dimensions in Solution: 2

- ▶ Select *Age in years* and *Marital status* as variables for the first set.
- ▶ Select *age* and click Define Range and Scale.

Figure 11-3
Define Range and Scale dialog box

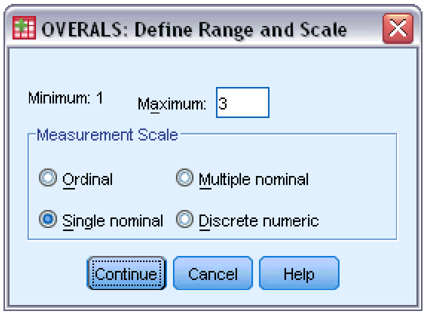


Minimum: 1 Maximum: 10

● Ordinal ○ Multiple nominal
○ Single nominal ○ Discrete numeric

- ▶ Type 10 as the maximum value for this variable.
- ▶ Click Continue.
- ▶ In the Nonlinear Canonical Correlation Analysis dialog box, select *marital* and click Define Range and Scale.

Figure 11-4
Define Range and Scale dialog box



OVERALS: Define Range and Scale

Minimum: 1

Maximum: 3

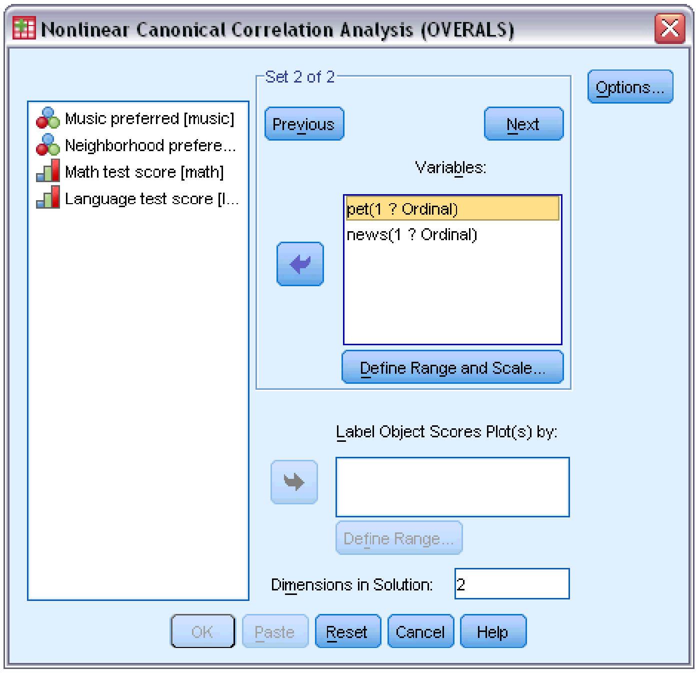
Measurement Scale

○ Ordinal ○ Multiple nominal

● Single nominal ○ Discrete numeric

- ▶ Type 3 as the maximum value for this variable.
- ▶ Select Single nominal as the measurement scale.
- ▶ Click Continue.
- ▶ In the Nonlinear Canonical Correlation Analysis dialog box, click Next to define the next variable set.

Figure 11-5
Nonlinear Canonical Correlation Analysis dialog box



Nonlinear Canonical Correlation Analysis (OVERALS)

Music preferred [music]
Neighborhood prefere...
Math test score [math]
Language test score [l...

Set 2 of 2

Variables:

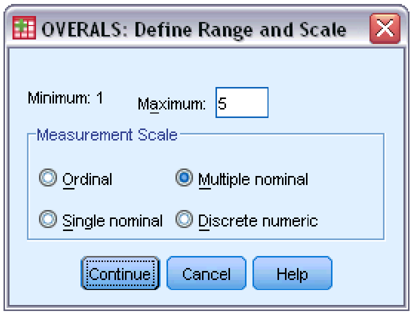
pet(1 ? Ordinal)
news(1 ? Ordinal)

Label Object Scores Plot(s) by:

Dimensions in Solution: 2

- ▶ Select *Pets owned* and *Newspaper read most often* as variables for the second set.
- ▶ Select *pet* and click Define Range and Scale.

Figure 11-6
Define Range and Scale dialog box



OVERALS: Define Range and Scale

Minimum: 1 Maximum: 5

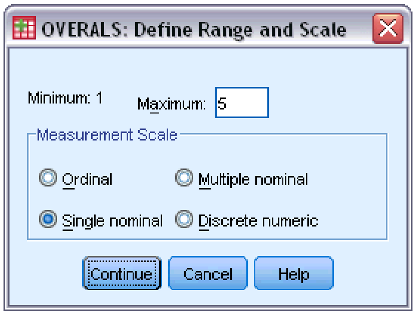
Measurement Scale

○ Ordinal ● Multiple nominal

○ Single nominal ○ Discrete numeric

- ▶ Type 5 as the maximum value for this variable.
- ▶ Select Multiple nominal as the measurement scale.
- ▶ Click Continue.
- ▶ In the Nonlinear Canonical Correlation Analysis dialog box, select *news* and click Define Range and Scale.

Figure 11-7
Define Range and Scale dialog box



OVERALS: Define Range and Scale

Minimum: 1 Maximum: 5

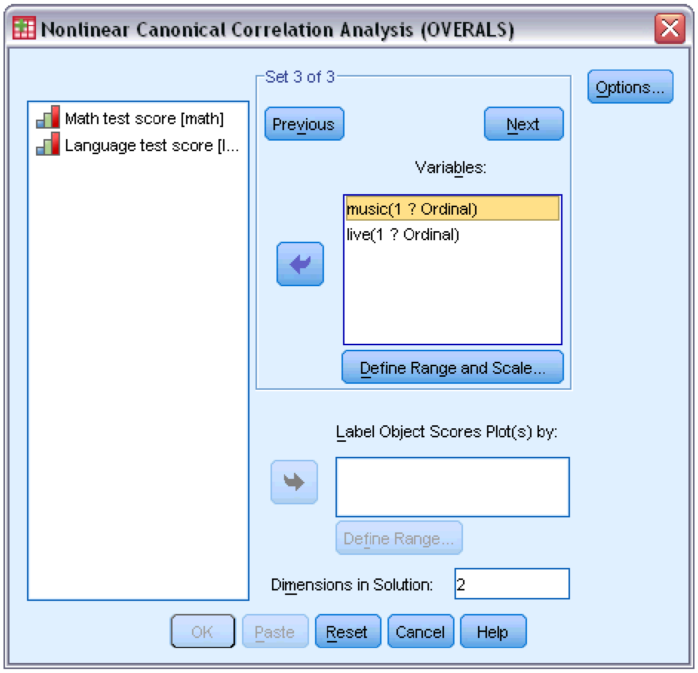
Measurement Scale

○ Ordinal ○ Multiple nominal

● Single nominal ○ Discrete numeric

- ▶ Type 5 as the maximum value for this variable.
- ▶ Select Single nominal as the measurement scale.
- ▶ Click Continue.
- ▶ In the Nonlinear Canonical Correlation Analysis dialog box, click Next to define the last variable set.

Figure 11-8
Nonlinear Canonical Correlation Analysis dialog box


Nonlinear Canonical Correlation Analysis (OVERALS)
Options...
Previous
Next
Define Range and Scale...
Define Range...
OK
Paste
Reset
Cancel
Help

Math test score [math]
Language test score [l...

Set 3 of 3

Variables:

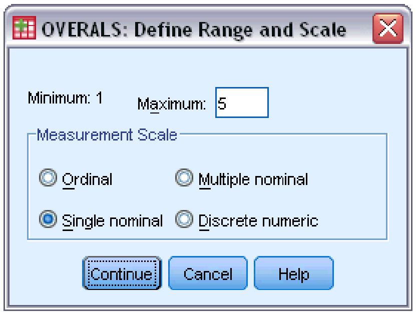
music(1 ? Ordinal)
live(1 ? Ordinal)

Label Object Scores Plot(s) by:

Dimensions in Solution: 2

- ▶ Select *Music preferred* and *Neighborhood preference* as variables for the third set.
- ▶ Select *music* and click Define Range and Scale.

Figure 11-9
Define Range and Scale dialog box



Minimum: 1 Maximum: 5

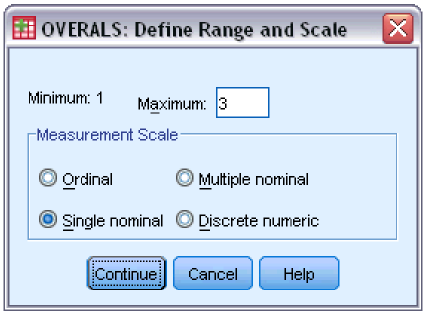
Measurement Scale

○ Ordinal ○ Multiple nominal

● Single nominal ○ Discrete numeric

- ▶ Type 5 as the maximum value for this variable.
- ▶ Select Single nominal as the measurement scale.
- ▶ Click Continue.
- ▶ In the Nonlinear Canonical Correlation Analysis dialog box, select *live* and click Define Range and Scale.

Figure 11-10
Define Range and Scale dialog box



OVERALS: Define Range and Scale

Minimum: 1 Maximum: 3

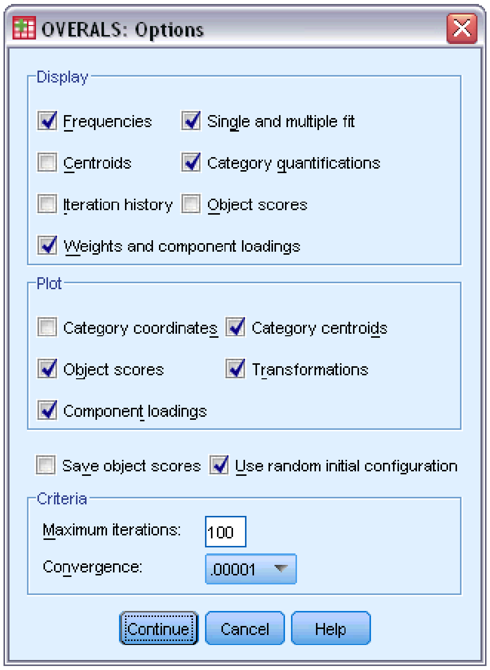
Measurement Scale

○ Ordinal ○ Multiple nominal

● Single nominal ○ Discrete numeric

- ▶ Type 3 as the maximum value for this variable.
- ▶ Select Single nominal as the measurement scale.
- ▶ Click Continue.
- ▶ In the Nonlinear Canonical Correlation Analysis dialog box, click Options.

Figure 11-11
Options dialog box



OVERALS: Options

Display

☑ Frequencies ☑ Single and multiple fit

□ Centroids ☑ Category quantifications

□ Iteration history □ Object scores

☑ Weights and component loadings

Plot

□ Category coordinates ☑ Category centroids

☑ Object scores ☑ Transformations

☑ Component loadings

□ Save object scores ☑ Use random initial configuration

Criteria

Maximum iterations: 100

Convergence: .00001

- ▶ Deselect Centroids and select Weights and component loadings in the Display group.
- ▶ Select Category centroids and Transformations in the Plot group.
- ▶ Select Use random initial configuration.
- ▶ Click Continue.
- ▶ In the Nonlinear Canonical Correlation Analysis dialog box, click OK.

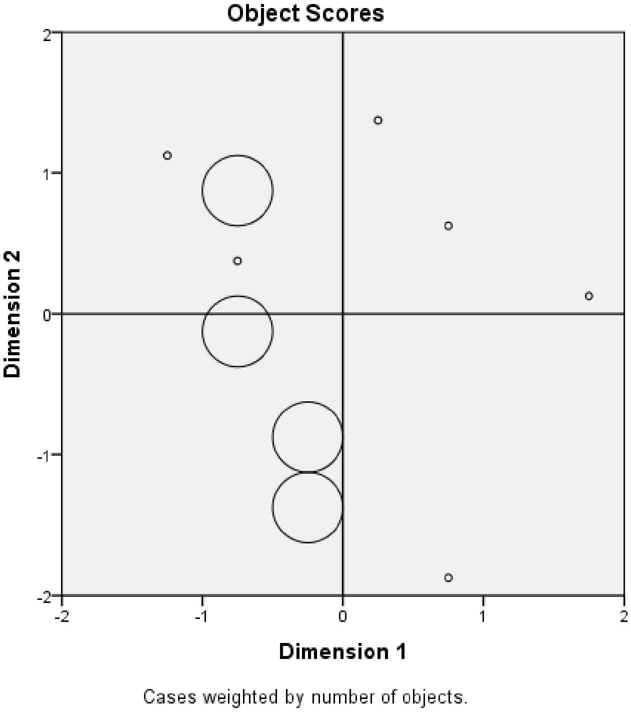
After a list of the variables with their levels of optimal scaling, categorical canonical correlation analysis with optimal scaling produces tables showing the frequencies of objects by category for each variable in the analysis. These tables are especially important if there are missing data, since almost-empty categories are more likely to dominate the solution. In this example, there are no missing data.

A second preliminary check is to examine the plot of object scores for outliers. Outliers have such different quantifications from the other objects that they will be at the boundaries of the plot, thus dominating one or more dimensions.

If you find outliers, you can handle them in one of two ways. You can simply eliminate them from the data and run the nonlinear canonical correlation analysis again. Alternatively, you can try recoding the extreme responses of the outlying objects by collapsing (merging) some categories.

As shown in the plot of object scores, there were no outliers for the survey data.

Figure 11-12
Object scores



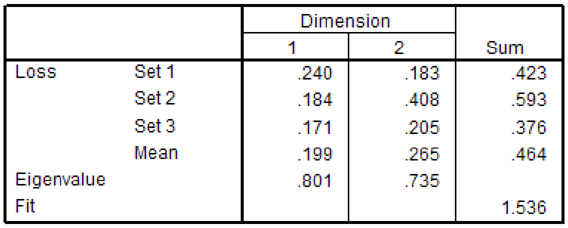
Accounting for Similarity between Sets

There are several ways to measure the association between sets in a nonlinear canonical correlation analysis (each way being detailed in a separate table or set of tables).

Summary of Analysis

The fit and loss values tell you how well the nonlinear canonical correlation analysis solution fits the optimally quantified data with respect to the association between the sets. The summary of analysis table shows the fit value, loss values, and eigenvalues for the survey example.

Figure 11-13
Summary of analysis



		Dimension		Sum
		1	2	
Loss	Set 1	.240	.183	.423
	Set 2	.184	.408	.593
	Set 3	.171	.205	.376
	Mean	.199	.265	.464
Eigenvalue		.801	.735	
Fit				1.536

Loss is partitioned across dimensions and sets. For each dimension and set, loss represents the proportion of variation in the object scores that cannot be accounted for by the weighted combination of variables in the set. The average loss is labeled Mean. In this example, the average loss over sets is 0.464. Notice that more loss occurs for the second dimension than for the first dimension.

The eigenvalue for each dimension equals 1 minus the average loss for the dimension and indicates how much of the relationship is shown by each dimension. The eigenvalues add up to the total fit. For Verdegaal's data, 0.801 / 1.536 = 52% of the actual fit is accounted for by the first dimension.

The maximum fit value equals the number of dimensions and, if obtained, indicates that the relationship is perfect. The average loss value over sets and dimensions tells you the difference between the maximum fit and the actual fit. Fit plus the average loss equals the number of dimensions. Perfect similarity rarely happens and usually capitalizes on trivial aspects in the data.

Another popular statistic with two sets of variables is the canonical correlation. Since the canonical correlation is related to the eigenvalue and thus provides no additional information, it is not included in the nonlinear canonical correlation analysis output. For two sets of variables, the canonical correlation per dimension is obtained by the following formula:

$$\rho_d = 2 \times E_d - 1$$

where *d* is the dimension number and *E* is the eigenvalue.

You can generalize the canonical correlation for more than two sets with the following formula:

$$\rho_d = ((K \times E_d) - 1)/(K - 1)$$

where *d* is the dimension number, *K* is the number of sets, and *E* is the eigenvalue. For our example,

$$\rho_1 = ((3 \times 0.801) - 1)/2 = 0.702$$

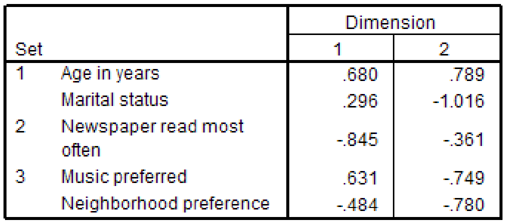
and

$$\rho_2 = ((3 \times 0.735) - 1)/2 = 0.603$$

Weights and Component Loadings

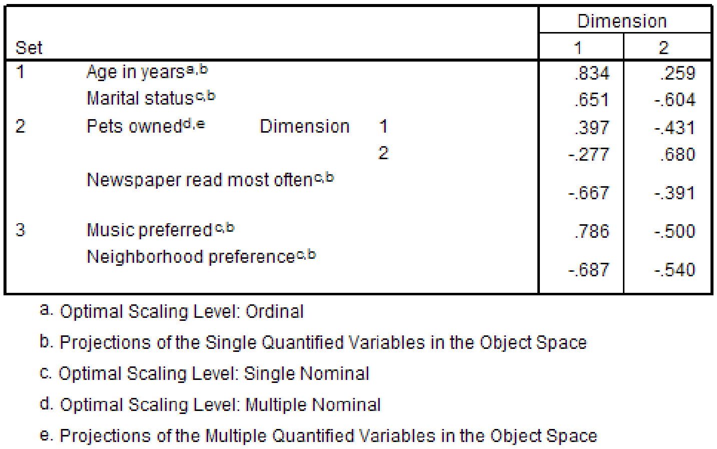
Another measure of association is the multiple correlation between linear combinations from each set and the object scores. If no variables in a set are multiple nominal, you can compute this measure by multiplying the weight and component loading of each variable within the set, adding these products, and taking the square root of the sum.

Figure 11-14
Weights



Set		Dimension	
		1	2
1	Age in years	.680	.789
	Marital status	.296	-1.016
2	Newspaper read most often	-.845	-.361
3	Music preferred	.631	-.749
	Neighborhood preference	-.484	-.780

Figure 11-15
Component loadings



Set				Dimension	
				1	2
1	Age in years[a,b]			.834	.259
	Marital status[c,b]			.651	-.604
2	Pets owned[d,e]	Dimension	1	.397	-.431
			2	-.277	.680
	Newspaper read most often[c,b]			-.667	-.391
3	Music preferred[c,b]			.786	-.500
	Neighborhood preference[c,b]			-.687	-.540

a. Optimal Scaling Level: Ordinal

b. Projections of the Single Quantified Variables in the Object Space

c. Optimal Scaling Level: Single Nominal

d. Optimal Scaling Level: Multiple Nominal

e. Projections of the Multiple Quantified Variables in the Object Space

These figures give the weights and component loadings for the variables in this example. The multiple correlation (R) is as follows for the first weighted sum of optimally scaled variables (*Age in years* and *Marital status*) with the first dimension of object scores:

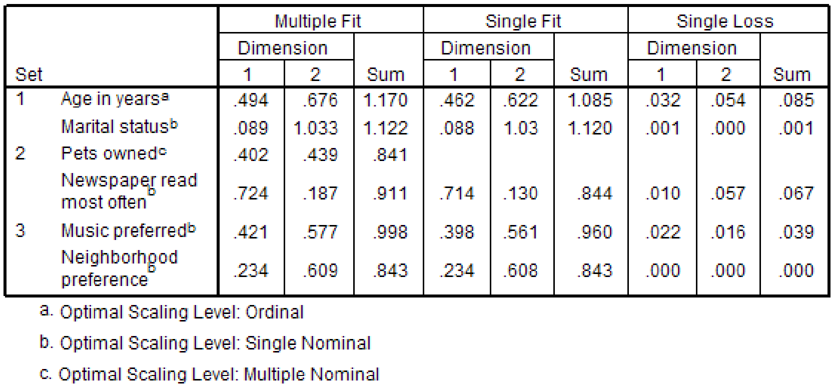
$$\begin{aligned} R &= \sqrt{(0.701 \times 0.841 + (-0.273 \times -0.631))} \\ &= \sqrt{(0.5895 + 0.1723)} \\ &= 0.873 \end{aligned}$$

For each dimension, 1 – loss = R^2. For example, from the Summary of analysis table, 1 – 0.238 = 0.762, which is 0.873 squared (plus some rounding error). Consequently, small loss values indicate large multiple correlations between weighted sums of optimally scaled variables and dimensions. Weights are not unique for multiple nominal variables. For multiple nominal variables, use 1 – loss per set.

Partitioning Fit and Loss

The loss of each set is partitioned by the nonlinear canonical correlation analysis in several ways. The fit table presents the multiple fit, single fit, and single loss tables produced by the nonlinear canonical correlation analysis for the survey example. Note that multiple fit minus single fit equals single loss.

Figure 11-16
Partitioning fit and loss



		Multiple Fit			Single Fit			Single Loss		
		Dimension			Dimension			Dimension		
Set		1	2	Sum	1	2	Sum	1	2	Sum
1	Age in years[a]	.494	.676	1.170	.462	.622	1.085	.032	.054	.085
	Marital status[b]	.089	1.033	1.122	.088	1.03	1.120	.001	.000	.001
2	Pets owned[c]	.402	.439	.841						
	Newspaper read most often[b]	.724	.187	.911	.714	.130	.844	.010	.057	.067
3	Music preferred[b]	.421	.577	.998	.398	.561	.960	.022	.016	.039
	Neighborhood preference[b]	.234	.609	.843	.234	.608	.843	.000	.000	.000

a. Optimal Scaling Level: Ordinal

b. Optimal Scaling Level: Single Nominal

c. Optimal Scaling Level: Multiple Nominal

Single loss indicates the loss resulting from restricting variables to one set of quantifications (that is, single nominal, ordinal, or nominal). If single loss is large, it is better to treat the variables as multiple nominal. In this example, however, single fit and multiple fit are almost equal, which means that the multiple coordinates are almost on a straight line in the direction given by the weights.

Multiple fit equals the variance of the multiple category coordinates for each variable. These measures are analogous to the discrimination measures that are found in homogeneity analysis. You can examine the multiple fit table to see which variables discriminate best. For example, look at the multiple fit table for *Marital status* and *Newspaper read most often*. The fit values, summed across the two dimensions, are 1.122 for *Marital status* and 0.911 for *Newspaper read most often*. This information tells us that a person's marital status provides greater discriminatory power than the newspaper they subscribe to.

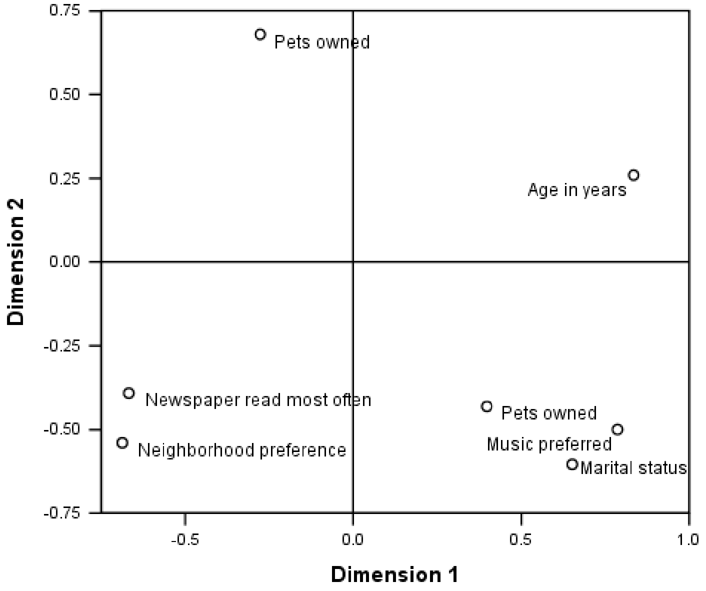
Single fit corresponds to the squared weight for each variable and equals the variance of the single category coordinates. As a result, the weights equal the standard deviations of the single category coordinates. By examining how the single fit is broken down across dimensions, we see that the variable *Newspaper read most often* discriminates mainly on the first dimension, and we see that the variable *Marital status* discriminates almost totally on the second dimension. In other words, the categories of *Newspaper read most often* are further apart in the first dimension than in the second, whereas the pattern is reversed for *Marital status*. In contrast, *Age in years* discriminates in both the first and second dimensions; thus, the spread of the categories is equal along both dimensions.

Component Loadings

The following figure shows the plot of component loadings for the survey data. When there are no missing data, the component loadings are equivalent to the Pearson correlations between the quantified variables and the object scores.

The distance from the origin to each variable point approximates the importance of that variable. The canonical variables are not plotted but can be represented by horizontal and vertical lines drawn through the origin.

Figure 11-17
Component loadings



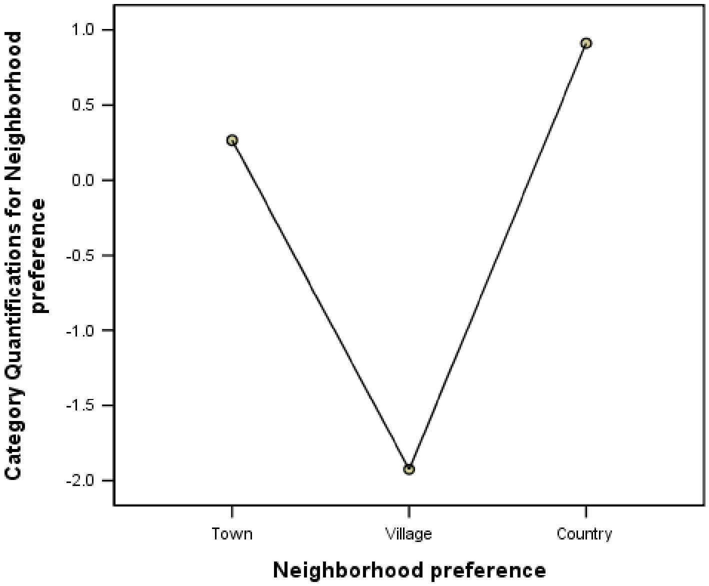
The relationships between variables are apparent. There are two directions that do not coincide with the horizontal and vertical axes. One direction is determined by *Age in years*, *Newspaper read most often*, and *Neighborhood preference*. The other direction is defined by the variables *Marital status*, *Music preferred*, and *Pets owned*. The *Pets owned* variable is a multiple nominal variable, so there are two points plotted for it. Each quantification is interpreted as a single variable.

Transformation Plots

The different levels at which each variable can be scaled impose restrictions on the quantifications. Transformation plots illustrate the relationship between the quantifications and the original categories resulting from the selected optimal scaling level.

The transformation plot for *Neighborhood preference*, which was treated as nominal, displays a U-shaped pattern, in which the middle category receives the lowest quantification, and the extreme categories receive values that are similar to each other. This pattern indicates a quadratic relationship between the original variable and the transformed variable. Using an alternative optimal scaling level is not suggested for *Neighborhood preference*.

Figure 11-18
Transformation plot for Neighborhood preference (nominal)



The quantifications for *Newspaper read most often*, in contrast, correspond to an increasing trend across the three categories that have observed cases. The first category receives the lowest quantification, the second category receives a higher value, and the third category receives the highest value. Although the variable is scaled as nominal, the category order is retrieved in the quantifications.

Figure 11-19
Transformation plot for Newspaper read most often (nominal)

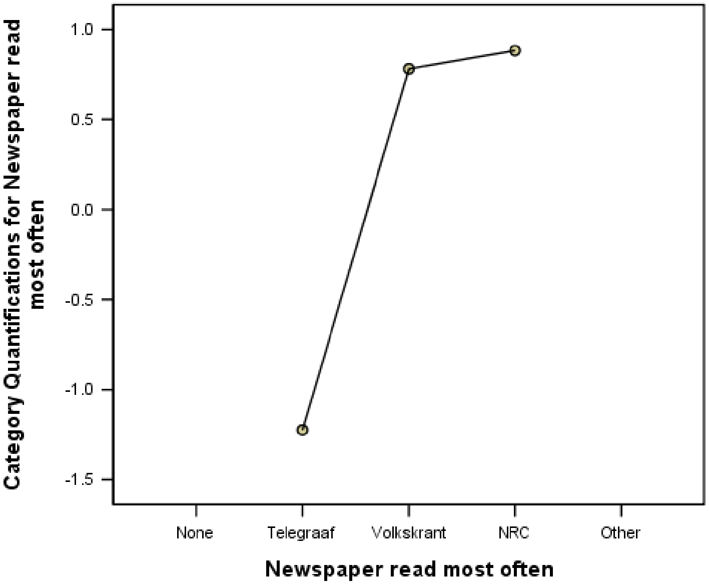
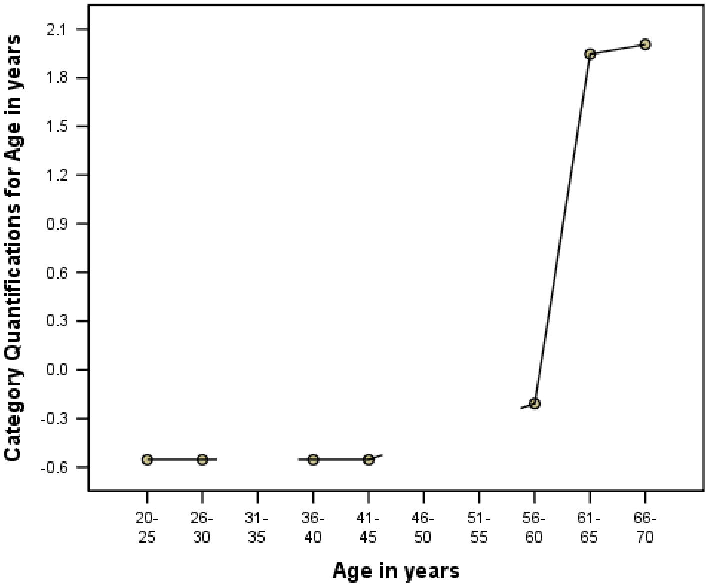


Figure 11-20
Transformation plot for Age in years (ordinal)



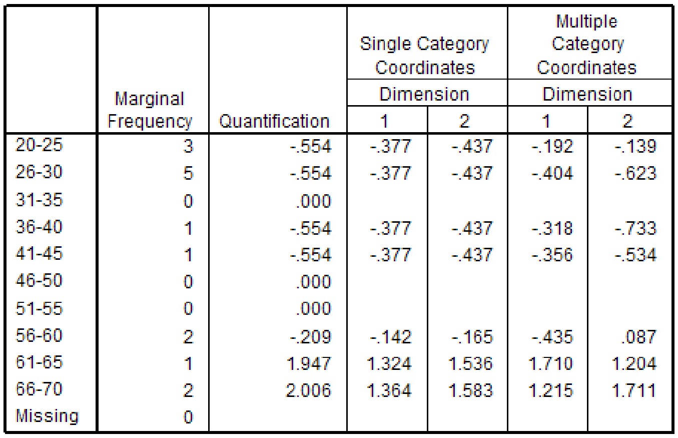
The transformation plot for *Age in years* displays an S-shaped curve. The four youngest observed categories all receive the same negative quantification, whereas the two oldest categories receive similar positive values. Consequently, collapsing all younger ages into one common category (that is, below 50) and collapsing the two oldest categories into one category may be attempted.

However, the exact equality of the quantifications for the younger groups indicates that restricting the order of the quantifications to the order of the original categories may not be desirable. Because the quantifications for the 26–30, 36–40, and 41–45 groups cannot be lower than the quantification for the 20–25 group, these values are set equal to the boundary value. Allowing these values to be smaller than the quantification for the youngest group (that is, treating age as nominal) may improve the fit. So although age may be considered an ordinal variable, treating it as such does not appear appropriate in this case. Moreover, treating age as numerical, and thus maintaining the distances between the categories, would substantially reduce the fit.

Single Category versus Multiple Category Coordinates

For every variable treated as single nominal, ordinal, or numerical, quantifications, single category coordinates, and multiple category coordinates are determined. These statistics for *Age in years* are presented.

Figure 11-21
Coordinates for Age in years



	Marginal Frequency	Quantification	Single Category Coordinates		Multiple Category Coordinates	
			Dimension		Dimension	
			1	2	1	2
20-25	3	-.554	-.377	-.437	-.192	-.139
26-30	5	-.554	-.377	-.437	-.404	-.623
31-35	0	.000				
36-40	1	-.554	-.377	-.437	-.318	-.733
41-45	1	-.554	-.377	-.437	-.356	-.534
46-50	0	.000				
51-55	0	.000				
56-60	2	-.209	-.142	-.165	-.435	.087
61-65	1	1.947	1.324	1.536	1.710	1.204
66-70	2	2.006	1.364	1.583	1.215	1.711
Missing	0					

Every category for which no cases were recorded receives a quantification of 0. For *Age in years*, this includes the 31–35, 46–50, and 51–55 categories. These categories are not restricted to be ordered with the other categories and do not affect any computations.

For multiple nominal variables, each category receives a different quantification on each dimension. For all other transformation types, a category has only one quantification, regardless of the dimensionality of the solution. Each set of single category coordinates represents the location of the category on a line in the object space. The coordinates for a given category equal the quantification multiplied by the variable dimension weights. For example, in the table for *Age in years*, the single category coordinates for category 56-60 (-0.142, -0.165) are the quantification (-0.209) multiplied by the dimension weights (0.680, 0.789).

