Chapter 40
Multivariate Analyses

Chapter Table of Contents

VARIABLES 648

METHOD 649
- Principal Component Analysis 651
- Principal Component Rotation 653
- Canonical Correlation Analysis 654
- Maximum Redundancy Analysis 655
- Canonical Discriminant Analysis 655

OUTPUT 657
- Principal Component Analysis 658
- Principal Component Rotation 659
- Canonical Correlation Analysis 660
- Maximum Redundancy Analysis 661
- Canonical Discriminant Analysis 662

TABLES 663
- Univariate Statistics 663
- Sums of Squares and Crossproducts 663
- Corrected Sums of Squares and Crossproducts 664
- Covariance Matrix 664
- Correlation Matrix 664
- P-Values of the Correlations 665
- Inverse Correlation Matrix 666
- Pairwise Correlations 667
- Principal Component Analysis 668
- Principal Components Rotation 671
- Canonical Correlation Analysis 673
- Maximum Redundancy Analysis 676
- Canonical Discriminant Analysis 678

GRAPHS 681
- Scatter Plot Matrix 681
- Principal Component Plots 682
- Component Rotation Plots 685
- Canonical Correlation Plots 686

Maximum Redundancy Plots 689
Canonical Discrimination Plots 691

CONFIDENCE ELLIPSES 694
Scatter Plot Confidence Ellipses 695
Canonical Discriminant Confidence Ellipses 696

OUTPUT VARIABLES 697
Principal Components 698
Principal Component Rotation 698
Canonical Variables 698
Maximum Redundancy 698
Canonical Discriminant 698

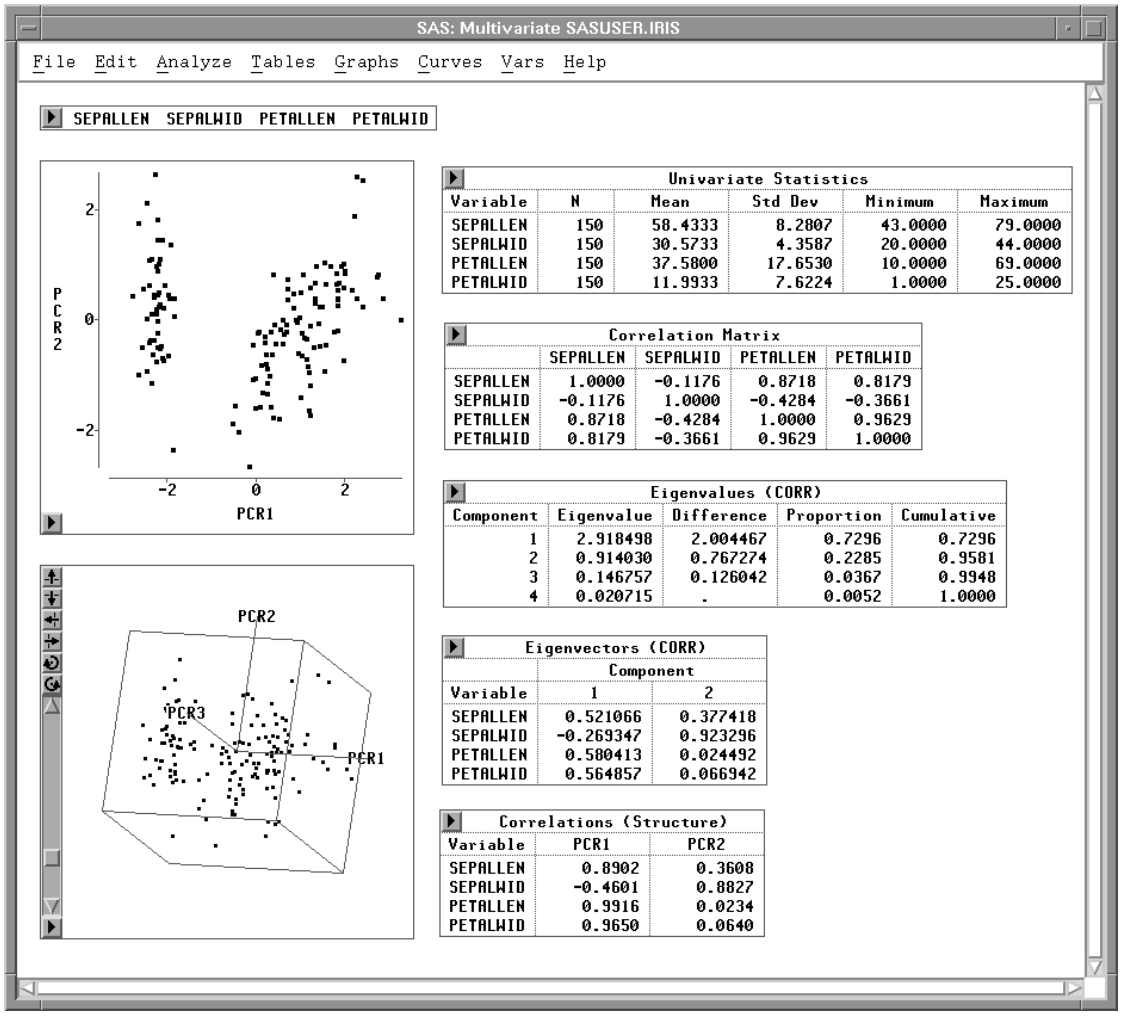
WEIGHTED ANALYSES 699

REFERENCES 700

Chapter 40
Multivariate Analyses

Choosing **Analyze:Multivariate (Y X)** gives you access to a variety of *multivariate analyses*. These provide methods for examining relationships among variables and between two sets of variables.

You can calculate correlation matrices and scatter plot matrices with confidence ellipses to explore relationships among pairs of variables. You can use principal component analysis to examine relationships among several variables, canonical correlation analysis and maximum redundancy analysis to examine relationships between two sets of interval variables, and canonical discriminant analysis to examine relationships between a nominal variable and a set of interval variables.



Univariate Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
SEPALLEN	150	58.4333	8.2807	43.0000	79.0000
SEPALWID	150	30.5733	4.3587	20.0000	44.0000
PETALLEN	150	37.5800	17.6530	10.0000	69.0000
PETALWID	150	11.9933	7.6224	1.0000	25.0000

Correlation Matrix

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.0000	-0.1176	0.8718	0.8179
SEPALWID	-0.1176	1.0000	-0.4284	-0.3661
PETALLEN	0.8718	-0.4284	1.0000	0.9629
PETALWID	0.8179	-0.3661	0.9629	1.0000

Eigenvalues (CORR)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	2.918498	2.004467	0.7296	0.7296
2	0.914030	0.767274	0.2285	0.9581
3	0.146757	0.126042	0.0367	0.9948
4	0.020715	.	0.0052	1.0000

Eigenvectors (CORR)

	Component	
Variable	1	2
SEPALLEN	0.521066	0.377418
SEPALWID	-0.269347	0.923296
PETALLEN	0.580413	0.024492
PETALWID	0.564857	0.066942

Correlations (Structure)

Variable	PCR1	PCR2
SEPALLEN	0.8902	0.3608
SEPALWID	-0.4601	0.8827
PETALLEN	0.9916	0.0234
PETALWID	0.9650	0.0640

Figure 40.1. Multivariate Analysis

Variables

To create a multivariate analysis, choose **Analyze:Multivariate (Y's)**. If you have already selected one or more interval variables, these selected variables are treated as **Y** variables and a multivariate analysis for the variables appears. If you have not selected any variables, a variables dialog appears.

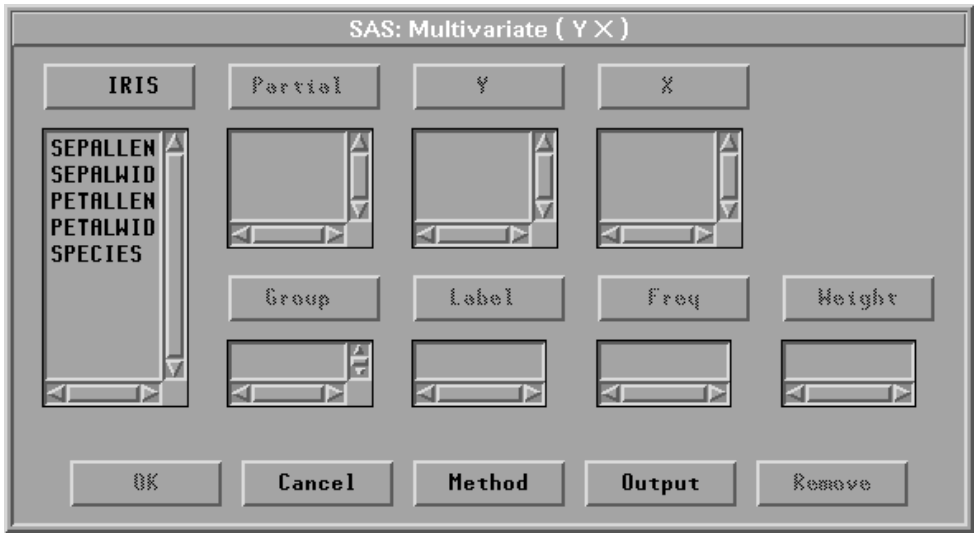


Figure 40.2. Multivariate Variables Dialog

Select at least one **Y** variable. With canonical correlation analysis and maximum redundancy analysis, you need to select a set of **X** variables. With canonical discriminant analysis, you need to select a nominal **Y** variable and a set of **X** variables.

Without **X** variables, sums of squares and crossproducts, corrected sums of squares and crossproducts, covariances, and correlations are displayed as symmetric matrices with **Y** variables as both the row variables and the column variables. With a nominal **Y** variable, these statistics are displayed as symmetric matrices with **X** variables as both the row variables and the column variables. When both interval **Y** variables and interval **X** variables are selected, these statistics are displayed as rectangular matrices with **Y** variables as the row variables and **X** variables as the column variables.

You can select one or more **Partial** variables. The multivariate analysis analyzes **Y** and **X** variables using their residuals after partialling out the **Partial** variables.

You can select one or more **Group** variables, if you have grouped data. This creates one multivariate analysis for each group. You can select a **Label** variable to label observations in the plots.

You can select a **Freq** variable. If you select a **Freq** variable, each observation is assumed to represent n_i observations, where n_i is the value of the **Freq** variable.

You can select a **Weight** variable to specify relative weights for each observation in the analysis. The details of weighted analyses are explained in the "Method" section, which follows, and the "Weighted Analyses" section at the end of this chapter.

Method

Observations with missing values for any of the **Partial** variables are not used. Observations with **Weight** or **Freq** values that are missing or that are less than or equal to 0 are not used. Only the integer part of **Freq** values is used.

Observations with missing values for **Y** or **X** variables are not used in the analysis except for the computation of pairwise correlations. Pairwise correlations are computed from all observations that have nonmissing values for any pair of variables.

The following notation is used in this chapter:

- n is the number of nonmissing observations.
- n_p, n_y, and n_x are the numbers of **Partial**, **Y**, and **X** variables.
- d is the variance divisor.
- w_i is the ith observation weight (values of the **Weight** variable).
- $\mathbf{y}_i$ and $\mathbf{x}_i$ are the ith observed nonmissing Y and X vectors.
- $\overline{\mathbf{y}}$ and $\overline{\mathbf{x}}$ are the sample mean vectors, $\sum_{i=1}^{n} \mathbf{y}_i/n$, $\sum_{i=1}^{n} \mathbf{x}_i/n$.

The sums of squares and crossproducts of the variables are

- $\mathbf{U}_{yy} = \sum_{i=1}^{n} \mathbf{y}_i\mathbf{y}_i'$
- $\mathbf{U}_{yx} = \sum_{i=1}^{n} \mathbf{y}_i\mathbf{x}_i'$
- $\mathbf{U}_{xx} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$

The corrected sums of squares and crossproducts of the variables are

- $\mathbf{C}_{yy} = \sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'$
- $\mathbf{C}_{yx} = \sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{x}_i - \overline{\mathbf{x}})'$
- $\mathbf{C}_{xx} = \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$

If you select a **Weight** variable, the sample mean vectors are

$$\overline{\mathbf{y}} = \sum_{i=1}^{n} w_i\mathbf{y}_i / \sum_{i=1}^{n} w_i \qquad \overline{\mathbf{x}} = \sum_{i=1}^{n} w_i\mathbf{x}_i / \sum_{i=1}^{n} w_i$$

The sums of squares and crossproducts with a **Weight** variable are

- $\mathbf{U}_{yy} = \sum_{i=1}^{n} w_i\mathbf{y}_i\mathbf{y}_i'$
- $\mathbf{U}_{yx} = \sum_{i=1}^{n} w_i\mathbf{y}_i\mathbf{x}_i'$
- $\mathbf{U}_{xx} = \sum_{i=1}^{n} w_i\mathbf{x}_i\mathbf{x}_i'$

The corrected sums of squares and crossproducts with a **Weight** variable are

- $\mathbf{C}_{yy} = \sum_{i=1}^{n} w_i(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'$
- $\mathbf{C}_{yx} = \sum_{i=1}^{n} w_i(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{x}_i - \overline{\mathbf{x}})'$
- $\mathbf{C}_{xx} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$

The covariance matrices are computed as

$$\mathbf{S}_{yy} = \mathbf{C}_{yy}/d \qquad \mathbf{S}_{yx} = \mathbf{C}_{yx}/d \qquad \mathbf{S}_{xx} = \mathbf{C}_{xx}/d$$

To view or change the variance divisor d used in the calculation of variances and covariances, or to view or change other method options in the multivariate analysis, click on the **Method** button from the variables dialog to display the method options dialog.

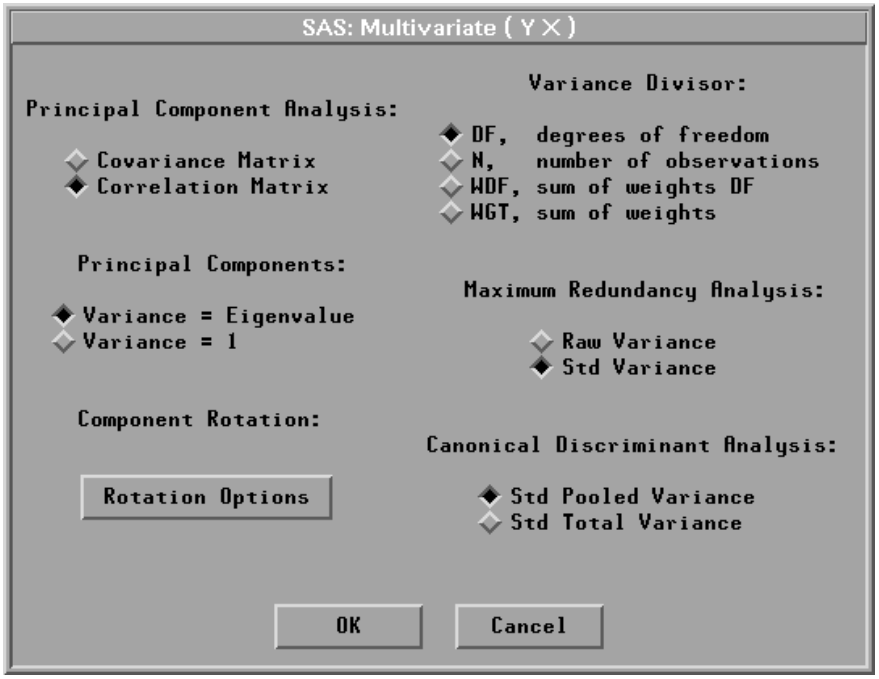


Figure 40.3. Multivariate Method Options Dialog

The variance divisor d is defined as

- $d = n - n_p - 1$ for vardef=**DF**, degrees of freedom
- $d = n$ for vardef=**N**, number of observations
- $d = \sum_i w_i - n_p - 1$ for vardef=**WDF**, sum of weights minus number of partial variables minus 1
- $d = \sum_i w_i$ for vardef=**WGT**, sum of weights

By default, SAS/INSIGHT software uses **DF, degrees of freedom**.

