IBM SPSS Categories 26

Note
Before using this information and the product it supports, read the information in "Notices" on page 47.

Product Information

This edition applies to version 26, release 0, modification 0 of IBM SPSS Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Categories

The following categories features are included in SPSS® Statistics Professional Edition or the Categories option.

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. 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.

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 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 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 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

Table 4. Combinations of region and job (continued)

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 Statistics Base Edition (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 Statistics Base Edition (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.

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.

Categorical Regression data considerations

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

1. From the menus choose:
   Analyze > **Regression** > **Optimal Scaling (CATREG)...**
2. Select the dependent variable and independent variable(s).
3. 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.

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.

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.

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.

Categorical Regression Discretization

The Discretize dialog 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.

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 allows you to choose the strategy for handling missing values in analysis variables and supplementary variables.

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

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). You cannot weight supplementary objects (specified weights are ignored).

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.

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 guaranteed, 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. See the topic "Categorical Regression Save" on page 14 for more information.

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.

Categorical Regression Regularization

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. See the topic "Categorical Regression Save" on page 14 for more information.

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.

Elastic Net Plots

For the Elastic Net method, separate regularization plots are produced by values of the Ridge regression penalty. **Produce all possible Elastic Net plots** uses every value in the range determined by the minimum and maximum Ridge regression penalty values specified. **Produce Elastic Net plots 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).

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.

Categorical Regression Output

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

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. The prediction error for each model is estimated by applying the fitted model 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 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.

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 allows you to specify the variables that will produce transformation and residual plots.

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.

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.

Categorical principal components analysis data considerations

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.

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.

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.

Degree
The degree of the polynomial.

Interior Knots
The number of interior knots.

Categorical Principal Components Analysis Discretize

The Discretization dialog 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.

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 Principal Components Analysis Missing Values

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

Missing Value Strategy

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

Strategy

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 provides controls to select the initial configuration, specify iteration and convergence criteria, select a normalization method, choose the method for labeling plots, and specify supplementary objects.

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. 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] in the **Custom value** field. 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.

Rotation
You can select a rotation method to obtain rotated results.

Varimax
An orthogonal rotation method that minimizes the number of variables that have high loadings on each component. It simplifies the interpretation of the components.

Oblimin
A method for oblique (non-orthogonal) rotation. When delta equals 0, components are most oblique. As delta becomes more negative, the components become less oblique. Positive values permit additional component correlation. The value of **Delta** must be less than or equal to 0.8.

Quartimax
A rotation method that minimizes the number of components that are needed to explain each variable. It simplifies the interpretation of the observed variables.

Equamax
A rotation method that is a combination of the Varimax method, which simplifies the components, and the Quartimax method, which simplifies the variables. The number of variables that load highly on a component and the number of components that are needed to explain a variable are minimized.

Promax
An oblique (non-orthogonal) rotation, which allows components to be correlated. It can be calculated more quickly than a direct Oblimin rotation, so it is useful for large

datasets. The amount of correlation (obliqueness) that is allowed is controlled by the kappa parameter. The value of **Kappa** must be greater than or equal to 1 and less 10,000.

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.

None A configuration file is not used.

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 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.

Tables

Object scores

Displays the object scores, including mass, inertia, and contributions, and enables the following options:

Object Scores 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. You can sort the component loadings by size.

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 (coordinates), including mass, and inertia for each dimension of the variable(s) selected.

Note: The coordinates (including the mass and inertia) are displayed in separate layers of the pivot table output, with the coordinates shown by default.

Descriptive Statistics

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

Categorical Principal Components Analysis Save

The Save dialog 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.
- File names 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.

Categorical Principal Components Analysis Object Plots

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

Plots

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 of variables.

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) 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 allows you to specify the types of plots you want and the variables for which plots will be produced.

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 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 controls the variables that are included in the plot, display of centroids in the loadings plot, and display of plots of variance accounted for.

Variance accounted for
For each dimension, displays a plot of variance accounted for.

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 can 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.

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.

Correspondence Analysis data considerations

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

1. From the menus choose:

   Analyze > **Dimension Reduction** > **Correspondence Analysis...**
2. Select a row variable.
3. Select a column variable.
4. Define the ranges for the variables.
5. 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.

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.

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

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 allows you to specify the numerical output produced.

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.

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.

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.