The multiple category coordinates for variables that are treated as single nominal, ordinal, or numerical represent the coordinates of the categories in the object space before ordinal or linear constraints are applied. These values are unconstrained minimizers of the loss. For multiple nominal variables, these coordinates represent the quantifications of the categories.

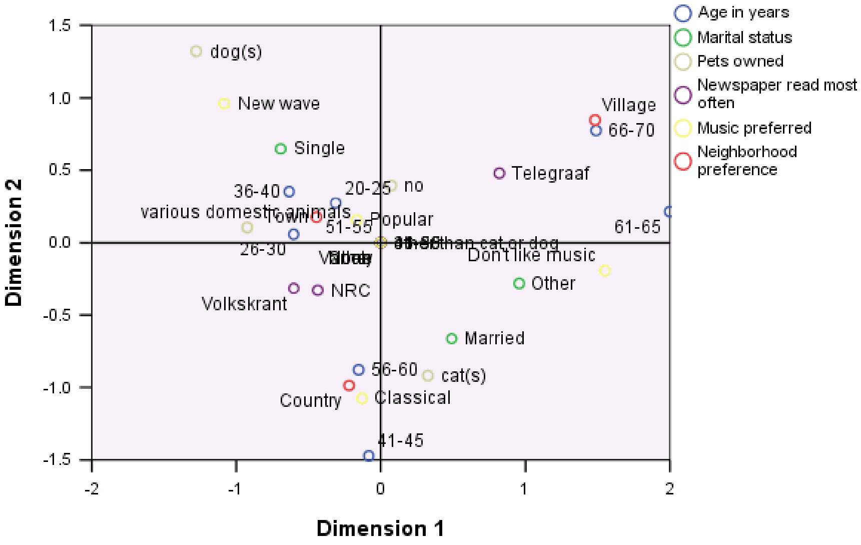
The effects of imposing constraints on the relationship between the categories and their quantifications are revealed by comparing the single category coordinates with the multiple category coordinates. On the first dimension, the multiple category coordinates for *Age in years* decrease to category 2 and remain relatively at the same level until category 9, at which point a dramatic increase occurs. A similar pattern is evidenced for the second dimension. These relationships are removed in the single category coordinates, in which the ordinal constraint is applied. On both dimensions, the coordinates are now nondecreasing. The differing structure of the two sets of coordinates suggests that a nominal treatment may be more appropriate.

Centroids and Projected Centroids

The plot of centroids labeled by variables should be interpreted in the same way as the category quantifications plot in homogeneity analysis or the multiple category coordinates in nonlinear principal components analysis. By itself, such a plot shows how well variables separate groups of objects (the centroids are in the center of gravity of the objects).

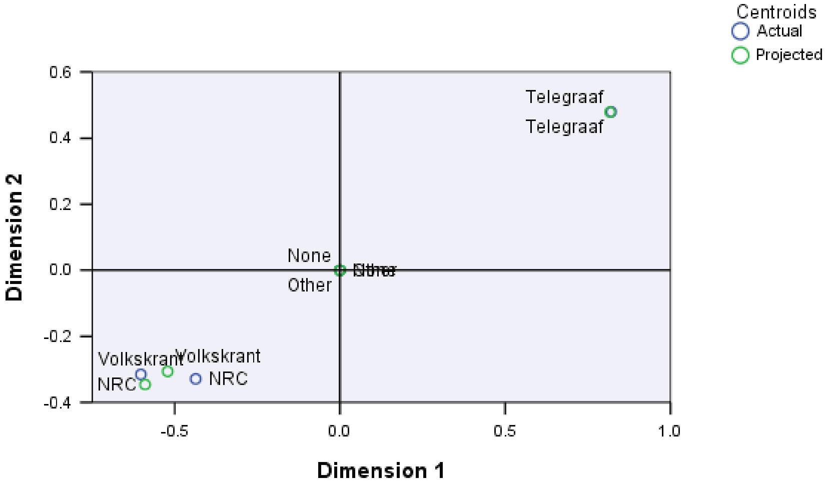
Notice that the categories for *Age in years* are not separated very clearly. The younger age categories are grouped together at the left of the plot. As suggested previously, ordinal may be too strict a scaling level to impose on *Age in years*.

Figure 11-22
Centroids labeled by variables



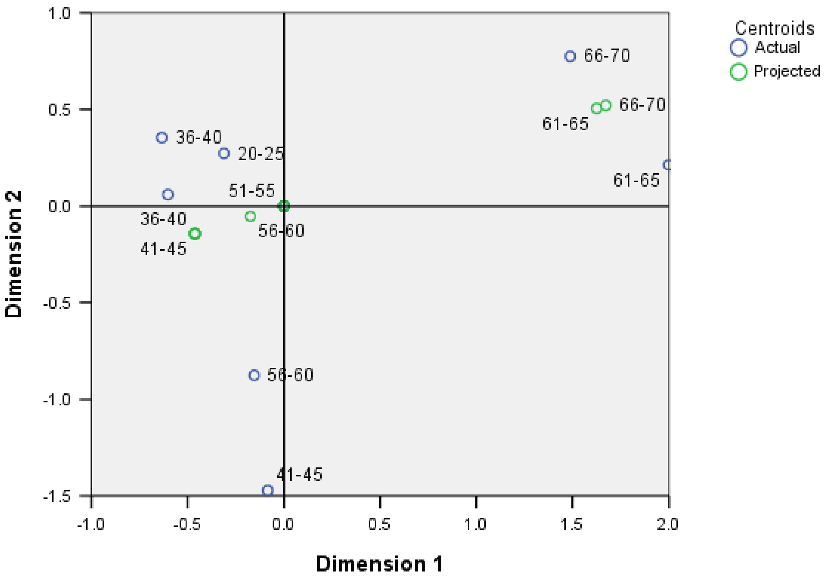
When you request centroid plots, individual centroid and projected centroid plots for each variable that is labeled by value labels are also produced. The projected centroids are on a line in the object space.

Figure 11-23
Centroids and projected centroids for Newspaper read most often



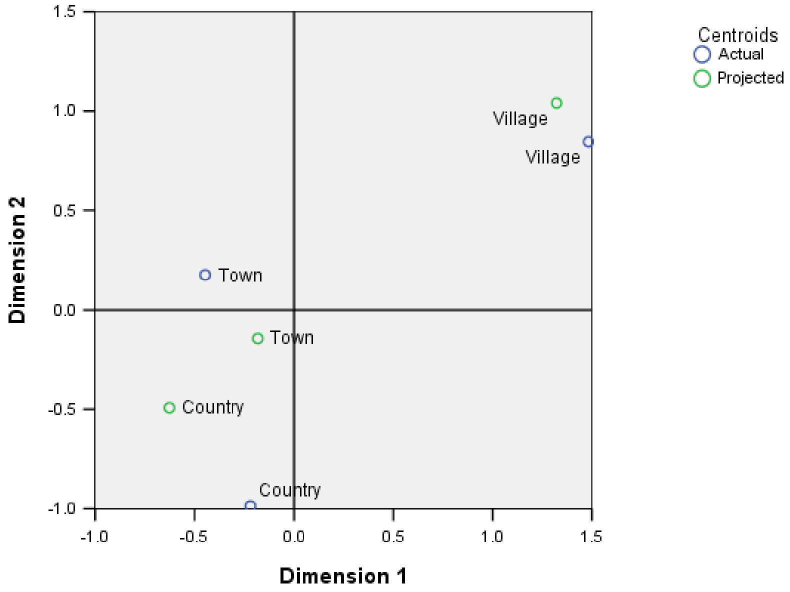
The actual centroids are projected onto the vectors that are defined by the component loadings. These vectors have been added to the centroid plots to aid in distinguishing the projected centroids from the actual centroids. The projected centroids fall into one of four quadrants formed by extending two perpendicular reference lines through the origin. The interpretation of the direction of single nominal, ordinal, or numerical variables is obtained from the position of the projected centroids. For example, the variable *Newspaper read most often* is specified as single nominal. The projected centroids show that *Volkskrant* and *NRC* are contrasted with *Telegraaf*.

Figure 11-24
Centroids and projected centroids for Age in years



The problem with *Age in years* is evident from the projected centroids. Treating *Age in years* as ordinal implies that the order of the age groups must be preserved. To satisfy this restriction, all age groups below age 45 are projected into the same point. Along the direction defined by *Age in years*, *Newspaper read most often*, and *Neighborhood preference*, there is no separation of the younger age groups. Such a finding suggests treating the variable as nominal.

Figure 11-25
Centroids and projected centroids for Neighborhood preference



To understand the relationships among variables, find out what the specific categories (values) are for clusters of categories in the centroid plots. The relationships among *Age in years*, *Newspaper read most often*, and *Neighborhood preference* can be described by looking at the upper right and lower left of the plots. In the upper right, the age groups are the older respondents; they read the newspaper Telegraaf and prefer living in a village. Looking at the lower left corner of each plot, you see that the younger to middle-aged respondents read the Volkskrant or NRC and want to live in the country or in a town. However, separating the younger groups is very difficult.

The same types of interpretations can be made about the other direction (*Music preferred*, *Marital status*, and *Pets owned*) by focusing on the upper left and the lower right of the centroid plots. In the upper left corner, we find that single people tend to have dogs and like new wave music. The married people and other categories for marital have cats; the former group prefers classical music, and the latter group does not like music.

An Alternative Analysis

The results of the analysis suggest that treating *Age in years* as ordinal does not appear appropriate. Although *Age in years* is measured at an ordinal level, its relationships with other variables are not monotonic. To investigate the effects of changing the optimal scaling level to single nominal, you may rerun the analysis.

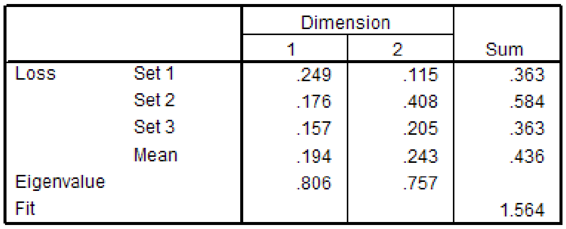
To Run the Analysis

- ▶ Recall the Nonlinear Canonical Correlation Analysis dialog box and navigate to the first set.
- ▶ Select *age* and click Define Range and Scale.

- In the Define Range and Scale dialog box, select Single nominal as the scaling range.
- Click Continue.
- In the Nonlinear Canonical Correlation Analysis dialog box, click OK.

The eigenvalues for a two-dimensional solution are 0.806 and 0.757, respectively, with a total fit of 1.564.

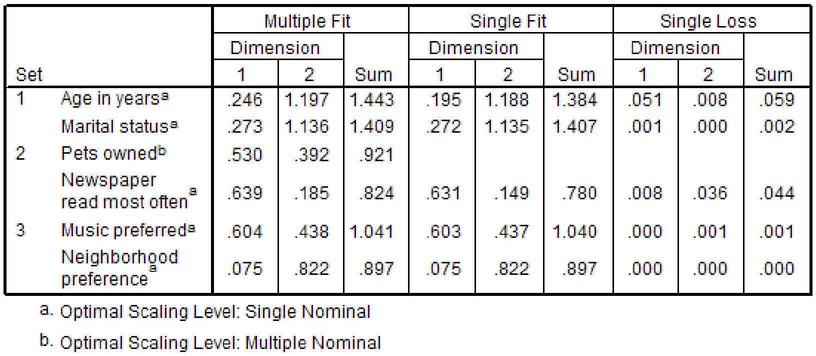
Figure 11-26
Eigenvalues for the two-dimensional solution



		Dimension		Sum
		1	2	
Loss	Set 1	.249	.115	.363
	Set 2	.176	.408	.584
	Set 3	.157	.205	.363
	Mean	.194	.243	.436
Eigenvalue		.806	.757	
Fit				1.564

The multiple fit and single fit tables show that *Age in years* is still a highly discriminating variable, as evidenced by the sum of the multiple fit values. In contrast to the earlier results, however, examination of the single fit values reveals the discrimination to be almost entirely along the second dimension.

Figure 11-27
Partitioning fit and loss



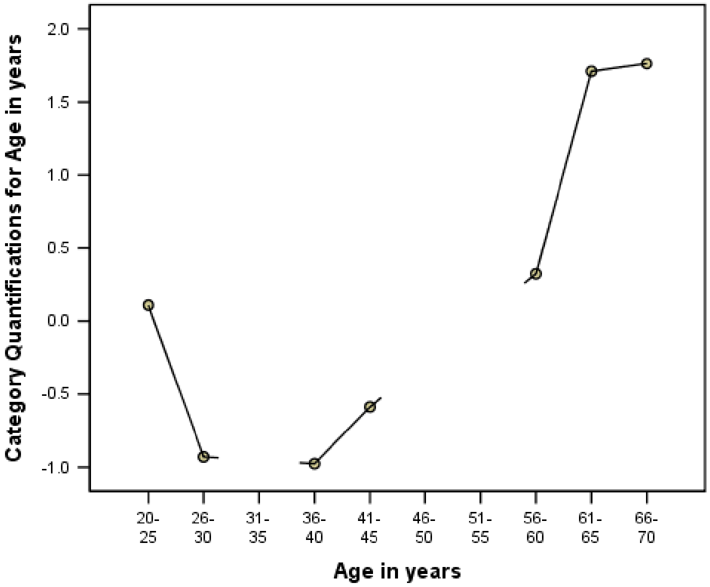
		Multiple Fit			Single Fit			Single Loss		
		Dimension		Sum	Dimension		Sum	Dimension		Sum
Set		1	2		1	2		1	2	
1	Age in years[a]	.246	1.197	1.443	.195	1.188	1.384	.051	.008	.059
	Marital status[a]	.273	1.136	1.409	.272	1.135	1.407	.001	.000	.002
2	Pets owned[b]	.530	.392	.921						
	Newspaper read most often[a]	.639	.185	.824	.631	.149	.780	.008	.036	.044
3	Music preferred[a]	.604	.438	1.041	.603	.437	1.040	.000	.001	.001
	Neighborhood preference[a]	.075	.822	.897	.075	.822	.897	.000	.000	.000

[a] Optimal Scaling Level: Single Nominal

[b] Optimal Scaling Level: Multiple Nominal

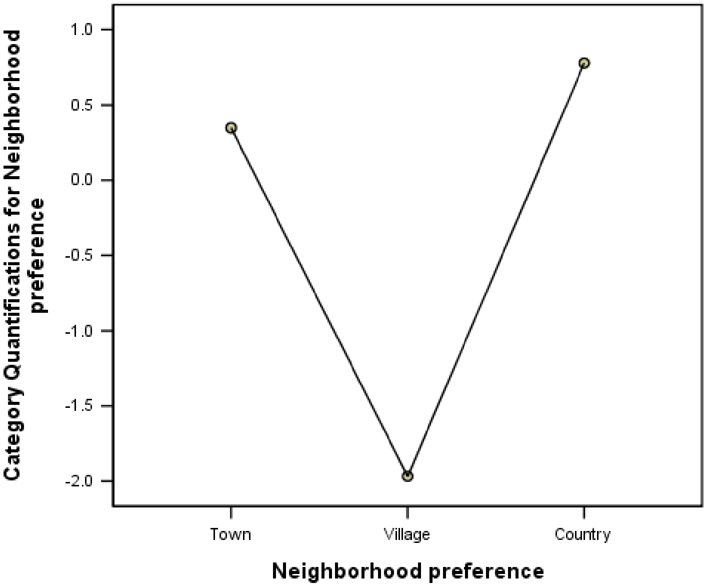
Turn to the transformation plot for *Age in years*. The quantifications for a nominal variable are unrestricted, so the nondecreasing trend that was displayed when *Age in years* was treated ordinally is no longer present. There is a decreasing trend until the age of 40 and an increasing trend thereafter, corresponding to a U-shaped (quadratic) relationship. The two older categories still receive similar scores, and subsequent analyses may involve combining these categories.

Figure 11-28
Transformation plot for Age in years (nominal)



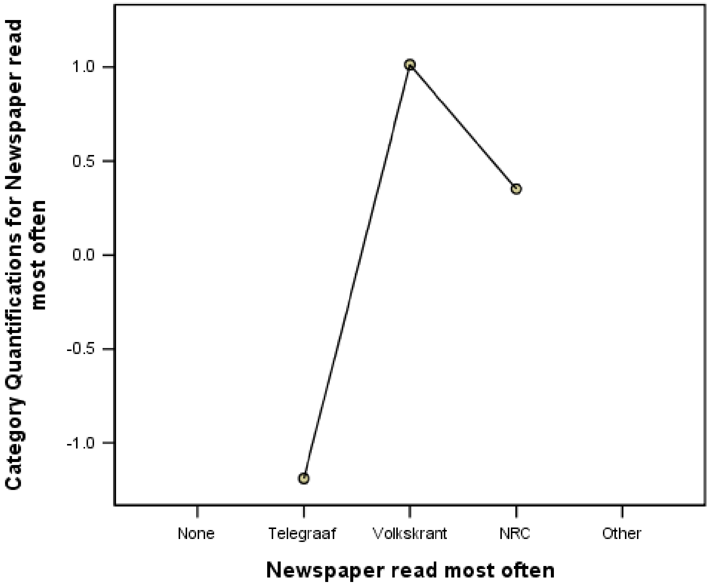
The transformation plot for *Neighborhood preference* is shown here. Treating *Age in years* as nominal does not affect the quantifications for *Neighborhood preference* to any significant degree. The middle category receives the smallest quantification, with the extreme categories receiving large positive values.

Figure 11-29
Transformation plot for Neighborhood preference (age nominal)



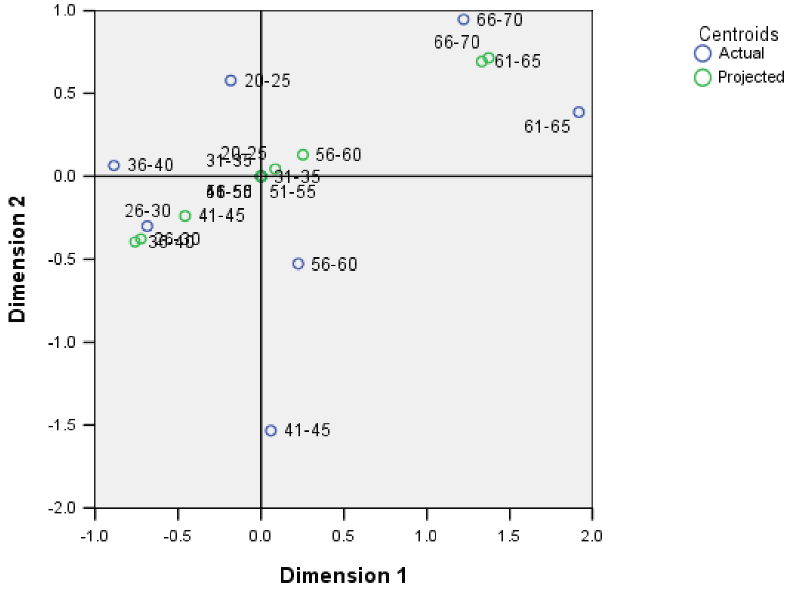
A change is found in the transformation plot for *Newspaper read most often*. Previously, an increasing trend was present in the quantifications, possibly suggesting an ordinal treatment for this variable. However, treating *Age in years* as nominal removes this trend from the news quantifications.

Figure 11-30
Transformation plot for Newspaper read most often (age nominal)



This plot is the centroid plot for *Age in years*. Notice that the categories do not fall in chronological order along the line joining the projected centroids. The 20–25 group is situated in the middle rather than at the end. The spread of the categories is much improved over the ordinal counterpart that was presented previously.

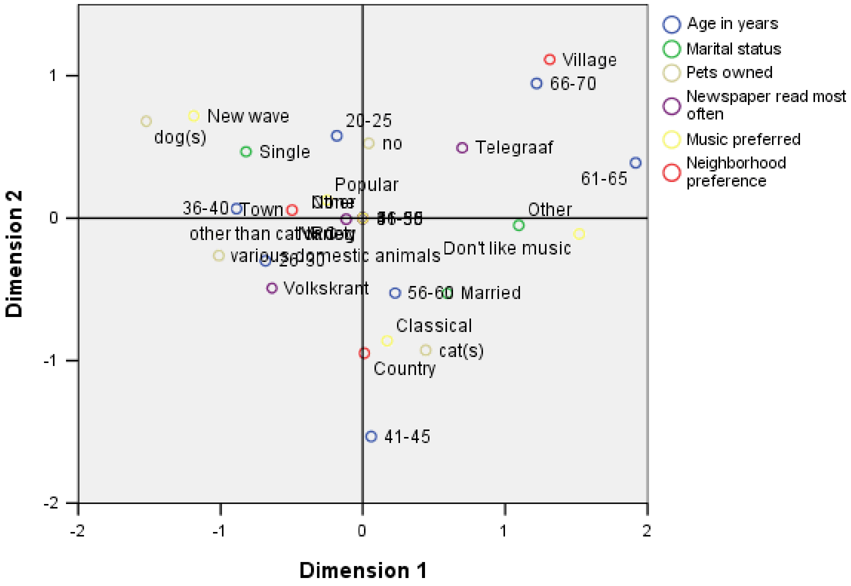
Figure 11-31
Centroids and projected centroids for Age in years (nominal)



Interpretation of the younger age groups is now possible from the centroid plot. The *Volkskrant* and *NRC* categories are also further apart than in the previous analysis, allowing for separate interpretations of each. The groups between the ages of 26 and 45 read the Volkskrant and prefer country living. The 20–25 and 56–60 age groups read the NRC; the former group prefers to live in a town, and the latter group prefers country living. The oldest groups read the Telegraaf and prefer village living.

Interpretation of the other direction (*Music preferred*, *Marital status*, and *Pets owned*) is basically unchanged from the previous analysis. The only obvious difference is that people with a marital status of *Other* have either cats or no pets.

Figure 11-32
Centroids labeled by variables (age nominal)



General Suggestions

After you have examined the initial results, you will probably want to refine your analysis by changing some of the specifications for the nonlinear canonical correlation analysis. Here are some tips for structuring your analysis:

- Create as many sets as possible. Put an important variable that you want to predict in a separate set by itself.
- Put variables that you consider predictors together in a single set. If there are many predictors, try to partition them into several sets.
- Put each multiple nominal variable in a separate set by itself.
- If variables are highly correlated to each other, and you don't want this relationship to dominate the solution, put those variables together in the same set.

Recommended Readings

See the following texts for more information about nonlinear canonical correlation analysis:

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association, 227–228.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd ed. Leiden: DSWO Press.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, 331–347.

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, 129–149.

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, 433–460.

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., and J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54–80.

Van der Burg, E., J. De Leeuw, and R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, 177–197.

Verboon, P., and I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, 485–507.

Chapter

12

Correspondence analysis

A **correspondence table** is any two-way table whose cells contain some measurement of correspondence between the rows and the columns. The measure of correspondence can be any indication of the similarity, affinity, confusion, association, or interaction between the row and column variables. A very common type of correspondence table is a crosstabulation, where the cells contain frequency counts.

Such tables can be obtained easily with the Crosstabs procedure. However, a crosstabulation does not always provide a clear picture of the nature of the relationship between the two variables. This is particularly true if the variables of interest are nominal (with no inherent order or rank) and contain numerous categories. Crosstabulation may tell you that the observed cell frequencies differ significantly from the expected values in a $10x9$ crosstabulation of *occupation* and *breakfast cereal*, but it may be difficult to discern which occupational groups have similar tastes or what those tastes are.

Correspondence Analysis allows you to examine the relationship between two nominal variables graphically in a multidimensional space. It computes row and column scores and produces plots based on the scores. Categories that are similar to each other appear close to each other in the plots. In this way, it is easy to see which categories of a variable are similar to each other or which categories of the two variables are related. The Correspondence Analysis procedure also allows you to fit supplementary points into the space defined by the active points.

If the ordering of the categories according to their scores is undesirable or counterintuitive, order restrictions can be imposed by constraining the scores for some categories to be equal. For example, suppose that you expect the variable *smoking behavior*, with categories *none*, *light*, *medium*, and *heavy*, to have scores that correspond to this ordering. However, if the analysis orders the categories *none*, *light*, *heavy*, and *medium*, constraining the scores for *heavy* and *medium* to be equal preserves the ordering of the categories in their scores.

The interpretation of correspondence analysis in terms of distances depends on the normalization method used. The Correspondence Analysis procedure can be used to analyze either the differences between categories of a variable or the differences between variables. With the default normalization, it analyzes the differences between the row and column variables.

The correspondence analysis algorithm is capable of many kinds of analyses. Centering the rows and columns and using chi-square distances corresponds to standard correspondence analysis. However, using alternative centering options combined with Euclidean distances allows for an alternative representation of a matrix in a low-dimensional space.

Three examples will be presented. The first employs a relatively small correspondence table and illustrates the concepts inherent in correspondence analysis. The second example demonstrates a practical marketing application. The final example uses a table of distances in a multidimensional scaling approach.

Normalization

Normalization is used to distribute the inertia over the row scores and column scores. Some aspects of the correspondence analysis solution, such as the singular values, the inertia per dimension, and the contributions, do not change under the various normalizations. The row and column scores and their variances are affected. Correspondence analysis has several ways to spread the inertia. The three most common include spreading the inertia over the row scores only, spreading the inertia over the column scores only, or spreading the inertia symmetrically over both the row scores and the column scores.

Row principal. In row principal normalization, the Euclidean distances between the row points approximate chi-square distances between the rows of the correspondence table. The row scores are the weighted average of the column scores. The column scores are standardized to have a weighted sum of squared distances to the centroid of 1. Since this method maximizes the distances between row categories, you should use row principal normalization if you are primarily interested in seeing how categories of the row variable differ from each other.

Column principal. On the other hand, you might want to approximate the chi-square distances between the columns of the correspondence table. In that case, the column scores should be the weighted average of the row scores. The row scores are standardized to have a weighted sum of squared distances to the centroid of 1. This method maximizes the distances between column categories and should be used if you are primarily concerned with how categories of the column variable differ from each other.

Symmetrical. You can also treat the rows and columns symmetrically. This normalization spreads inertia equally over the row and column scores. Note that neither the distances between the row points nor the distances between the column points are approximations of chi-square distances in this case. Use this method if you are primarily interested in the differences or similarities between the two variables. Usually, this is the preferred method to make biplots.

Principal. A fourth option is called principal normalization, in which the inertia is spread twice in the solution—once over the row scores and once over the column scores. You should use this method if you are interested in the distances between the row points and the distances between the column points separately but not in how the row and column points are related to each other. Biplots are not appropriate for this normalization option and are therefore not available if you have specified the principal normalization method.

Example: Perceptions of Coffee Brands

The previous example involved a small table of hypothetical data. Actual applications often involve much larger tables. In this example, you will use data pertaining to perceived images of six iced-coffee brands (Kennedy, Riquier, and Sharp, 1996). This dataset can be found in *coffee.sav*. For more information, see the topic Sample Files in Appendix A on p. 292.

For each of 23 iced-coffee image attributes, people selected all brands that were described by the attribute. The six brands are denoted as *AA*, *BB*, *CC*, *DD*, *EE*, and *FF* to preserve confidentiality.

Table 12-1
Iced-coffee attributes

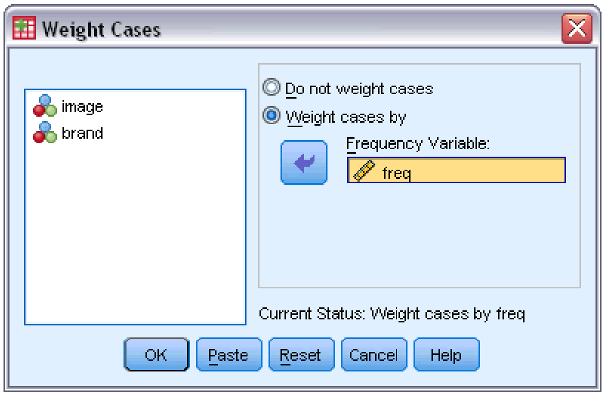
Image attribute	Label	Image attribute	Label
good hangover cure	*cure*	fattening brand	*fattening*
low fat/calorie brand	*low fat*	appeals to men	*men*
brand for children	*children*	South Australian brand	*South Australian*
working class brand	*working*	traditional/old fashioned brand	*traditional*
rich/sweet brand	*sweet*	premium quality brand	*premium*
unpopular brand	*unpopular*	healthy brand	*healthy*
brand for fat/ugly people	*ugly*	high caffeine brand	*caffeine*
very fresh	*fresh*	new brand	*new*
brand for yuppies	*yuppies*	brand for attractive people	*attractive*
nutritious brand	*nutritious*	tough brand	*tough*
brand for women	*women*	popular brand	*popular*
minor brand	*minor*		

Initially, you will focus on how the attributes are related to each other and how the brands are related to each other. Using principal normalization spreads the total inertia once over the rows and once over the columns. Although this prevents biplot interpretation, the distances between the categories for each variable can be examined.

Running the Analysis

▶ The setup of the data requires that the cases be weighted by the variable *freq*. To do this, from the menus choose:
Data > Weight Cases...

Figure 12-1
Weight Cases dialog box



○ Do not weight cases
● Weight cases by
Frequency Variable: freq

Current Status: Weight cases by freq

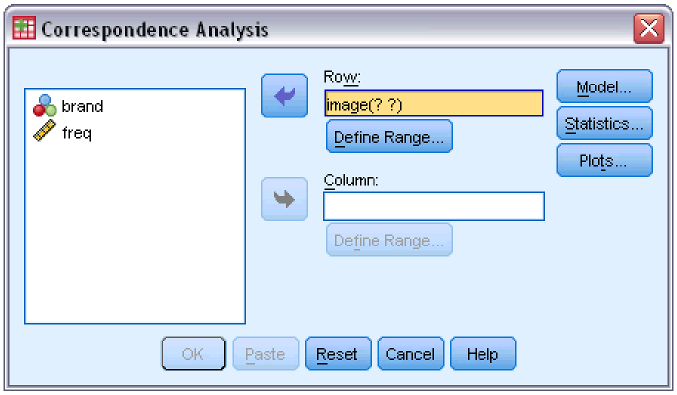
▶ Weight cases by *freq*.

▶ Click OK.

- ▶ To obtain an initial solution in five dimensions with principal normalization, from the menus choose:

  Analyze > Dimension Reduction > Correspondence Analysis...

Figure 12-2
Correspondence Analysis dialog box

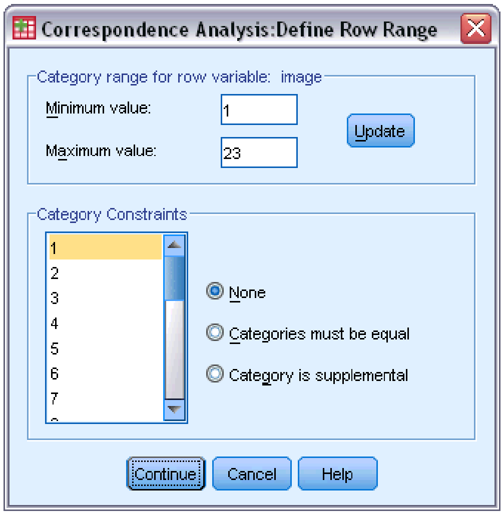


Row: image(? ?)

Column:

- ▶ Select *image* as the row variable.
- ▶ Click Define Range.

Figure 12-3
Define Row Range dialog box


Correspondence Analysis:Define Row Range
Category range for row variable: image
Update
Category Constraints
1
2
3
4
5
6
7
Continue
Cancel
Help

Minimum value: 1

Maximum value: 23

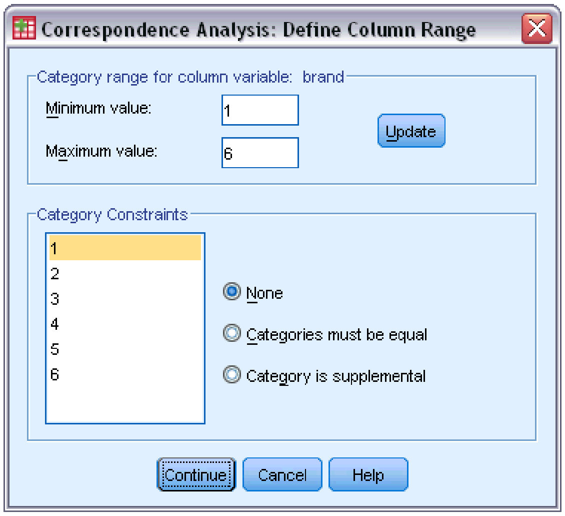
● None

○ Categories must be equal

○ Category is supplemental

- ▶ Type 1 as the minimum value.
- ▶ Type 23 as the maximum value.
- ▶ Click Update.
- ▶ Click Continue.
- ▶ Select *brand* as the column variable.
- ▶ Click Define Range in the Correspondence Analysis dialog box.