The correlation matrices $\mathbf{R}_{yy}$, $\mathbf{R}_{yx}$, and $\mathbf{R}_{xx}$, containing the Pearson product-moment correlations of the variables, are derived by scaling their corresponding covariance matrices:

- $\mathbf{R}_{yy} = \mathbf{D}_{yy}^{-1}\mathbf{S}_{yy}\mathbf{D}_{yy}^{-1}$
- $\mathbf{R}_{yx} = \mathbf{D}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{D}_{xx}^{-1}$
- $\mathbf{R}_{xx} = \mathbf{D}_{xx}^{-1}\mathbf{S}_{xx}\mathbf{D}_{xx}^{-1}$

where $\mathbf{D}_{yy}$ and $\mathbf{D}_{xx}$ are diagonal matrices whose diagonal elements are the square roots of the diagonal elements of $\mathbf{S}_{yy}$ and $\mathbf{S}_{xx}$:

- $\mathbf{D}_{yy} = (\mathrm{diag}(\mathbf{S}_{yy}))^{1/2}$
- $\mathbf{D}_{xx} = (\mathrm{diag}(\mathbf{S}_{xx}))^{1/2}$

Principal Component Analysis

Principal component analysis was originated by Pearson (1901) and later developed by Hotelling (1933). It is a multivariate technique for examining relationships among several quantitative variables. Principal component analysis can be used to summarize data and detect linear relationships. It can also be used for exploring polynomial relationships and for multivariate outlier detection (Gnanadesikan 1997).

Principal component analysis reduces the dimensionality of a set of data while trying to preserve the structure. Given a data set with n_y **Y** variables, n_y eigenvalues and their associated eigenvectors can be computed from its covariance or correlation matrix. The eigenvectors are standardized to unit length.

The principal components are linear combinations of the **Y** variables. The coefficients of the linear combinations are the eigenvectors of the covariance or correlation matrix. Principal components are formed as follows:

- The first principal component is the linear combination of the **Y** variables that accounts for the greatest possible variance.
- Each subsequent principal component is the linear combination of the **Y** variables that has the greatest possible variance and is uncorrelated with the previously defined components.

For a covariance or correlation matrix, the sum of its eigenvalues equals the *trace* of the matrix, that is, the sum of the variances of the n_y variables for a covariance matrix, and n_y for a correlation matrix. The principal components are sorted by descending order of their variances, which are equal to the associated eigenvalues.

Principal components can be used to reduce the number of variables in statistical analyses. Different methods for selecting the number of principal components to retain have been suggested. One simple criterion is to retain components with associated eigenvalues greater than the average eigenvalue (Kaiser 1958). SAS/INSIGHT software offers this criterion as an option for selecting the numbers of eigenvalues, eigenvectors, and principal components in the analysis.

Principal components have a variety of useful properties (Rao 1964; Kshirsagar 1972):

- The eigenvectors are orthogonal, so the principal components represent jointly perpendicular directions through the space of the original variables.
- The principal component scores are jointly uncorrelated. Note that this property is quite distinct from the previous one.
- The first principal component has the largest variance of any unit-length linear combination of the observed variables. The jth principal component has the largest variance of any unit-length linear combination orthogonal to the first $j - 1$ principal components. The last principal component has the smallest variance of any linear combination of the original variables.
- The scores on the first j principal components have the highest possible generalized variance of any set of unit-length linear combinations of the original variables.
- In geometric terms, the j-dimensional linear subspace spanned by the first j principal components gives the best possible fit to the data points as measured by the sum of squared perpendicular distances from each data point to the subspace. This is in contrast to the geometric interpretation of least squares regression, which minimizes the sum of squared vertical distances. For example, suppose you have two variables. Then, the first principal component minimizes the sum of squared perpendicular distances from the points to the first principal axis. This is in contrast to least squares, which would minimize the sum of squared vertical distances from the points to the fitted line.

SAS/INSIGHT software computes principal components from either the correlation or the covariance matrix. The covariance matrix can be used when the variables are measured on comparable scales. Otherwise, the correlation matrix should be used. The new variables with principal component scores have variances equal to corresponding eigenvalues (**Variance=Eigenvalues**) or one (**Variance=1**). You specify the computation method and type of output components in the method options dialog, as shown in Figure 40.3. By default, SAS/INSIGHT software uses the correlation matrix with new variable variances equal to corresponding eigenvalues.

Principal Component Rotation

Orthogonal transformations can be used on principal components to obtain factors that are more easily interpretable. The principal components are uncorrelated with each other, the rotated principal components are also uncorrelated after an orthogonal transformation. Different orthogonal transformations can be derived from maximizing the following quantity with respect to γ:

$$\sum_{j=1}^{n_f}\left(\sum_{i=1}^{n_y} b_{ij}^4 - \frac{\gamma}{n_y}\left(\sum_{i=1}^{n_y} b_{ij}^2\right)^2\right)$$

where n_f is the specified number of principal components to be rotated (number of factors), $b_{ij}^2 = r_{ij}^2 / \sum_{k=1}^{n_f} r_{ik}^2$, and r_{ij} is the correlation between the ith **Y** variable and the jth principal component.

SAS/INSIGHT software uses the following orthogonal transformations:

Equamax $\gamma = \frac{n_f}{2}$

Orthomax γ

Parsimax $\gamma = \frac{n_y(n_f-1)}{(n_y+n_f-2)}$

Quartimax $\gamma = 0$

Varimax $\gamma = 1$

To view or change the principal components rotation options, click on the **Rotation Options** button in the method options dialog shown in Figure 40.3 to display the Rotation Options dialog.

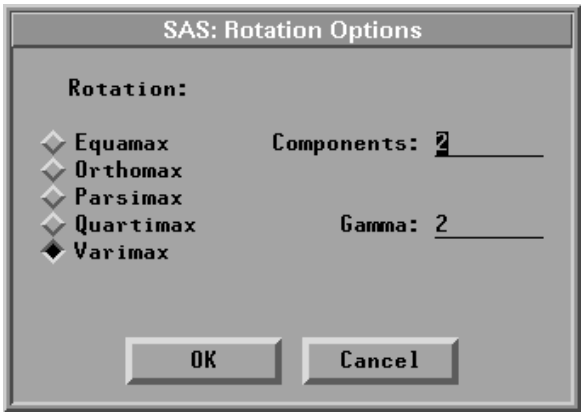


Figure 40.4. Rotation Options Dialog

You can specify the type of rotation and number of principal components to be rotated in the dialog. By default, SAS/INSIGHT software uses **Varimax** rotation on the first two components. If you specify **Orthomax**, you also need to enter the γ value for the rotation in the **Gamma:** field.

Canonical Correlation Analysis

Canonical correlation was developed by Hotelling (1935, 1936). Its application is discussed by Cooley and Lohnes (1971), Kshirsagar (1972), and Mardia, Kent, and Bibby (1979). It is a technique for analyzing the relationship between two sets of variables. Each set can contain several variables. Multiple and simple correlation are special cases of canonical correlation in which one or both sets contain a single variable, respectively.

Given two sets of variables, canonical correlation analysis finds a linear combination from each set, called a canonical variable, such that the correlation between the two canonical variables is maximized. This correlation between the two canonical variables is the first canonical correlation. The coefficients of the linear combinations are canonical coefficients or canonical weights. It is customary to normalize the canonical coefficients so that each canonical variable has a variance of 1.

The first canonical correlation is at least as large as the multiple correlation between any variable and the opposite set of variables. It is possible for the first canonical correlation to be very large while all the multiple correlations for predicting one of the original variables from the opposite set of canonical variables are small.

Canonical correlation analysis continues by finding a second set of canonical variables, uncorrelated with the first pair, that produces the second highest correlation coefficient. The process of constructing canonical variables continues until the number of pairs of canonical variables equals the number of variables in the smaller group.

Each canonical variable is uncorrelated with all the other canonical variables of either set except for the one corresponding canonical variable in the opposite set. The canonical coefficients are not generally orthogonal, however, so the canonical variables do not represent jointly perpendicular directions through the space of the original variables.

The canonical correlation analysis includes tests of a series of hypotheses that each canonical correlation and all smaller canonical correlations are zero in the population. SAS/INSIGHT software uses an F approximation (Rao 1973; Kshirsagar 1972) that gives better small sample results than the usual χ^2 approximation. At least one of the two sets of variables should have an approximately multivariate normal distribution in order for the probability levels to be valid.

Canonical redundancy analysis (Stewart and Love 1968; Cooley and Lohnes 1971; van den Wollenberg 1977) examines how well the original variables can be predicted from the canonical variables. The analysis includes the proportion and cumulative proportion of the variance of the set of **Y** and the set of **X** variables explained by their own canonical variables and explained by the opposite canonical variables. Either raw or standardized variance can be used in the analysis.

Maximum Redundancy Analysis

In contrast to canonical redundancy analysis, which examines how well the original variables can be predicted from the canonical variables, maximum redundancy analysis finds the variables that can best predict the original variables.

Given two sets of variables, maximum redundancy analysis finds a linear combination from one set of variables that best predicts the variables in the opposite set. SAS/INSIGHT software normalizes the coefficients of the linear combinations so that each maximum redundancy variable has a variance of 1.

Maximum redundancy analysis continues by finding a second maximum redundancy variable from one set of variables, uncorrelated with the first one, that produces the second highest prediction power for the variables in the opposite set. The process of constructing maximum redundancy variables continues until the number of maximum redundancy variables equals the number of variables in the smaller group.

Either raw variances (**Raw Variance**) or standardized variances (**Std Variance**) can be used in the analysis. You specify the selection in the method options dialog as shown in Figure 40.3. By default, standardized variances are used.

Canonical Discriminant Analysis

Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. Given a classification variable and several interval variables, canonical discriminant analysis derives *canonical variables* (linear combinations of the interval variables) that summarize between-class variation in much the same way that principal components summarize total variation.

Given two or more groups of observations with measurements on several interval variables, canonical discriminant analysis derives a linear combination of the variables that has the highest possible multiple correlation with the groups. This maximal multiple correlation is called the first canonical correlation. The coefficients of the linear combination are the canonical coefficients or canonical weights. The variable defined by the linear combination is the first canonical variable or canonical component. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the groups. The process of extracting canonical variables can be repeated until the number of canonical variables equals the number of original variables or the number of classes minus one, whichever is smaller.

The first canonical correlation is at least as large as the multiple correlation between the groups and any of the original variables. If the original variables have high within-group correlations, the first canonical correlation can be large even if all the multiple correlations are small. In other words, the first canonical variable can show substantial differences among the classes, even if none of the original variables does.

For each canonical correlation, canonical discriminant analysis tests the hypothesis that it and all smaller canonical correlations are zero in the population. An F approximation is used that gives better small-sample results than the usual χ^2 approximation. The variables should have an approximate multivariate normal distribution within each class, with a common covariance matrix in order for the probability levels to be valid.

The new variables with canonical variable scores in canonical discriminant analysis have either pooled within-class variances equal to one (**Std Pooled Variance**) or total-sample variances equal to one (**Std Total Variance**). You specify the selection in the method options dialog as shown in Figure 40.3. By default, canonical variable scores have pooled within-class variances equal to one.

Output

To view or change the output options associated with your multivariate analysis, click on the **Output** button from the variables dialog. This displays the output options dialog.

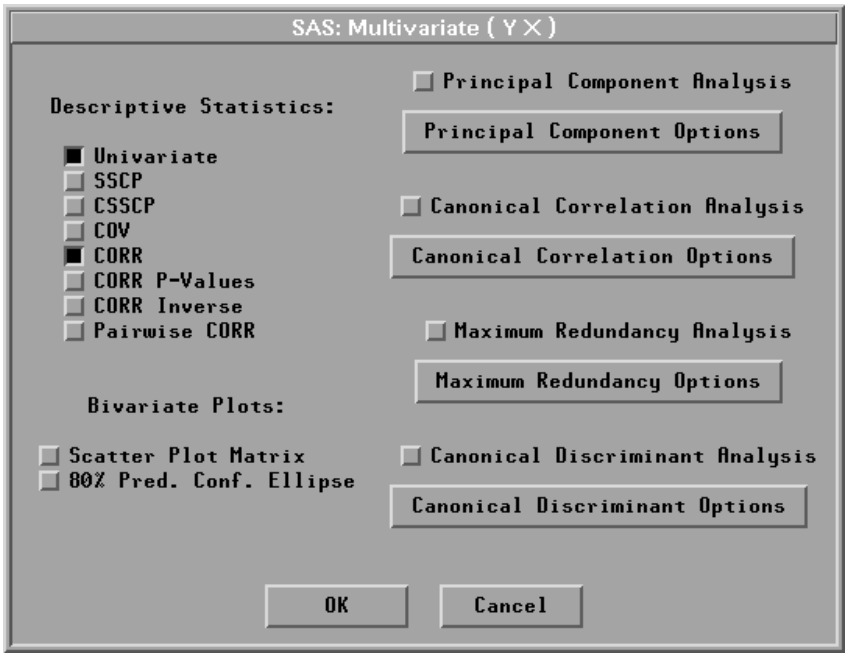


Figure 40.5. Multivariate Output Options Dialog

The options you set in this dialog determine which tables and graphs appear in the multivariate window. SAS/INSIGHT software provides univariate statistics and correlation matrix tables by default.

Descriptive statistics provide tables for examining the relationships among quantitative variables from univariate, bivariate, and multivariate perspectives.

Plots can be more informative than tables when you are trying to understand multivariate data. You can display a matrix of scatter plots for the analyzing variables. You can also add a bivariate confidence ellipse for mean or predicted values to the scatter plots. Using the confidence ellipses assumes each pair of variables has a bivariate normal distribution.

With appropriate variables chosen, you can generate principal component analysis (interval Y variables), canonical correlation analysis (interval Y, X variables), maximum redundancy analysis (interval Y, X variables), and canonical discriminant analysis (one nominal Y variable, interval X variables) by selecting the corresponding checkbox in the Output Options dialog.

Principal Component Analysis

Clicking the **Principal Component Options** button in the Output Options dialog shown in Figure 40.5 displays the dialog shown in Figure 40.6.