Confidence Statistics for

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

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

Correspondence Analysis Plots

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

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.

ID label width for scatterplots
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.

ID label width for line plots
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.

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.

Multiple Correspondence Analysis data considerations

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 non-missing.

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

1. From the menus choose:

   Analyze > **Dimension Reduction** > **Optimal Scaling...**
2. Select **All variables multiple nominal**.
3. Select **One set**.
4. Click **Define**.

5. Select at least two analysis variables and specify the number of dimensions in the solution.
6. 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.

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 Discretize

The Discretization dialog 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.

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 allows you to choose the strategy for handling missing values in analysis variables and supplementary variables.

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

Strategy

  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 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.

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. 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] in the **Custom value** field. 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.

None A configuration file is not used.

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 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.

Tables

Object scores

Displays the object scores, including mass, inertia, and contributions, and enables the following options:

Object Scores 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 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.
- File names 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.

Multiple Correspondence Analysis Object Plots

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

Plots

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) or with their case numbers. One plot is produced per variable if **Variable** is selected.

Multiple Correspondence Analysis Variable Plots

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

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; 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.

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.

Multidimensional Scaling data considerations

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

1. From the menus choose:

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

   This opens the Data Format dialog box.
2. 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.

Proximities in Matrices across Columns

If you select the proximities in matrices data model for either one source or multiple sources in the Variables dialog, then do the following:

1. 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.)
2. 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.)
3. 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 Variables dialog, then do the following:

1. Select two or more proximities variables. (Each variable is assumed to be a matrix of proximities from a separate source.)
2. Select a rows variable to define the row locations for the proximities in each proximities variable.
3. 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.)
4. 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 Variables dialog, then do the following:

1. Select a proximities variable. (t is assumed to be one or more matrices of proximities.)
2. Select a rows variable to define the row locations for the proximities in the proximities variable.
3. Select a columns variable to define the column locations for the proximities in the proximities variable.
4. 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.)
5. 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 Variables dialog, then do the following:

1. If you create distances between variables, 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.
2. If there are multiple sources, select a sources variable.
3. 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

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.

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**).

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

Define a Multidimensional Scaling Model

The Model dialog 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.

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.

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.

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.

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

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 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 allows you to place restrictions on the common space.

Restrictions on Common Space
Specify the type of restriction you want.

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

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.

Torgerson
A classical scaling solution is used as the initial configuration.

Single random start
A configuration is chosen at random.

Custom
You select variables that contain the coordinates of your own initial configuration (in the **Custom Configuration** section). 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.

Multiple random starts
Several configurations are chosen at random, and the configuration with the lowest normalized raw stress is used as the initial configuration.

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 allows you to specify which plots will be produced. This topic describes the Plots dialog if you have the Proximities in Columns data format. 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 Variables dialog.

Plots

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 allows you to specify which plots will be produced. This topic describes the Plots dialog if your data format is anything other than Proximities in Columns. 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.

Multidimensional Scaling Output

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

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.

Common space coordinates
Displays the coordinates of the common space. You can save the common space coordinates to separate IBM SPSS Statistics data files.

Distances
Displays the distances between the objects in the configuration. You can save the distances to separate IBM SPSS Statistics data files.

Transformed independent variables
If a linear combination restriction was selected, the transformed independent variables and the corresponding regression weights are displayed. You can save the transformed independent variables to separate IBM SPSS Statistics data files.

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. You can save the individual space weights to separate IBM SPSS Statistics data files.

Transformed proximities
Displays the transformed proximities between the objects in the configuration. You can save the transformed proximities 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.

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.

Multidimensional Unfolding data considerations

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

1. From the menus choose:

   Analyze > **Scale** > **Multidimensional Unfolding (PREFSCAL)...**
2. Select two or more variables that identify the columns in the rectangular proximity matrix. Each variable represents a separate column object.
3. 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.
4. Optionally, select a rows variable. The values (or value labels) of this variable are used to label row objects in the output.
5. 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 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.

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 allows you to place restrictions on the common space.

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

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.

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 allows you to specify which plots will be produced.

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.

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

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.

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.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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, MD-NC119
Armonk, NY 10504-1785
US

For license inquiries regarding double-byte (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.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION 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 jurisdictions 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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide 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 Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

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.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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-IBMproducts. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at 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.

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.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

N

O

P

R

S

T

V

W

Printed in USA