Figure 12-4
Define Column Range dialog box



Correspondence Analysis: Define Column Range

Category range for column variable: brand

Minimum value: 1

Maximum value: 6

Category Constraints

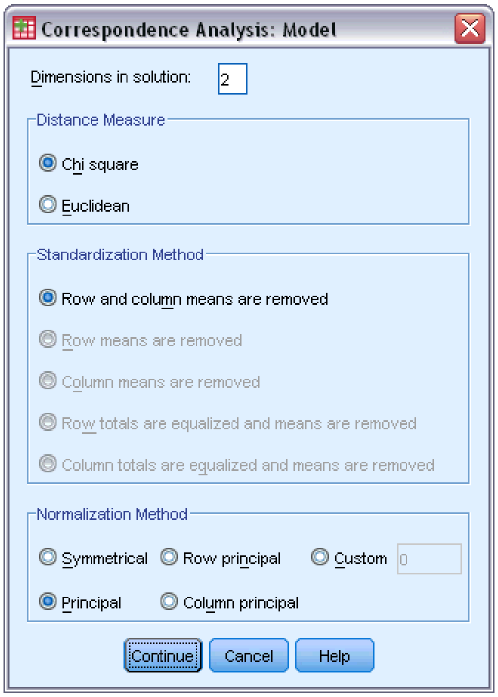
● None

○ Categories must be equal

○ Category is supplemental

- ▶ Type 1 as the minimum value.
- ▶ Type 6 as the maximum value.
- ▶ Click Update.
- ▶ Click Continue.
- ▶ Click Model in the Correspondence Analysis dialog box.

Figure 12-5
Model dialog box



Correspondence Analysis: Model

Dimensions in solution: 2

Distance Measure

● Chi square

○ Euclidean

Standardization Method

● Row and column means are removed

○ Row means are removed

○ Column means are removed

○ Row totals are equalized and means are removed

○ Column totals are equalized and means are removed

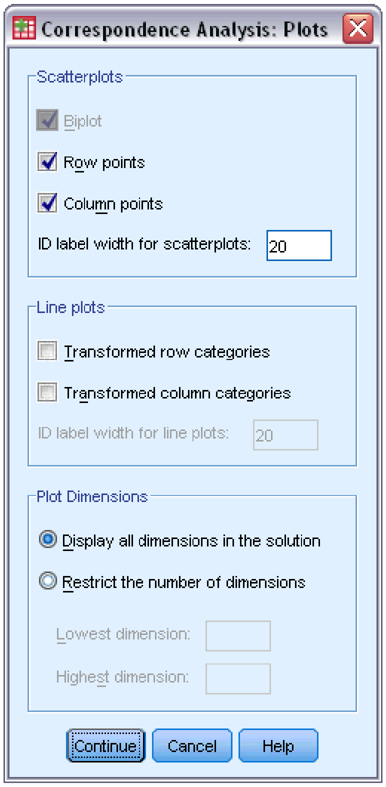
Normalization Method

○ Symmetrical ○ Row principal ○ Custom 0

● Principal ○ Column principal

- ▶ Select Principal as the normalization method.
- ▶ Click Continue.
- ▶ Click Plots in the Correspondence Analysis dialog box.

Figure 12-6
Plots dialog box



Correspondence Analysis: Plots

Scatterplots

☑ Biplot

☑ Row points

☑ Column points

ID label width for scatterplots: 20

Line plots

□ Transformed row categories

□ Transformed column categories

ID label width for line plots: 20

Plot Dimensions

● Display all dimensions in the solution

○ Restrict the number of dimensions

Lowest dimension:

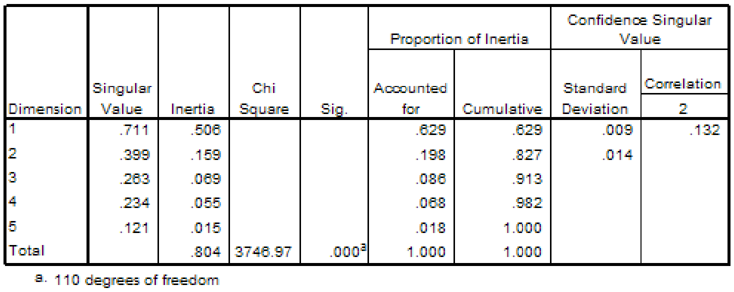
Highest dimension:

- ▶ Select Row points and Column points in the Scatterplots group.
- ▶ Click Continue.
- ▶ Click OK in the Correspondence Analysis dialog box.

Dimensionality

The inertia per dimension shows the decomposition of the total inertia along each dimension. Two dimensions account for 83% of the total inertia. Adding a third dimension adds only 8.6% to the accounted-for inertia. Thus, you elect to use a two-dimensional representation.

Figure 12-7
Inertia per dimension



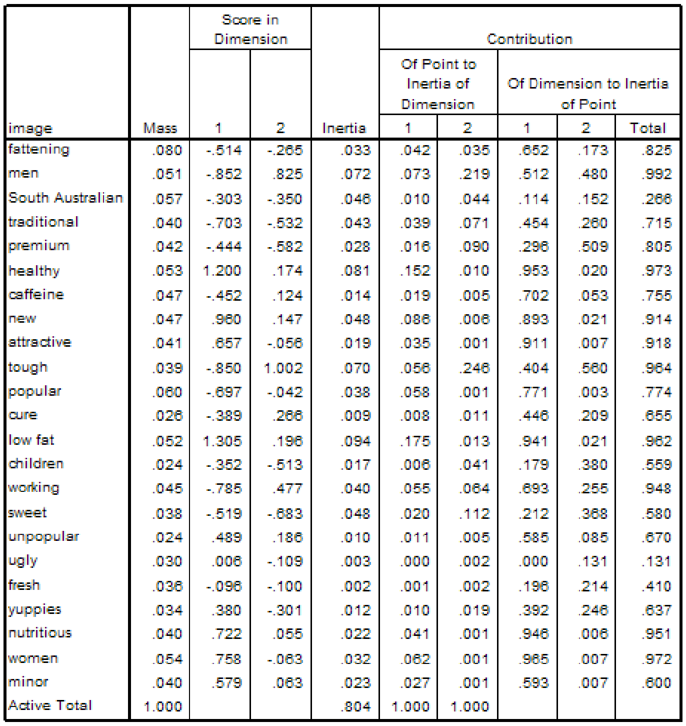
					Proportion of Inertia		Confidence Singular Value	
								Correlation
Dimension	Singular Value	Inertia	Chi Square	Sig.	Accounted for	Cumulative	Standard Deviation	2
1	.711	.506			.629	.629	.009	.132
2	.399	.159			.198	.827	.014	
3	.263	.069			.086	.913		
4	.234	.055			.068	.982		
5	.121	.015			.018	1.000		
Total		.804	3746.97	.000[a]	1.000	1.000		

[a] 110 degrees of freedom

Contributions

The row points overview shows the contributions of the row points to the inertia of the dimensions and the contributions of the dimensions to the inertia of the row points. If all points contributed equally to the inertia, the contributions would be 0.043. *Healthy* and *low fat* both contribute a substantial portion to the inertia of the first dimension. *Men* and *tough* contribute the largest amounts to the inertia of the second dimension. Both *ugly* and *fresh* contribute very little to either dimension.

Figure 12-8
Attribute contributions

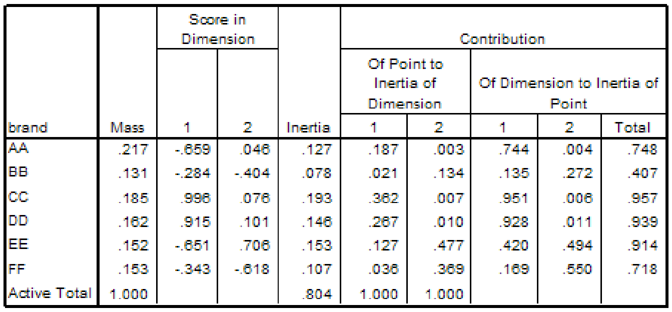


		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
image	Mass	1	2	Inertia	1	2	1	2	Total
fattening	.080	-.514	-.265	.033	.042	.035	.652	.173	.825
men	.051	-.852	.825	.072	.073	.219	.512	.480	.992
South Australian	.057	-.303	-.350	.046	.010	.044	.114	.152	.266
traditional	.040	-.703	-.532	.043	.039	.071	.454	.260	.715
premium	.042	-.444	-.582	.028	.016	.090	.296	.509	.805
healthy	.053	1.200	.174	.081	.152	.010	.953	.020	.973
caffeine	.047	-.452	.124	.014	.019	.005	.702	.053	.755
new	.047	.960	.147	.048	.086	.006	.893	.021	.914
attractive	.041	.657	-.056	.019	.035	.001	.911	.007	.918
tough	.039	-.850	1.002	.070	.056	.246	.404	.560	.964
popular	.060	-.697	-.042	.038	.058	.001	.771	.003	.774
cure	.026	-.389	.266	.009	.008	.011	.446	.209	.655
low fat	.052	1.305	.196	.094	.175	.013	.941	.021	.962
children	.024	-.352	-.513	.017	.006	.041	.179	.380	.559
working	.045	-.785	.477	.040	.055	.064	.693	.255	.948
sweet	.038	-.519	-.683	.048	.020	.112	.212	.368	.580
unpopular	.024	.489	.186	.010	.011	.005	.585	.085	.670
ugly	.030	.006	-.109	.003	.000	.002	.000	.131	.131
fresh	.036	-.096	-.100	.002	.001	.002	.196	.214	.410
yuppies	.034	.380	-.301	.012	.010	.019	.392	.246	.637
nutritious	.040	.722	.055	.022	.041	.001	.946	.006	.951
women	.054	.758	-.063	.032	.062	.001	.965	.007	.972
minor	.040	.579	.063	.023	.027	.001	.593	.007	.600
Active Total	1.000			.804	1.000	1.000			

Two dimensions contribute a large amount to the inertia for most row points. The large contributions of the first dimension to *healthy*, *new*, *attractive*, *low fat*, *nutritious*, and *women* indicate that these points are very well represented in one dimension. Consequently, the higher dimensions contribute little to the inertia of these points, which will lie very near the horizontal axis. The second dimension contributes most to *men*, *premium*, and *tough*. Both dimensions contribute very little to the inertia for *South Australian* and *ugly*, so these points are poorly represented.

The column points overview displays the contributions involving the column points. Brands *CC* and *DD* contribute the most to the first dimension, whereas *EE* and *FF* explain a large amount of the inertia for the second dimension. *AA* and *BB* contribute very little to either dimension.

Figure 12-9
Brand contributions



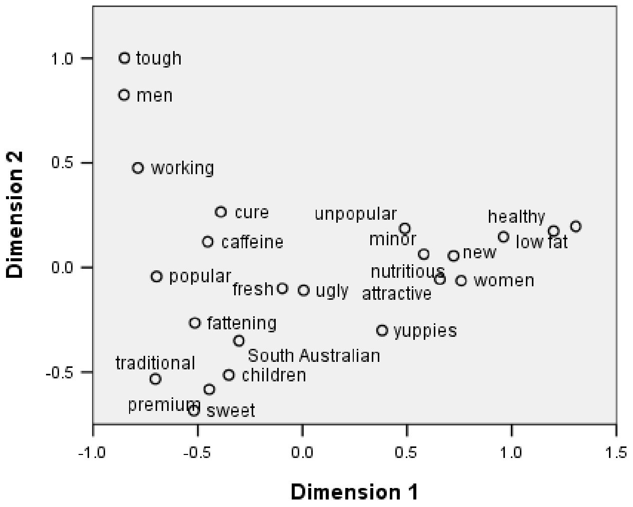
		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
brand	Mass	1	2	Inertia	1	2	1	2	Total
AA	.217	-.659	.046	.127	.187	.003	.744	.004	.748
BB	.131	-.284	-.404	.078	.021	.134	.135	.272	.407
CC	.185	.996	.076	.193	.362	.007	.951	.006	.957
DD	.162	.915	.101	.146	.267	.010	.928	.011	.939
EE	.152	-.651	.706	.153	.127	.477	.420	.494	.914
FF	.153	-.343	-.618	.107	.036	.369	.169	.550	.718
Active Total	1.000			.804	1.000	1.000			

In two dimensions, all brands but *BB* are well represented. *CC* and *DD* are represented well in one dimension. The second dimension contributes the largest amounts for *EE* and *FF*. Notice that *AA* is represented well in the first dimension but does not have a very high contribution to that dimension.

Plots

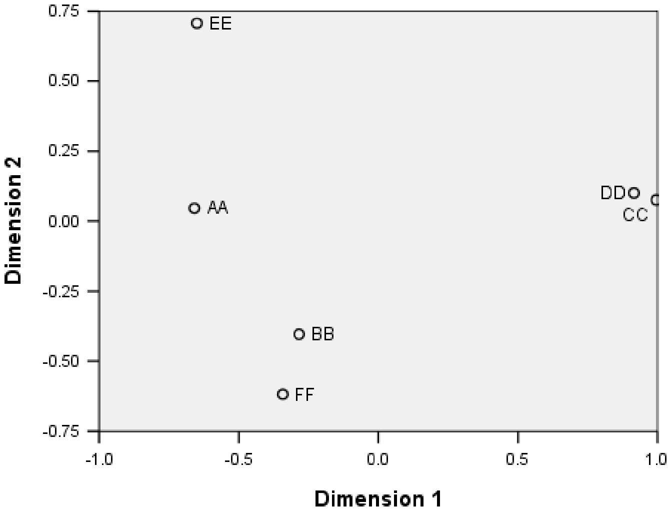
The row points plot shows that *fresh* and *ugly* are both very close to the origin, indicating that they differ little from the average row profile. Three general classifications emerge. Located in the upper left of the plot, *tough*, *men*, and *working* are all similar to each other. The lower left contains *sweet*, *fattening*, *children*, and *premium*. In contrast, *healthy*, *low fat*, *nutritious*, and *new* cluster on the right side of the plot.

Figure 12-10
Plot of image attributes (principal normalization)



Notice in the column points plot that all brands are far from the origin, so no brand is similar to the overall centroid. Brands *CC* and *DD* group together at the right, whereas brands *BB* and *FF* cluster in the lower half of the plot. Brands *AA* and *EE* are not similar to any other brand.

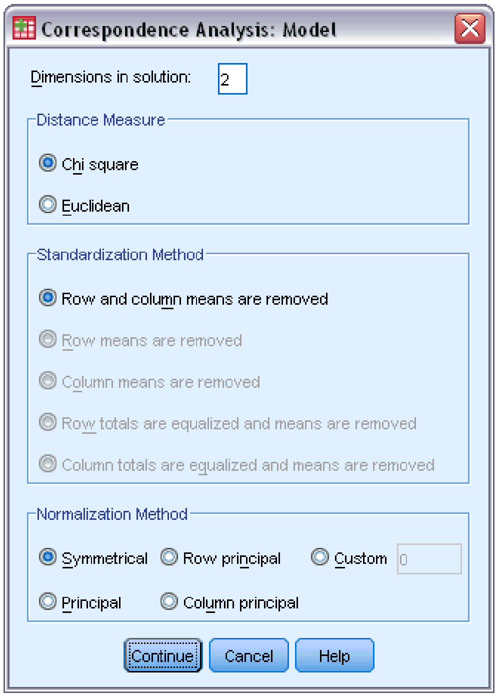
Figure 12-11
Plot of brands (principal normalization)



Symmetrical Normalization

How are the brands related to the image attributes? Principal normalization cannot address these relationships. To focus on how the variables are related to each other, use symmetrical normalization. Rather than spread the inertia twice (as in principal normalization), symmetrical normalization divides the inertia equally over both the rows and columns. Distances between categories for a single variable cannot be interpreted, but distances between the categories for different variables are meaningful.

Figure 12-12
Model dialog box



Correspondence Analysis: Model

Dimensions in solution: 2

Distance Measure
● Chi square
○ Euclidean

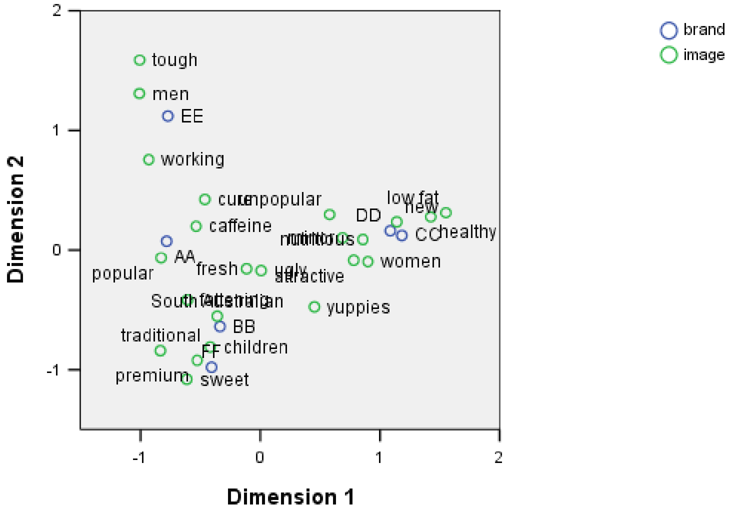
Standardization Method
● Row and column means are removed
○ Row means are removed
○ Column means are removed
○ Row totals are equalized and means are removed
○ Column totals are equalized and means are removed

Normalization Method
● Symmetrical ○ Row principal ○ Custom 0
○ Principal ○ Column principal

- ▶ To produce the following solution with symmetrical normalization, recall the Correspondence Analysis dialog box and click Model.
- ▶ Select Symmetrical as the normalization method.
- ▶ Click Continue.
- ▶ Click OK in the Correspondence Analysis dialog box.

In the upper left of the resulting biplot, brand *EE* is the only tough, working brand and appeals to men. Brand *AA* is the most popular and also viewed as the most highly caffeinated. The sweet, fattening brands include *BB* and *FF*. Brands *CC* and *DD*, while perceived as new and healthy, are also the most unpopular.

Figure 12-13
Biplot of the brands and the attributes (symmetrical normalization)



For further interpretation, you can draw a line through the origin and the two image attributes *men* and *yuppies*, and project the brands onto this line. The two attributes are opposed to each other, indicating that the association pattern of brands for *men* is reversed compared to the pattern for *yuppies*. That is, men are most frequently associated with brand *EE* and least frequently with brand *CC*, whereas yuppies are most frequently associated with brand *CC* and least frequently with brand *EE*.

Recommended readings

See the following texts for more information on correspondence analysis:

Fisher, R. A. 1938. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, 422–429.

Gilula, Z., and S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, 760–771.

Chapter

13

Multiple Correspondence Analysis

The purpose of multiple correspondence analysis, also known as homogeneity analysis, is to find quantifications that are optimal in the sense that the categories are separated from each other as much as possible. This implies that objects in the same category are plotted close to each other and objects in different categories are plotted as far apart as possible. The term **homogeneity** also refers to the fact that the analysis will be most successful when the variables are homogeneous; that is, when they partition the objects into clusters with the same or similar categories.

Example: Characteristics of Hardware

To explore how multiple correspondence analysis works, you will use data from Hartigan (Hartigan, 1975), which can be found in *screws.sav*. For more information, see the topic Sample Files in Appendix A on p. 292. This dataset contains information on the characteristics of screws, bolts, nuts, and tacks. The following table shows the variables, along with their variable labels, and the value labels assigned to the categories of each variable in the Hartigan hardware dataset.

Table 13-1
Hartigan hardware dataset

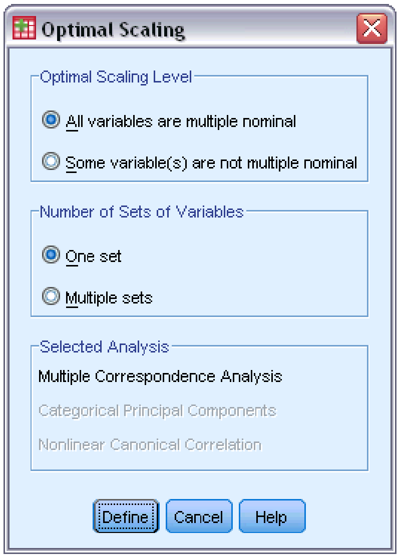
Variable name	**Variable label**	**Value label**
thread	*Thread*	*Yes_Thread*, *No_Thread*
head	*Head form*	*Flat*, *Cup*, *Cone*, *Round*, *Cylinder*
indhead	*Indentation of head*	*None*, *Star*, *Slit*
bottom	*Bottom shape*	*sharp*, *flat*
length	*Length in half inches*	*1/2_in*, *1_in*, *1_1/2_ in*, *2_in*, *2_1/2_in*
brass	*Brass*	*Yes_Br*, *Not_Br*
object	*Object*	*tack, nail1, nail2, nail3, nail4, nail5, nail6, nail7, nail8, screw1, screw2, screw3, screw4, screw5, bolt1, bolt2, bolt3, bolt4, bolt5, bolt6, tack1, tack2, nailb, screwb*

Running the Analysis

▶ To obtain a Multiple Correspondence Analysis, from the menus choose:

Analyze > Dimension Reduction > Optimal Scaling...

Figure 13-1
Optimal Scaling dialog box



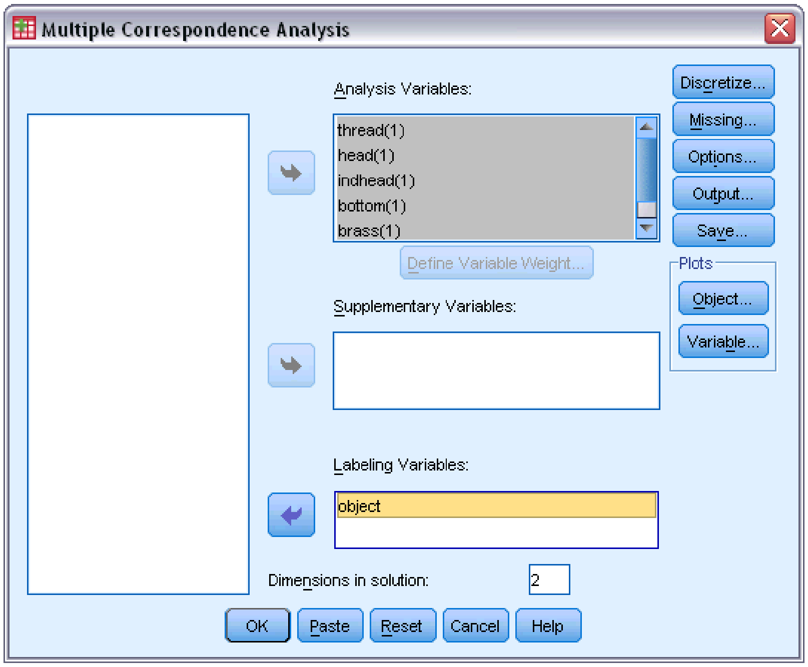
Optimal Scaling Level
● All variables are multiple nominal
○ Some variable(s) are not multiple nominal

Number of Sets of Variables
● One set
○ Multiple sets

Selected Analysis
Multiple Correspondence Analysis
Categorical Principal Components
Nonlinear Canonical Correlation

- ▶ Make sure All variables are multiple nominal and One set are selected, and click Define.

Figure 13-2
Multiple Correspondence Analysis dialog box



Analysis Variables:
thread(1)
head(1)
indhead(1)
bottom(1)
brass(1)

Supplementary Variables:

Labeling Variables:
object

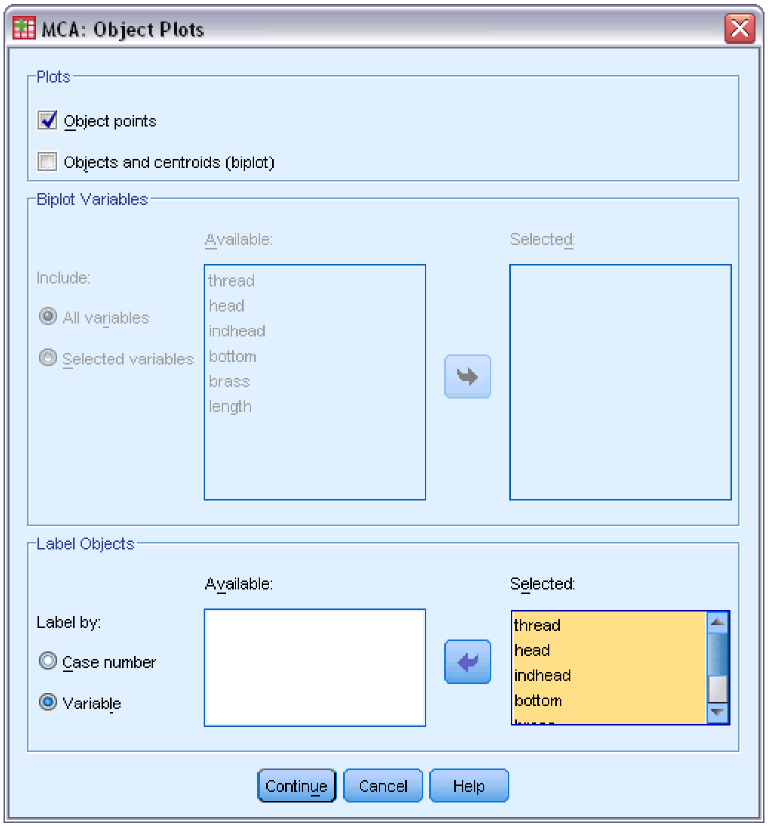
Dimensions in solution: 2

- ▶ Select *Thread* through *Length in half-inches* as analysis variables.

- ▶ Select *object* as a labeling variable.

- ▶ Click Object in the Plots group.

Figure 13-3
Object Plots dialog box



MCA: Object Plots

Plots

☑ Object points

□ Objects and centroids (biplot)

Biplot Variables

Include:

● All variables

○ Selected variables

Available: thread, head, indhead, bottom, brass, length

Selected:

Label Objects

Label by:

○ Case number

● Variable

Available:

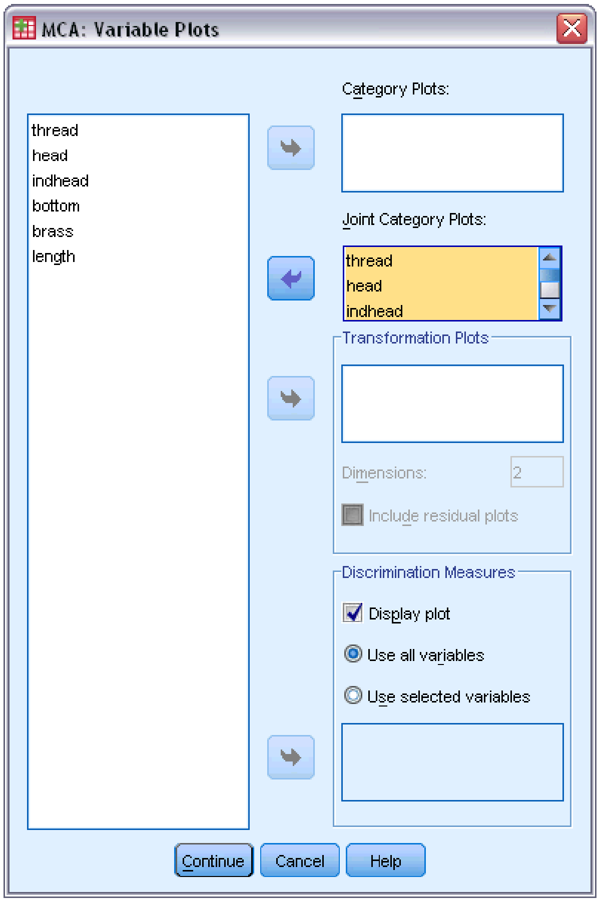
Selected: thread, head, indhead, bottom

▶ Choose to label objects by Variable.

▶ Select *thread* through *object* as labeling variables.

▶ Click Continue, and then click Variable in the Plots group of the Multiple Correspondence Analysis dialog box.

Figure 13-4
Variable Plots dialog box



thread
head
indhead
bottom
brass
length

Joint Category Plots: thread, head, indhead

Dimensions: 2

□ Include residual plots

☑ Display plot

● Use all variables

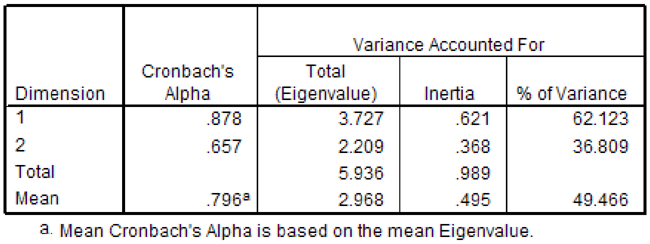
○ Use selected variables

- ▶ Choose to produce a joint category plot for *thread* through *length*.
- ▶ Click Continue.
- ▶ Click OK in the Multiple Correspondence Analysis dialog box.

Model Summary

Homogeneity analysis can compute a solution for several dimensions. The maximum number of dimensions equals either the number of categories minus the number of variables with no missing data or the number of observations minus one, whichever is smaller. However, you should rarely use the maximum number of dimensions. A smaller number of dimensions is easier to interpret, and after a certain number of dimensions, the amount of additional association accounted for becomes negligible. A one-, two-, or three-dimensional solution in homogeneity analysis is very common.

Figure 13-5
Model summary



		Variance Accounted For		
Dimension	Cronbach's Alpha	Total (Eigenvalue)	Inertia	% of Variance
1	.878	3.727	.621	62.123
2	.657	2.209	.368	36.809
Total		5.936	.989	
Mean	.796[a]	2.968	.495	49.466

[a] Mean Cronbach's Alpha is based on the mean Eigenvalue.

Nearly all of the variance in the data is accounted for by the solution, 62.1% by the first dimension and 36.8% by the second.

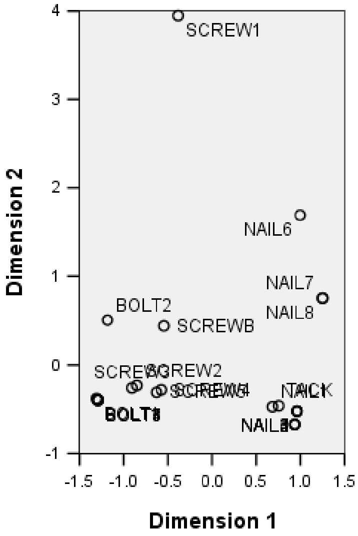
The two dimensions together provide an interpretation in terms of distances. If a variable discriminates well, the objects will be close to the categories to which they belong. Ideally, objects in the same category will be close to each other (that is, they should have similar scores), and categories of different variables will be close if they belong to the same objects (that is, two objects that have similar scores for one variable should also score close to each other for the other variables in the solution).

Object Scores

After examining the model summary, you should look at the object scores. You can specify one or more variables to label the object scores plot. Each labeling variable produces a separate plot labeled with the values of that variable. We'll take a look at the plot of object scores labeled by the variable object. This is just a case-identification variable and was not used in any computations.

The distance from an object to the origin reflects variation from the "average" response pattern. This average response pattern corresponds to the most frequent category for each variable. Objects with many characteristics corresponding to the most frequent categories lie near the origin. In contrast, objects with unique characteristics are located far from the origin.

Figure 13-6
Object scores plot labeled by object



Examining the plot, you see that the first dimension (the horizontal axis) discriminates the screws and bolts (which have threads) from the nails and tacks (which don't have threads). This is easily seen on the plot since screws and bolts are on one end of the horizontal axis and tacks and nails are on the other. To a lesser extent, the first dimension also separates the bolts (which have flat bottoms) from all the others (which have sharp bottoms).

The second dimension (the vertical axis) seems to separate *SCREW1* and *NAIL6* from all other objects. What *SCREW1* and *NAIL6* have in common are their values on variable length—they are the longest objects in the data. Moreover, *SCREW1* lies much farther from the origin than the other objects, suggesting that, taken as a whole, many of the characteristics of this object are not shared by the other objects.

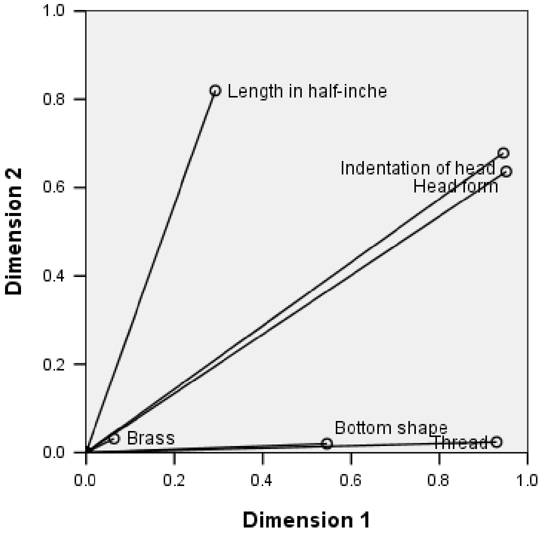
The object scores plot is particularly useful for spotting outliers. *SCREW1* might be considered an outlier. Later, we'll consider what happens if you drop this object.

Discrimination Measures

Before examining the rest of the object scores plots, let's see if the discrimination measures agree with what we've said so far. For each variable, a discrimination measure, which can be regarded as a squared component loading, is computed for each dimension. This measure is also the variance of the quantified variable in that dimension. It has a maximum value of 1, which is achieved if the object scores fall into mutually exclusive groups and all object scores within a category are identical. (*Note*: This measure may have a value greater than 1 if there are missing data.) Large discrimination measures correspond to a large spread among the categories of the variable and, consequently, indicate a high degree of discrimination between the categories of a variable along that dimension.

The average of the discrimination measures for any dimension equals the percentage of variance accounted for that dimension. Consequently, the dimensions are ordered according to average discrimination. The first dimension has the largest average discrimination, the second dimension has the second largest average discrimination, and so on, for all dimensions in the solution.