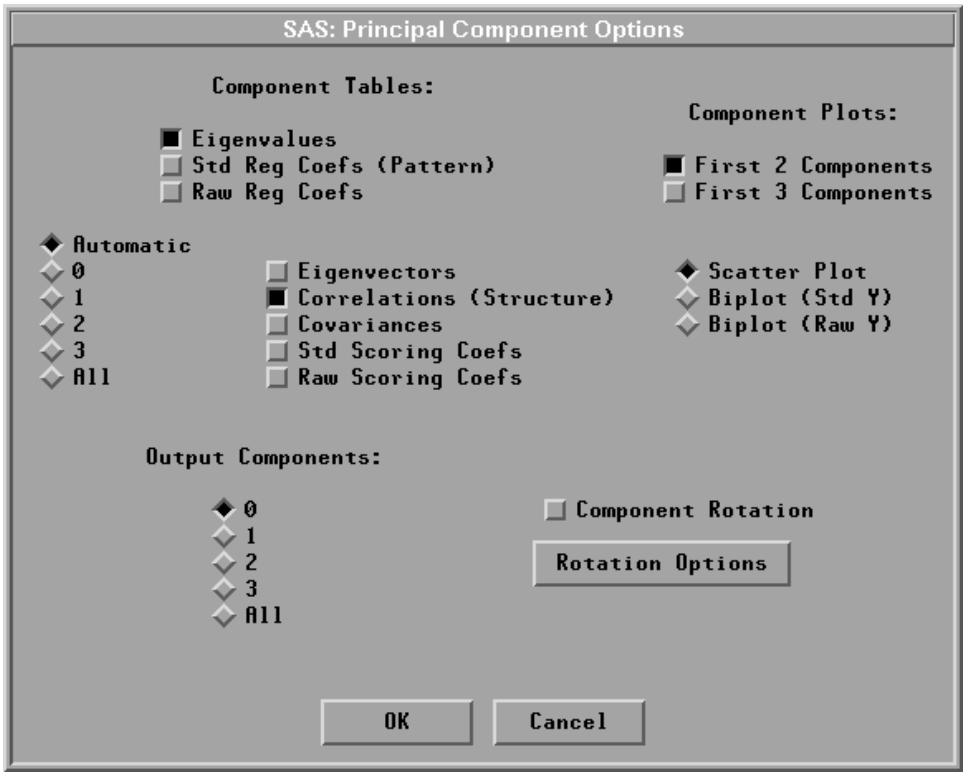


Figure 40.6. Principal Components Options Dialog

The dialog enables you to view or change the output options associated with principal component analyses and save principal component scores in the data window.

In the dialog, you need to specify the number of components when selecting tables of **Eigenvectors**, **Correlations (Structure)**, **Covariances**, **Std Scoring Coefs**, and **Raw Scoring Coefs**. **Automatic** uses principal components with corresponding eigenvalues greater than the average eigenvalue. By default, SAS/INSIGHT software displays a plot of the first two principal components, a table of all the eigenvalues, and a table of correlations between the **Y** variables and principal components with corresponding eigenvalues greater than the average eigenvalue.

You can generate principal component rotation analysis by selecting the **Component Rotation** checkbox in the dialog.

Principal Component Rotation

Clicking the **Rotation Options** button in the **Principal Components Options** dialog shown in Figure 40.6 displays the **Rotation Options** dialog shown in Figure 40.7.

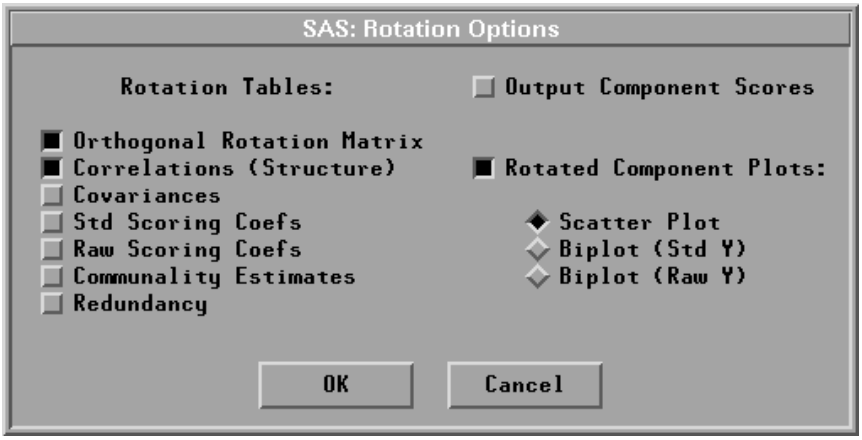


Figure 40.7. Principal Components Rotation Options Dialog

The number of components rotated is specified in the **Principal Components Rotation Options** dialog shown in Figure 40.4. By default, SAS/INSIGHT software displays a plot of the rotated components (when the specified number is two or three), a rotation matrix table, and a table of correlations between the **Y** variables and rotated principal components.

Canonical Correlation Analysis

Clicking the **Canonical Correlation Options** button in the Output Options dialog shown in Figure 40.5 displays the dialog shown in Figure 40.8.

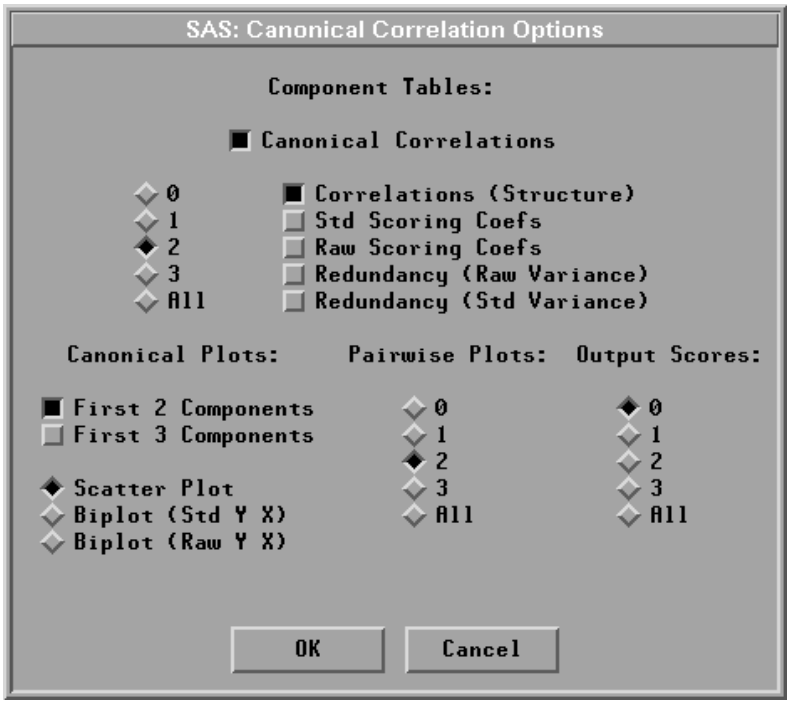


Figure 40.8. Canonical Correlation Options Dialog

This dialog enables you to view or change the options associated with canonical correlation analyses and save maximum redundancy variable scores in the data window. You specify the number of components when selecting tables of **Correlations (Structure)**, **Std Scoring Coefs**, **Raw Scoring Coefs**, **Redundancy (Raw Variance)**, and **Redundancy (Std Variance)**.

By default, SAS/INSIGHT software displays a plot of the first two canonical variables, plots of the first two pairs of canonical variables, a canonical correlations table, and a table of correlations between the **Y**, **X** variables and the first two canonical variables from both **Y** variables and **X** variables.

Maximum Redundancy Analysis

Clicking the **Maximum Redundancy Options** button in the Output Options dialog shown in Figure 40.5 displays the dialog shown in Figure 40.9.

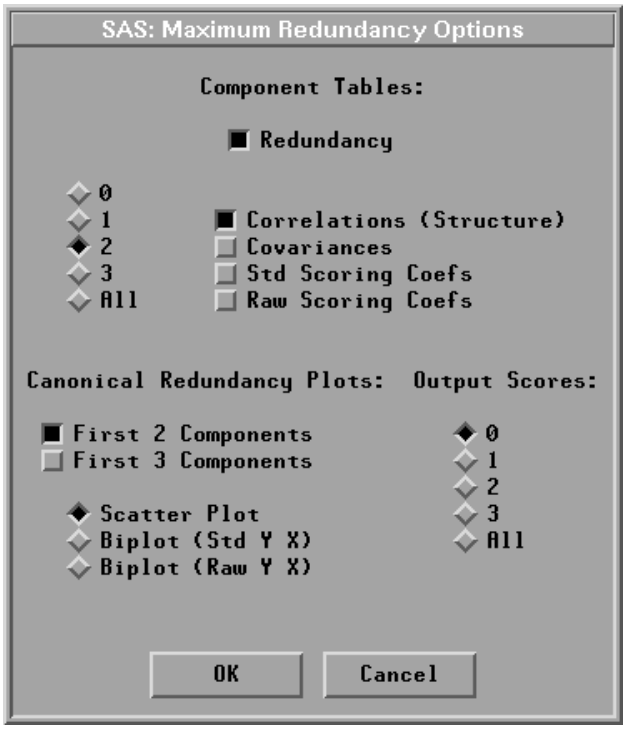


Figure 40.9. Maximum Redundancy Options Dialog

This dialog enables you to view or change the options associated with canonical correlation analyses and save maximum redundancy variable scores in the data window. You specify the number of components when selecting tables of **Correlations (Structure)**, **Covariances**, **Std Scoring Coefs**, and **Raw Scoring Coefs**.

By default, SAS/INSIGHT software displays a plot of the first two canonical redundancy variables, a canonical redundancy table, and a table of correlations between the **Y**, **X** variables and the first two canonical redundancy variables from both **Y** variables and **X** variables.

Canonical Discriminant Analysis

Clicking the **Canonical Discriminant Options** button in the Output Options dialog shown in Figure 40.5 displays the dialog shown in Figure 40.10.

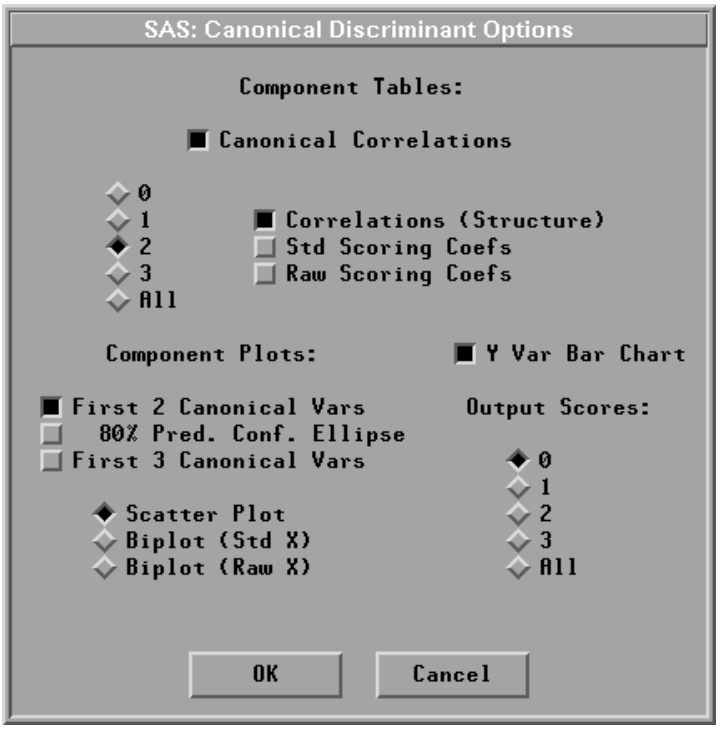


Figure 40.10. Canonical Discriminant Options Dialog

You specify the number of components when selecting tables of **Correlations (Structure)**, **Std Scoring Coefs**, and **Raw Scoring Coefs**.

By default, SAS/INSIGHT software displays a plot of the first two canonical variables, a bar chart for the nominal **Y** variable, a canonical correlation table, and a table of correlations between the **X** variables and the first two canonical variables.

Tables

You can generate tables of descriptive statistics and output from multivariate analyses by setting options in output options dialogs, as shown in Figure 40.5 to Figure 40.10, or by choosing from the **Tables** menu shown in Figure 40.11.

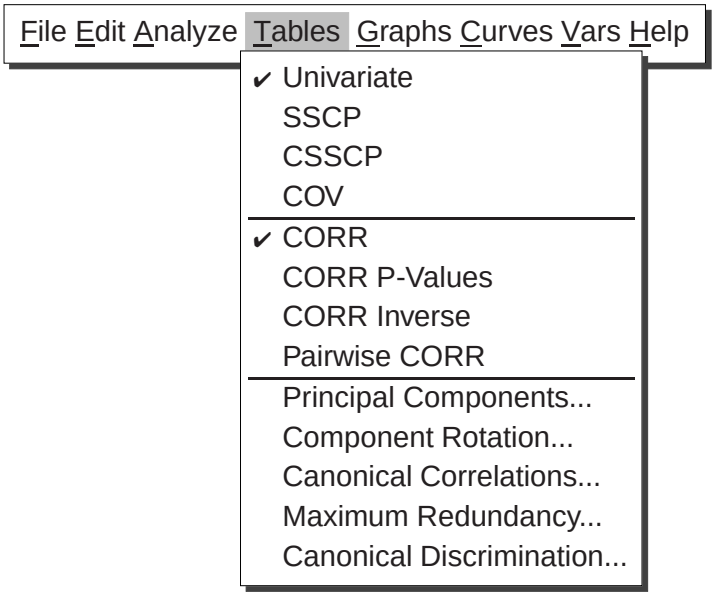


Figure 40.11. Tables Menu

Univariate Statistics

The **Univariate Statistics** table, as shown in Figure 40.12 contains the following information:

- **Variable** is the variable name.
- **N** is the number of nonmissing observations, n.
- **Mean** is the variable mean, $\overline{\mathbf{y}}$ or $\overline{\mathbf{x}}$.
- **Std Dev** is the standard deviation of the variable, the square root of the corresponding diagonal element of $\mathbf{S}_{yy}$ or $\mathbf{S}_{xx}$.
- **Minimum** is the minimum value.
- **Maximum** is the maximum value.
- **Partial Std Dev** (with selected **Partial** variables) is the partial standard deviation of the variable after partialling out the **Partial** variables.

Sums of Squares and Crossproducts

The **Sums of Squares and Crossproducts** (SSCP) table, as illustrated by Figure 40.12, contains the sums of squares and crossproducts of the variables.

Corrected Sums of Squares and Crossproducts

The **Corrected Sums of Squares and Crossproducts** (CSSCP) table, as shown in Figure 40.12, contains the sums of squares and crossproducts of the variables corrected for the mean.