Figure 13-7
Plot of discrimination measures



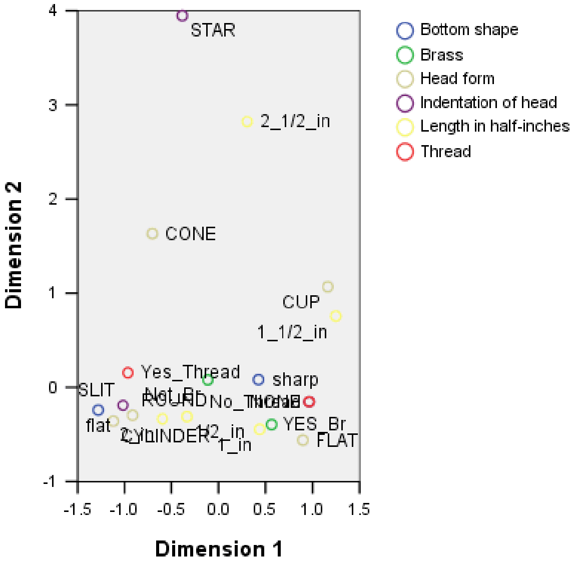
As noted on the object scores plot, the discrimination measures plot shows that the first dimension is related to variables *Thread* and *Bottom shape*. These variables have large discrimination measures on the first dimension and small discrimination measures on the second dimension. Thus, for both of these variables, the categories are spread far apart along the first dimension only. *Length in half-inches* has a large value on the second dimension but a small value on the first dimension. As a result, *length* is closest to the second dimension, agreeing with the observation from the object scores plot that the second dimension seems to separate the longest objects from the rest. *Indentation of head* and *Head form* have relatively large values on both dimensions, indicating discrimination in both the first and second dimensions. The variable *Brass*, located very close to the origin, does not discriminate at all in the first two dimensions. This makes sense, since all of the objects can be made of brass or not made of brass.

Category Quantifications

Recall that a discrimination measure is the variance of the quantified variable along a particular dimension. The discrimination measures plot contains these variances, indicating which variables discriminate along which dimension. However, the same variance could correspond to all of the categories being spread moderately far apart or to most of the categories being close together, with a few categories differing from this group. The discrimination plot cannot differentiate between these two conditions.

Category quantification plots provide an alternative method of displaying discrimination of variables that can identify category relationships. In this plot, the coordinates of each category on each dimension are displayed. Thus, you can determine which categories are similar for each variable.

Figure 13-8
Category quantifications



Length in half-inches has five categories, three of which group together near the top of the plot. The remaining two categories are in the lower half of the plot, with the *2_1/2_in* category very far from the group. The large discrimination for length along dimension 2 is a result of this one category being very different from the other categories of length. Similarly, for *Head form*, the category *STAR* is very far from the other categories and yields a large discrimination measure along the second dimension. These patterns cannot be illustrated in a plot of discrimination measures.

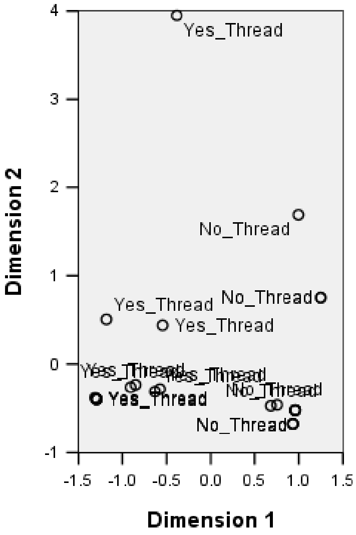
The spread of the category quantifications for a variable reflects the variance and thus indicates how well that variable is discriminated in each dimension. Focusing on dimension 1, the categories for *Thread* are far apart. However, along dimension 2, the categories for this variable are very close. Thus, *Thread* discriminates better in dimension 1 than in dimension 2. In contrast, the categories for *Head form* are spread far apart along both dimensions, suggesting that this variable discriminates well in both dimensions.

In addition to determining the dimensions along which a variable discriminates and how that variable discriminates, the category quantification plot also compares variable discrimination. A variable with categories that are far apart discriminates better than a variable with categories that are close together. For example, along dimension 1, the two categories of *Brass* are much closer to each other than the two categories of *Thread*, indicating that *Thread* discriminates better than *Brass* along this dimension. However, along dimension 2, the distances are very similar, suggesting that these variables discriminate to the same degree along this dimension. The discrimination measures plot discussed previously identifies these same relationships by using variances to reflect the spread of the categories.

A More Detailed Look at Object Scores

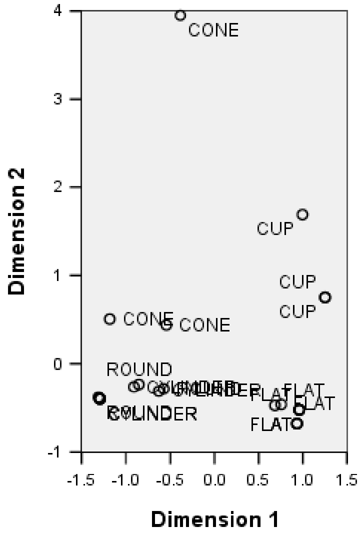
A greater insight into the data can be gained by examining the object scores plots labeled by each variable. Ideally, similar objects should form exclusive groups, and these groups should be far from each other.

Figure 13-9
Object scores labeled with Thread



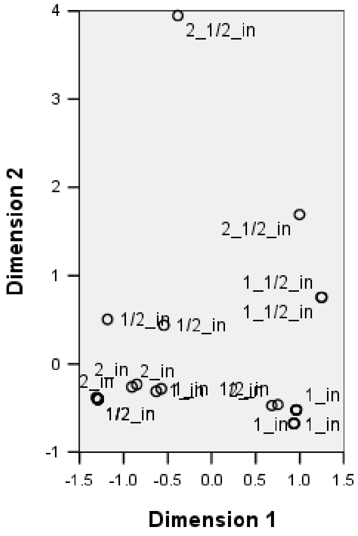
The plot labeled with *Thread* shows that the first dimension separates *Yes_Thread* and *No_Thread* perfectly. All of the objects with threads have negative object scores, whereas all of the nonthreaded objects have positive scores. Although the two categories do not form compact groups, the perfect differentiation between the categories is generally considered a good result.

Figure 13-10
Object scores labeled with Head form



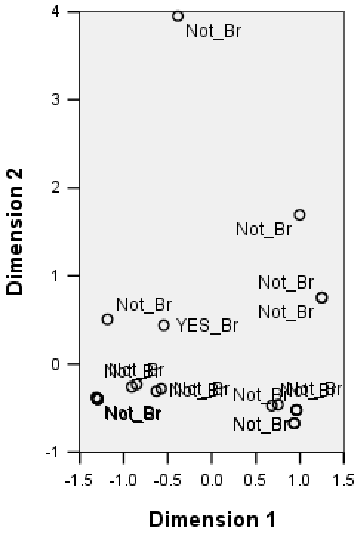
The plot labeled with *Head form* shows that this variable discriminates in both dimensions. The *FLAT* objects group together in the lower right corner of the plot, whereas the *CUP* objects group together in the upper right. *CONE* objects all lie in the upper left. However, these objects are more spread out than the other groups and, thus, are not as homogeneous. Finally, *CYLINDER* objects cannot be separated from *ROUND* objects, both of which lie in the lower left corner of the plot.

Figure 13-11
Object scores labeled with Length in half-inches



The plot labeled with *Length in half-inches* shows that this variable does not discriminate in the first dimension. Its categories display no grouping when projected onto a horizontal line. However, *Length in half-inches* does discriminate in the second dimension. The shorter objects correspond to positive scores, and the longer objects correspond to large negative scores.

Figure 13-12
Object scores labeled with Brass



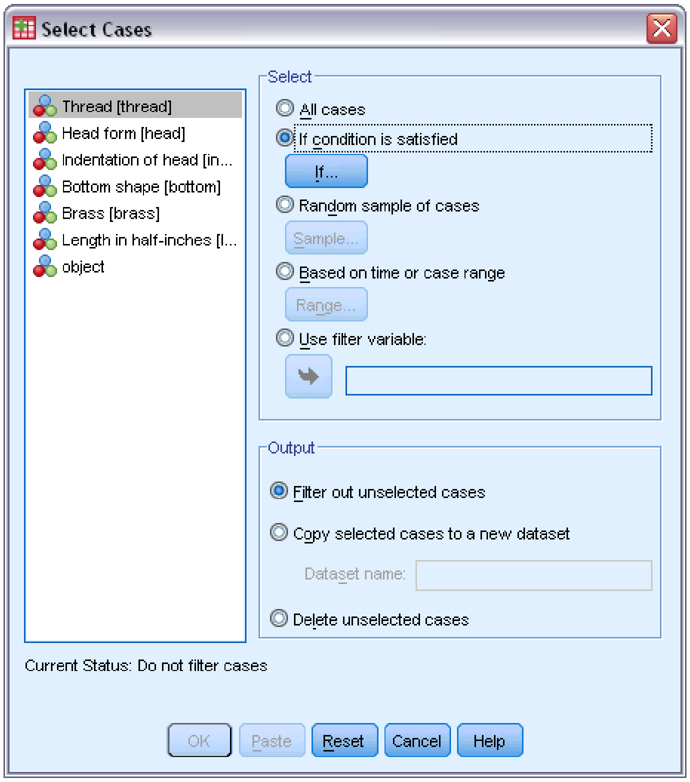
The plot labeled with *Brass* shows that this variable has categories that cannot be separated very well in the first or second dimensions. The object scores are widely spread throughout the space. The brass objects cannot be differentiated from the nonbrass objects.

Omission of Outliers

In homogeneity analysis, outliers are objects that have too many unique features. As noted earlier, *SCREW1* might be considered an outlier.

To delete this object and run the analysis again, from the menus choose:
Data > Select Cases...

Figure 13-13
Select Cases dialog box



Select Cases

Thread [thread]
Head form [head]
Indentation of head [in...
Bottom shape [bottom]
Brass [brass]
Length in half-inches [l...
object

Select

○ All cases
● If condition is satisfied
If...
○ Random sample of cases
Sample...
○ Based on time or case range
Range...
○ Use filter variable:

Output

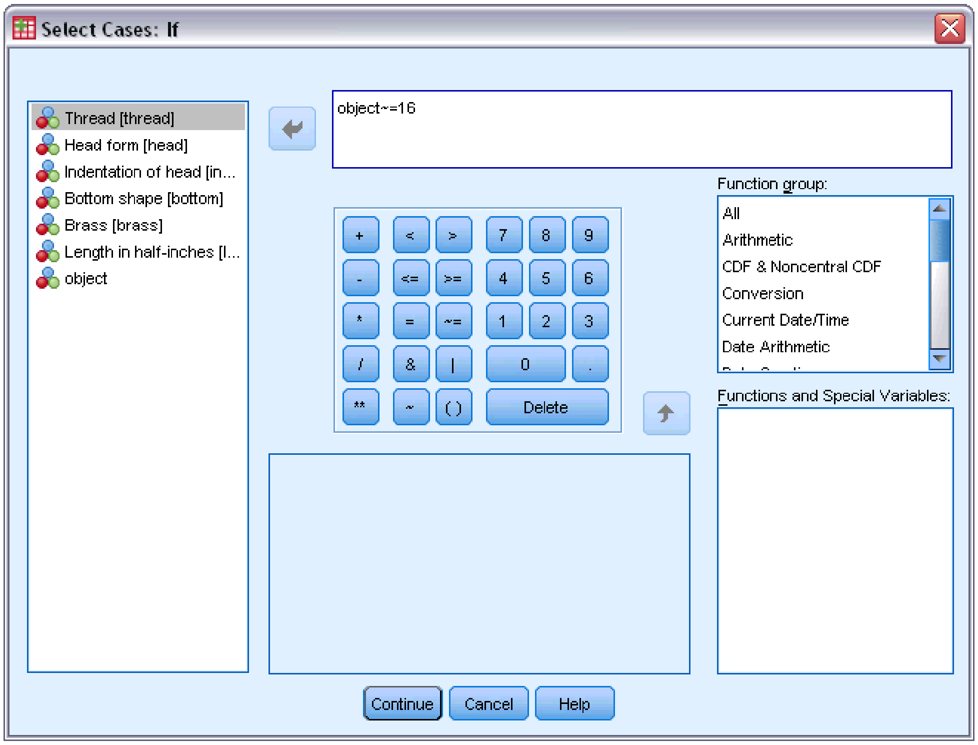
● Filter out unselected cases
○ Copy selected cases to a new dataset
Dataset name:
○ Delete unselected cases

Current Status: Do not filter cases

OK Paste Reset Cancel Help

- ▶ Select If condition is satisfied.
- ▶ Click If.

Figure 13-14
If dialog box



Select Cases: If

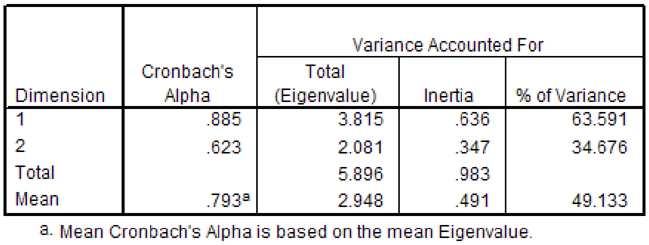
Variables:
- Thread [thread]
- Head form [head]
- Indentation of head [in...
- Bottom shape [bottom]
- Brass [brass]
- Length in half-inches [l...
- object

object~=16

Function group:
- All
- Arithmetic
- CDF & Noncentral CDF
- Conversion
- Current Date/Time
- Date Arithmetic

- ▶ Type object ~= 16 as the condition.
- ▶ Click Continue.
- ▶ Click OK in the Select Cases dialog box.
- ▶ Finally, recall the Multiple Correspondence Analysis dialog box, and click OK.

Figure 13-15
Model summary (outlier removed)

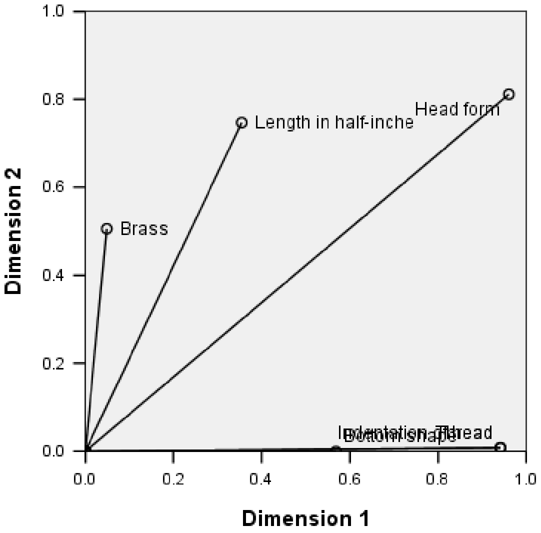


Dimension	Cronbach's Alpha	Variance Accounted For		
		Total (Eigenvalue)	Inertia	% of Variance
1	.885	3.815	.636	63.591
2	.623	2.081	.347	34.676
Total		5.896	.983	
Mean	.793[a]	2.948	.491	49.133

[a] Mean Cronbach's Alpha is based on the mean Eigenvalue.

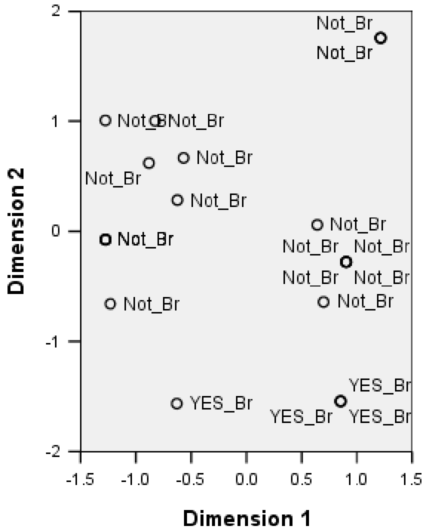
The eigenvalues shift slightly. The first dimension now accounts for a little more of the variance.

Figure 13-16
Discrimination measures



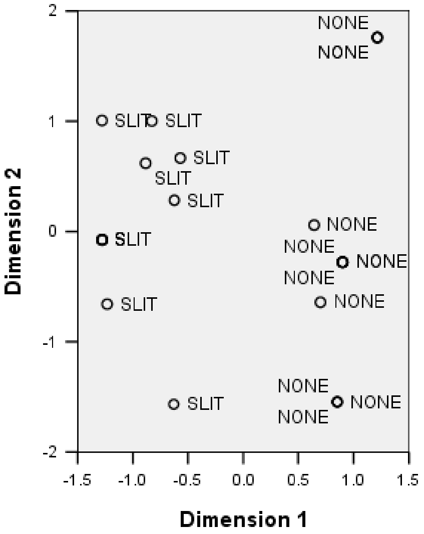
As shown in the discrimination plot, *Indentation of head* no longer discriminates in the second dimension, whereas *Brass* changes from no discrimination in either dimension to discrimination in the second dimension. Discrimination for the other variables is largely unchanged.

Figure 13-17
Object scores labeled with Brass (outlier removed)



The object scores plot labeled by *Brass* shows that the four brass objects all appear near the bottom of the plot (three objects occupy identical locations), indicating high discrimination along the second dimension. As was the case for *Thread* in the previous analysis, the objects do not form compact groups, but the differentiation of objects by categories is perfect.

Figure 13-18
Object scores labeled with Indentation of head (outlier removed)



The object scores plot labeled by *Indentation of head* shows that the first dimension discriminates perfectly between the non-indented objects and the indented objects, as in the previous analysis. In contrast to the previous analysis, however, the second dimension cannot now distinguish the two categories.

Thus, the omission of *SCREW1*, which is the only object with a star-shaped head, dramatically affects the interpretation of the second dimension. This dimension now differentiates objects based on *Brass*, *Head form*, and *Length in half-inches*.

Recommended Readings

See the following texts for more information on multiple correspondence analysis:

Benzécri, J. P. 1992. *Correspondence analysis handbook.* New York: Marcel Dekker.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: *The Prediction of Personal Adjustment,* P. Horst, ed. New York: Social Science Research Council, 319–348.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, 249–266.

Meulman, J. J., and W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. In: *Visual Display of Categorical Data,* M. Greenacre, and J. Blasius, eds. New York: Academic Press, 277–296.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, 25–36.

Tenenhaus, M., and F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 91–119.

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Chapter
14

Multidimensional Scaling

Given a set of objects, the goal of multidimensional scaling is to find a representation of the objects in a low-dimensional space. This solution is found by using the **proximities** between the objects. The procedure minimizes the squared deviations between the original, possibly transformed, object proximities and their Euclidean distances in the low-dimensional space.

The purpose of the low-dimensional space is to uncover relationships between the objects. By restricting the solution to be a linear combination of independent variables, you may be able to interpret the dimensions of the solution in terms of these variables. In the following example, you will see how 15 different kinship terms can be represented in three dimensions and how that space can be interpreted with respect to the gender, generation, and degree of separation of each of the terms.

Example: An Examination of Kinship Terms

Rosenberg and Kim (Rosenberg and Kim, 1975) set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criteria from the first sort. Thus, a total of six “sources” were obtained, as outlined in the following table.

Table 14-1
Source structure of kinship data

Source	Gender	Condition	Sample size
1	Female	Single sort	85
2	Male	Single sort	85
3	Female	First sort	80
4	Female	Second sort	80
5	Male	First sort	80
6	Male	Second sort	80

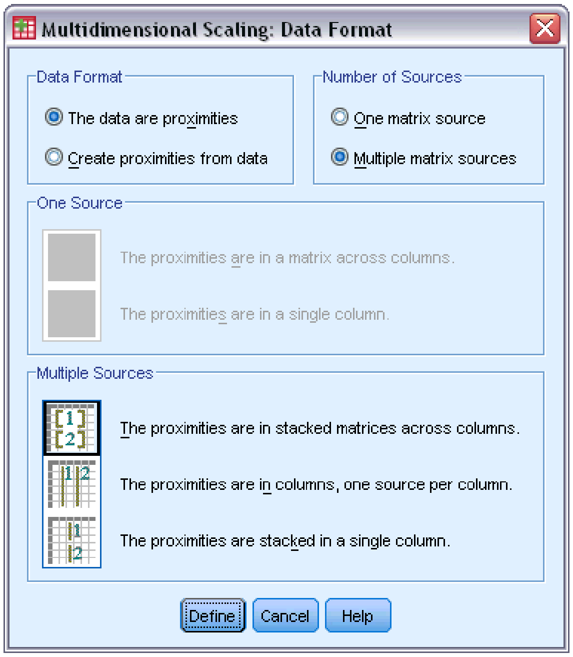
Each source corresponds to a 15 × 15 proximity matrix, whose cells are equal to the number of people in a source minus the number of times that the objects were partitioned together in that source. This dataset can be found in *kinship_dat.sav*. For more information, see the topic Sample Files in Appendix A on p. 292.

Choosing the Number of Dimensions

It is up to you to decide how many dimensions the solution should have. The scree plot can help you make this decision.

- ▶ To create a scree plot, from the menus choose:
  Analyze > Scale > Multidimensional Scaling (PROXSCAL)...

Figure 14-1
Data Format dialog box



Multidimensional Scaling: Data Format

Data Format
● The data are proximities
○ Create proximities from data

Number of Sources
○ One matrix source
● Multiple matrix sources

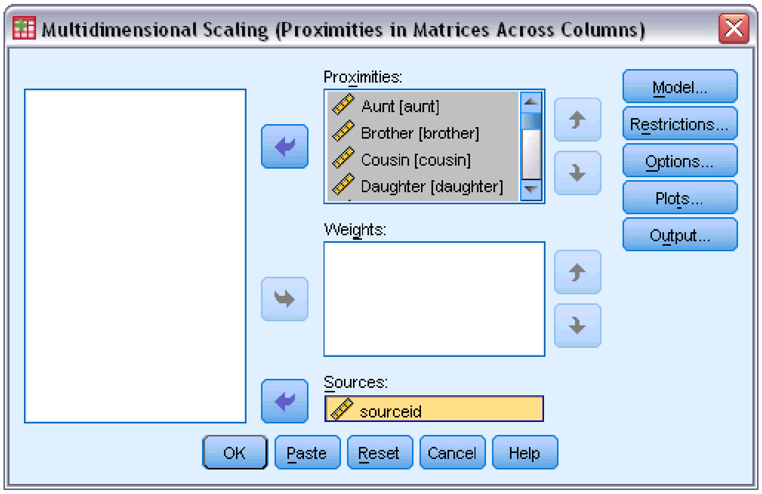
One Source
The proximities are in a matrix across columns.
The proximities are in a single column.

Multiple Sources
The proximities are in stacked matrices across columns.
The proximities are in columns, one source per column.
The proximities are stacked in a single column.

- ▶ Select Multiple matrix sources in the Number of Sources group.

- ▶ Click Define.

Figure 14-2
Multidimensional Scaling dialog box



Multidimensional Scaling (Proximities in Matrices Across Columns)

Proximities:
Aunt [aunt]
Brother [brother]
Cousin [cousin]
Daughter [daughter]

Weights:

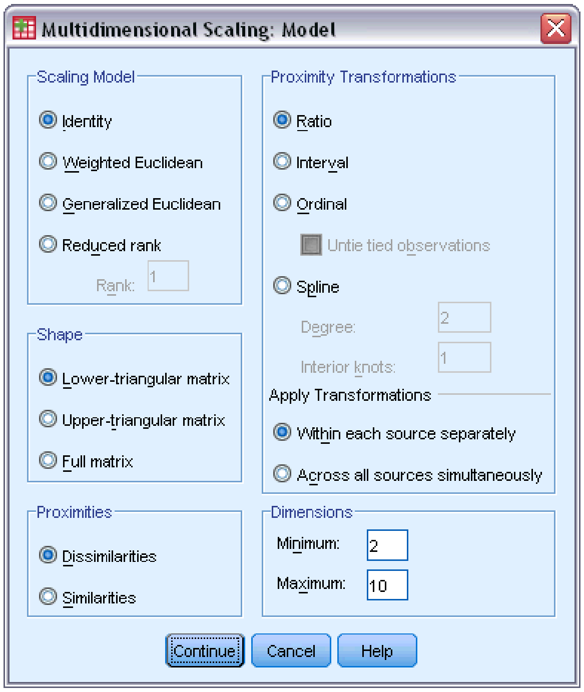
Sources:
sourceid

- ▶ Select *Aunt* through *Uncle* as proximities variables.

- ▶ Select *sourceid* as the variable identifying the source.

▶ Click Model.

Figure 14-3
Model dialog box



Multidimensional Scaling: Model

Scaling Model

● Identity
○ Weighted Euclidean
○ Generalized Euclidean
○ Reduced rank
Rank: 1

Shape

● Lower-triangular matrix
○ Upper-triangular matrix
○ Full matrix

Proximities

● Dissimilarities
○ Similarities

Proximity Transformations

● Ratio
○ Interval
○ Ordinal
☐ Untie tied observations
○ Spline
Degree: 2
Interior knots: 1

Apply Transformations

● Within each source separately
○ Across all sources simultaneously

Dimensions

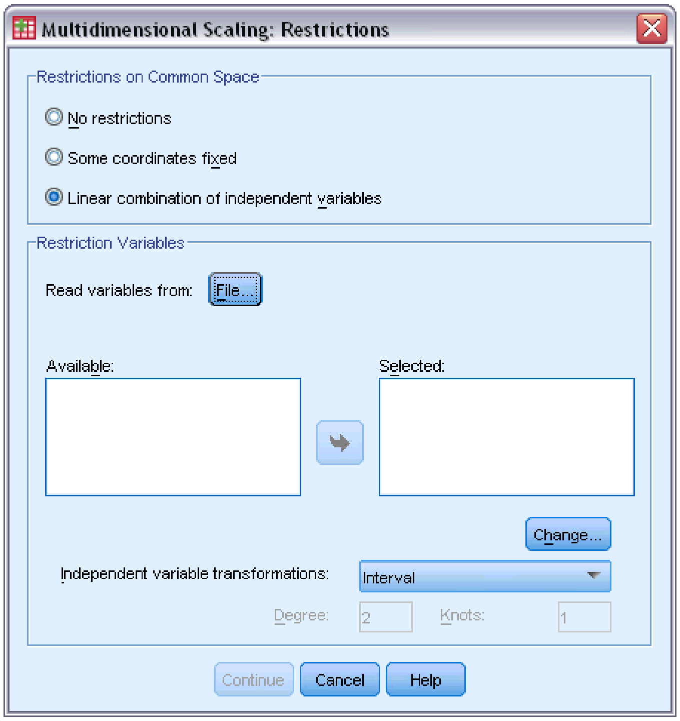
Minimum: 2
Maximum: 10

▶ Type 10 as the maximum number of dimensions.

▶ Click Continue.

▶ Click Restrictions in the Multidimensional Scaling dialog box.

Figure 14-4
Restrictions dialog box



Multidimensional Scaling: Restrictions

Restrictions on Common Space

○ No restrictions

○ Some coordinates fixed

● Linear combination of independent variables

Restriction Variables

Read variables from: File...

Available:

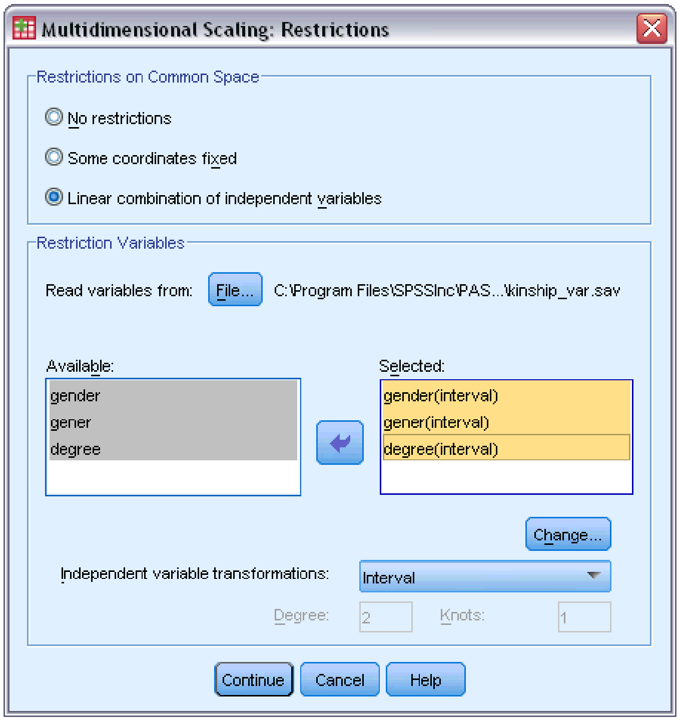
Selected:

Independent variable transformations: Interval

Degree: 2 Knots: 1

- ▶ Select Linear combination of independent variables.
- ▶ Click File to select the source of the independent variables.
- ▶ Select *kinship_var.sav*.

Figure 14-5
Restrictions dialog box



Multidimensional Scaling: Restrictions

Restrictions on Common Space

○ No restrictions

○ Some coordinates fixed

● Linear combination of independent variables

Restriction Variables

Read variables from: File... C:\Program Files\SPSSInc\PAS...\kinship_var.sav

Available:
gender
gener
degree

Selected:
gender(interval)
gener(interval)
degree(interval)

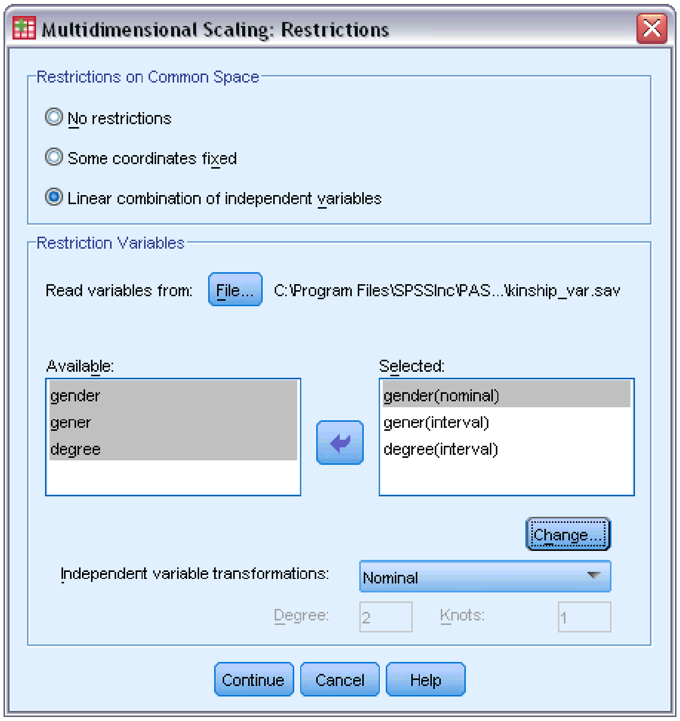
Independent variable transformations: Interval

Degree: 2 Knots: 1

▶ Select *gender*, *gener*, and *degree* as restriction variables.

Note that the variable *gender* has a user-missing value—9 = missing (for cousin). The procedure treats this as a valid category. Thus, the default linear transformation is unlikely to be appropriate. Use a nominal transformation instead.

Figure 14-6
Restrictions dialog box


Multidimensional Scaling: Restrictions
Restrictions on Common Space
Restriction Variables
File...
Change...
Continue
Cancel
Help

○ No restrictions

○ Some coordinates fixed

● Linear combination of independent variables

Read variables from: C:\Program Files\SPSSInc\PAS...\kinship_var.sav

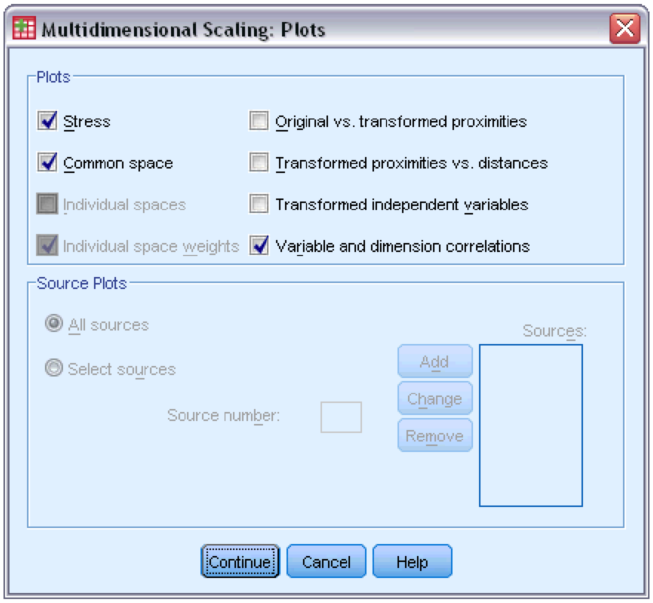
Available:	Selected:
gender	gender(nominal)
gener	gener(interval)
degree	degree(interval)

Independent variable transformations: Nominal

Degree: 2 Knots: 1

- ▶ Select *gender*.
- ▶ Select Nominal from the Independent Variable Transformations drop-down list.
- ▶ Click Change.
- ▶ Click Continue.
- ▶ Click Plots in the Multidimensional Scaling dialog box.

Figure 14-7
Plots dialog box



Plots

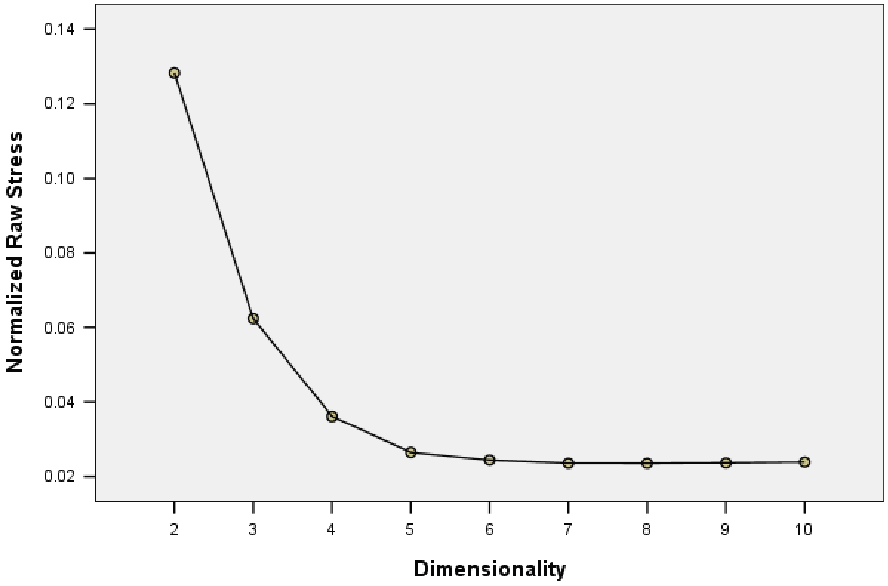
☑ Stress
☑ Common space
□ Individual spaces
☑ Individual space weights
□ Original vs. transformed proximities
□ Transformed proximities vs. distances
□ Transformed independent variables
☑ Variable and dimension correlations

Source Plots

● All sources
○ Select sources
Source number:
Sources:

- ▶ Select **Stress** in the Plots group.
- ▶ Click **Continue**.
- ▶ Click **OK** in the Multidimensional Scaling dialog box.

Figure 14-8
Scree plot



The procedure begins with a 10-dimensional solution and works down to a 2-dimensional solution. The scree plot shows the normalized raw stress of the solution at each dimension. You can see from the plot that increasing the dimensionality from 2 to 3 and from 3 to 4 offers large improvements in the stress. After 4, the improvements are rather small. You will choose to analyze the data by using a 3-dimensional solution, because the results are easier to interpret.

A Three-Dimensional Solution

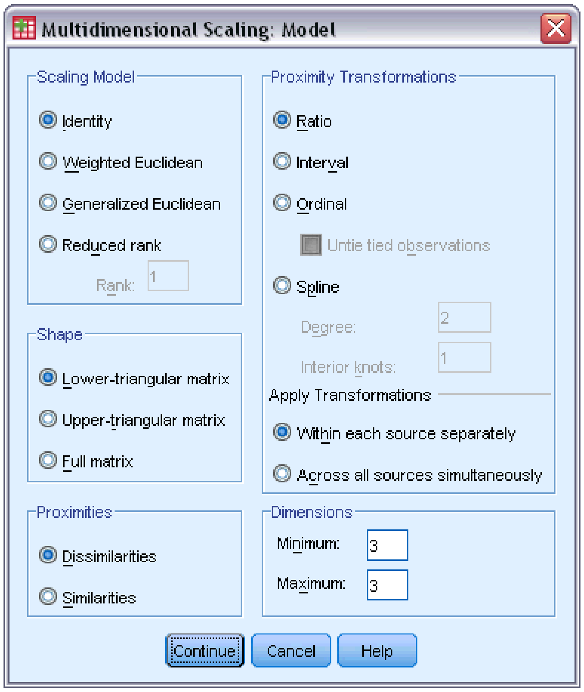
The independent variables *gender*, *gener* (generation), and *degree* (of separation) were constructed with the intention of using them to interpret the dimensions of the solution. The independent variables were constructed as follows:

gender	1 = male, 2 = female, 9 = missing (for cousin)
gener	The number of generations from you if the term refers to your kin, with lower numbers corresponding to older generations. Thus, grandparents are –2, grandchildren are 2, and siblings are 0.
degree	The number of degrees of separation along your family tree. Thus, your parents are up 1 node, while your children are down 1 node. Your siblings are up 1 node to your parents and then down 1 node to them, for 2 degrees of separation. Your cousin is 4 degrees away—2 up to your grandparents and then 2 down through your aunt/uncle to them.

The external variables can be found in *kinship_var.sav*. Additionally, an initial configuration from an earlier analysis is supplied in *kinship_ini.sav*. For more information, see the topic Sample Files in Appendix A on p. 292.

Running the Analysis

Figure 14-9
Model dialog box



Multidimensional Scaling: Model

Scaling Model

- ● Identity
- ○ Weighted Euclidean
- ○ Generalized Euclidean
- ○ Reduced rank

Rank: 1

Shape

- ● Lower-triangular matrix
- ○ Upper-triangular matrix
- ○ Full matrix

Proximities

- ● Dissimilarities
- ○ Similarities

Proximity Transformations

- ● Ratio
- ○ Interval
- ○ Ordinal
  - ☐ Untie tied observations
- ○ Spline

Degree: 2

Interior knots: 1

Apply Transformations

- ● Within each source separately
- ○ Across all sources simultaneously

Dimensions

Minimum: 3

Maximum: 3

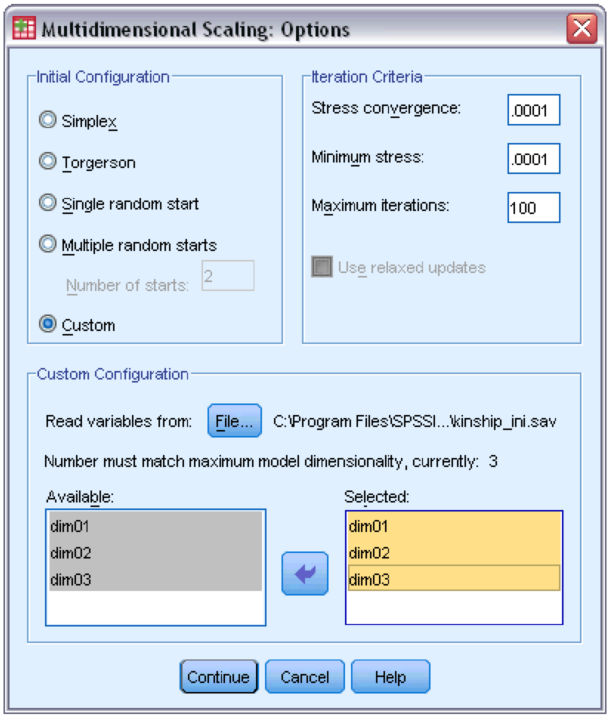
▶ To obtain a three-dimensional solution, recall the Multidimensional Scaling dialog box and click Model.

▶ Type 3 as the minimum and maximum number of dimensions.

▶ Click Continue.

▶ Click Options in the Multidimensional Scaling dialog box.

Figure 14-10
Options dialog box



Multidimensional Scaling: Options

Initial Configuration

○ Simplex

○ Torgerson

○ Single random start

○ Multiple random starts

Number of starts: 2

● Custom

Iteration Criteria

Stress convergence: .0001

Minimum stress: .0001

Maximum iterations: 100

□ Use relaxed updates

Custom Configuration

Read variables from: File... C:\Program Files\SPSSI...\kinship_ini.sav

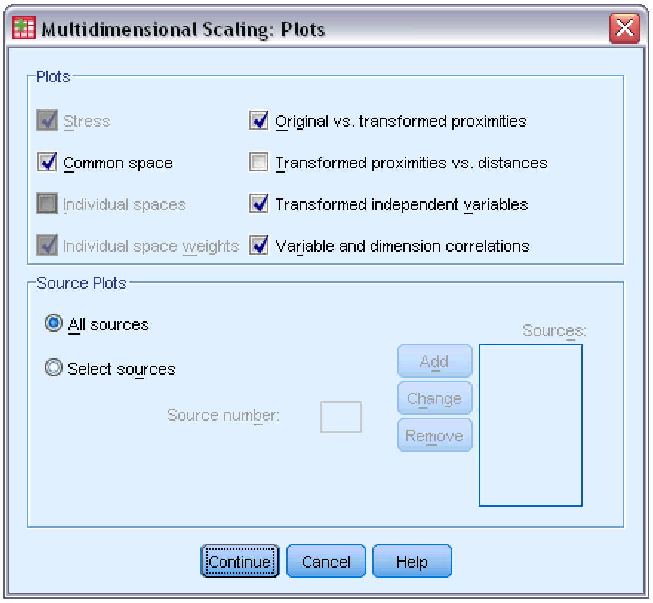
Number must match maximum model dimensionality, currently: 3

Available: dim01 dim02 dim03

Selected: dim01 dim02 dim03

- ▶ Select Custom as the initial configuration.
- ▶ Select *kinship_ini.sav* as the file to read variables from.
- ▶ Select *dim01*, *dim02*, and *dim03* as variables.
- ▶ Click Continue.
- ▶ Click Plots in the Multidimensional Scaling dialog box.

Figure 14-11
Plots dialog box



Multidimensional Scaling: Plots

Plots

☑ Stress
☑ Original vs. transformed proximities
☑ Common space
□ Transformed proximities vs. distances
□ Individual spaces
☑ Transformed independent variables
☑ Individual space weights
☑ Variable and dimension correlations

Source Plots

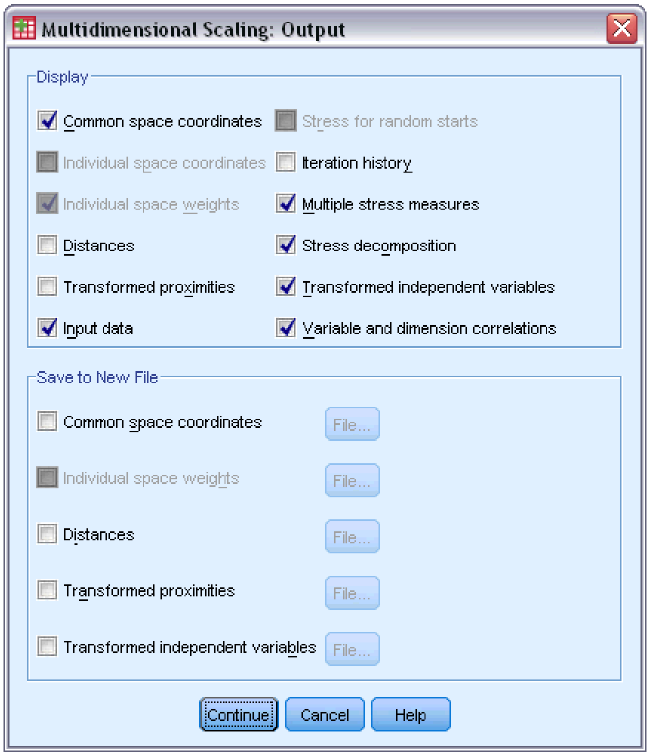
● All sources
○ Select sources

Source number:

Sources:

- ▶ Select Original vs. transformed proximities and Transformed independent variables.
- ▶ Click Continue.
- ▶ Click Output in the Multidimensional Scaling dialog box.

Figure 14-12
Output dialog box



Multidimensional Scaling: Output

Display

☑ Common space coordinates
□ Individual space coordinates
☑ Individual space weights
□ Distances
□ Transformed proximities
☑ Input data
□ Stress for random starts
□ Iteration history
☑ Multiple stress measures
☑ Stress decomposition
☑ Transformed independent variables
☑ Variable and dimension correlations

Save to New File

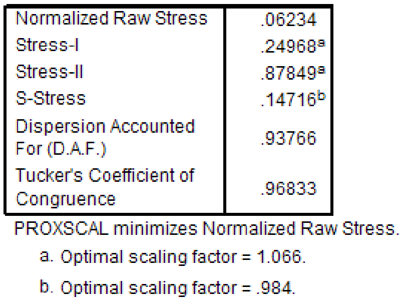
□ Common space coordinates
□ Individual space weights
□ Distances
□ Transformed proximities
□ Transformed independent variables

- ▶ Select Input data, Stress decomposition, and Variable and dimension correlations.
- ▶ Click Continue.
- ▶ Click OK in the Multidimensional Scaling dialog box.

Stress Measures

The stress and fit measures give an indication of how well the distances in the solution approximate the original distances.

Figure 14-13
Stress and fit measures



Normalized Raw Stress	.06234
Stress-I	.24968[a]
Stress-II	.87849[a]
S-Stress	.14716[b]
Dispersion Accounted For (D.A.F.)	.93766
Tucker's Coefficient of Congruence	.96833

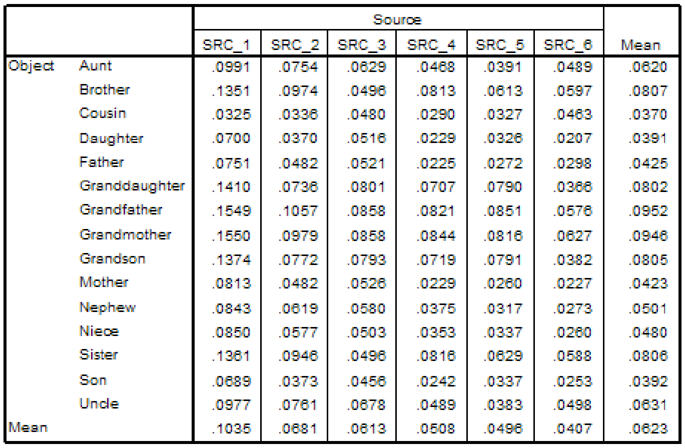
PROXSCAL minimizes Normalized Raw Stress.

a. Optimal scaling factor = 1.066.

b. Optimal scaling factor = .984.

Each of the four stress statistics measures the misfit of the data, while the dispersion accounted for and Tucker's coefficient of congruence measure the fit. Lower stress measures (to a minimum of 0) and higher fit measures (to a maximum of 1) indicate better solutions.

Figure 14-14
Decomposition of normalized raw stress



		Source						
		SRC_1	SRC_2	SRC_3	SRC_4	SRC_5	SRC_6	Mean
Object	Aunt	.0991	.0754	.0629	.0468	.0391	.0489	.0620
	Brother	.1351	.0974	.0496	.0813	.0613	.0597	.0807
	Cousin	.0325	.0336	.0480	.0290	.0327	.0463	.0370
	Daughter	.0700	.0370	.0516	.0229	.0326	.0207	.0391
	Father	.0751	.0482	.0521	.0225	.0272	.0298	.0425
	Granddaughter	.1410	.0736	.0801	.0707	.0790	.0366	.0802
	Grandfather	.1549	.1057	.0858	.0821	.0851	.0576	.0952
	Grandmother	.1550	.0979	.0858	.0844	.0816	.0627	.0946
	Grandson	.1374	.0772	.0793	.0719	.0791	.0382	.0805
	Mother	.0813	.0482	.0526	.0229	.0260	.0227	.0423
	Nephew	.0843	.0619	.0580	.0375	.0317	.0273	.0501
	Niece	.0850	.0577	.0503	.0353	.0337	.0260	.0480
	Sister	.1361	.0946	.0496	.0816	.0629	.0588	.0806
	Son	.0689	.0373	.0456	.0242	.0337	.0253	.0392
	Uncle	.0977	.0761	.0678	.0489	.0383	.0498	.0631
Mean		.1035	.0681	.0613	.0508	.0496	.0407	.0623

The decomposition of stress helps you identify which sources and objects contribute the most to the overall stress of the solution. In this case, most of the stress among the sources is attributable to sources 1 and 2, while among the objects, most of the stress is attributable to *Brother*, *Granddaughter*, *Grandfather*, *Grandmother*, *Grandson*, and *Sister*.

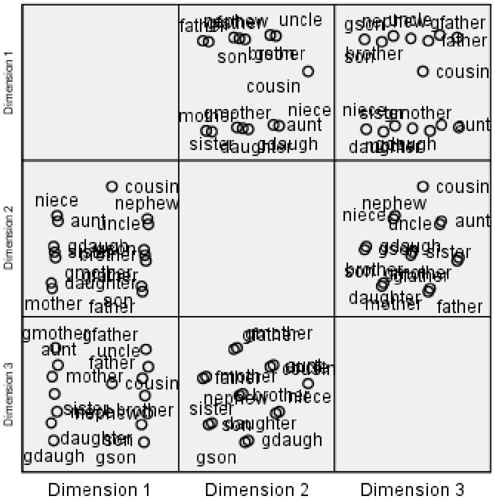
The two sources that are accountable for most of the stress are the two groups that sorted the terms only once. This information suggests that the students considered multiple factors when sorting the terms, and those students who were allowed to sort twice focused on a portion of those factors for the first sort and then considered the remaining factors during the second sort.

The objects that account for most of the stress are those objects with a *degree* of 2. These people are relations who are not part of the "nuclear" family (*Mother*, *Father*, *Daughter*, *Son*) but are nonetheless closer than other relations. This middle position could easily cause some differential sorting of these terms.

Final Coordinates of the Common Space

The common space plot gives a visual representation of the relationships between the objects.

Figure 14-15
Common space coordinates



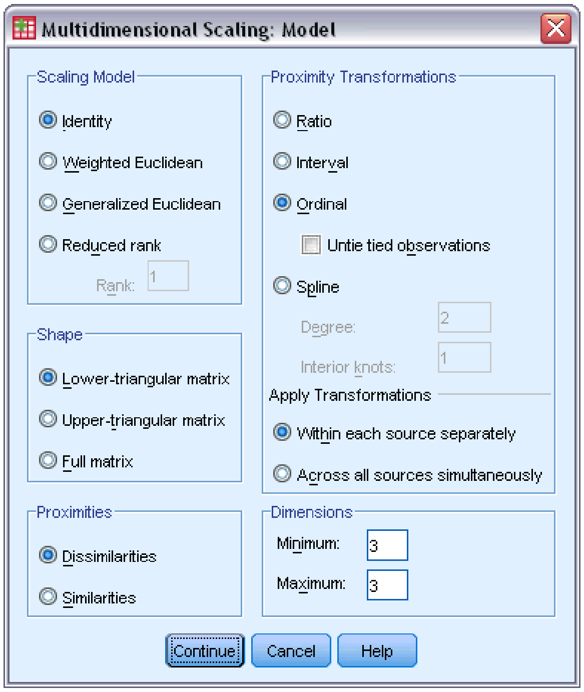
Look at the final coordinates for the objects in dimensions 1 and 3; this is the plot in the lower left corner of the scatterplot matrix. This plot shows that dimension 1 (on the *x* axis) is correlated with the variable *gender*, and dimension 3 (on the *y* axis) is correlated with *gener*. From left to right, you see that dimension 1 separates the female and male terms, with the genderless term *Cousin* in the middle. From the bottom of the plot to the top, increasing values along the axis correspond to terms that are older.

Now look at the final coordinates for the objects in dimensions 2 and 3; this plot is the plot on the middle right side of the scatterplot matrix. From this plot, you can see that the second dimension (along the *y* axis) corresponds to the variable *degree*, with larger values along the axis corresponding to terms that are further from the “nuclear” family.

A Three-Dimensional Solution with Nondefault Transformations

The previous solution was computed by using the default ratio transformation for proximities and interval transformations for the independent variables *gener* and *degree*. The results are pretty good, but you may be able to do better by using other transformations. For example, the proximities, *gener*, and *degree* all have natural orderings, but they may be better modeled by an ordinal transformation than a linear transformation.

Figure 14-16
Model dialog box



Multidimensional Scaling: Model

Scaling Model
- ● Identity
- ○ Weighted Euclidean
- ○ Generalized Euclidean
- ○ Reduced rank
  - Rank: 1

Shape
- ● Lower-triangular matrix
- ○ Upper-triangular matrix
- ○ Full matrix

Proximities
- ● Dissimilarities
- ○ Similarities

Proximity Transformations
- ○ Ratio
- ○ Interval
- ● Ordinal
  - □ Untie tied observations
- ○ Spline
  - Degree: 2
  - Interior knots: 1

Apply Transformations
- ● Within each source separately
- ○ Across all sources simultaneously

Dimensions
- Minimum: 3
- Maximum: 3

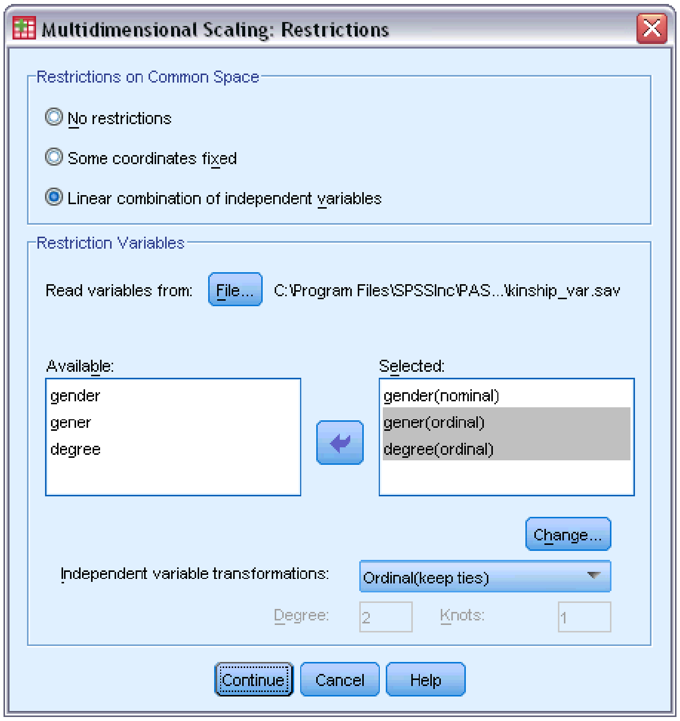
▶ To rerun the analysis, scaling the proximities, *gener*, and *degree* at the ordinal level (keeping ties), recall the Multidimensional Scaling dialog box and click Model.

▶ Select Ordinal as the proximity transformation.

▶ Click Continue.

▶ Click Restrictions in the Multidimensional Scaling dialog box.

Figure 14-17
Restrictions dialog box



Multidimensional Scaling: Restrictions

Restrictions on Common Space

○ No restrictions

○ Some coordinates fixed

● Linear combination of independent variables

Restriction Variables

Read variables from: File... C:\Program Files\SPSSInc\PAS...\kinship_var.sav

Available:	Selected:
gender	gender(nominal)
gener	gener(ordinal)
degree	degree(ordinal)

Independent variable transformations: Ordinal(keep ties)

Degree: 2 Knots: 1

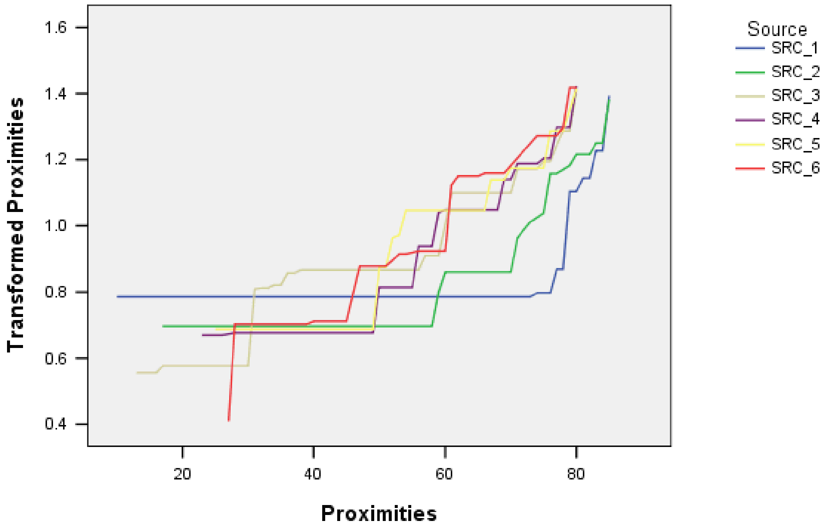
- ▶ Select *gener* and *degree*.
- ▶ Select Ordinal (keep ties) from the Independent Variable Transformations drop-down list.
- ▶ Click Change.
- ▶ Click Continue.
- ▶ Click OK in the Multidimensional Scaling dialog box.

Transformation Plots

The transformation plots are a good first check to see whether the original transformations were appropriate. If the plots are approximately linear, the linear assumption is appropriate. If not, check the stress measures to see whether there is an improvement in fit and check the common space plot to see whether the interpretation is more useful.

The independent variables each obtain approximately linear transformations, so it may be appropriate to interpret them as numerical. However, the proximities do not obtain a linear transformation, so it is possible that the ordinal transformation is more appropriate for the proximities.

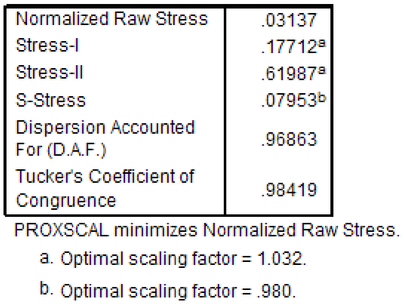
Figure 14-18
Transformed proximities



Stress Measures

The stress for the current solution supports the argument for scaling the proximities at the ordinal level.

Figure 14-19
Stress and fit measures



Normalized Raw Stress	.03137
Stress-I	.17712[a]
Stress-II	.61987[a]
S-Stress	.07953[b]
Dispersion Accounted For (D.A.F.)	.96863
Tucker's Coefficient of Congruence	.98419

PROXSCAL minimizes Normalized Raw Stress.

a. Optimal scaling factor = 1.032.

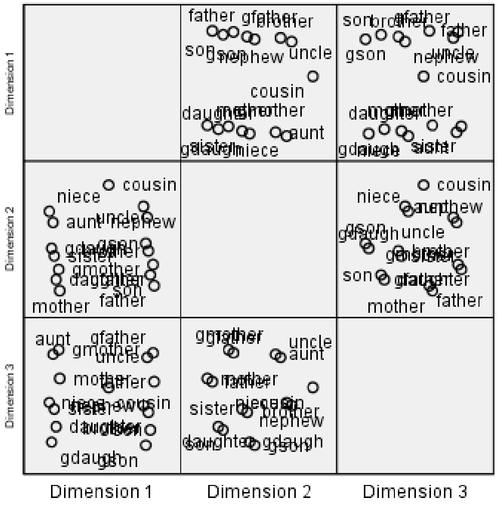
b. Optimal scaling factor = .980.

The normalized raw stress for the previous solution is 0.06234. Scaling the variables by using nondefault transformations halves the stress to 0.03137.

Final Coordinates of the Common Space

The common space plots offer essentially the same interpretation of the dimensions as the previous solution.

Figure 14-20
Common space coordinates



Discussion

It is best to treat the proximities as ordinal variables, because there is great improvement in the stress measures. As a next step, you may want to “untie” the ordinal variables—that is, allow equivalent values of the original variables to obtain different transformed values. For example, in the first source, the proximities between *Aunt* and *Son*, and *Aunt* and *Grandson*, are 85. The “tied” approach to ordinal variables forces the transformed values of these proximities to be equivalent, but there is no particular reason for you to assume that they should be. In this case, allowing the proximities to become untied frees you from an unnecessary restriction.

Recommended Readings

See the following texts for more information on multidimensional scaling:

Commandeur, J. J. F., and W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, University of Leiden.

De Leeuw, J., and W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. In: *Multivariate Analysis, Vol. V,* P. R. Krishnaiah, ed. Amsterdam: North-Holland, 501–522.

Heiser, W. J. 1981. *Unfolding analysis of proximity data*. Leiden: Department of Data Theory, University of Leiden.

Heiser, W. J., and F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: SagePublications, Inc., 25–48.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–28.

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, 125–140.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, 219–246.

Chapter

15

Multidimensional Unfolding

The Multidimensional Unfolding procedure attempts to find a common quantitative scale that allows you to visually examine the relationships between two sets of objects.

Example: Breakfast Item Preferences

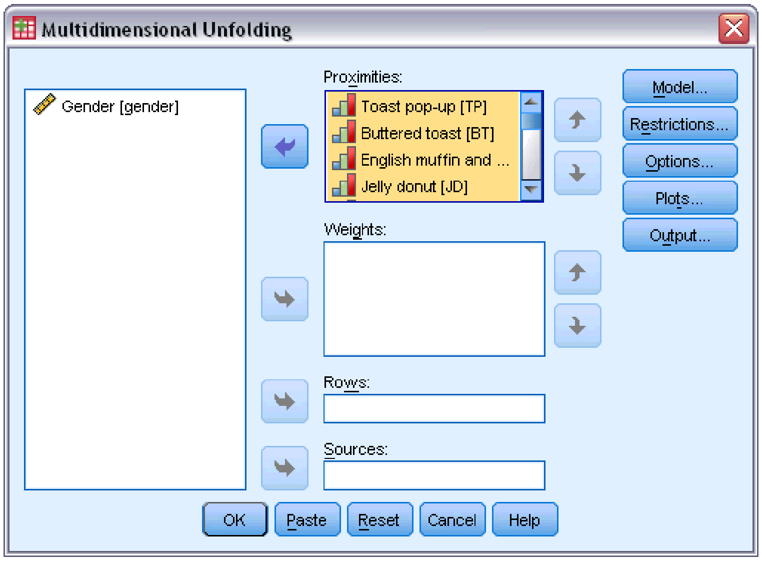
In a classic study (Green and Rao, 1972), 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference, from 1 = "most preferred" to 15 = "least preferred." This information is collected in *breakfast_overall.sav*. For more information, see the topic Sample Files in Appendix A on p. 292.

The results of the study provide a typical example of the degeneracy problem inherent in most multidimensional unfolding algorithms that is solved by penalizing the coefficient of variation of the transformed proximities (Busing, Groenen, and Heiser, 2005). You will see a degenerate solution and will see how to solve the problem using Multidimensional Unfolding, allowing you to determine how individuals discriminate between breakfast items. Syntax for reproducing these analyses can be found in *prefscal_breakfast-overall.sps*.

Producing a Degenerate Solution

▶ To run a Multidimensional Unfolding analysis, from the menus choose:
Analyze > Scale > Multidimensional Unfolding (PREFSCAL)...

Figure 15-1
Multidimensional Unfolding main dialog box



Multidimensional Unfolding

Gender [gender]

Proximities:
- Toast pop-up [TP]
- Buttered toast [BT]
- English muffin and ...
- Jelly donut [JD]

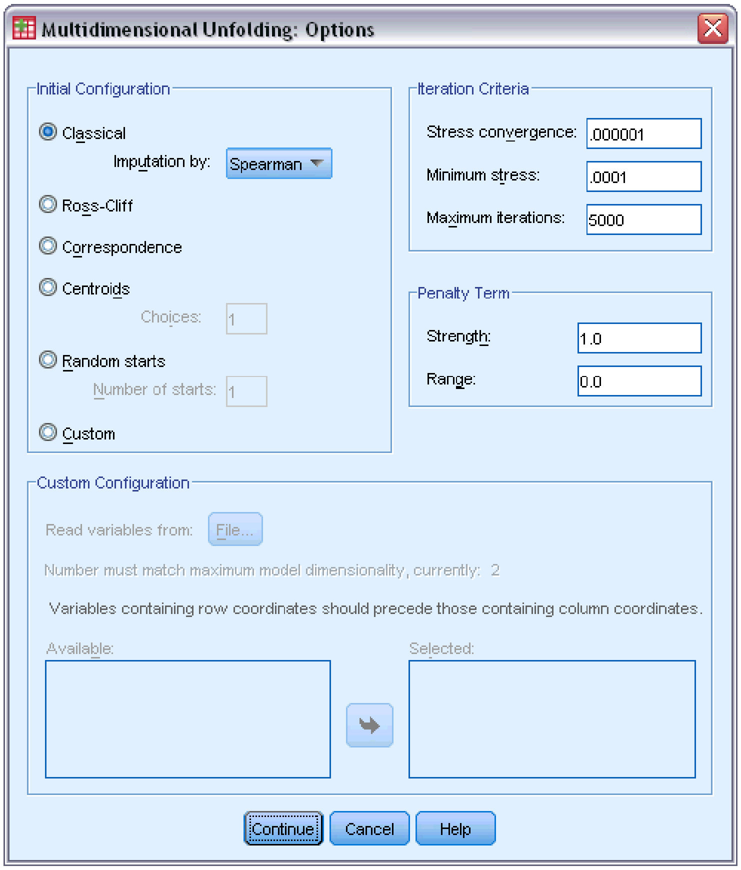
Weights:

Rows:

Sources:

- ▶ Select *Toast pop-up* through *Corn muffin and butter* as proximities variables.
- ▶ Click Options.

Figure 15-2
Options dialog box



Multidimensional Unfolding: Options

Initial Configuration

● Classical

Imputation by: Spearman

○ Ross-Cliff

○ Correspondence

○ Centroids

Choices: 1

○ Random starts

Number of starts: 1

○ Custom

Iteration Criteria

Stress convergence: .000001

Minimum stress: .0001

Maximum iterations: 5000

Penalty Term

Strength: 1.0

Range: 0.0

Custom Configuration

Read variables from:

Number must match maximum model dimensionality, currently: 2

Variables containing row coordinates should precede those containing column coordinates.

Available:

Selected:

- Select Spearman as the imputation method for the Classical start.
- In the Penalty Term group, type 1.0 as the value of the Strength parameter and 0.0 as the value of the Range parameter. This turns off the penalty term.
- Click Continue.
- Click OK in the Multidimensional Unfolding dialog box.