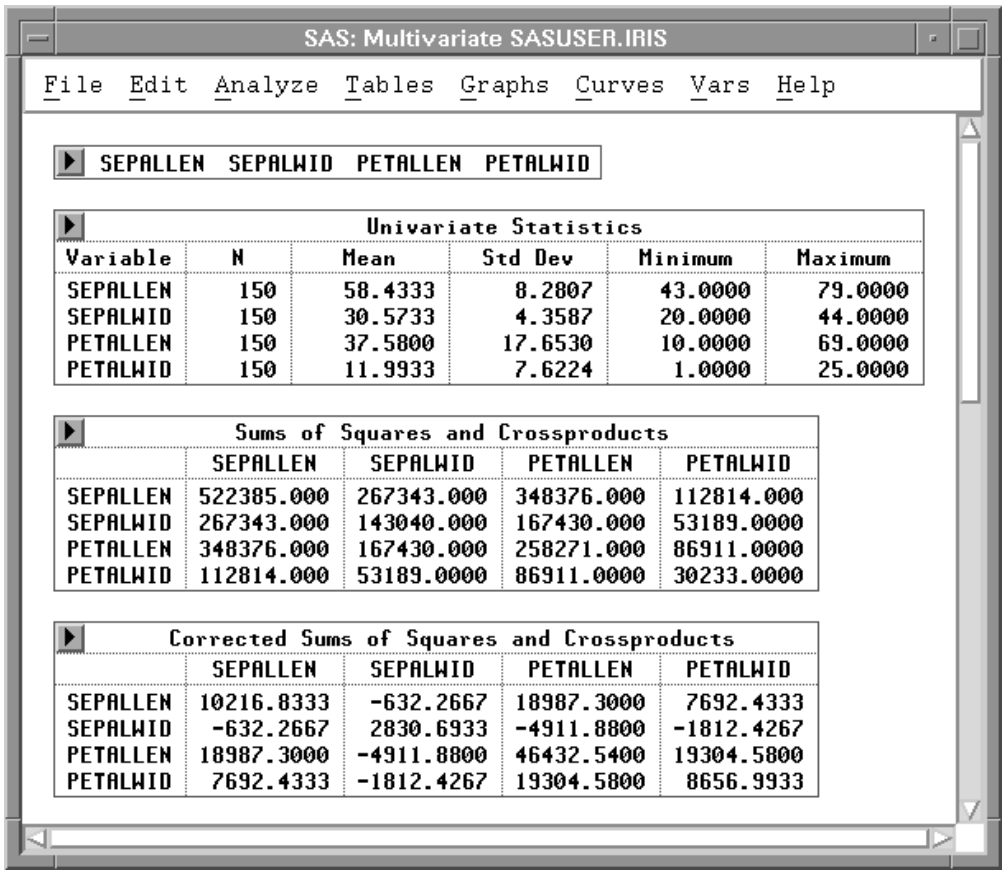


Figure 40.12. Univariate Statistics, SSCP, and CSSCP Tables

Covariance Matrix

The **Covariance Matrix** (COV) table, as shown in Figure 40.13, contains the estimated variances and covariances of the variables, with their associated degrees of freedom. The variance measures the spread of the distribution around the mean, and the covariance measures the tendency of two variables to linearly increase or decrease together

Correlation Matrix

The **Correlation Matrix** (CORR) table contains the Pearson product-moment correlations of the **Y** variables, as shown in Figure 40.13. Correlation measures the strength of the linear relationship between two variables. A correlation of 0 means that there is no linear association between two variables. A correlation of 1 (-1) means that there is an exact positive (negative) linear association between the two variables.

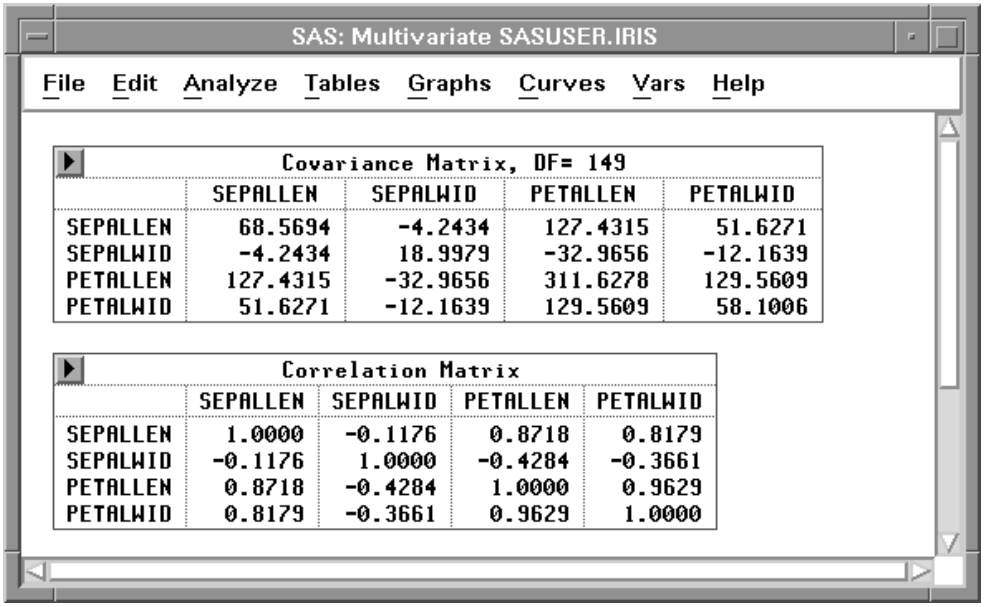


Figure 40.13. COV and CORR Tables

P-Values of the Correlations

The **P-Values of the Correlations** table contains the p-value of each correlation under the null hypothesis that the correlation is 0, assuming independent and identically distributed (unless weights are specified) observations from a bivariate distribution with at least one variable normally distributed. This table is shown in Figure 40.14. Each p-value in this table can be used to assess the significance of the corresponding correlation coefficient.

The p-value of a correlation *r* is obtained by treating the statistic

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

as having a Student's t distribution with $n-2$ degrees of freedom. The p-value of the correlation *r* is the probability of obtaining a Student's t statistic greater in absolute value than the absolute value of the observed statistic t.

With partial variables, the p-value of a correlation is obtained by treating the statistic

$$t = \sqrt{n-n_p-2}\frac{r}{\sqrt{1-r^2}}$$

as having a Student's t distribution with $n-n_p-2$ degrees of freedom.

Inverse Correlation Matrix

For a symmetric correlation matrix, the **Inverse Correlation Matrix** table contains the inverse of the correlation matrix, as shown in Figure 40.14.

The diagonal elements of the inverse correlation matrix, sometimes referred to as *variance inflation factors*, measure the extent to which the variables are linear combinations of other variables. The jth diagonal element of the inverse correlation matrix is $1/(1 - \mathrm{R}_j^2)$, where R_j^2 is the squared multiple correlation of the jth variable with the other variables. Large diagonal elements indicate that variables are highly correlated.

When a correlation matrix is singular (less than full rank), some variables are linear functions of other variables, and a g2 inverse for the matrix is displayed. The g2 inverse depends on the order in which you select the variables. A value of 0 in the jth diagonal indicates that the jth variable is a linear function of the previous variables.

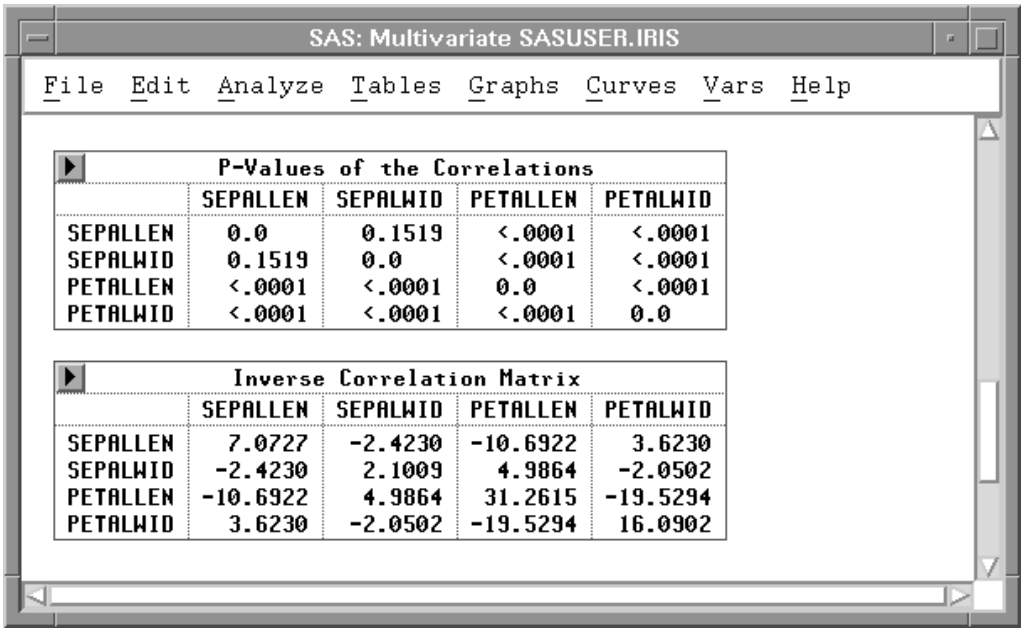
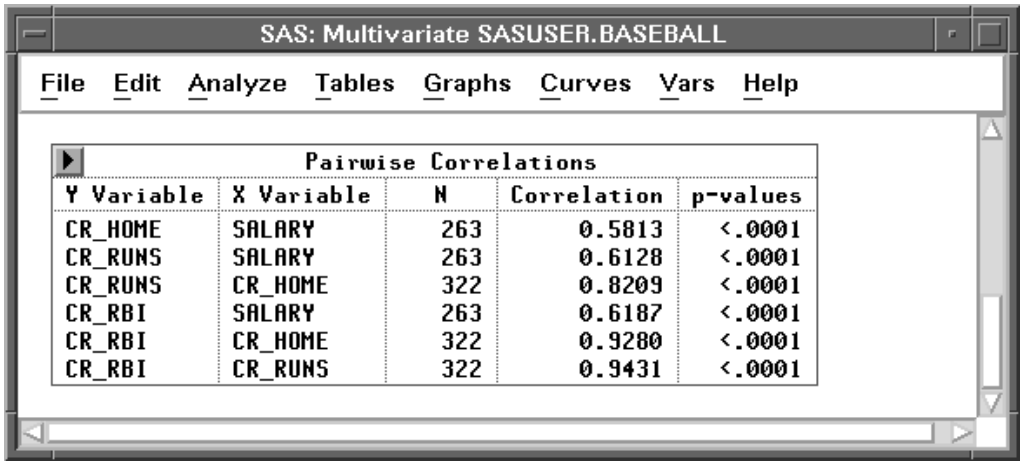


Figure 40.14. P-values of Correlations and Inverse Correlation Matrix

Pairwise Correlations

SAS/INSIGHT software drops an observation with a missing value for any variable used in the analysis from all calculations. The **Pairwise CORR** table gives correlations that are computed from all observations that have nonmissing values for any pair of variables. Figure 40.15 shows a table of pairwise correlations.



Pairwise Correlations

Y Variable	X Variable	N	Correlation	p-values
CR_HOME	SALARY	263	0.5813	<.0001
CR_RUNS	SALARY	263	0.6128	<.0001
CR_RUNS	CR_HOME	322	0.8209	<.0001
CR_RBI	SALARY	263	0.6187	<.0001
CR_RBI	CR_HOME	322	0.9280	<.0001
CR_RBI	CR_RUNS	322	0.9431	<.0001

Figure 40.15. Pairwise CORR Table

Principal Component Analysis

You can generate tables of output from principal component analyses by setting options in the principal component options dialog shown in Figure 40.6 or from the **Tables** menu shown in Figure 40.11. Select **Principal Components** from the **Tables** menu to display the principal component tables dialog shown in Figure 40.16.

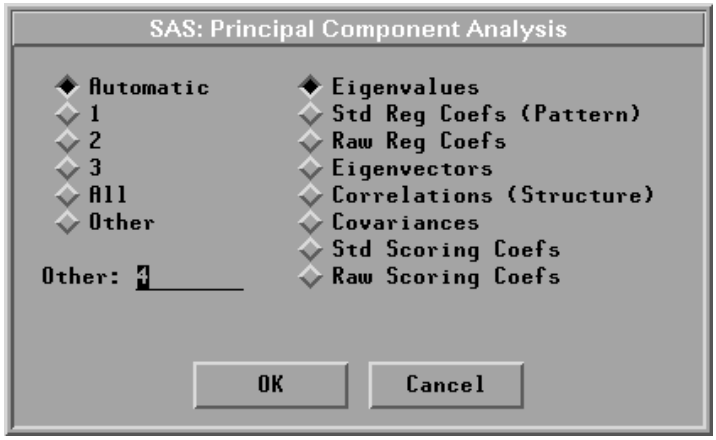


Figure 40.16. Principal Component Tables Dialog

Choose **Automatic** to display principal components with eigenvalues greater than the average eigenvalue. Selecting **1**, **2**, or **3** gives you 1, 2, or 3 principal components. **All** gives you all eigenvalues. Selecting **0** in the principal component options dialog suppresses the principal component tables.

The **Eigenvalues (COV)** or **Eigenvalues (CORR)** table includes the eigenvalues of the covariance or correlation matrix, the difference between successive eigenvalues, the proportion of variance explained by each eigenvalue, and the cumulative proportion of variance explained.

Eigenvalues correspond to each of the principal components and represent a partitioning of the total variation in the sample. The sum of all eigenvalues is equal to the sum of all variable variances if the covariance matrix is used or to the number of variables, p, if the correlation matrix is used.

The **Eigenvectors (COV)** or **Eigenvectors (CORR)** table includes the eigenvectors of the covariance or correlation matrix. Eigenvectors correspond to each of the principal components and are used as the coefficients to form linear combinations of the **Y** variables (principal components).

Figure 40.17 shows tables of all eigenvalues and eigenvectors for the first two principal components.

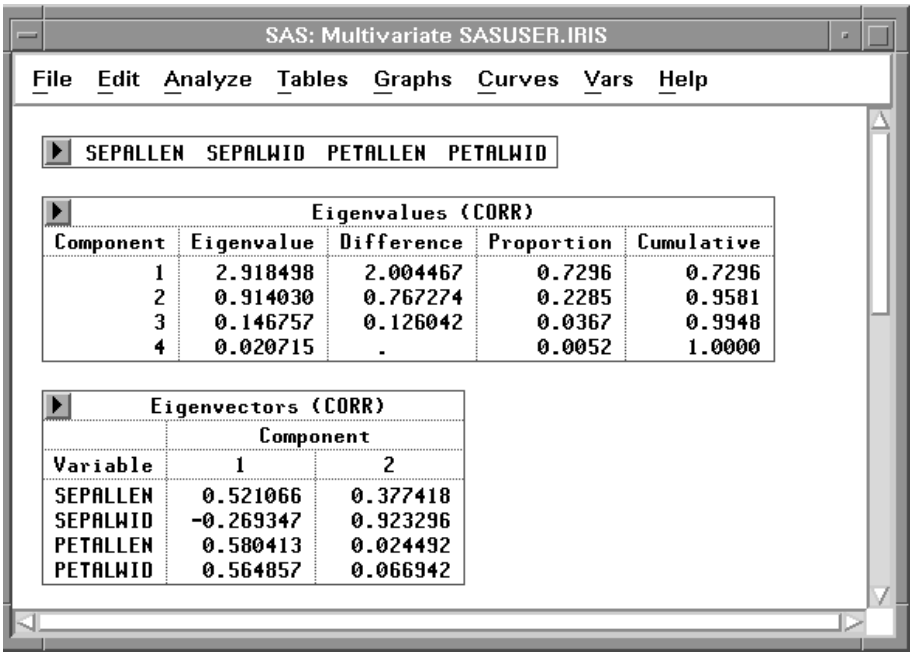


Figure 40.17. Eigenvalues and Eigenvectors Tables

The **Correlations (Structure)** and **Covariances** tables include the correlations and covariances, respectively, between the **Y** variables and principal components. The correlation and covariance matrices measure the strength of the linear relationship between the derived principal components and each of the **Y** variables. Figure 40.18 shows the correlations and covariances between the **Y** variables and the first two principal components.