Following is the command syntax generated by these selections:

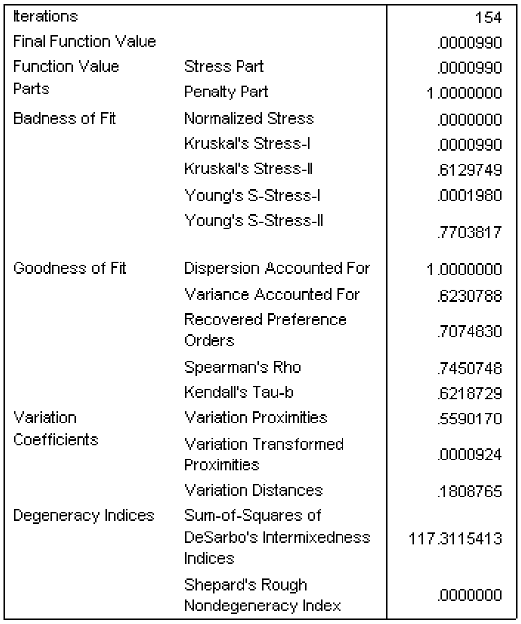
```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(1.0) OMEGA(0.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON .
```

- This syntax specifies an analysis on variables *tp (Toast pop-up)* through *cmb (Corn muffin and butter)*.

- The INITIAL subcommand specifies that the starting values be imputed using Spearman distances.
- The specified values on the PENALTY subcommand essentially turn off the penalty term, and as a result, the procedure will minimize Kruskal's Stress-I. This will result in a degenerate solution.
- The PLOT subcommand requests plots of the common space.
- All other parameters fall back to their default values.

Measures

Figure 15-3
Measures for degenerate solution



Iterations		154
Final Function Value		.0000990
Function Value Parts	Stress Part	.0000990
	Penalty Part	1.0000000
Badness of Fit	Normalized Stress	.0000000
	Kruskal's Stress-I	.0000990
	Kruskal's Stress-II	.6129749
	Young's S-Stress-I	.0001980
	Young's S-Stress-II	.7703817
Goodness of Fit	Dispersion Accounted For	1.0000000
	Variance Accounted For	.6230788
	Recovered Preference Orders	.7074830
	Spearman's Rho	.7450748
	Kendall's Tau-b	.6218729
Variation Coefficients	Variation Proximities	.5590170
	Variation Transformed Proximities	.0000924
	Variation Distances	.1808765
Degeneracy Indices	Sum-of-Squares of DeSarbo's Intermixedness Indices	117.3115413
	Shepard's Rough Nondegeneracy Index	.0000000

The algorithm converges to a solution after 154 iterations, with a penalized stress (marked final function value) of 0.0000990. Since the penalty term has been turned off, penalized stress is equal to Kruskal's Stress-I (the stress part of the function value is equivalent to Kruskal's badness-of-fit measure). Low stress values generally indicate that the solution fits the data well, but there are several warning signs of a degenerate solution:

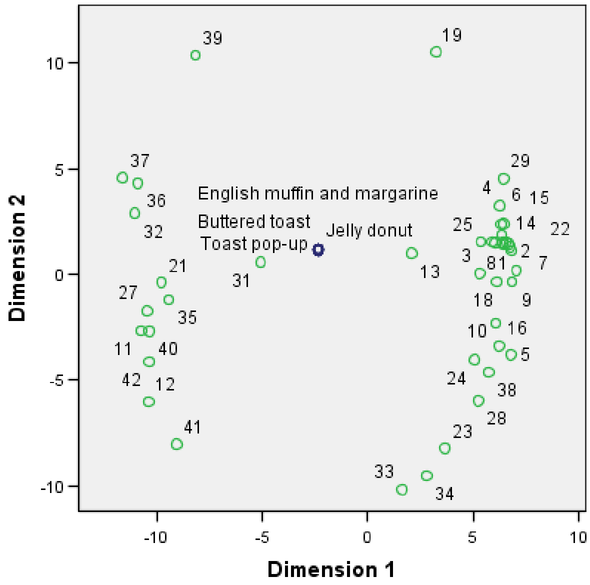
- The coefficient of variation for the transformed proximities is very small relative to the coefficient of variation for the original proximities. This suggests that the transformed proximities for each row are near-constant, and thus the solution will not provide any discrimination between objects.

- The sum-of-squares of DeSarbo's intermixedness indices are a measure of how well the points of the different sets are intermixed. If they are not intermixed, this is a warning sign that the solution may be degenerate. The closer to 0, the more intermixed the solution. The reported value is very large, indicating that the solution is not intermixed.
- Shepard's rough nondegeneracy index, which is reported as a percentage of different distances, is equal to 0. This is a clear numerical indication that there are insufficiently different distances and that the solution is probably degenerate.

Common Space

Figure 15-4

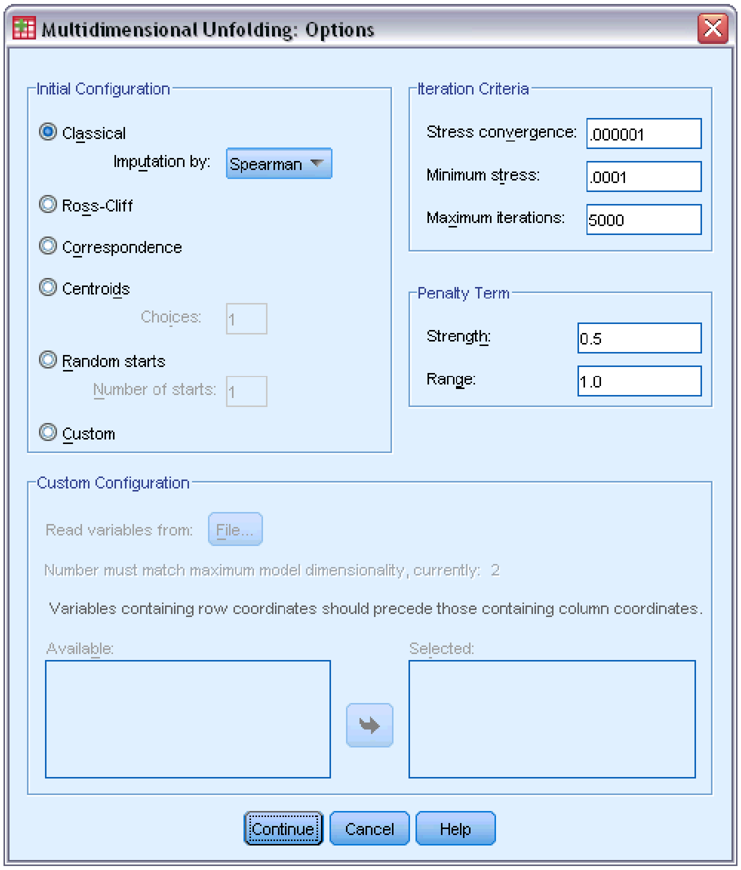
Joint plot of common space for degenerate solution



Visual confirmation that the solution is degenerate is found in the joint plot of the common space of row and column objects. The row objects (individuals) are situated on the circumference of a circle centered on the column objects (breakfast items), whose coordinates have collapsed to a single point.

Running a Nondegenerate Analysis

Figure 15-5
Options dialog box



Multidimensional Unfolding: Options

Initial Configuration

● Classical

Imputation by: Spearman

○ Ross-Cliff

○ Correspondence

○ Centroids

Choices: 1

○ Random starts

Number of starts: 1

○ Custom

Iteration Criteria

Stress convergence: .000001

Minimum stress: .0001

Maximum iterations: 5000

Penalty Term

Strength: 0.5

Range: 1.0

Custom Configuration

Read variables from: File...

Number must match maximum model dimensionality, currently: 2

Variables containing row coordinates should precede those containing column coordinates.

Available:

Selected:

- ▶ To produce a nondegenerate solution, click the Dialog Recall tool and select Multidimensional Unfolding.
- ▶ Click Options in the Multidimensional Unfolding dialog box.
- ▶ In the Penalty Term group, type 0.5 as the value of the Strength parameter and 1.0 as the value of the Range parameter. This turns off the penalty term.
- ▶ Click Continue.
- ▶ Click OK in the Multidimensional Unfolding dialog box.

Following is the command syntax generated by these selections:

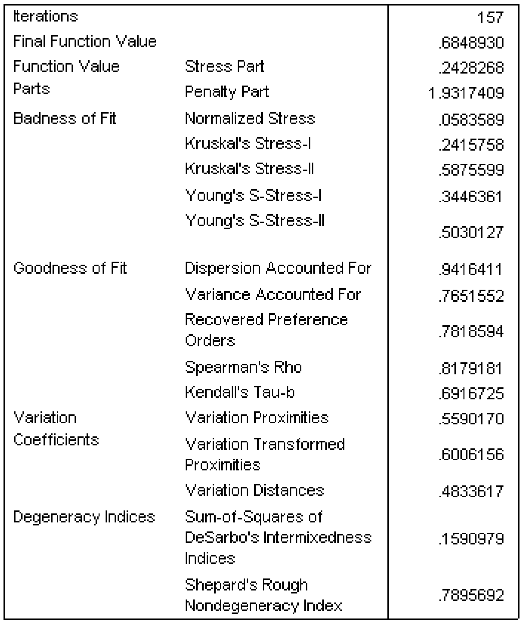
```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
```

```
/PRINT=MEASURES COMMON
/PLOT=COMMON .
```

- The only change is on the `PENALTY` subcommand. `LAMBDA` has been set to 0.5, and `OMEGA` has been set to 1.0, their default values.

Measures

Figure 15-6
Measures for nondegenerate solution



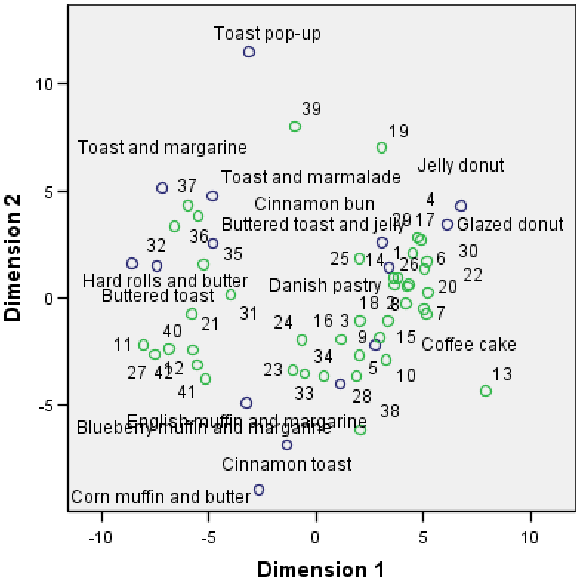
Iterations		157
Final Function Value		.6848930
Function Value Parts	Stress Part	.2428268
	Penalty Part	1.9317409
Badness of Fit	Normalized Stress	.0583589
	Kruskal's Stress-I	.2415758
	Kruskal's Stress-II	.5875599
	Young's S-Stress-I	.3446361
	Young's S-Stress-II	.5030127
Goodness of Fit	Dispersion Accounted For	.9416411
	Variance Accounted For	.7651552
	Recovered Preference Orders	.7818594
	Spearman's Rho	.8179181
	Kendall's Tau-b	.6916725
Variation Coefficients	Variation Proximities	.5590170
	Variation Transformed Proximities	.6006156
	Variation Distances	.4833617
Degeneracy Indices	Sum-of-Squares of DeSarbo's Intermixedness Indices	.1590979
	Shepard's Rough Nondegeneracy Index	.7895692

The problems noted in the measures for the degenerate solution have been corrected here.

- The normalized stress is no longer 0.
- The coefficient of variation for the transformed proximities now has a similar value to the coefficient of variation for the original proximities.
- DeSarbo's intermixedness indices are much closer to 0, indicating that the solution is much better intermixed.
- Shepard's rough nondegeneracy index, which is reported as a percentage of different distances, is now nearly 80%. There are sufficiently different distances, and the solution is probably nondegenerate.

Common Space

Figure 15-7
Joint plot of common space for nondegenerate solution



The joint plot of the common space allows for an interpretation of the dimensions. The horizontal dimension appears to discriminate between soft and hard bread or toast, with softer items as you move right along the axis. The vertical dimension does not have a clear interpretation, although it perhaps discriminates based on convenience, with more "formal" items as you move down along the axis.

This creates several clusters of breakfast items. For example, the donuts, cinnamon buns, and Danish pastry form a cluster of soft and somewhat informal items. The muffins and cinnamon toast form a cluster of harder but more formal items. The other toasts and hard rolls form a cluster of hard and somewhat informal items. The toast pop-up is a hard item that is extremely informal.

The individuals represented by the row objects are clearly split into clusters according to preference for hard or soft items, with considerable within-cluster variation along the vertical dimension.

Example: Three-Way Unfolding of Breakfast Item Preferences

In a classic study (Green et al., 1972), 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference, from 1 = "most preferred" to 15 = "least preferred." Their preferences were recorded under six different scenarios, from "Overall preference" to "Snack, with beverage only." This information is collected in *breakfast.sav*. For more information, see the topic Sample Files in Appendix A on p. 292.

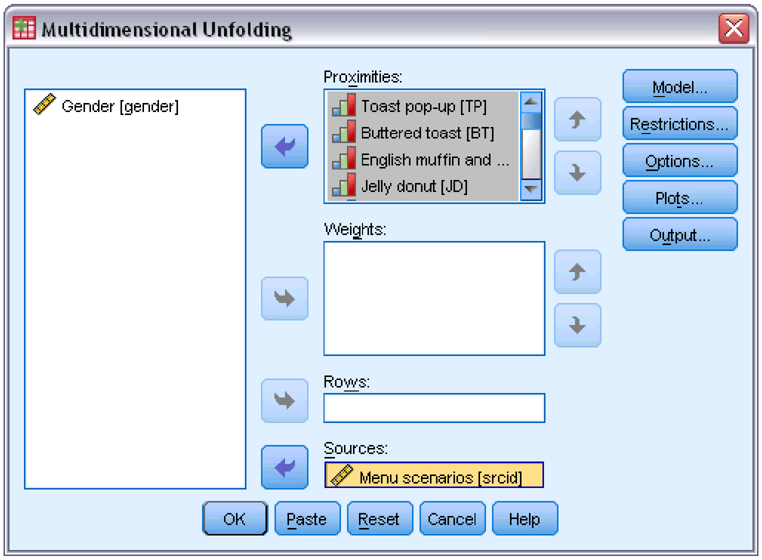
The six scenarios can be treated as separate sources. Use PREFSCAL to perform a three-way unfolding of the rows, columns, and sources. Syntax for reproducing these analyses can be found in *prefscal_breakfast.sps*.

Running the Analysis

▶ To run a Multidimensional Unfolding analysis, from the menus choose:

Analyze > Scale > Multidimensional Unfolding (PREFSCAL)...

Figure 15-8
Multidimensional Unfolding main dialog box



Multidimensional Unfolding

Gender [gender]

Proximities:
Toast pop-up [TP]
Buttered toast [BT]
English muffin and ...
Jelly donut [JD]

Weights:

Rows:

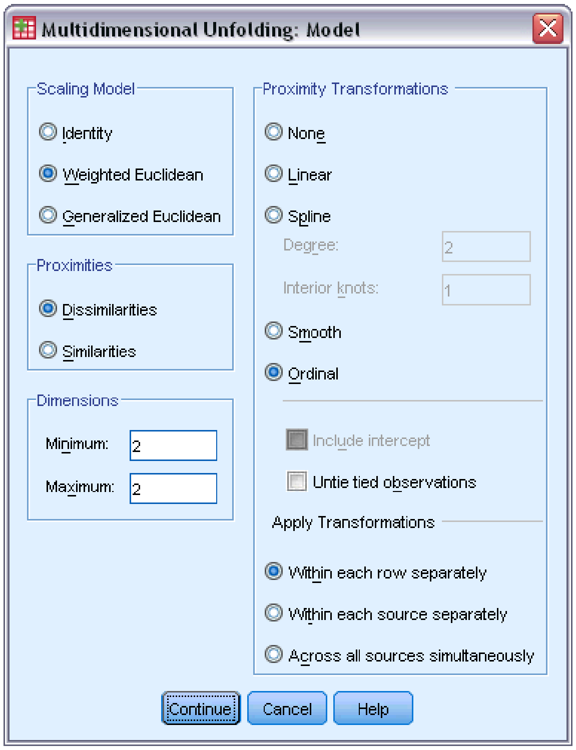
Sources:
Menu scenarios [srcid]

▶ Select *Toast pop-up* through *Corn muffin and butter* as proximities variables.

▶ Select *Menu scenarios* as the source variable.

▶ Click Model.

Figure 15-9
Model dialog box



Multidimensional Unfolding: Model

Scaling Model
- ○ Identity
- ● Weighted Euclidean
- ○ Generalized Euclidean

Proximities
- ● Dissimilarities
- ○ Similarities

Dimensions
- Minimum: 2
- Maximum: 2

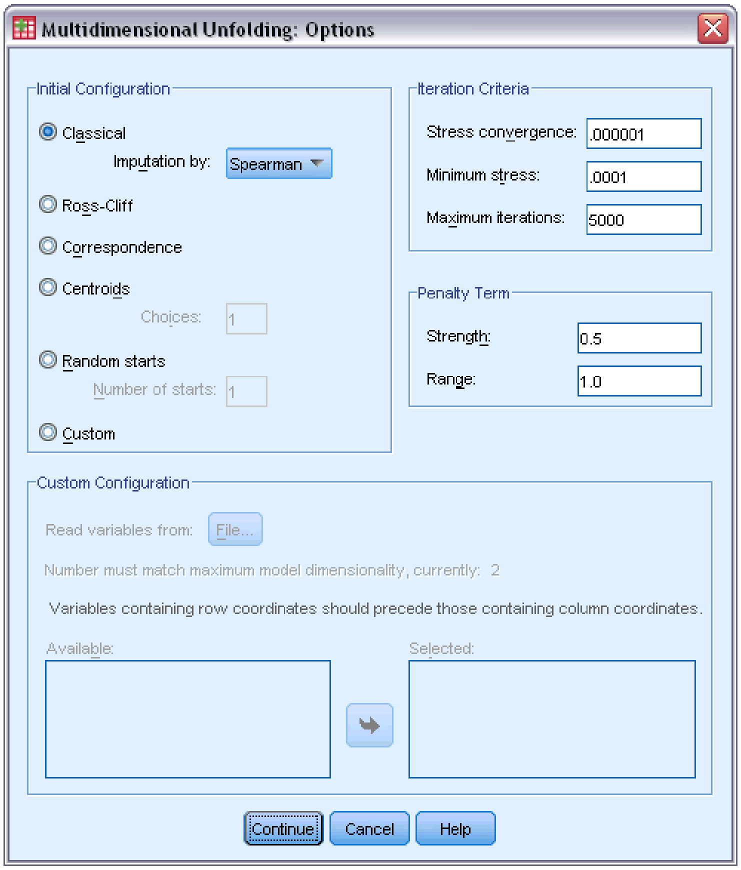
Proximity Transformations
- ○ None
- ○ Linear
- ○ Spline
  - Degree: 2
  - Interior knots: 1
- ○ Smooth
- ● Ordinal

- ☐ Include intercept
- ☐ Untie tied observations

Apply Transformations
- ● Within each row separately
- ○ Within each source separately
- ○ Across all sources simultaneously

- ▶ Select Weighted Euclidean as the scaling model.
- ▶ Click Continue.
- ▶ Click Options in the Multidimensional Unfolding dialog box.

Figure 15-10
Options dialog box



Multidimensional Unfolding: Options

Initial Configuration

● Classical

Imputation by: Spearman

○ Ross-Cliff

○ Correspondence

○ Centroids

Choices: 1

○ Random starts

Number of starts: 1

○ Custom

Iteration Criteria

Stress convergence: .000001

Minimum stress: .0001

Maximum iterations: 5000

Penalty Term

Strength: 0.5

Range: 1.0

Custom Configuration

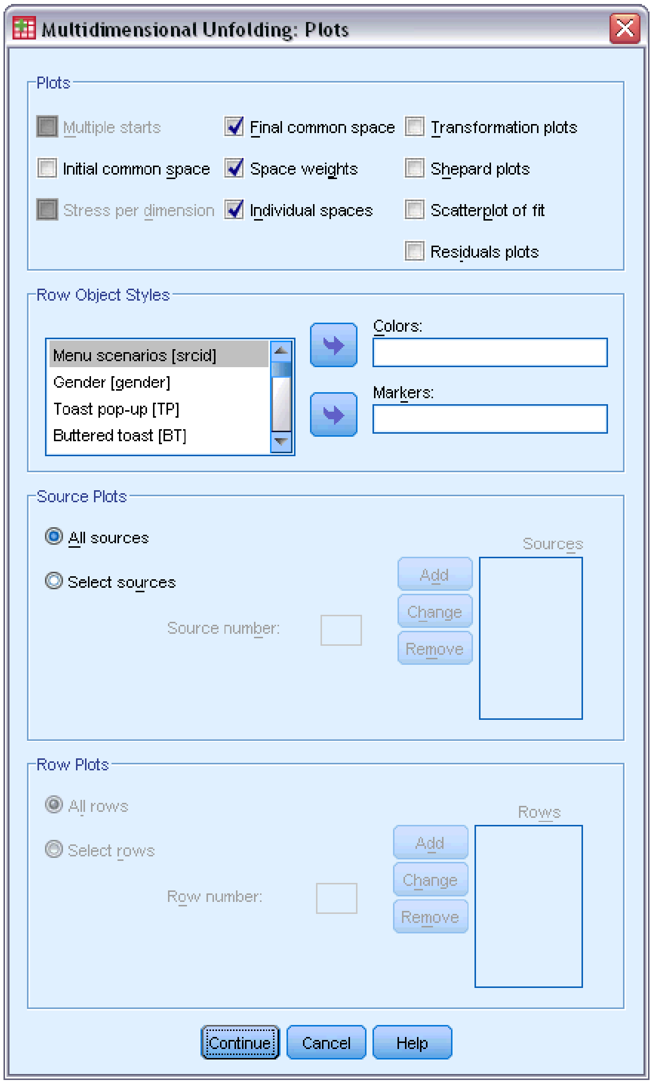
Read variables from: File...

Number must match maximum model dimensionality, currently: 2

Variables containing row coordinates should precede those containing column coordinates.

- ▶ Select Spearman as the imputation method for the Classical start.
- ▶ Click Continue.
- ▶ Click Plots in the Multidimensional Unfolding dialog box.

Figure 15-11
Plots dialog box



Multidimensional Unfolding: Plots

Plots

☐ Multiple starts ☑ Final common space ☐ Transformation plots

☐ Initial common space ☑ Space weights ☐ Shepard plots

☐ Stress per dimension ☑ Individual spaces ☐ Scatterplot of fit

☐ Residuals plots

Row Object Styles

Menu scenarios [srcid]
Gender [gender]
Toast pop-up [TP]
Buttered toast [BT]

Colors:

Markers:

Source Plots

● All sources

○ Select sources

Source number:

Sources

Add Change Remove

Row Plots

● All rows

○ Select rows

Row number:

Rows

Add Change Remove

Continue Cancel Help

- ▶ Select Individual spaces in the Plots group.
- ▶ Click Continue.
- ▶ Click OK in the Multidimensional Unfolding dialog box.

Following is the command syntax generated by these selections:

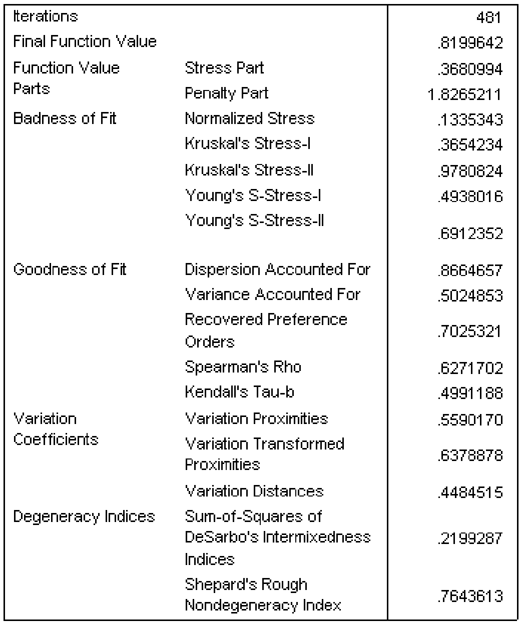
```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CLASSICAL (SPEARMAN)
  /CONDITION=ROW
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=WEIGHTED
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
```

```
    MAXITER(5000)
    /PENALTY=LAMBDA(0.5) OMEGA(1.0)
    /PRINT=MEASURES COMMON
    /PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- This syntax specifies an analysis on variables *tp (Toast pop-up)* through *cmb (Corn muffin and butter)*. The variable *srcid* is used to identify the sources.
- The `INITIAL` subcommand specifies that the starting values be imputed using Spearman distances.
- The `MODEL` subcommand specifies a weighted Euclidean model, which allows each individual space to weight the dimensions of the common space differently.
- The `PLOT` subcommand requests plots of the common space, individual spaces, and individual space weights.
- All other parameters fall back to their default values.

Measures

Figure 15-12
Measures

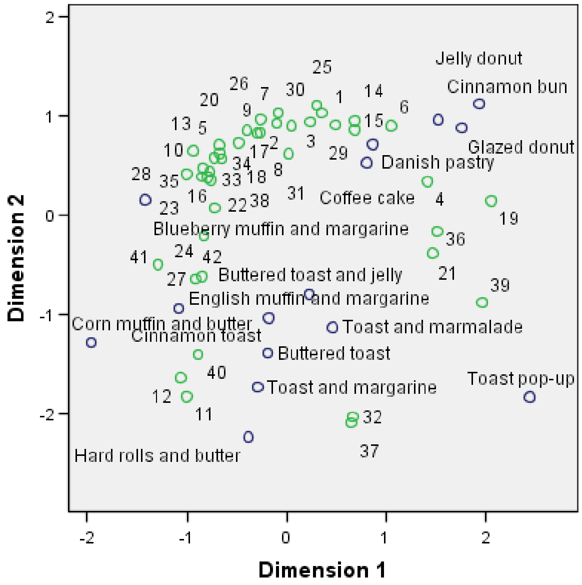


Iterations		481
Final Function Value		.8199642
Function Value Parts	Stress Part	.3680994
	Penalty Part	1.8265211
Badness of Fit	Normalized Stress	.1335343
	Kruskal's Stress-I	.3654234
	Kruskal's Stress-II	.9780824
	Young's S-Stress-I	.4938016
	Young's S-Stress-II	.6912352
Goodness of Fit	Dispersion Accounted For	.8664657
	Variance Accounted For	.5024853
	Recovered Preference Orders	.7025321
	Spearman's Rho	.6271702
	Kendall's Tau-b	.4991188
Variation Coefficients	Variation Proximities	.5590170
	Variation Transformed Proximities	.6378878
	Variation Distances	.4484515
Degeneracy Indices	Sum-of-Squares of DeSarbo's Intermixedness Indices	.2199287
	Shepard's Rough Nondegeneracy Index	.7643613

The algorithm converges after 481 iterations, with a final penalized stress of 0.8199642. The variation coefficients and Shepard's index are sufficiently large, and DeSarbo's indices are sufficiently low, to suggest that there are no problems with degeneracy.

Common Space

Figure 15-13
Joint plot of common space

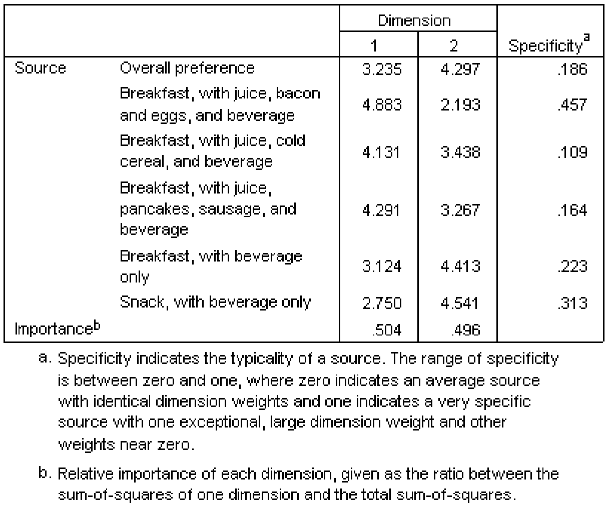


The joint plot of the common space shows a final configuration that is very similar to the two-way analysis on the overall preferences, with the solution flipped over the 45-degree line. Thus, the vertical dimension now appears to discriminate between soft and hard bread or toast, with softer items as you move up along the axis. The horizontal dimension now does not have a clear interpretation, though perhaps it discriminates based on convenience, with more “formal” items as you move left along the axis.

The individuals represented by the row objects are still clearly split into clusters according to preference for “hard” or “soft” items, with considerable within-cluster variation along the horizontal dimension.

Individual Spaces

Figure 15-14
Dimension weights



		Dimension		Specificity[a]
		1	2	
Source	Overall preference	3.235	4.297	.186
	Breakfast, with juice, bacon and eggs, and beverage	4.883	2.193	.457
	Breakfast, with juice, cold cereal, and beverage	4.131	3.438	.109
	Breakfast, with juice, pancakes, sausage, and beverage	4.291	3.267	.164
	Breakfast, with beverage only	3.124	4.413	.223
	Snack, with beverage only	2.750	4.541	.313
Importance[b]		.504	.496	

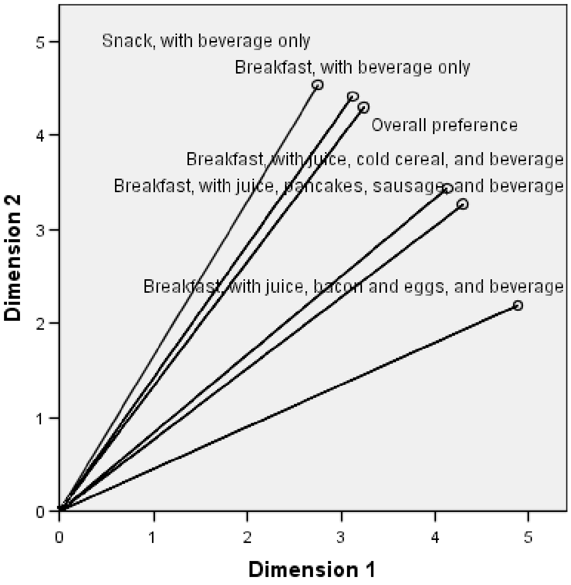
a. Specificity indicates the typicality of a source. The range of specificity is between zero and one, where zero indicates an average source with identical dimension weights and one indicates a very specific source with one exceptional, large dimension weight and other weights near zero.

b. Relative importance of each dimension, given as the ratio between the sum-of-squares of one dimension and the total sum-of-squares.

An individual space is computed for each source. The dimension weights show how the individual spaces load on the dimensions of the common space. A larger weight indicates a larger distance in the individual space and thus greater discrimination between the objects on that dimension for that individual space.

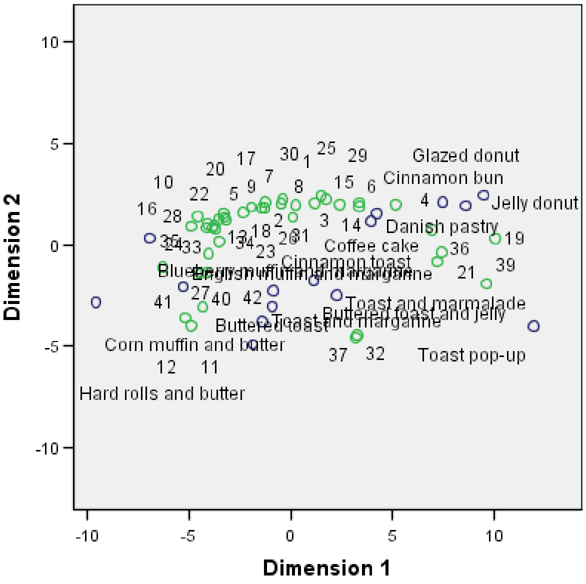
- **Specificity** is a measure of how different an individual space is from the common space. An individual space that was identical to the common space would have identical dimension weights and a specificity of 0, while an individual space that was specific to a particular dimension would have a single large dimension weight and a specificity of 1. In this case, the most divergent sources are *Breakfast, with juice, bacon and eggs, and beverage*, and *Snack, with beverage only*.
- **Importance** is a measure of the relative contribution of each dimension to the solution. In this case, the dimensions are equally important.

Figure 15-15
Dimension weights



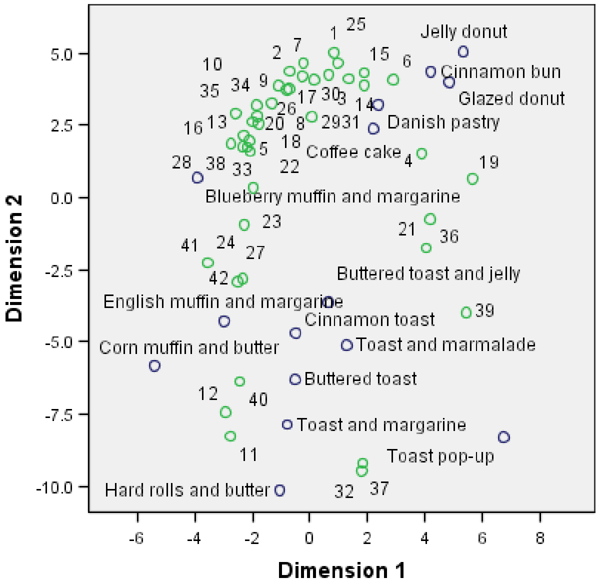
The dimension weights chart provides a visualization of the weights table. *Breakfast, with juice, bacon and eggs, and beverage* and *Snack, with beverage only* are the nearest to the dimension axes, but neither are strongly specific to a particular dimension.

Figure 15-16
Joint plot of individual space "Breakfast, with juice, bacon and eggs, and beverage"



The joint plot of the individual space *Breakfast, with juice, bacon and eggs, and beverage* shows the effect of this scenario on the preferences. This source loads more heavily on the first dimension, so the differentiation between items is mostly due to the first dimension.

Figure 15-17
Joint plot of individual space "Snack, with beverage only"



The joint plot of the individual space *Snack, with beverage only* shows the effect of this scenario on the preferences. This source loads more heavily on the second dimension, so the differentiation between items is mostly due to the second dimension. However, there is still quite a bit of differentiation along the first dimension because of the fairly low specificity of this source.

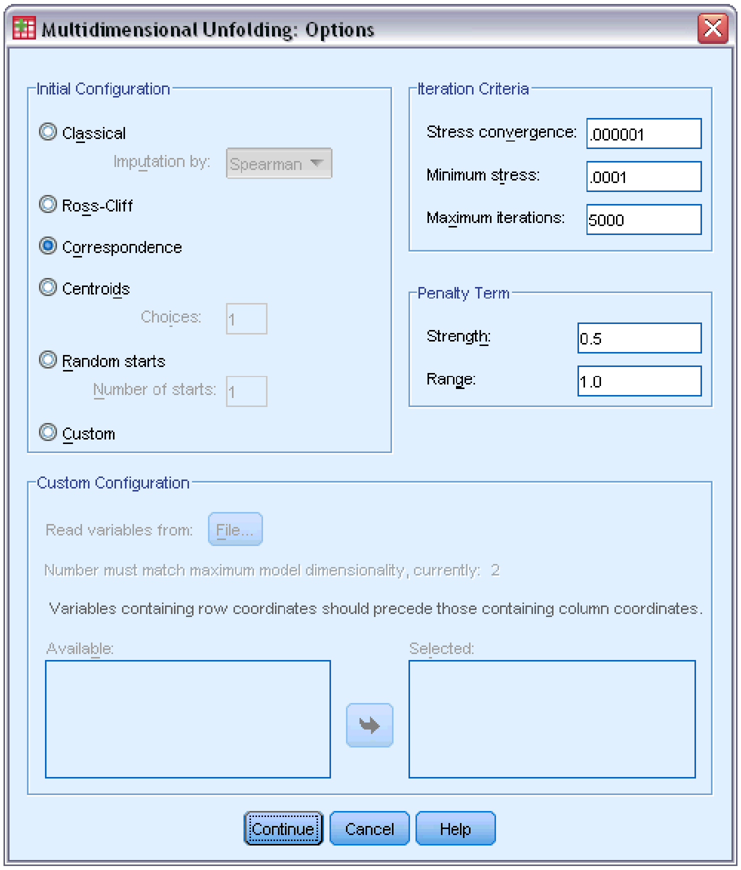
Using a Different Initial Configuration

The final configuration can depend on the starting points given to the algorithm. Ideally, the general structure of the solution should remain the same; otherwise, it can be difficult to ascertain which is correct. However, details may come into sharper focus as you try different initial configurations, such as using a correspondence start on the three-way analysis of the breakfast data.

▶ To produce a solution with a correspondence start, click the Dialog Recall tool and select Multidimensional Unfolding.

▶ Click Options in the Multidimensional Unfolding dialog box.

Figure 15-18
Options dialog box



Multidimensional Unfolding: Options

Initial Configuration

○ Classical

Imputation by: Spearman

○ Ross-Cliff

● Correspondence

○ Centroids

Choices: 1

○ Random starts

Number of starts: 1

○ Custom

Iteration Criteria

Stress convergence: .000001

Minimum stress: .0001

Maximum iterations: 5000

Penalty Term

Strength: 0.5

Range: 1.0

Custom Configuration

Read variables from: File...

Number must match maximum model dimensionality, currently: 2

Variables containing row coordinates should precede those containing column coordinates.

Available:

Selected:

▶ Select Correspondence in the Initial Configuration group.

▶ Click Continue.

▶ Click OK in the Multidimensional Unfolding dialog box.

Following is the command syntax generated by these selections:

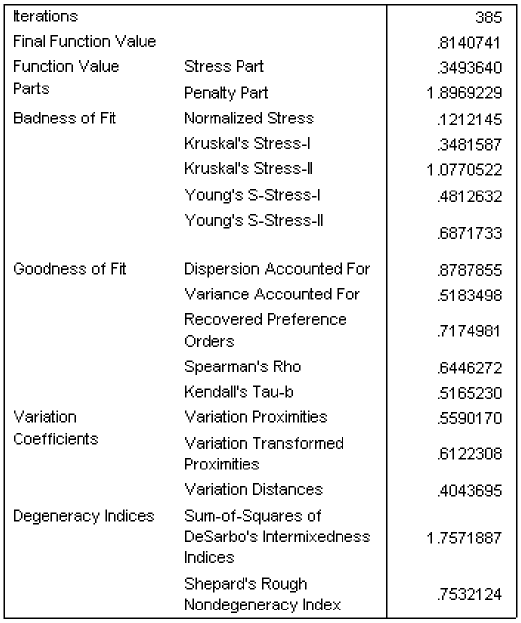
```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CORRESPONDENCE
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
```

```
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- The only change is on the `INITIAL` subcommand. The starting configuration has been set to `CORRESPONDENCE`, which uses the results of a correspondence analysis on the reversed data (similarities instead of dissimilarities), with a symmetric normalization of row and column scores.

Measures

Figure 15-19
Measures for correspondence initial configuration

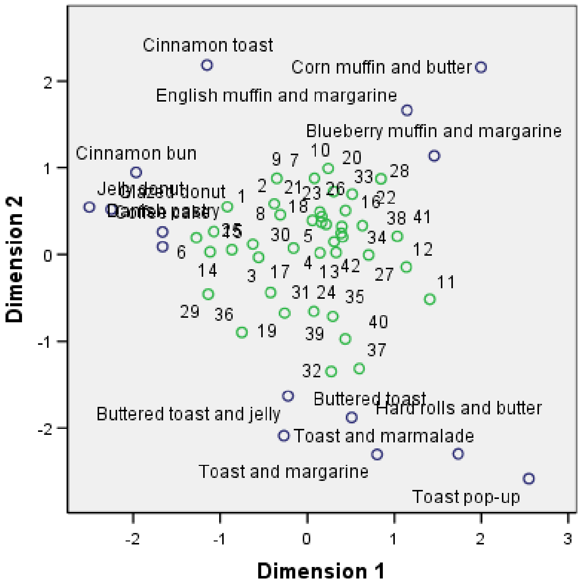


Iterations		385
Final Function Value		.8140741
Function Value Parts	Stress Part	.3493640
	Penalty Part	1.8969229
Badness of Fit	Normalized Stress	.1212145
	Kruskal's Stress-I	.3481587
	Kruskal's Stress-II	1.0770522
	Young's S-Stress-I	.4812632
	Young's S-Stress-II	.6871733
Goodness of Fit	Dispersion Accounted For	.8787855
	Variance Accounted For	.5183498
	Recovered Preference Orders	.7174981
	Spearman's Rho	.6446272
	Kendall's Tau-b	.5165230
Variation Coefficients	Variation Proximities	.5590170
	Variation Transformed Proximities	.6122308
	Variation Distances	.4043695
Degeneracy Indices	Sum-of-Squares of DeSarbo's Intermixedness Indices	1.7571887
	Shepard's Rough Nondegeneracy Index	.7532124

The algorithm converges after 385 iterations, with a final penalized stress of 0.8140741. This statistic, the badness of fit, the goodness of fit, variation coefficients, and Shepard's index are all very similar to those for the solution using the classical Spearman start. DeSarbo's indices is somewhat different, with a value of 1.7571887 instead of 0.2199287, which suggests that the solution using the correspondence start is not as well mixed. To see how this affects the solution, look at the joint plot of the common space.

Common Space

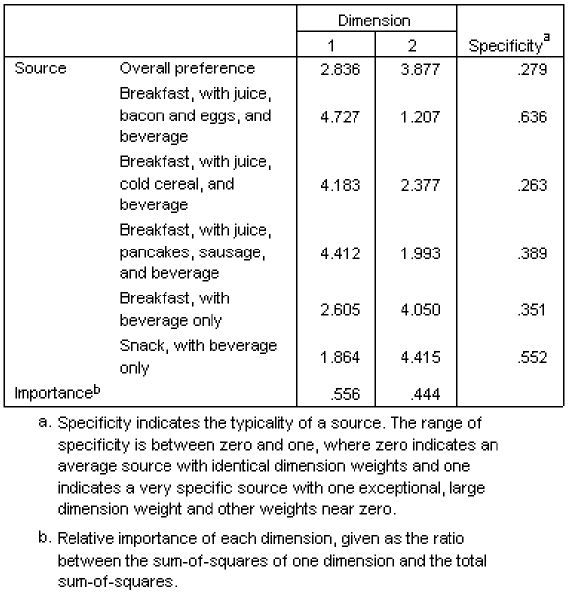
Figure 15-20
Joint plot of common space for correspondence initial configuration



The joint plot of the common space shows a final configuration that is similar to the analysis with the classical Spearman initial configuration; however, the column objects (breakfast items) are situated around the row objects (individuals) rather than intermixed with them.

Individual Spaces

Figure 15-21
Dimension weights for correspondence initial configuration



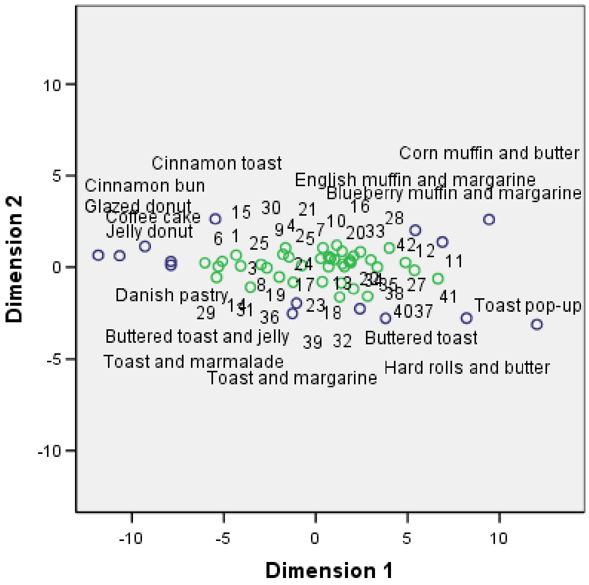
		Dimension		
		1	2	Specificity[a]
Source	Overall preference	2.836	3.877	.279
	Breakfast, with juice, bacon and eggs, and beverage	4.727	1.207	.636
	Breakfast, with juice, cold cereal, and beverage	4.183	2.377	.263
	Breakfast, with juice, pancakes, sausage, and beverage	4.412	1.993	.389
	Breakfast, with beverage only	2.605	4.050	.351
	Snack, with beverage only	1.864	4.415	.552
Importance[b]		.556	.444	

a. Specificity indicates the typicality of a source. The range of specificity is between zero and one, where zero indicates an average source with identical dimension weights and one indicates a very specific source with one exceptional, large dimension weight and other weights near zero.

b. Relative importance of each dimension, given as the ratio between the sum-of-squares of one dimension and the total sum-of-squares.

Under the correspondence initial configuration, each of the individual spaces has a higher specificity; that is, each situation under which the participants ranked the breakfast items is more strongly associated with a specific dimension. The most divergent sources are still *Breakfast, with juice, bacon and eggs, and beverage*, and *Snack, with beverage only*.

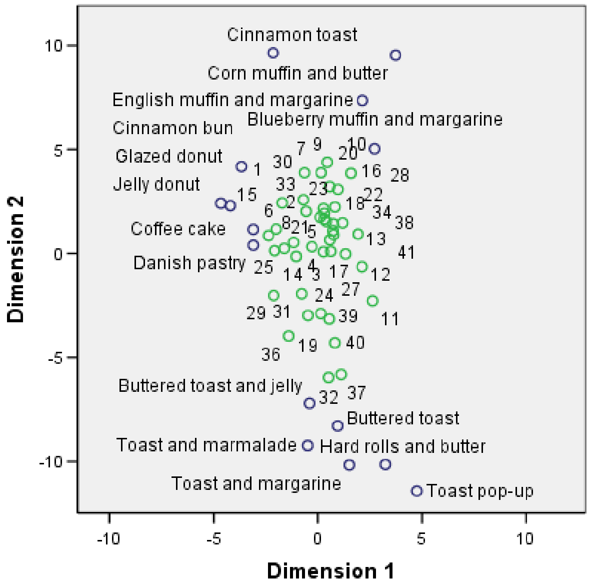
Figure 15-22
Joint plot of individual space "Breakfast, with juice, bacon and eggs, and beverage" for correspondence initial configuration



The higher specificity is evident in the joint plot of the individual space *Breakfast, with juice, bacon and eggs, and beverage*. The source loads even more heavily on the first dimension than under the classical Spearman start, so the row and column objects show a little less variation on the vertical axis and a little more variation on the horizontal axis.

Figure 15-23

Joint plot of individual space "Snack, with beverage only" for correspondence initial configuration



The joint plot of the individual space *Snack, with beverage only* shows that the row and column objects lie more closely to a vertical line than under the classical Spearman start.

Example: Examining Behavior-Situation Appropriateness

In a classic example (Price and Bouffard, 1974), 52 students were asked to rate the combinations of 15 situations and 15 behaviors on a 10-point scale ranging from 0 = "extremely appropriate" to 9 = "extremely inappropriate." Averaged over individuals, the values are taken as dissimilarities.

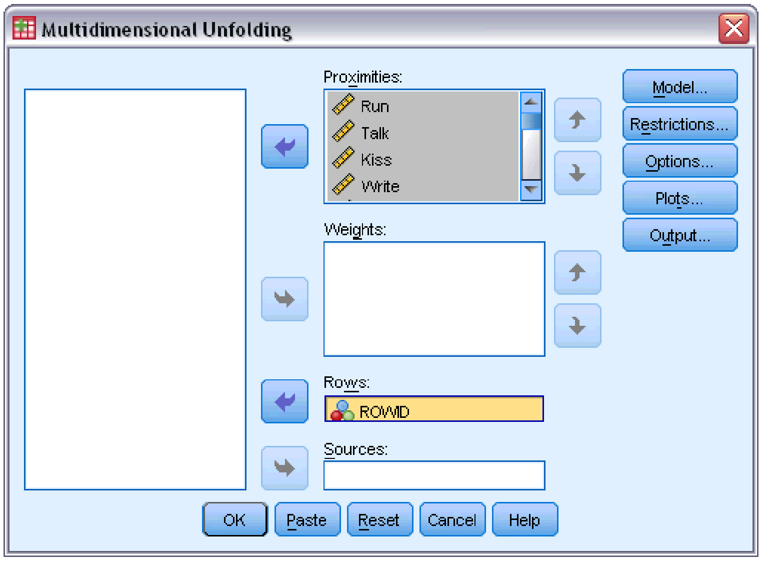
This information is collected in *behavior.sav*. For more information, see the topic Sample Files in Appendix A on p. 292. Use Multidimensional Unfolding to find clusterings of similar situations and the behaviors with which they are most closely associated. Syntax for reproducing these analyses can be found in *prefscal_behavior.sps*.

Running the Analysis

▶ To run a Multidimensional Unfolding analysis, from the menus choose:

Analyze > Scale > Multidimensional Unfolding (PREFSCAL)...

Figure 15-24
Multidimensional Unfolding main dialog box



Proximities: Run, Talk, Kiss, Write

Weights:

Rows: ROWID

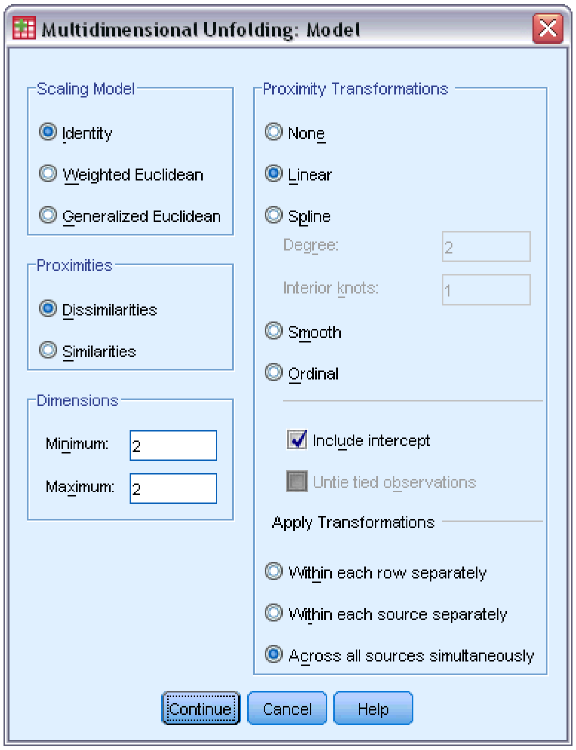
Sources:

- ▶ Select *Run* through *Shout* as proximities variables.

- ▶ Select *ROWID* as the row variable.

- ▶ Click Model.

Figure 15-25
Model dialog box



Multidimensional Unfolding: Model

Scaling Model
● Identity
○ Weighted Euclidean
○ Generalized Euclidean

Proximities
● Dissimilarities
○ Similarities

Dimensions
Minimum: 2
Maximum: 2

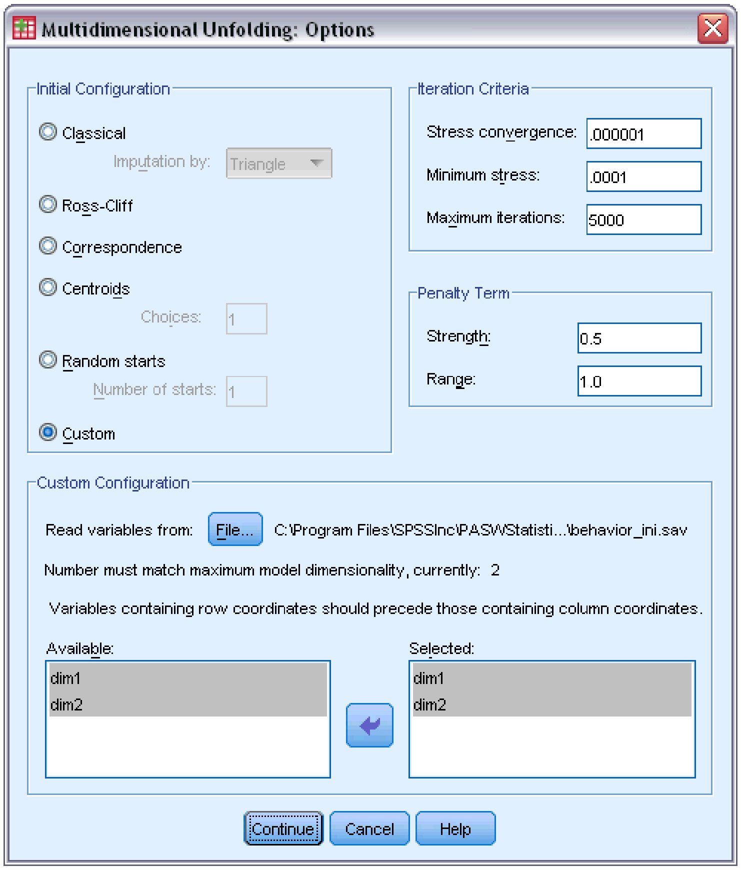
Proximity Transformations
○ None
● Linear
○ Spline
Degree: 2
Interior knots: 1
○ Smooth
○ Ordinal

☑ Include intercept
□ Untie tied observations

Apply Transformations
○ Within each row separately
○ Within each source separately
● Across all sources simultaneously

- ▶ Select Linear as the proximity transformation, and choose to Include intercept.
- ▶ Choose to apply transformations Across all sources simultaneously.
- ▶ Click Continue.
- ▶ Click Options in the Multidimensional Unfolding dialog box.

Figure 15-26
Options dialog box



Multidimensional Unfolding: Options

Initial Configuration

○ Classical

Imputation by: Triangle

○ Ross-Cliff

○ Correspondence

○ Centroids

Choices: 1

○ Random starts

Number of starts: 1

● Custom

Iteration Criteria

Stress convergence: .000001

Minimum stress: .0001

Maximum iterations: 5000

Penalty Term

Strength: 0.5

Range: 1.0

Custom Configuration

Read variables from: File... C:\Program Files\SPSSInc\PASWStatisti...\behavior_ini.sav

Number must match maximum model dimensionality, currently: 2

Variables containing row coordinates should precede those containing column coordinates.

Available: dim1, dim2

Selected: dim1, dim2

- Select Custom in the Initial Configuration group.

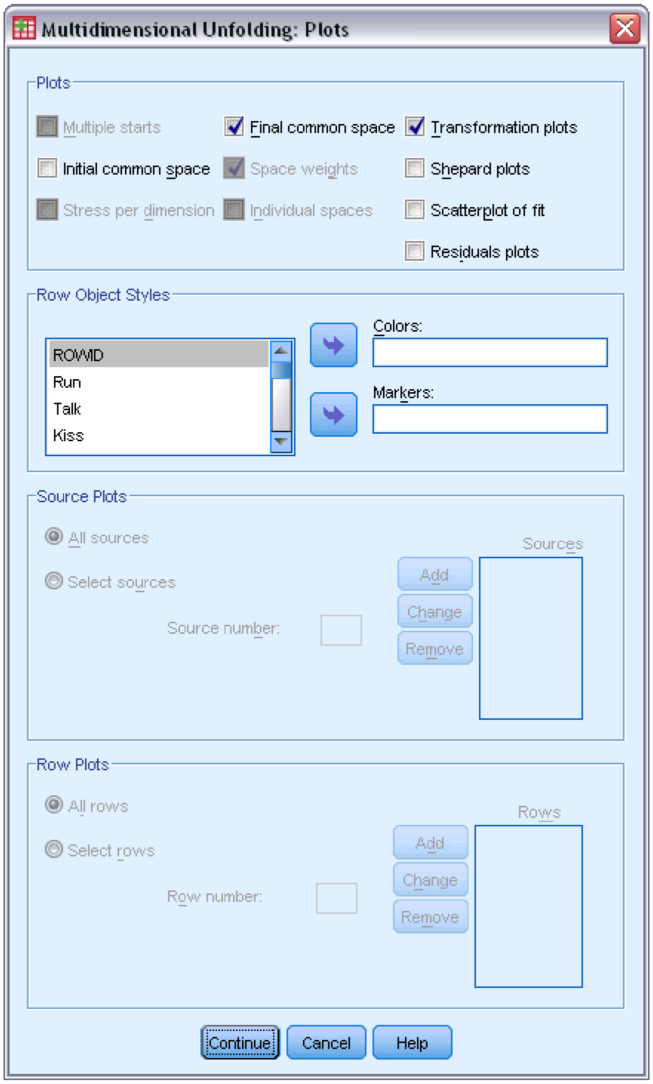
- Browse to and choose *behavior_ini.sav* as the file containing the custom initial configuration. For more information, see the topic Sample Files in Appendix A on p. 292.

- Select *dim1* and *dim2* as the variables specifying the initial configuration.

- Click Continue.

- Click Plots in the Multidimensional Unfolding dialog box.

Figure 15-27
Plots dialog box



Multidimensional Unfolding: Plots

Plots

☐ Multiple starts ☑ Final common space ☑ Transformation plots
☐ Initial common space ☑ Space weights ☐ Shepard plots
☐ Stress per dimension ☐ Individual spaces ☐ Scatterplot of fit
☐ Residuals plots

Row Object Styles

ROWID
Run
Talk
Kiss

Colors:
Markers:

Source Plots

● All sources
○ Select sources
Source number:

Row Plots

● All rows
○ Select rows
Row number:

- ▶ Select Transformation plots in the Plots group.
- ▶ Click Continue.
- ▶ Click OK in the Multidimensional Unfolding dialog box.

Following is the command syntax generated by these selections:

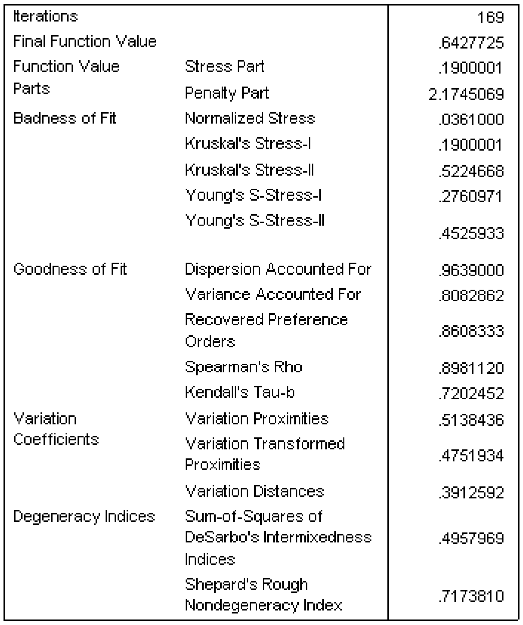
```
PREFSCAL
  VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
  Cry Laugh Shout
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=LINEAR (INTERCEPT)
  /PROXIMITIES=DISSIMILARITIES
```

```
/MODEL=IDENTITY
/CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON TRANSFORMATIONS .
```

- This syntax specifies an analysis on variables *run* through *shout*. The variable *rowid* is used to identify the rows.
- The INITIAL subcommand specifies that the starting values be taken from the file *behavior_ini.sav*. The row and column coordinates are stacked, with the column coordinates following the row coordinates.
- The CONDITION subcommand specifies that all proximities can be compared with each other. This is true in this analysis, since you should be able to compare the proxmities for running in a park and running in church and see that one behavior is considered less appropriate than the other.
- The TRANSFORMATION subcommand specifies a linear transformation of the proximities, with intercept. This is appropriate if a 1-point difference in proximities is equivalent across the range of the 10-point scale. That is, if the students have assigned their scores so that the difference between 0 and 1 is the same as the difference between 5 and 6, then a linear transformation is appropriate.
- The PLOT subcommand requests plots of the common space and transformation plots.
- All other parameters fall back to their default values.

Measures

Figure 15-28
Measures

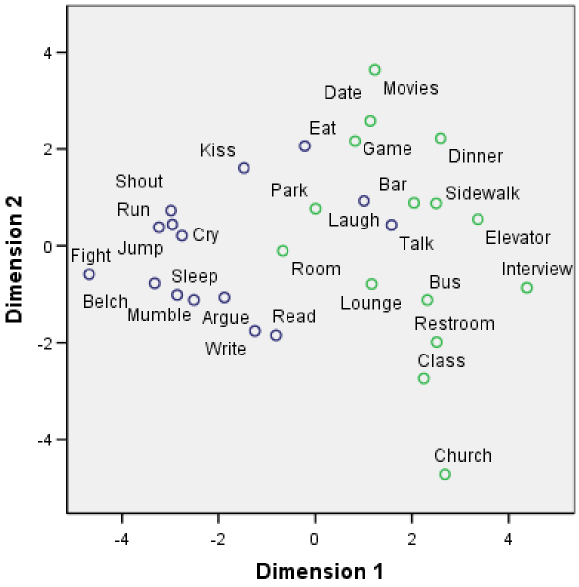


Iterations		169
Final Function Value		.6427725
Function Value Parts	Stress Part	.1900001
	Penalty Part	2.1745069
Badness of Fit	Normalized Stress	.0361000
	Kruskal's Stress-I	.1900001
	Kruskal's Stress-II	.5224668
	Young's S-Stress-I	.2760971
	Young's S-Stress-II	.4525933
Goodness of Fit	Dispersion Accounted For	.9639000
	Variance Accounted For	.8082862
	Recovered Preference Orders	.8608333
	Spearman's Rho	.8981120
	Kendall's Tau-b	.7202452
Variation Coefficients	Variation Proximities	.5138436
	Variation Transformed Proximities	.4751934
	Variation Distances	.3912592
Degeneracy Indices	Sum-of-Squares of DeSarbo's Intermixedness Indices	.4957969
	Shepard's Rough Nondegeneracy Index	.7173810

The algorithm converges after 169 iterations, with a final penalized stress of 0.6427725. The variation coefficients and Shepard's index are sufficiently large, and DeSarbo's indices are sufficiently low, to suggest that there are no problems with degeneracy.

Common Space

Figure 15-29
Joint plot of common space

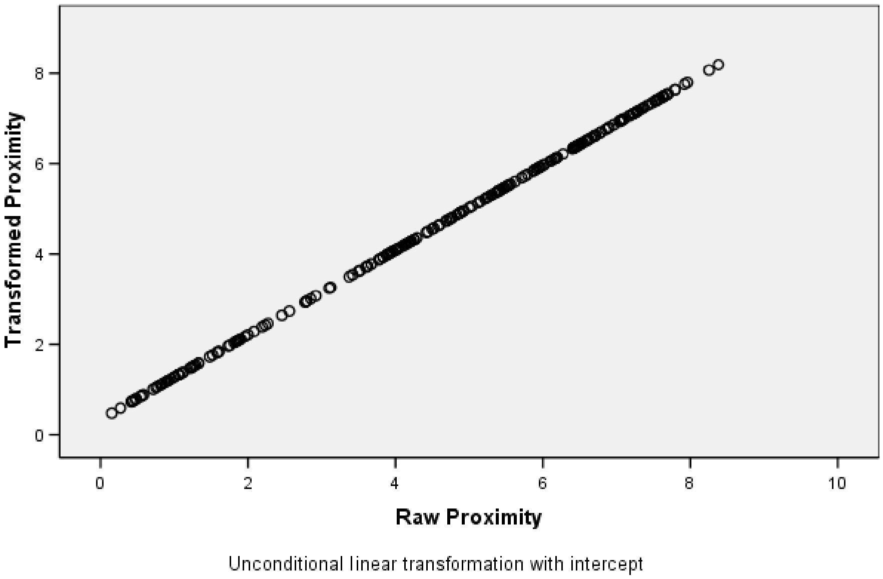


The horizontal dimension appears to be more strongly associated with the column objects (behaviors) and discriminates between inappropriate behaviors (fighting, belching) and more appropriate behaviors. The vertical dimension appears to be more strongly associated with the row objects (situations) and defines different situational-behavior restrictions.

- Toward the bottom of the vertical dimension are situations (church, class) that restrict behavior to the quieter/introspective types of behaviors (read, write). Thus, these behaviors are pulled down the vertical axis.
- Toward the top of the vertical dimension are situations (movies, game, date) that restrict behavior to the social/extroverted types of behaviors (eat, kiss, laugh). Thus, these behaviors are pulled up the vertical axis.
- At the center of the vertical dimension, situations are separated on the horizontal dimension based on the general restrictiveness of the situation. Those further from the behaviors (interview) are the most restricted, while those closer to the behaviors (room, park) are generally less restricted.

Proximity Transformations

Figure 15-30
Transformation plot



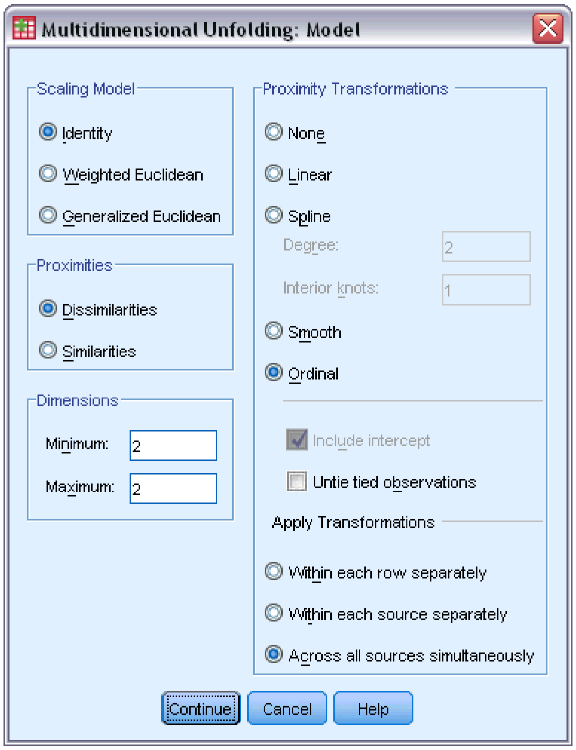
The proximities were treated as linear in this analysis, so the plot of the transformed values versus the original proximities forms a straight line. The fit of this solution is good, but perhaps a better fit can be achieved with a different transformation of the proximities.

Changing the Proximities Transformation (Ordinal)

▶ To produce a solution with an ordinal transformation of the proximities, click the Dialog Recall tool and select Multidimensional Unfolding.

▶ Click Model in the Multidimensional Unfolding dialog box.