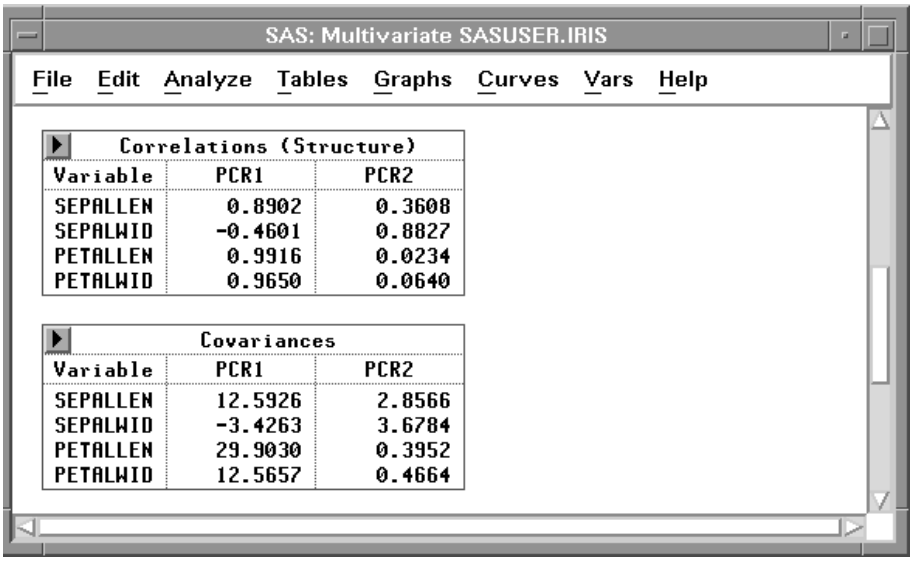


Figure 40.18. Correlations and Covariances Tables

The scoring coefficients are the coefficients of the **Y** variables used to generate principal components. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** variables.

The regression coefficients are the coefficients of principal components used to generate estimated **Y** variables. The **Std Reg Coefs (Pattern)** and **Raw Reg Coefs** tables include the regression coefficients of principal components used to generate estimated standardized and centered **Y** variables. Figure 40.19 shows the regression coefficients of the principal components for the standardized **Y** variables, as well as the scoring coefficients of the standardized **Y** variables for the first two principal components.

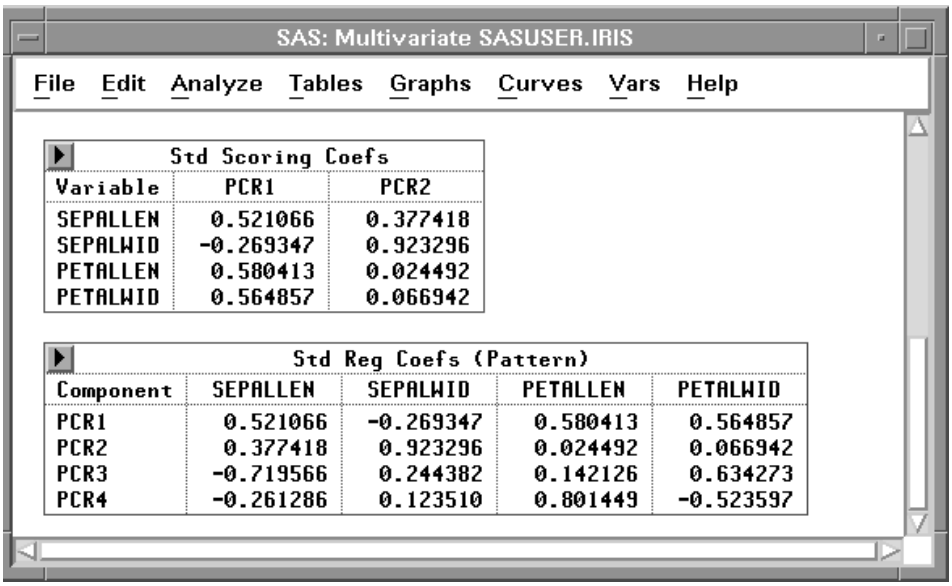


Figure 40.19. Regression Coefficients and Scoring Coefficients Tables

Principal Components Rotation

You can generate tables of output from principal component rotation by setting options in the **Rotation Options** dialog shown in Figure 40.7 or from the **Tables** menu shown in Figure 40.11. Select **Component Rotation** from the **Tables** menu to display the principal component rotation dialog shown in Figure 40.20.

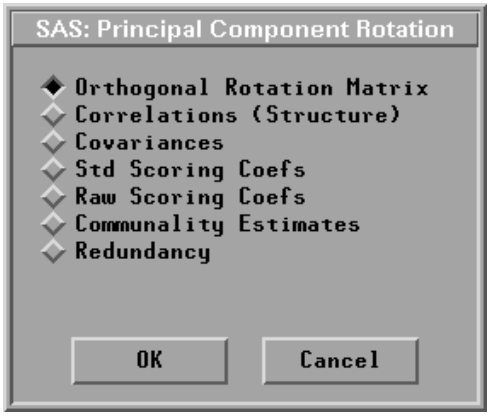


Figure 40.20. Principal Components Rotation Dialog

You specify the number of components and type of rotation in the **Rotation Options** dialog, as shown in Figure 40.4.

The **Orthogonal Rotation Matrix** is the orthogonal rotation matrix used to compute the rotated principal components from the standardized principal components.

The **Correlations (Structure)** and **Covariances** tables include the correlations and covariances between the **Y** variables and the rotated principal components.

Figure 40.21 shows the rotation matrix and correlations and covariances between the **Y** variables and the first two rotated principal components.

The scoring coefficients are the coefficients of the **Y** variables used to generate rotated principal components. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** variables.

The **Communality Estimates** table gives the standardized variance of each **Y** variable explained by the rotated principal components.

The **Redundancy** table gives the variances of the standardized **Y** variables explained by each rotated principal component.

Figure 40.22 shows the scoring coefficients of the standardized **Y** variables, communality estimates for the **Y** variables, and redundancy for each rotated component.

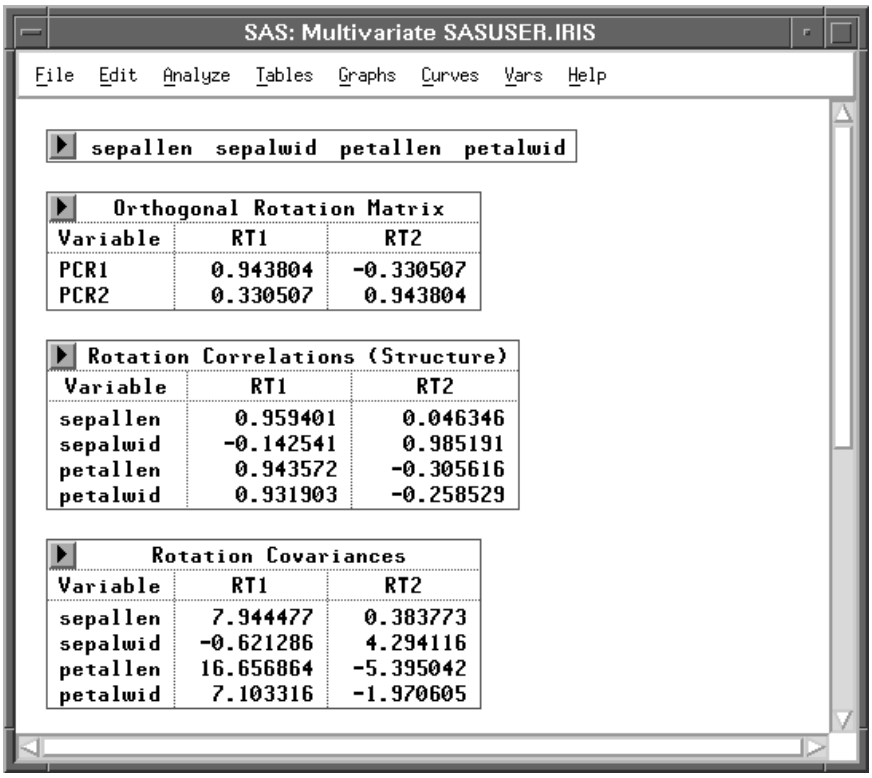


Figure 40.21. Rotation Matrix, Correlation, and Covariance Tables

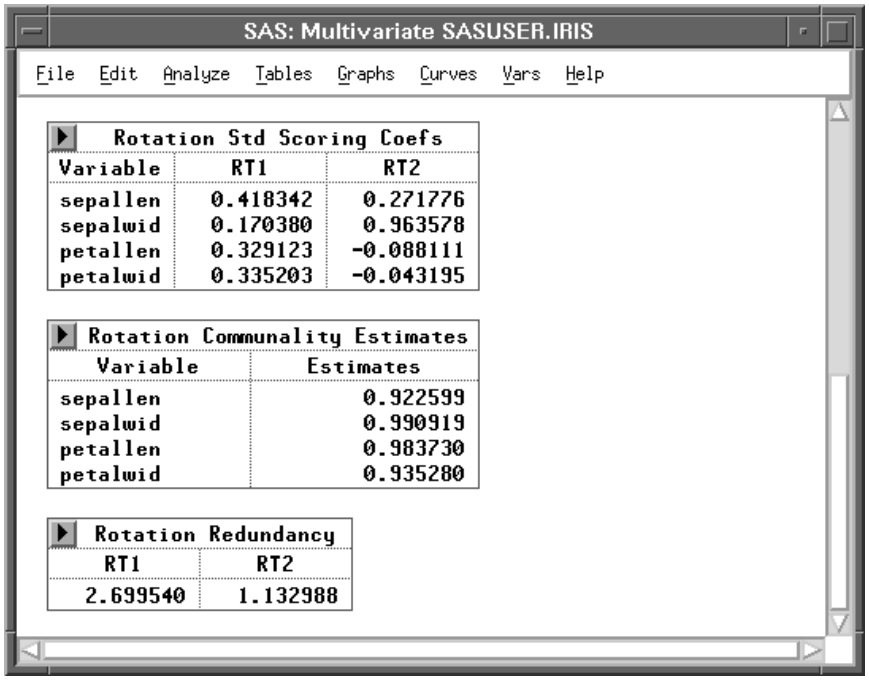


Figure 40.22. Scoring Coefficients, Communality, and Redundancy Tables

Canonical Correlation Analysis

You can generate tables of output from canonical correlation analyses by setting options in the Canonical Correlation Options dialog shown in Figure 40.8 or from the **Tables** menu shown in Figure 40.11. Select **Canonical Correlations** from the **Tables** menu to display the canonical correlation dialog shown in Figure 40.23.

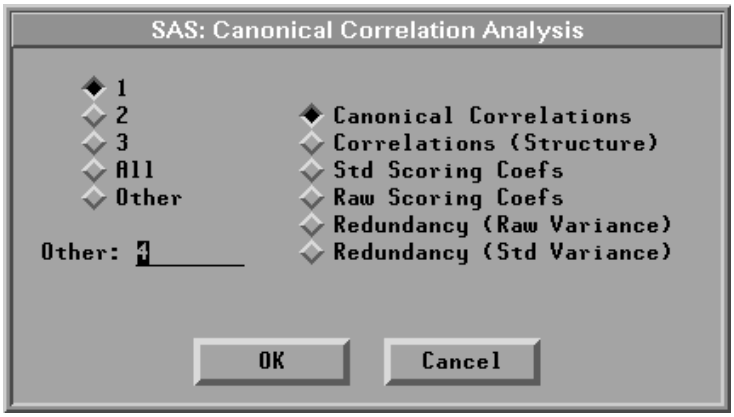


Figure 40.23. Canonical Correlation Dialog

The **Canonical Correlations** table contains the following:

- **CanCorr**, the canonical correlations, which are always nonnegative
- **Adj. CanCorr**, the adjusted canonical correlations, which are asymptotically less biased than the raw correlations and may be negative. The adjusted canonical correlations may not be computable, and they are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.
- **Approx Std. Error**, the approximate standard errors of the canonical correlations
- **CanRsq**, the squared canonical correlations
- **Eigenvalues**, the eigenvalues of the matrix $R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{yx}'$. These eigenvalues are equal to $\mathrm{CanRsq}/(1-\mathrm{CanRsq})$, where CanRsq is the corresponding squared canonical correlation. Also printed for each eigenvalue is the difference from the next eigenvalue, the proportion of the sum of the eigenvalues, and the cumulative proportion.
- **Test for H0: CanCorrj=0, j>=k**, the likelihood ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population
- **Approx F** based on Rao's approximation to the distribution of the likelihood ratio
- **Num DF** and **Den DF** (numerator and denominator degrees of freedom) and **Pr > F** (probability level) associated with the F statistic

Figure 40.24 shows tables of canonical correlations.

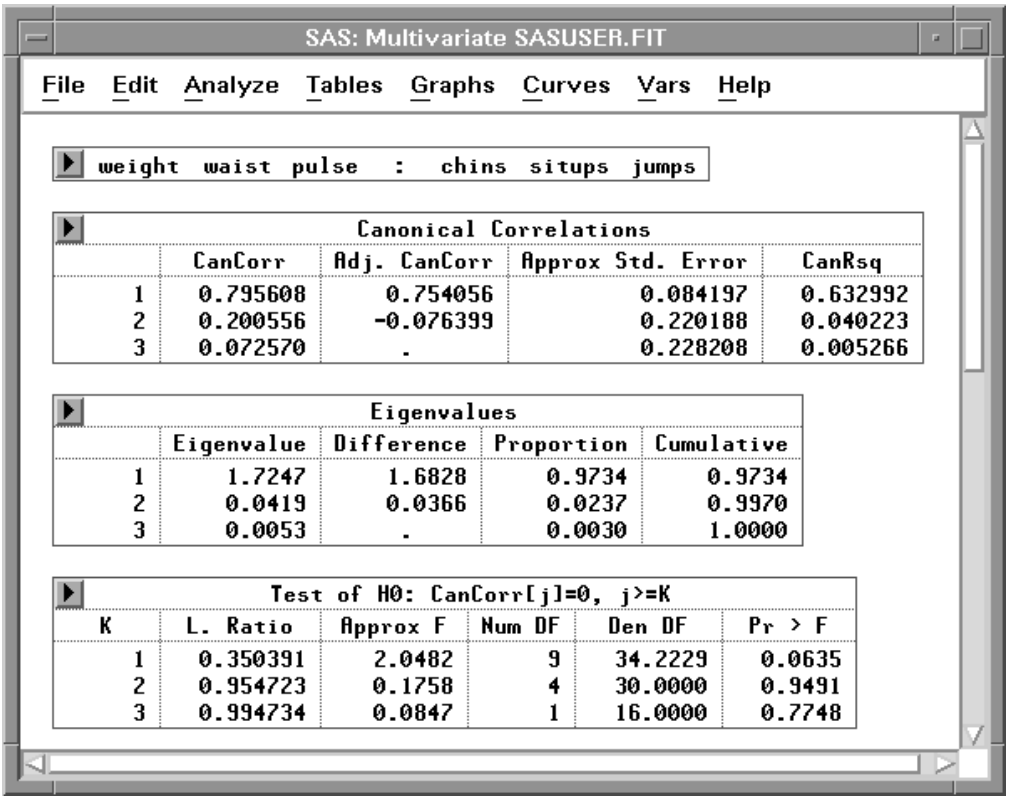


Figure 40.24. Canonical Correlations Tables

The **Correlations (Structure)** table includes the correlations between the input **Y**, **X** variables and canonical variables.

The scoring coefficients are the coefficients of the **Y** or **X** variables that are used to compute canonical variable scores. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** or **X** variables and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** or **X** variables.

Figure 40.25 shows a table of correlations between the **Y**, **X** variables and the first two canonical variables from the **Y** and **X** variables and the tables of scoring coefficients of the standardized **Y** and **X** variables.

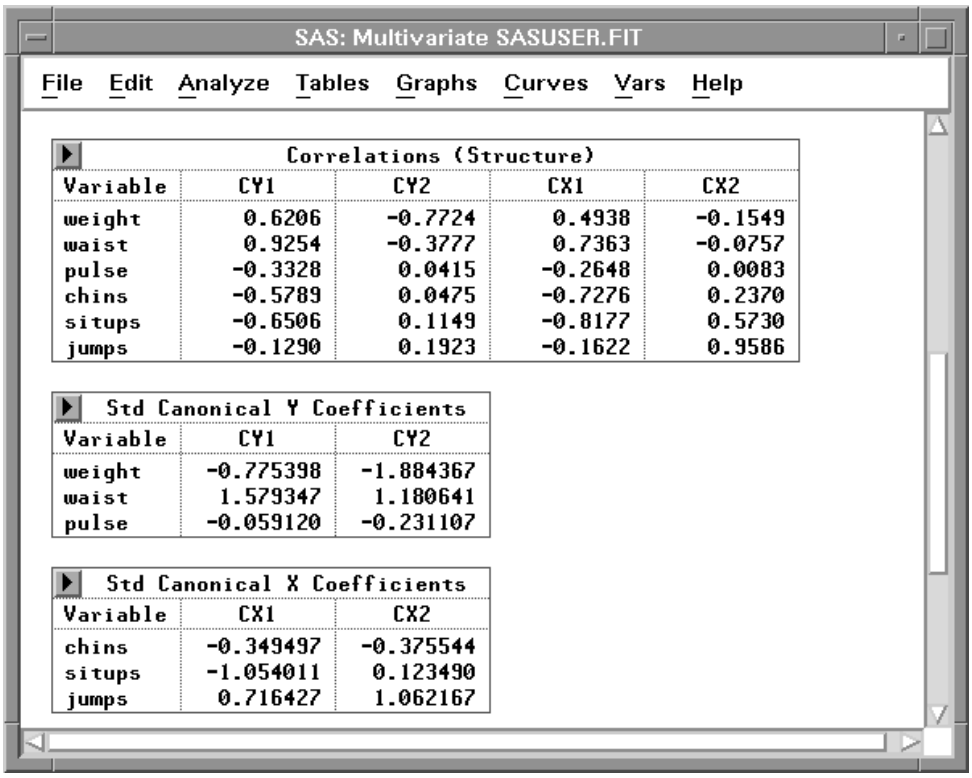


Figure 40.25. Correlations and Scoring Coefficients Tables

The **Redundancy** table gives the canonical redundancy analysis, which includes the proportion and cumulative proportion of the raw (unstandardized) and the standardized variance of the set of **Y** and the set of **X** variables explained by their own canonical variables and explained by the opposite canonical variables. Figure 40.26 shows tables of redundancy of standardized **Y** and **X** variables.

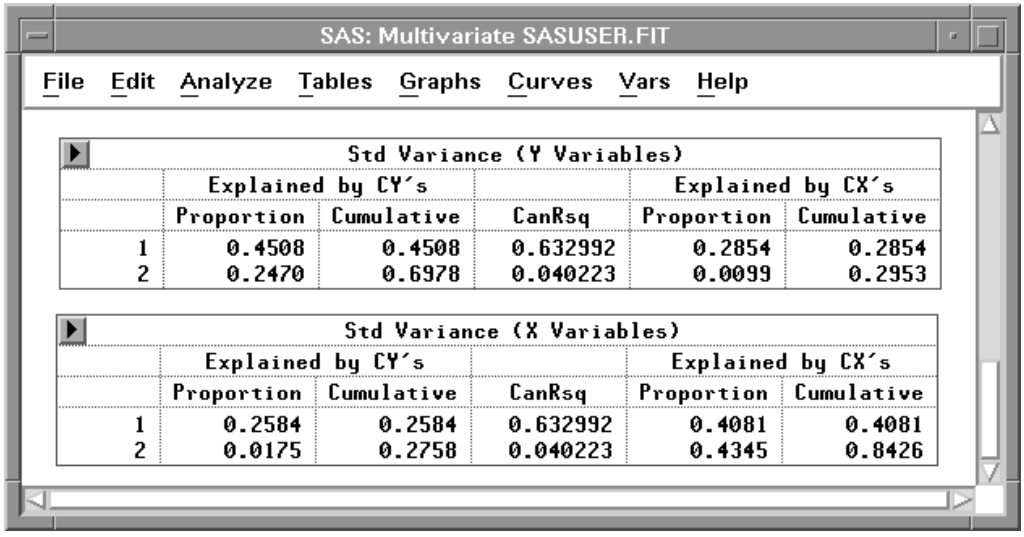


Figure 40.26. Redundancy Tables

Maximum Redundancy Analysis

You can generate tables of output from maximum redundancy analysis by setting options in the Maximum Redundancy Options dialog shown in Figure 40.9 or from the **Tables** menu shown in Figure 40.11. Select **Maximum Redundancy** from the **Tables** menu to display the maximum redundancy dialog shown in Figure 40.27.

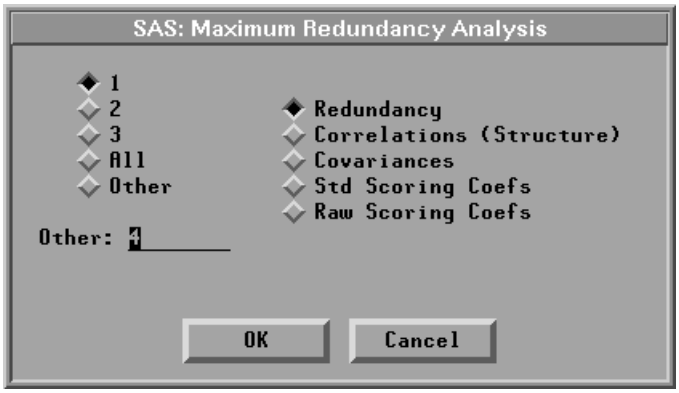


Figure 40.27. Maximum Redundancy Dialog

Either the raw (centered) or standardized variance is used in the maximum redundancy analysis, and it is specified in the Multivariate Method Options dialog in Figure 40.3. The **Redundancy** table includes the proportion and cumulative proportion of the variance of the set of **Y** variables and the set of **X** variables explained by the opposite canonical variables. Figure 40.28 shows tables of redundancy of the standardized **Y** and **X** variables.

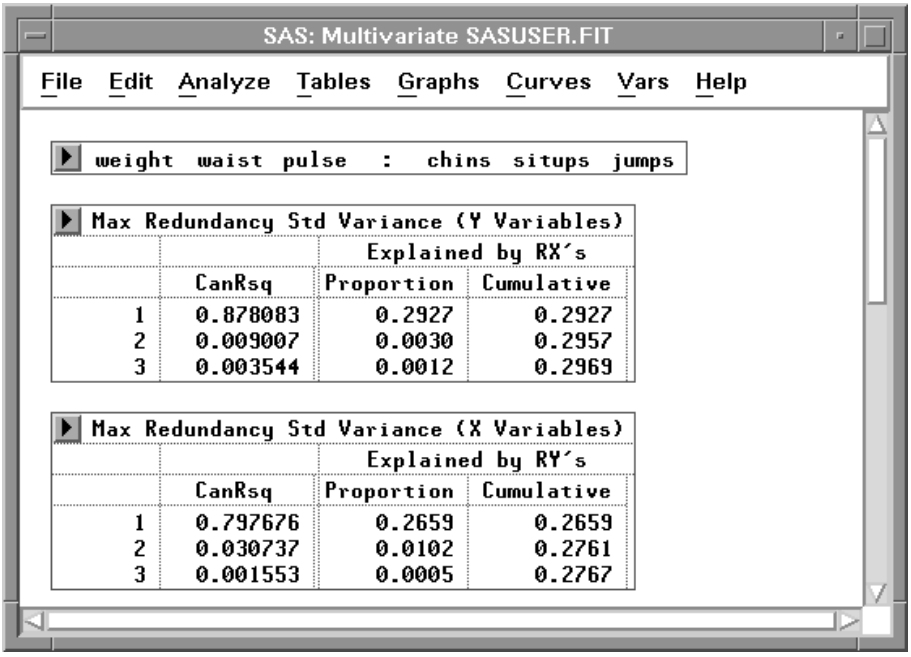


Figure 40.28. Maximum Redundancy Tables

The **Correlations (Structure)** or **Covariances** table includes the correlations or covariances between the **Y**, **X** variables and the maximum redundancy variables. Figure 40.29 shows the correlations and covariances between the **Y**, **X** variables and the first two maximum redundancy variables from the **Y** variables and the **X** variables.

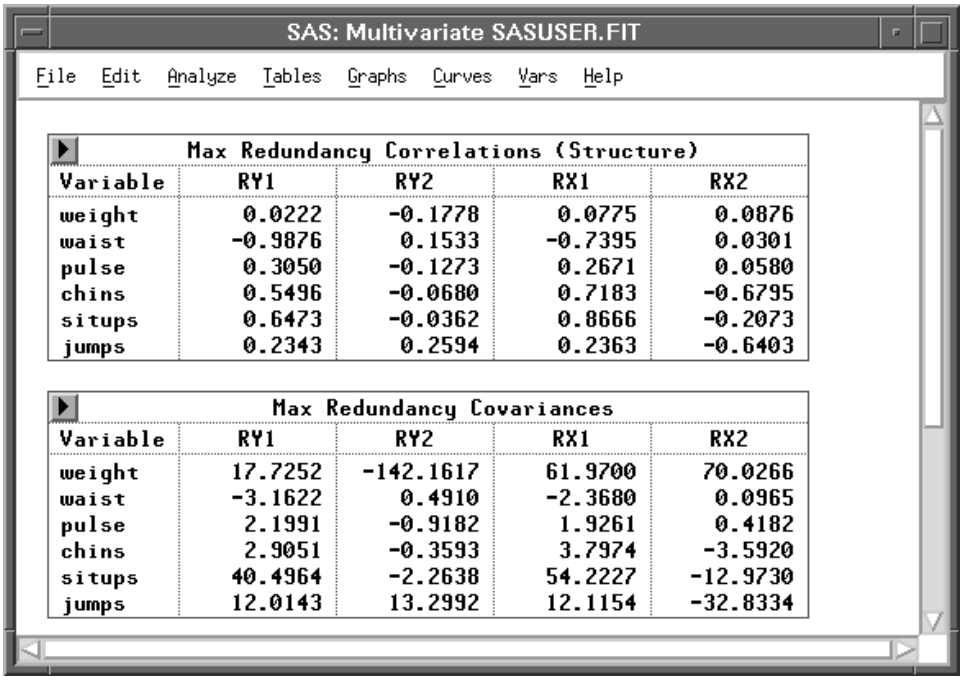


Figure 40.29. Correlation and Covariance Tables

The scoring coefficients are the coefficients of the **Y** or **X** variables that are used to compute maximum redundancy variables. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **Y** or **X** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **Y** or **X** variables. Figure 40.30 shows tables of the scoring coefficients of the standardized **Y** and **X** variables.

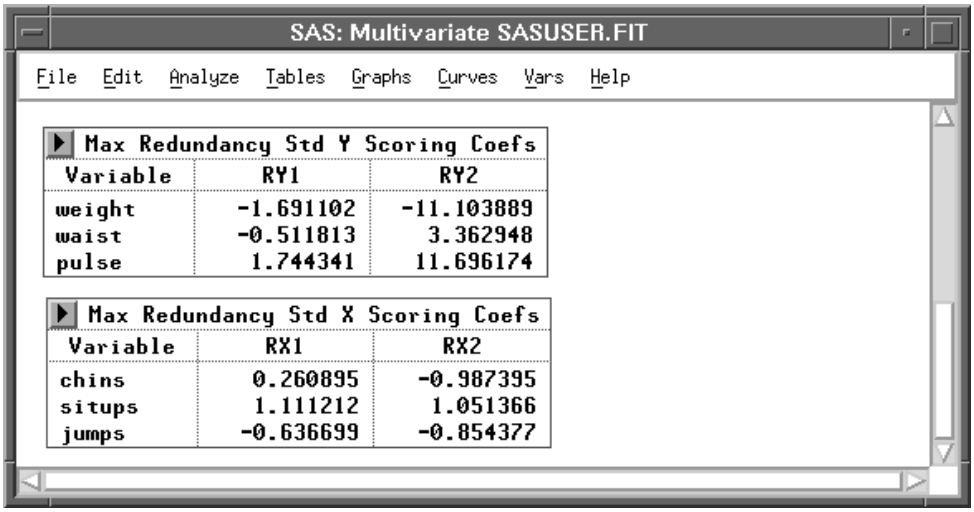


Figure 40.30. Standardized Scoring Coefficients Tables

Canonical Discriminant Analysis

You can generate tables of output from canonical discriminant analyses by setting options in the Canonical Discriminant Options dialog shown in Figure 40.10 or from the **Tables** menu shown in Figure 40.11. Select **Canonical Discrimination** from the **Tables** menu to display the canonical discriminant analysis dialog shown in Figure 40.31.

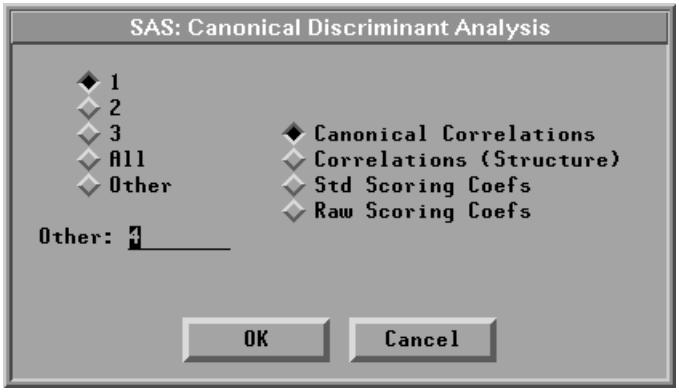


Figure 40.31. Canonical Discriminant Analysis Dialog

The **Canonical Correlations** table, as shown in Figure 40.32, contains the following:

- **CanCorr**, the canonical correlations, which are always nonnegative
- **Adj. CanCorr**, the adjusted canonical correlations, which are asymptotically less biased than the raw correlations and may be negative. The adjusted canonical correlations may not be computable and are displayed as missing values if two canonical correlations are nearly equal or if some are close to zero. A missing value is also displayed if an adjusted canonical correlation is larger than a previous adjusted canonical correlation.
- **Approx Std. Error**, the approximate standard errors of the canonical correlations
- **CanRsq**, the squared canonical correlations
- **Eigenvalues**, eigenvalues of the matrix $E^{-1}H$, where E is the matrix of the within-class sums of squares and crossproducts and H is the matrix of the between-class sums of squares and crossproducts. These eigenvalues are equal to $\mathrm{CanRsq}/(1-\mathrm{CanRsq})$, where CanRsq is the corresponding squared canonical correlation. Also displayed for each eigenvalue is the difference from the next eigenvalue, the proportion of the sum of the eigenvalues, and the cumulative proportion.