Figure 15-31
Model dialog box



Multidimensional Unfolding: Model

Scaling Model
- ● Identity
- ○ Weighted Euclidean
- ○ Generalized Euclidean

Proximities
- ● Dissimilarities
- ○ Similarities

Dimensions
- Minimum: 2
- Maximum: 2

Proximity Transformations
- ○ None
- ○ Linear
- ○ Spline
  - Degree: 2
  - Interior knots: 1
- ○ Smooth
- ● Ordinal
- ☑ Include intercept
- □ Untie tied observations

Apply Transformations
- ○ Within each row separately
- ○ Within each source separately
- ● Across all sources simultaneously

▶ Select Ordinal as the proximity transformation.

▶ Click Continue.

▶ Click OK in the Multidimensional Unfolding dialog box.

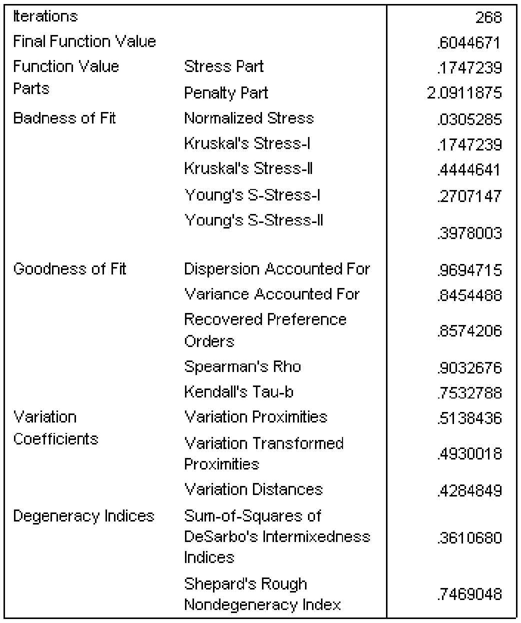
Following is the command syntax generated by these selections:

```
PREFSCAL
  VARIABLES=Run Talk Kiss Write Eat Sleep Mumble Read Fight Belch Argue Jump
  Cry Laugh Shout
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=ORDINAL (KEEPTIES)
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=IDENTITY
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON TRANSFORMATIONS .
```

- The only change is on the `TRANSFORMATION` subcommand. The transformation has been set to `ORDINAL`, which preserves the order of proximities but does not require that the transformed values be proportional to the original values.

Measures

Figure 15-32
Measures for solution with ordinal transformation

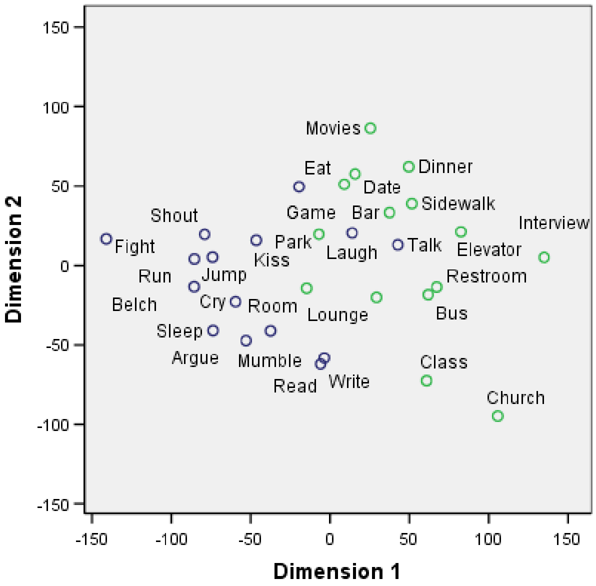


Iterations		268
Final Function Value		.6044671
Function Value Parts	Stress Part	.1747239
	Penalty Part	2.0911875
Badness of Fit	Normalized Stress	.0305285
	Kruskal's Stress-I	.1747239
	Kruskal's Stress-II	.4444641
	Young's S-Stress-I	.2707147
	Young's S-Stress-II	.3978003
Goodness of Fit	Dispersion Accounted For	.9694715
	Variance Accounted For	.8454488
	Recovered Preference Orders	.8574206
	Spearman's Rho	.9032676
	Kendall's Tau-b	.7532788
Variation Coefficients	Variation Proximities	.5138436
	Variation Transformed Proximities	.4930018
	Variation Distances	.4284849
Degeneracy Indices	Sum-of-Squares of DeSarbo's Intermixedness Indices	.3610680
	Shepard's Rough Nondegeneracy Index	.7469048

The algorithm converges after 268 iterations, with a final penalized stress of 0.6044671. This statistic and the other measures are slightly better for this solution than the one with a linear transformation of the proximities.

Common Space

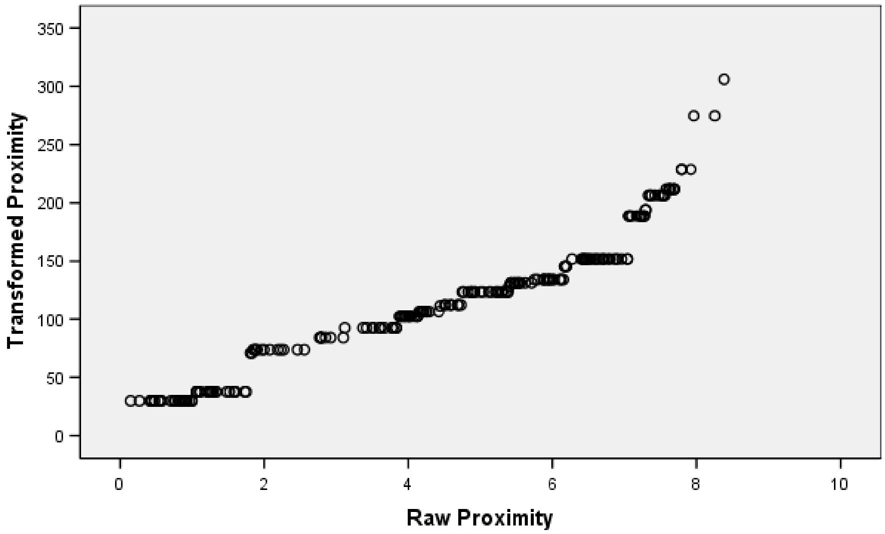
Figure 15-33
Joint plot of common space for solution with ordinal transformation



The interpretation of the common space is the same under both solutions. Perhaps this solution (with the ordinal transformation) has relatively less variation on the vertical dimension than on the horizontal dimension than is evident in the solution with the linear transformation.

Proximity Transformations

Figure 15-34
Transformation plot for solution with ordinal transformation



Unconditional ordinal transformation with ties kept tied

Aside from the values with the largest proximities, which bend up from the rest of the values, the ordinal transformation of proximities is fairly linear. These proximities likely account for most of the differences between the ordinal and linear solutions; however, there isn't enough information here to determine whether this nonlinear trend in the higher values is a true trend or an anomaly.

Recommended Readings

See the following texts for more information:

Busing, F. M. T. A., P. J. F. Groenen, and W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, 71–98.

Green, P. E., and V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Price, R. H., and D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, 579–586.

Appendix

A

Sample Files

The sample files installed with the product can be found in the *Samples* subdirectory of the installation directory. There is a separate folder within the Samples subdirectory for each of the following languages: English, French, German, Italian, Japanese, Korean, Polish, Russian, Simplified Chinese, Spanish, and Traditional Chinese.

Not all sample files are available in all languages. If a sample file is not available in a language, that language folder contains an English version of the sample file.

Descriptions

Following are brief descriptions of the sample files used in various examples throughout the documentation.

- **accidents.sav.** This is a hypothetical data file that concerns an insurance company that is studying age and gender risk factors for automobile accidents in a given region. Each case corresponds to a cross-classification of age category and gender.
- **adl.sav.** This is a hypothetical data file that concerns efforts to determine the benefits of a proposed type of therapy for stroke patients. Physicians randomly assigned female stroke patients to one of two groups. The first received the standard physical therapy, and the second received an additional emotional therapy. Three months following the treatments, each patient's abilities to perform common activities of daily life were scored as ordinal variables.
- **advert.sav.** This is a hypothetical data file that concerns a retailer's efforts to examine the relationship between money spent on advertising and the resulting sales. To this end, they have collected past sales figures and the associated advertising costs..
- **aflatoxin.sav.** This is a hypothetical data file that concerns the testing of corn crops for aflatoxin, a poison whose concentration varies widely between and within crop yields. A grain processor has received 16 samples from each of 8 crop yields and measured the alfatoxin levels in parts per billion (PPB).
- **anorectic.sav.** While working toward a standardized symptomatology of anorectic/bulimic behavior, researchers (Van der Ham, Meulman, Van Strien, and Van Engeland, 1997) made a study of 55 adolescents with known eating disorders. Each patient was seen four times over four years, for a total of 220 observations. At each observation, the patients were scored for each of 16 symptoms. Symptom scores are missing for patient 71 at time 2, patient 76 at time 2, and patient 47 at time 3, leaving 217 valid observations.
- **bankloan.sav.** This is a hypothetical data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks.
- **bankloan_binning.sav.** This is a hypothetical data file containing financial and demographic information on 5,000 past customers.

- **behavior.sav.** In a classic example (Price and Bouffard, 1974), 52 students were asked to rate the combinations of 15 situations and 15 behaviors on a 10-point scale ranging from 0="extremely appropriate" to 9="extremely inappropriate." Averaged over individuals, the values are taken as dissimilarities.
- **behavior_ini.sav.** This data file contains an initial configuration for a two-dimensional solution for *behavior.sav*.
- **brakes.sav.** This is a hypothetical data file that concerns quality control at a factory that produces disc brakes for high-performance automobiles. The data file contains diameter measurements of 16 discs from each of 8 production machines. The target diameter for the brakes is 322 millimeters.
- **breakfast.sav.** In a classic study (Green and Rao, 1972), 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference with 1="most preferred" to 15="least preferred." Their preferences were recorded under six different scenarios, from "Overall preference" to "Snack, with beverage only."
- **breakfast-overall.sav.** This data file contains the breakfast item preferences for the first scenario, "Overall preference," only.
- **broadband_1.sav.** This is a hypothetical data file containing the number of subscribers, by region, to a national broadband service. The data file contains monthly subscriber numbers for 85 regions over a four-year period.
- **broadband_2.sav.** This data file is identical to *broadband_1.sav* but contains data for three additional months.
- **car_insurance_claims.sav.** A dataset presented and analyzed elsewhere (McCullagh and Nelder, 1989) concerns damage claims for cars. The average claim amount can be modeled as having a gamma distribution, using an inverse link function to relate the mean of the dependent variable to a linear combination of the policyholder age, vehicle type, and vehicle age. The number of claims filed can be used as a scaling weight.
- **car_sales.sav.** This data file contains hypothetical sales estimates, list prices, and physical specifications for various makes and models of vehicles. The list prices and physical specifications were obtained alternately from *edmunds.com* and manufacturer sites.
- **car_sales_uprepared.sav.** This is a modified version of *car_sales.sav* that does not include any transformed versions of the fields.
- **carpet.sav.** In a popular example (Green and Wind, 1973), a company interested in marketing a new carpet cleaner wants to examine the influence of five factors on consumer preference—package design, brand name, price, a *Good Housekeeping* seal, and a money-back guarantee. There are three factor levels for package design, each one differing in the location of the applicator brush; three brand names (*K2R*, *Glory*, and *Bissell*); three price levels; and two levels (either no or yes) for each of the last two factors. Ten consumers rank 22 profiles defined by these factors. The variable *Preference* contains the rank of the average rankings for each profile. Low rankings correspond to high preference. This variable reflects an overall measure of preference for each profile.
- **carpet_prefs.sav.** This data file is based on the same example as described for *carpet.sav*, but it contains the actual rankings collected from each of the 10 consumers. The consumers were asked to rank the 22 product profiles from the most to the least preferred. The variables *PREF1* through *PREF22* contain the identifiers of the associated profiles, as defined in *carpet_plan.sav*.

- **catalog.sav.** This data file contains hypothetical monthly sales figures for three products sold by a catalog company. Data for five possible predictor variables are also included.
- **catalog_seasfac.sav.** This data file is the same as *catalog.sav* except for the addition of a set of seasonal factors calculated from the Seasonal Decomposition procedure along with the accompanying date variables.
- **cellular.sav.** This is a hypothetical data file that concerns a cellular phone company's efforts to reduce churn. Churn propensity scores are applied to accounts, ranging from 0 to 100. Accounts scoring 50 or above may be looking to change providers.
- **ceramics.sav.** This is a hypothetical data file that concerns a manufacturer's efforts to determine whether a new premium alloy has a greater heat resistance than a standard alloy. Each case represents a separate test of one of the alloys; the heat at which the bearing failed is recorded.
- **cereal.sav.** This is a hypothetical data file that concerns a poll of 880 people about their breakfast preferences, also noting their age, gender, marital status, and whether or not they have an active lifestyle (based on whether they exercise at least twice a week). Each case represents a separate respondent.
- **clothing_defects.sav.** This is a hypothetical data file that concerns the quality control process at a clothing factory. From each lot produced at the factory, the inspectors take a sample of clothes and count the number of clothes that are unacceptable.
- **coffee.sav.** This data file pertains to perceived images of six iced-coffee brands (Kennedy, Riquier, and Sharp, 1996) . For each of 23 iced-coffee image attributes, people selected all brands that were described by the attribute. The six brands are denoted AA, BB, CC, DD, EE, and FF to preserve confidentiality.
- **contacts.sav.** This is a hypothetical data file that concerns the contact lists for a group of corporate computer sales representatives. Each contact is categorized by the department of the company in which they work and their company ranks. Also recorded are the amount of the last sale made, the time since the last sale, and the size of the contact's company.
- **creditpromo.sav.** This is a hypothetical data file that concerns a department store's efforts to evaluate the effectiveness of a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months. Half received a standard seasonal ad.
- **customer_dbase.sav.** This is a hypothetical data file that concerns a company's efforts to use the information in its data warehouse to make special offers to customers who are most likely to reply. A subset of the customer base was selected at random and given the special offers, and their responses were recorded.
- **customer_information.sav.** A hypothetical data file containing customer mailing information, such as name and address.
- **customer_subset.sav.** A subset of 80 cases from *customer_dbase.sav*.
- **debate.sav.** This is a hypothetical data file that concerns paired responses to a survey from attendees of a political debate before and after the debate. Each case corresponds to a separate respondent.
- **debate_aggregate.sav.** This is a hypothetical data file that aggregates the responses in *debate.sav*. Each case corresponds to a cross-classification of preference before and after the debate.

- **demo.sav.** This is a hypothetical data file that concerns a purchased customer database, for the purpose of mailing monthly offers. Whether or not the customer responded to the offer is recorded, along with various demographic information.
- **demo_cs_1.sav.** This is a hypothetical data file that concerns the first step of a company's efforts to compile a database of survey information. Each case corresponds to a different city, and the region, province, district, and city identification are recorded.
- **demo_cs_2.sav.** This is a hypothetical data file that concerns the second step of a company's efforts to compile a database of survey information. Each case corresponds to a different household unit from cities selected in the first step, and the region, province, district, city, subdivision, and unit identification are recorded. The sampling information from the first two stages of the design is also included.
- **demo_cs.sav.** This is a hypothetical data file that contains survey information collected using a complex sampling design. Each case corresponds to a different household unit, and various demographic and sampling information is recorded.
- **dmdata.sav.** This is a hypothetical data file that contains demographic and purchasing information for a direct marketing company. *dmdata2.sav* contains information for a subset of contacts that received a test mailing, and *dmdata3.sav* contains information on the remaining contacts who did not receive the test mailing.
- **dietstudy.sav.** This hypothetical data file contains the results of a study of the "Stillman diet" (Rickman, Mitchell, Dingman, and Dalen, 1974). Each case corresponds to a separate subject and records his or her pre- and post-diet weights in pounds and triglyceride levels in mg/100 ml.
- **dvdplayer.sav.** This is a hypothetical data file that concerns the development of a new DVD player. Using a prototype, the marketing team has collected focus group data. Each case corresponds to a separate surveyed user and records some demographic information about them and their responses to questions about the prototype.
- **german_credit.sav.** This data file is taken from the "German credit" dataset in the Repository of Machine Learning Databases (Blake and Merz, 1998) at the University of California, Irvine.
- **grocery_1month.sav.** This hypothetical data file is the *grocery_coupons.sav* data file with the weekly purchases "rolled-up" so that each case corresponds to a separate customer. Some of the variables that changed weekly disappear as a result, and the amount spent recorded is now the sum of the amounts spent during the four weeks of the study.
- **grocery_coupons.sav.** This is a hypothetical data file that contains survey data collected by a grocery store chain interested in the purchasing habits of their customers. Each customer is followed for four weeks, and each case corresponds to a separate customer-week and records information about where and how the customer shops, including how much was spent on groceries during that week.
- **guttman.sav.** Bell (Bell, 1961) presented a table to illustrate possible social groups. Guttman (Guttman, 1968) used a portion of this table, in which five variables describing such things as social interaction, feelings of belonging to a group, physical proximity of members, and formality of the relationship were crossed with seven theoretical social groups, including crowds (for example, people at a football game), audiences (for example, people at a theater or classroom lecture), public (for example, newspaper or television audiences), mobs (like a crowd but with much more intense interaction), primary groups (intimate), secondary groups

(voluntary), and the modern community (loose confederation resulting from close physical proximity and a need for specialized services).

- **health_funding.sav.** This is a hypothetical data file that contains data on health care funding (amount per 100 population), disease rates (rate per 10,000 population), and visits to health care providers (rate per 10,000 population). Each case represents a different city.
- **hivassay.sav.** This is a hypothetical data file that concerns the efforts of a pharmaceutical lab to develop a rapid assay for detecting HIV infection. The results of the assay are eight deepening shades of red, with deeper shades indicating greater likelihood of infection. A laboratory trial was conducted on 2,000 blood samples, half of which were infected with HIV and half of which were clean.
- **hourlywagedata.sav.** This is a hypothetical data file that concerns the hourly wages of nurses from office and hospital positions and with varying levels of experience.
- **insurance_claims.sav.** This is a hypothetical data file that concerns an insurance company that wants to build a model for flagging suspicious, potentially fraudulent claims. Each case represents a separate claim.
- **insure.sav.** This is a hypothetical data file that concerns an insurance company that is studying the risk factors that indicate whether a client will have to make a claim on a 10-year term life insurance contract. Each case in the data file represents a pair of contracts, one of which recorded a claim and the other didn't, matched on age and gender.
- **judges.sav.** This is a hypothetical data file that concerns the scores given by trained judges (plus one enthusiast) to 300 gymnastics performances. Each row represents a separate performance; the judges viewed the same performances.
- **kinship_dat.sav.** Rosenberg and Kim (Rosenberg and Kim, 1975) set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criterion from the first sort. Thus, a total of six "sources" were obtained. Each source corresponds to a 15 × 15 proximity matrix, whose cells are equal to the number of people in a source minus the number of times the objects were partitioned together in that source.
- **kinship_ini.sav.** This data file contains an initial configuration for a three-dimensional solution for *kinship_dat.sav*.
- **kinship_var.sav.** This data file contains independent variables *gender*, *gener*(ation), and *degree* (of separation) that can be used to interpret the dimensions of a solution for *kinship_dat.sav*. Specifically, they can be used to restrict the space of the solution to a linear combination of these variables.
- **marketvalues.sav.** This data file concerns home sales in a new housing development in Algonquin, Ill., during the years from 1999–2000. These sales are a matter of public record.
- **nhis2000_subset.sav.** The National Health Interview Survey (NHIS) is a large, population-based survey of the U.S. civilian population. Interviews are carried out face-to-face in a nationally representative sample of households. Demographic information and observations about health behaviors and status are obtained for members of each household. This data file contains a subset of information from the 2000 survey. National Center for Health Statistics. National Health Interview Survey, 2000. Public-use data file and documentation. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Accessed 2003.

- **ozone.sav.** The data include 330 observations on six meteorological variables for predicting ozone concentration from the remaining variables. Previous researchers (Breiman and Friedman, 1985), (Hastie and Tibshirani, 1990), among others found nonlinearities among these variables, which hinder standard regression approaches.
- **pain_medication.sav.** This hypothetical data file contains the results of a clinical trial for anti-inflammatory medication for treating chronic arthritic pain. Of particular interest is the time it takes for the drug to take effect and how it compares to an existing medication.
- **patient_los.sav.** This hypothetical data file contains the treatment records of patients who were admitted to the hospital for suspected myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **patlos_sample.sav.** This hypothetical data file contains the treatment records of a sample of patients who received thrombolytics during treatment for myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.
- **poll_cs.sav.** This is a hypothetical data file that concerns pollsters' efforts to determine the level of public support for a bill before the legislature. The cases correspond to registered voters. Each case records the county, township, and neighborhood in which the voter lives.
- **poll_cs_sample.sav.** This hypothetical data file contains a sample of the voters listed in *poll_cs.sav*. The sample was taken according to the design specified in the *poll.csplan* plan file, and this data file records the inclusion probabilities and sample weights. Note, however, that because the sampling plan makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*poll_jointprob.sav*). The additional variables corresponding to voter demographics and their opinion on the proposed bill were collected and added the data file after the sample as taken.
- **property_assess.sav.** This is a hypothetical data file that concerns a county assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties sold in the county in the past year. Each case in the data file records the township in which the property lies, the assessor who last visited the property, the time since that assessment, the valuation made at that time, and the sale value of the property.
- **property_assess_cs.sav.** This is a hypothetical data file that concerns a state assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties in the state. Each case in the data file records the county, township, and neighborhood in which the property lies, the time since the last assessment, and the valuation made at that time.
- **property_assess_cs_sample.sav.** This hypothetical data file contains a sample of the properties listed in *property_assess_cs.sav*. The sample was taken according to the design specified in the *property_assess.csplan* plan file, and this data file records the inclusion probabilities and sample weights. The additional variable *Current value* was collected and added to the data file after the sample was taken.
- **recidivism.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender and records their demographic information, some details of their first crime, and then the time until their second arrest, if it occurred within two years of the first arrest.

- **recidivism_cs_sample.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender, released from their first arrest during the month of June, 2003, and records their demographic information, some details of their first crime, and the data of their second arrest, if it occurred by the end of June, 2006. Offenders were selected from sampled departments according to the sampling plan specified in *recidivism_cs.csplan*; because it makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** A hypothetical data file containing purchase transaction data, including date of purchase, item(s) purchased, and monetary amount of each transaction.
- **salesperformance.sav.** This is a hypothetical data file that concerns the evaluation of two new sales training courses. Sixty employees, divided into three groups, all receive standard training. In addition, group 2 gets technical training; group 3, a hands-on tutorial. Each employee was tested at the end of the training course and their score recorded. Each case in the data file represents a separate trainee and records the group to which they were assigned and the score they received on the exam.
- **satisf.sav.** This is a hypothetical data file that concerns a satisfaction survey conducted by a retail company at 4 store locations. 582 customers were surveyed in all, and each case represents the responses from a single customer.
- **screws.sav.** This data file contains information on the characteristics of screws, bolts, nuts, and tacks (Hartigan, 1975).
- **shampoo_ph.sav.** This is a hypothetical data file that concerns the quality control at a factory for hair products. At regular time intervals, six separate output batches are measured and their pH recorded. The target range is 4.5–5.5.
- **ships.sav.** A dataset presented and analyzed elsewhere (McCullagh et al., 1989) that concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the ship type, construction period, and service period. The aggregate months of service for each cell of the table formed by the cross-classification of factors provides values for the exposure to risk.
- **site.sav.** This is a hypothetical data file that concerns a company's efforts to choose new sites for their expanding business. They have hired two consultants to separately evaluate the sites, who, in addition to an extended report, summarized each site as a "good," "fair," or "poor" prospect.
- **smokers.sav.** This data file is abstracted from the 1998 National Household Survey of Drug Abuse and is a probability sample of American households. (*http://dx.doi.org/10.3886/ICPSR02934*) Thus, the first step in an analysis of this data file should be to weight the data to reflect population trends.
- **stocks.sav** This hypothetical data file contains stocks prices and volume for one year.
- **stroke_clean.sav.** This hypothetical data file contains the state of a medical database after it has been cleaned using procedures in the Data Preparation option.
- **stroke_invalid.sav.** This hypothetical data file contains the initial state of a medical database and contains several data entry errors.

- **stroke_survival.** This hypothetical data file concerns survival times for patients exiting a rehabilitation program post-ischemic stroke face a number of challenges. Post-stroke, the occurrence of myocardial infarction, ischemic stroke, or hemorrhagic stroke is noted and the time of the event recorded. The sample is left-truncated because it only includes patients who survived through the end of the rehabilitation program administered post-stroke.
- **stroke_valid.sav.** This hypothetical data file contains the state of a medical database after the values have been checked using the Validate Data procedure. It still contains potentially anomalous cases.
- **survey_sample.sav.** This data file contains survey data, including demographic data and various attitude measures. It is based on a subset of variables from the 1998 NORC General Social Survey, although some data values have been modified and additional fictitious variables have been added for demonstration purposes.
- **telco.sav.** This is a hypothetical data file that concerns a telecommunications company's efforts to reduce churn in their customer base. Each case corresponds to a separate customer and records various demographic and service usage information.
- **telco_extra.sav.** This data file is similar to the *telco.sav* data file, but the "tenure" and log-transformed customer spending variables have been removed and replaced by standardized log-transformed customer spending variables.
- **telco_missing.sav.** This data file is a subset of the *telco.sav* data file, but some of the demographic data values have been replaced with missing values.
- **testmarket.sav.** This hypothetical data file concerns a fast food chain's plans to add a new item to its menu. There are three possible campaigns for promoting the new product, so the new item is introduced at locations in several randomly selected markets. A different promotion is used at each location, and the weekly sales of the new item are recorded for the first four weeks. Each case corresponds to a separate location-week.
- **testmarket_1month.sav.** This hypothetical data file is the *testmarket.sav* data file with the weekly sales "rolled-up" so that each case corresponds to a separate location. Some of the variables that changed weekly disappear as a result, and the sales recorded is now the sum of the sales during the four weeks of the study.
- **tree_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree_credit.sav.** This is a hypothetical data file containing demographic and bank loan history data.
- **tree_missing_data.sav** This is a hypothetical data file containing demographic and bank loan history data with a large number of missing values.
- **tree_score_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.
- **tree_textdata.sav.** A simple data file with only two variables intended primarily to show the default state of variables prior to assignment of measurement level and value labels.
- **tv-survey.sav.** This is a hypothetical data file that concerns a survey conducted by a TV studio that is considering whether to extend the run of a successful program. 906 respondents were asked whether they would watch the program under various conditions. Each row represents a separate respondent; each column is a separate condition.

- **ulcer_recurrence.sav.** This file contains partial information from a study designed to compare the efficacy of two therapies for preventing the recurrence of ulcers. It provides a good example of interval-censored data and has been presented and analyzed elsewhere (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** This file reorganizes the information in *ulcer_recurrence.sav* to allow you model the event probability for each interval of the study rather than simply the end-of-study event probability. It has been presented and analyzed elsewhere (Collett et al., 2003).
- **verd1985.sav.** This data file concerns a survey (Verdegaal, 1985). The responses of 15 subjects to 8 variables were recorded. The variables of interest are divided into three sets. Set 1 includes *age* and *marital*, set 2 includes *pet* and *news*, and set 3 includes *music* and *live*. *Pet* is scaled as multiple nominal and *age* is scaled as ordinal; all of the other variables are scaled as single nominal.
- **virus.sav.** This is a hypothetical data file that concerns the efforts of an Internet service provider (ISP) to determine the effects of a virus on its networks. They have tracked the (approximate) percentage of infected e-mail traffic on its networks over time, from the moment of discovery until the threat was contained.
- **wheeze_steubenville.sav.** This is a subset from a longitudinal study of the health effects of air pollution on children (Ware, Dockery, Spiro III, Speizer, and Ferris Jr., 1984). The data contain repeated binary measures of the wheezing status for children from Steubenville, Ohio, at ages 7, 8, 9 and 10 years, along with a fixed recording of whether or not the mother was a smoker during the first year of the study.
- **workprog.sav.** This is a hypothetical data file that concerns a government works program that tries to place disadvantaged people into better jobs. A sample of potential program participants were followed, some of whom were randomly selected for enrollment in the program, while others were not. Each case represents a separate program participant.
- **worldsales.sav** This hypothetical data file contains sales revenue by continent and product.

Appendix

B

Notices

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, ibm.com, and SPSS are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shtml*.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Java
COMPATIBLE

Bibliography

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, and H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: *Methodologies of Pattern Recognition,* S. Watanabe, ed. New York: Academic Press, 35–74.

Benzécri, J. P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Bishop, Y. M., S. E. Feinberg, and P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

Blake, C. L., and C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., and J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580–598.

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, 1032–1069.

Busing, F. M. T. A., P. J. F. Groenen, and W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, 71–98.

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association, 227–228.

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

Commandeur, J. J. F., and W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, University of Leiden.

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, and J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, 579–595.

De Leeuw, J. 1982. Nonlinear principal components analysis. In: *COMPSTAT Proceedings in Computational Statistics,* Vienna: Physica Verlag, 77–89.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd ed. Leiden: DSWO Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. In: *Data Analysis and Informatics III,* E. Diday, et al., ed., 415–424.

De Leeuw, J., and W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. In: *Multivariate Analysis, Vol. V,* P. R. Krishnaiah, ed. Amsterdam: North-Holland, 501–522.

De Leeuw, J., and J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. In: *Data Analysis and Informatics,* E. Diday,et al., ed. Amsterdam: North-Holland, 231–242.

De Leeuw, J., F. W. Young, and Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 471–503.

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. In: *Progress in Multivariate Analysis,* S. Das Gupta, and J. Sethuraman, eds. Calcutta: Indian Statistical Institute.

Eckart, C., and G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, 211–218.

Fisher, R. A. 1938. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, 422–429.

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, 453–467.

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02*. Leiden: Department of Data Theory, University of Leiden.

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Gilula, Z., and S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, 760–771.

Gower, J. C., and J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, 43–51.

Green, P. E., and V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., and Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: *The Prediction of Personal Adjustment,* P. Horst, ed. New York: Social Science Research Council, 319–348.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, 469–506.

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., and R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Hastie, T., R. Tibshirani, and A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, 1255–1270.

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, 93–96.

Heiser, W. J. 1981. *Unfolding analysis of proximity data*. Leiden: Department of Data Theory, University of Leiden.

Heiser, W. J., and F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: SagePublications, Inc., 25–48.

Heiser, W. J., and J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, and J. Blasius, eds. New York: Academic Press, 179–209.

Heiser, W. J., and J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. In: *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski, ed. Oxford: Oxford UniversityPress, 51–89.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, 331–347.

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, 129–149.

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Kennedy, R., C. Riquier, and B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56–70.

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, 433–460.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–28.

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, 251–263.

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. In: *International Encyclopedia of Statistics,* W. H. Kruskal, and J. M. Tanur, eds. New York: The Free Press, 307–330.

Kruskal, J. B., and R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, 123–157.

Krzanowski, W. J., and F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. London: Edward Arnold.

Lebart, L., A. Morineau, and K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Lingoes, J. C. 1968. The multivariate analysis of qualitative data. *Multivariate Behavioral Research*, 3, 61–94.

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, 7–12.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, 539–565.

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, 287–300.

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, 249–266.

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, 493–517.

Meulman, J. J., and W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. In: *Visual Display of Categorical Data,* M. Greenacre, and J. Blasius, eds. New York: Academic Press, 277–296.

Meulman, J. J., and P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, 7–35.

Meulman, J. J., A. J. Van der Kooij, and A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. In: *Classification, Automation and New Media,* W. Gaul, and G. Ritter, eds. Berlin: Springer-Verlag, 207–217.

Meulman, J. J., A. J. Van der Kooij, and W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan, ed. Thousand Oaks, Calif.: Sage Publications, Inc., 49–70.

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, 25–36.

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Pratt, J. W. 1987. Dividing the indivisible: Using simple symmetry to partition variance explained. In: *Proceedings of the Second International Conference in Statistics,* T. Pukkila, and S. Puntanen, eds. Tampere, Finland: Universityof Tampere, 245–260.

Price, R. H., and D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, 579–586.

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, 425–441.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In: *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah, ed. Amsterdam: North-Holland, 3–22.

Rickman, R., N. Mitchell, J. Dingman, and J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, 54–58.

Rosenberg, S., and M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, 489–502.

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, 125–140.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, 219–246.

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, 287–315.

Tenenhaus, M., and F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 91–119.

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, and J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, 109–126.

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. In: *Psychological Scaling: Theory & Applications,* H. Gulliksen, and S. Messick, eds. NewYork: John Wiley and Sons, 155–167.

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., and J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54–80.

Van der Burg, E., J. De Leeuw, and R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, 177–197.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, and H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, 363–368.

Van der Kooij, A. J., and J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. In: *Softstat '97,* F. Faulbaum, and W. Bandilla, eds. Stuttgart: Gustav Fisher, 99–106.

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Verboon, P., and I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, 485–507.

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in Dutch)*. Leiden: Department of Data Theory, University of Leiden.

Vlek, C., and P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, 235–271.

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. London: Lawrence Erlbaum Associates, Inc.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, and B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, 366–374.

Winsberg, S., and J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, 669–674.

Winsberg, S., and J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, 575–595.

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, 357–387.

Young, F. W., J. De Leeuw, and Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 505–528.

Young, F. W., Y. Takane, and J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, 279–281.

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, and J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, 281–302.

Index