- **Test for H0: CanCorrj=0, j>=k**, the likelihood ratio for the hypothesis that the current canonical correlation and all smaller ones are zero in the population
- **Approx F** based on Rao's approximation to the distribution of the likelihood ratio
- **Num DF** and **Den DF** (numerator and denominator degrees of freedom) and **Pr > F** (probability level) associated with the F statistic

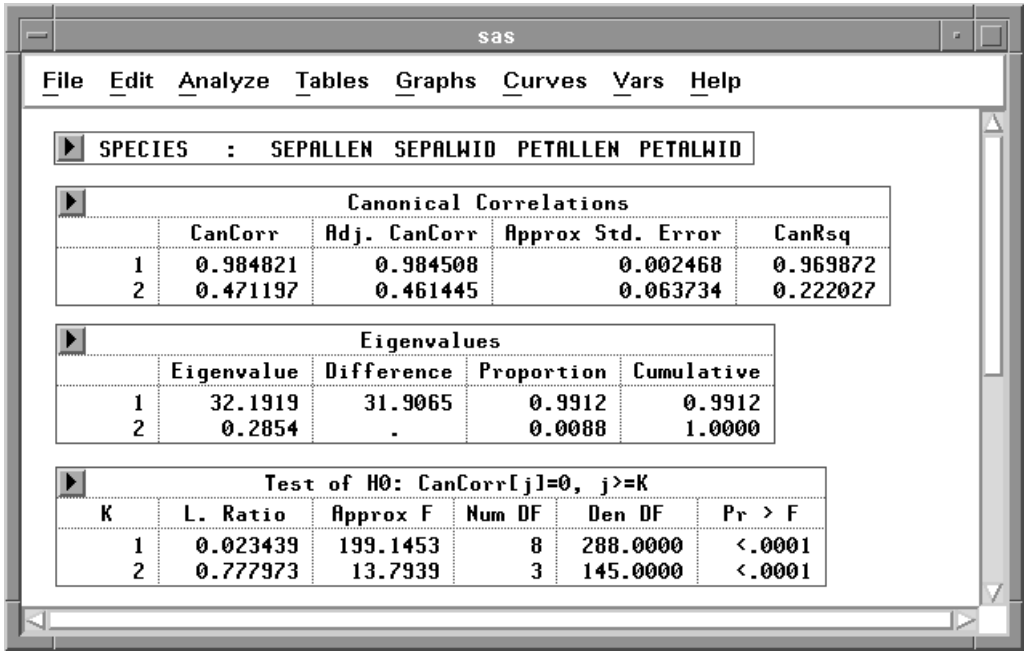


Figure 40.32. Canonical Correlations Tables

The **Correlations (Structure)** table includes the correlations between the input **X** variables and the canonical variables. The scoring coefficients are the coefficients of the **X** variables that are used to compute canonical variable scores. The **Std Scoring Coefs** table includes the scoring coefficients of the standardized **X** variables, and the **Raw Scoring Coefs** table includes the scoring coefficients of the centered **X** variables.

Figure 40.33 shows tables of correlations between the **X** variables and the first two canonical variables, and the scoring coefficients of the standardized **X** variables.

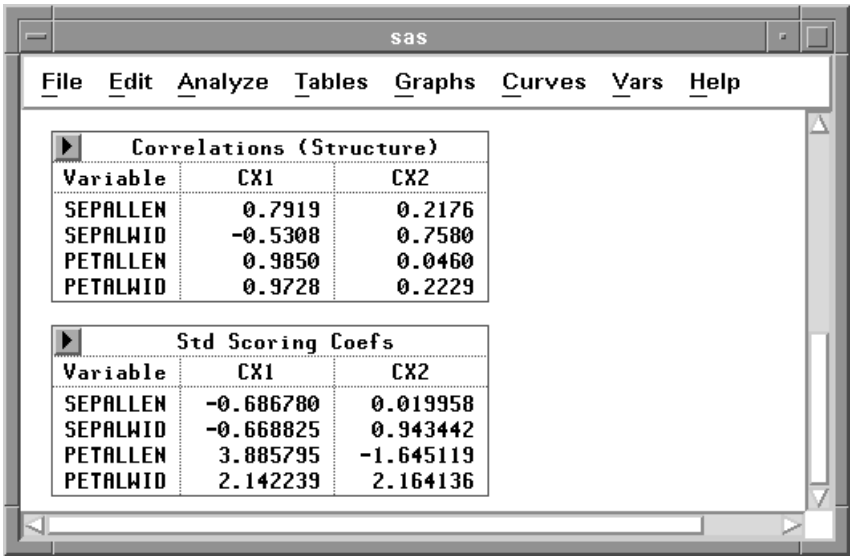


Figure 40.33. Correlations and Scoring Coefficients Tables

Graphs

You can create a scatter plot matrix and plots corresponding to various multivariate analyses by setting options in the Output Options dialogs, as shown in Figure 40.5 to Figure 40.10, or by choosing from the **Graphs** menu, as shown in Figure 40.34.

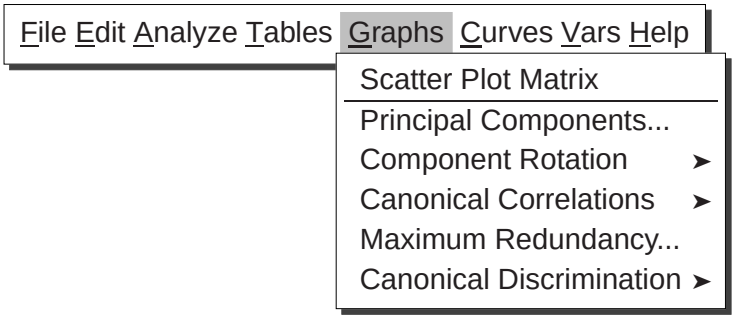


Figure 40.34. Graphs Menu

Scatter Plot Matrix

Scatter plots are displayed for pairs of variables. Without **X** variables, scatter plots are displayed as a symmetric matrix containing each pair of **Y** variables. With a nominal **Y** variable, scatter plots are displayed as a symmetric matrix containing each pair of **X** variables. When both interval **Y** variables and interval **X** variables are selected, scatter plots are displayed as a rectangular matrix with **Y** variables as the row variables and **X** variables as the column variables.

Figure 40.35 displays part of a scatter plot matrix with 80% prediction confidence ellipses.

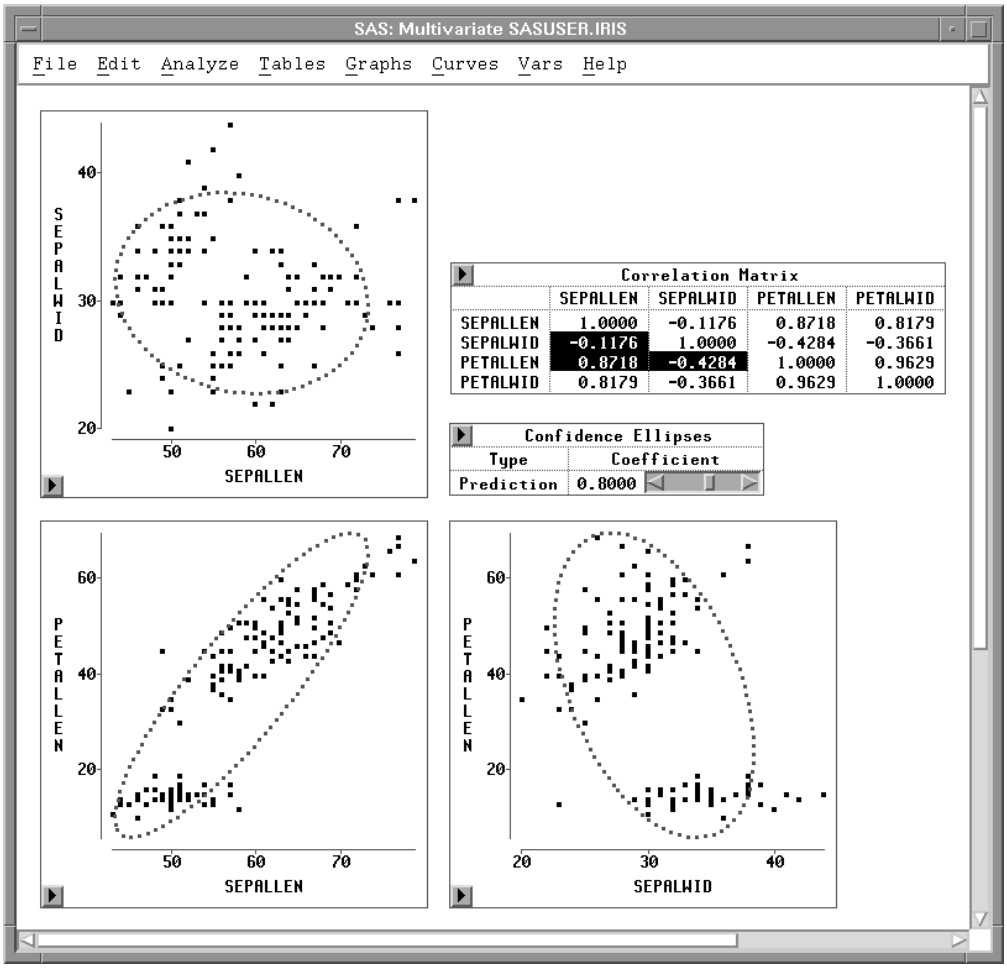


Figure 40.35. Scatter Plot Matrix with 80% Prediction Confidence Ellipses

Principal Component Plots

You can use principal component analysis to transform the **Y** variables into a smaller number of principal components that account for most of the variance of the **Y** variables. The plots of the first few components can reveal useful information about the distribution of the data, such as identifying different groups of the data or identifying observations with extreme values (possible outliers).

You can request a plot of the first two principal components or the first three principal components from the Principal Components Options dialog, shown in Figure 40.6, or from the **Graphs** menu, shown in Figure 40.34. Select **Principal Components** from the **Graphs** menu to display the **Principal Component Plots** dialog.

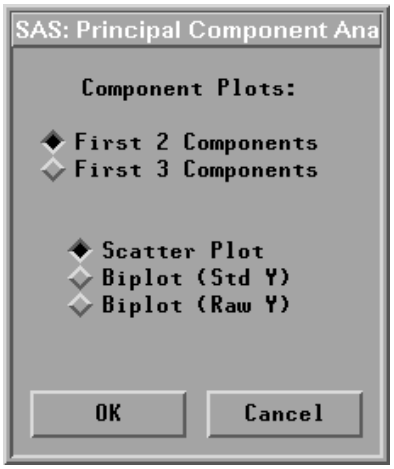


Figure 40.36. Principal Component Plots Dialog

In the dialog, you choose a principal component scatter plot (**Scatter Plot**), a principal component biplot with standardized **Y** variables (**Biplot (Std Y)**), or a principal component biplot with centered **Y** variables (**Biplot (Raw Y)**).

A *biplot* is a joint display of two sets of variables. The data points are first displayed in a scatter plot of principal components. With the approximated **Y** variable axes also displayed in the scatter plot, the data values of the **Y** variables are graphically estimated.

The **Y** variable axes are generated from the regression coefficients of the **Y** variables on the principal components. The lengths of the axes are approximately proportional to the standard deviations of the variables. A closer parallel between a **Y** variable axis and a principal component axis indicates a higher correlation between the two variables.

For a **Y** variable **Y1**, the **Y1** variable value of a data point y in a principal component biplot is geometrically evaluated as follows:

- A perpendicular is dropped from point y onto the **Y1** axis.
- The distance from the origin to this perpendicular is measured.
- The distance is multiplied by the length of the **Y1** axis; this gives an approximation of the **Y1** variable value for point y.

Two sets of variables are used in creating principal component biplots. One set is the **Y** variables. Either standardized or centered **Y** variables are used, as specified in the Principal Component Plots dialog, shown in Figure 40.36.

The other set is the principal component variables. These variables have variances either equal to one or equal to corresponding eigenvalues. You specify the principal component variable variance in the Multivariate Method Options dialog, shown in Figure 40.3.

† **Note:** A biplot with principal component variable variances equal to one is called a **GH'** biplot, and a biplot with principal component variable variances equal to corresponding eigenvalues is called a **JK'** biplot.

A biplot is a useful tool for examining data patterns and outliers. Figure 40.37 shows a biplot of the first two principal components from the correlation matrix and a rotating plot of the first three principal components. The biplot shows that the variable SEPALWID (highlighted axis) has a moderate negative correlation with PCR1 and a high correlation with PCR2.

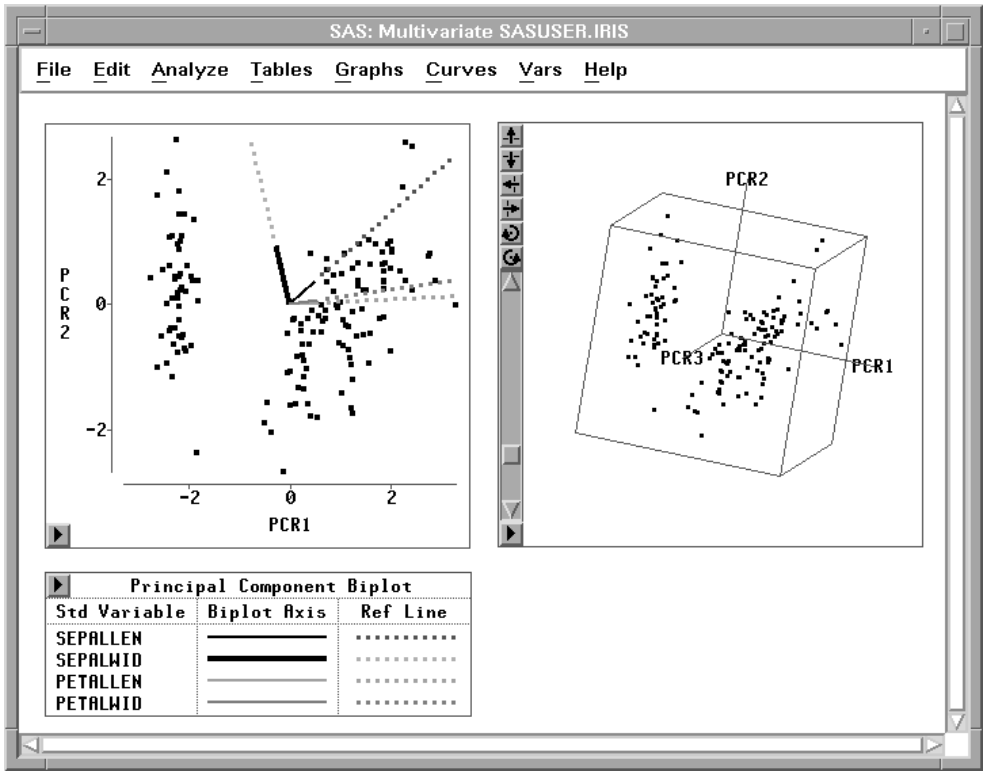


Figure 40.37. Principal Component Plots

Component Rotation Plots

You can request a plot of the rotated principal components from the Principal Components Rotation Options dialog, shown in Figure 40.7, or from the **Component Rotation** menu, shown in Figure 40.38.

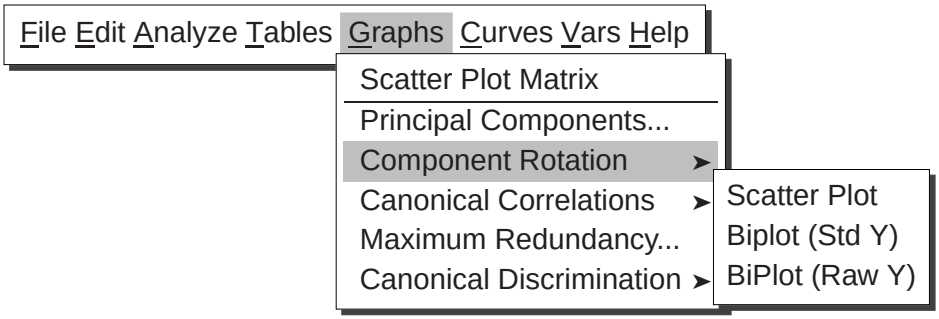


Figure 40.38. Component Rotation Menu

In the menu, you select a rotated component scatter plot (**Scatter Plot**), a rotated component biplot with standardized **Y** variables (**Biplot (Std Y)**), or a rotated component biplot with centered **Y** variables (**Biplot (Raw Y)**).

In a component rotation plot, the data points are displayed in a scatter plot of rotated principal components. With the approximated **Y** variable axes also displayed in the scatter plot, the data values of the **Y** variables are graphically estimated, as described previously in the "Principal Component Plots" section.

Figure 40.39 shows a biplot of the rotated first two principal components with standardized **Y** variables. The biplot shows that the variable **SEPALWID** (highlighted axis) has a high correlation with **RT2** and that the other three **Y** variables all have high correlations with **RT1**.

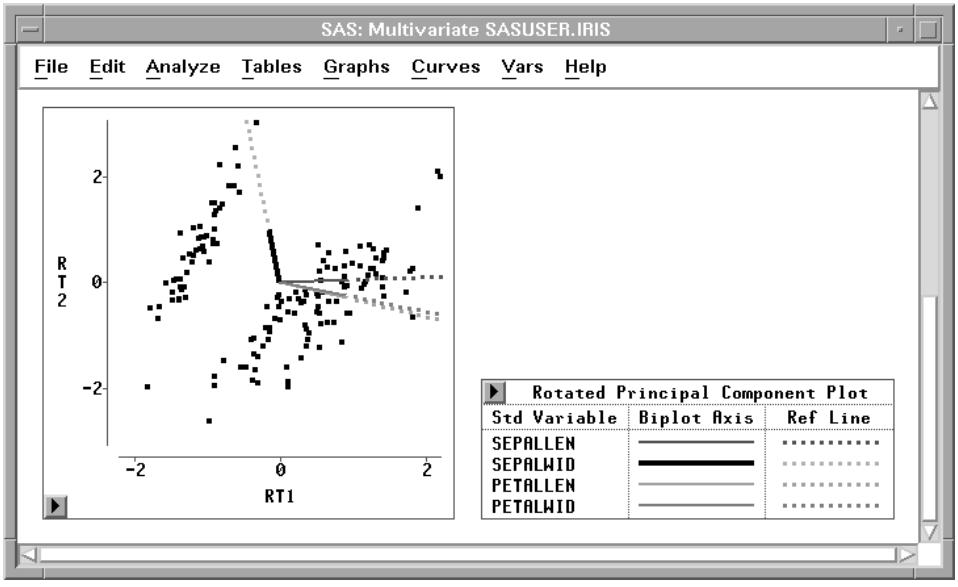


Figure 40.39. Rotated Principal Component Biplots

Canonical Correlation Plots

You can request pairwise canonical variable plots and a plot of the first two canonical variables or the first three canonical variables from each variable set from the Canonical Correlation Options dialog, shown in Figure 40.8, or from the **Graphs** menu, shown in Figure 40.40.

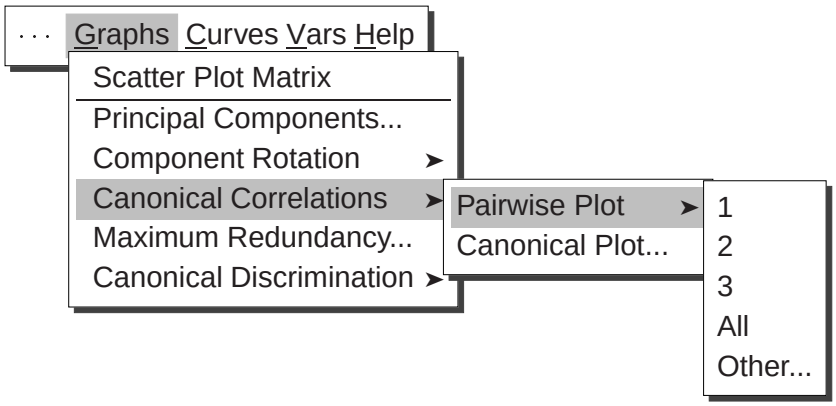


Figure 40.40. Canonical Correlations Menu

Figure 40.41 shows scatter plots of the first two pairs of canonical variables. The first scatter plot shows a high canonical correlation (0.7956) between canonical variables **CX1** and **CY1** and the second scatter plot shows a low correlation (0.2005) between **CX2** and **CY2**.

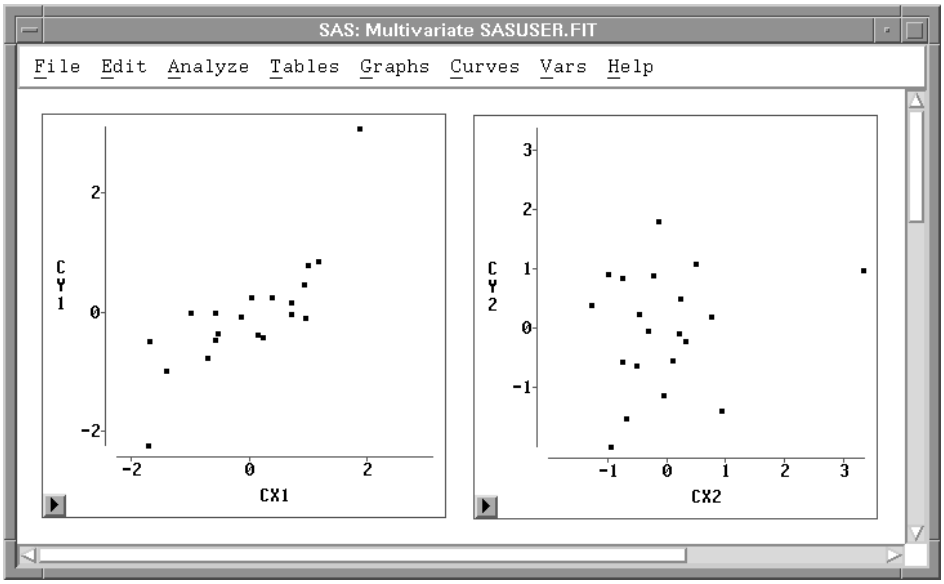


Figure 40.41. Canonical Correlation Pairwise Plots

Select **Canonical Plot** from the **Canonical Correlations** menu in Figure 40.40 to display a Canonical Correlation Component Plots dialog.

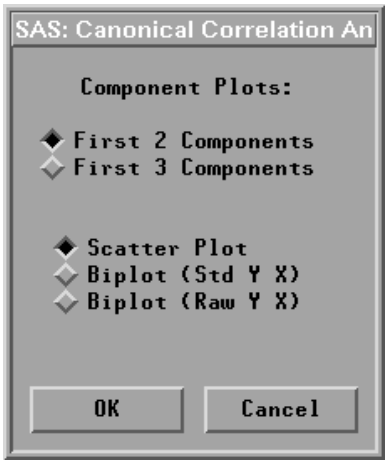


Figure 40.42. Canonical Correlation Component Plots Dialog

In the dialog, you choose a canonical correlation component scatter plot (**Scatter Plot**), a component biplot with standardized **Y** and **X** variables (**Biplot (Std Y X)**), or a component biplot with centered **Y** and **X** variables (**Biplot (Raw Y X)**).

In a canonical correlation component biplot, the data points are displayed in a scatter plot of canonical correlation components. With the approximated **Y** and **X** variable axes also displayed in the scatter plot, the data values of the **Y** and **X** variables are graphically estimated, as described previously in the "Principal Component Plots" section.

Figure 40.43 shows a biplot of the first two canonical variables from the **Y** variable sets with standardized **Y** and **X** variables. The biplot shows that the variables WEIGHT and WAIST (highlighted axes) have positive correlations with CY1 and negative correlations with CY2. The other four variables have negative correlations with CY1 and positive correlations with CY2.

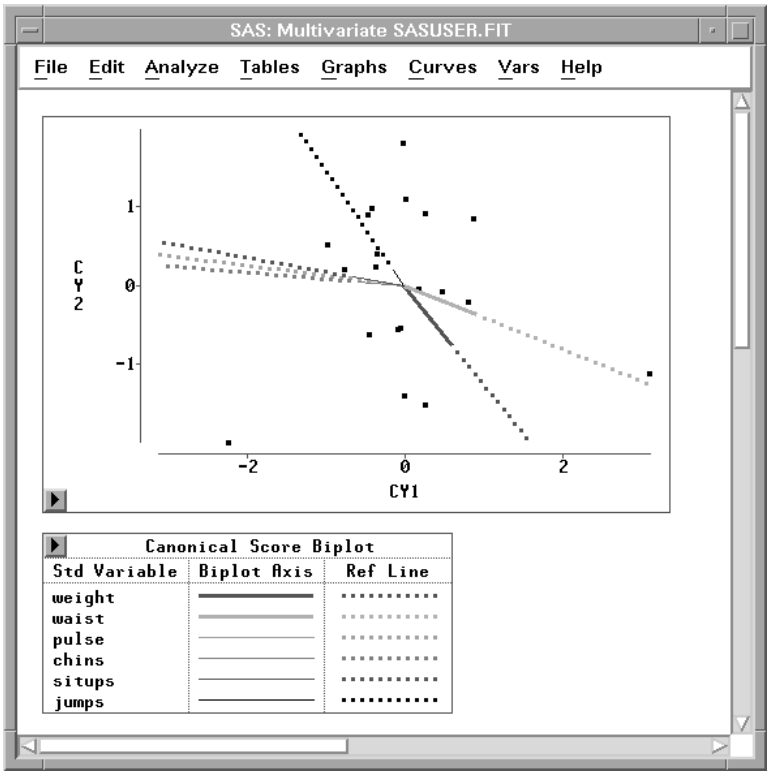


Figure 40.43. Canonical Correlation Component Biplot

Maximum Redundancy Plots

You can request a plot of the first two canonical variables or the first three canonical variables from each variable set from the Maximum Redundancy Options dialog, shown in Figure 40.9, or from the **Graphs** menu, shown in Figure 40.34. Select **Maximum Redundancy** from the **Graphs** menu to display a Maximum Redundancy Component Plots dialog.

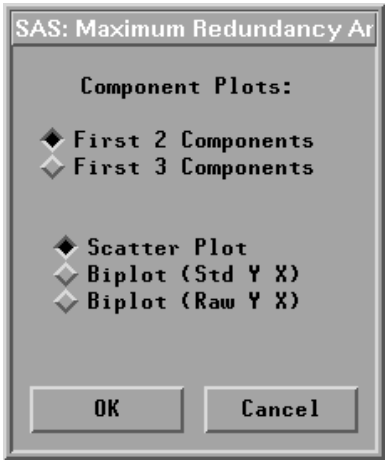


Figure 40.44. Maximum Redundancy Component Plots Dialog

In the dialog, you choose a maximum redundancy component scatter plot (**Scatter Plot**), a component biplot with standardized **Y** and **X** variables (**Biplot (Std Y X)**), or a component biplot with centered **Y** and **X** variables (**Biplot (Raw Y X)**).

In a maximum redundancy component biplot, the data points are displayed in a scatter plot of maximum redundancy components. With the approximated **Y** and **X** variable axes also displayed in the scatter plot, the data values of the **Y** and **X** variables are graphically estimated, as described previously in the "Principal Component Plots" section.

Figure 40.45 shows scatter plots of the first two canonical variables from each set of variables. The canonical variables in each plot are uncorrelated.

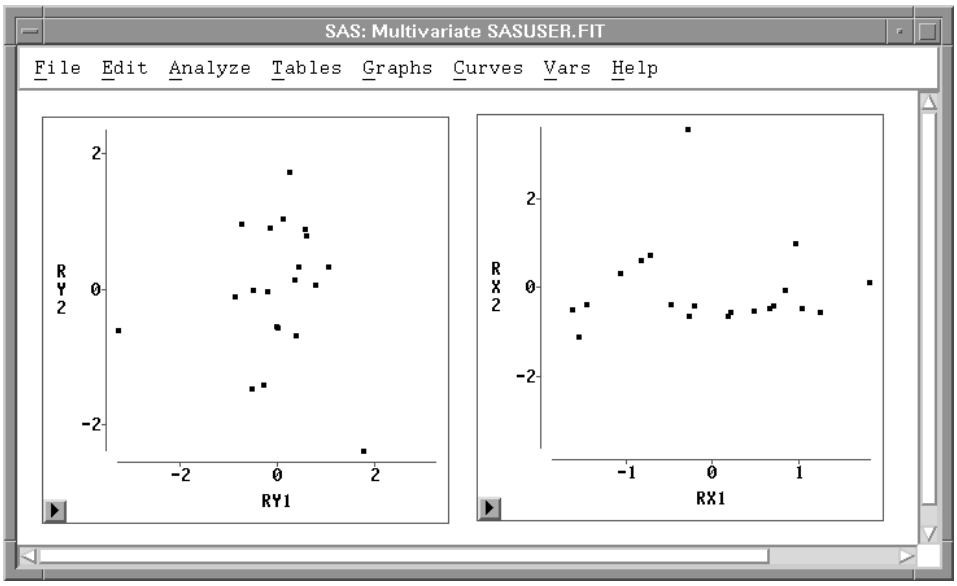


Figure 40.45. Maximum Redundancy Component Scatter Plots

Canonical Discrimination Plots

You can request a bar chart for the **Y** variable and a plot of the first two canonical variables or the first three canonical variables from the canonical discriminant options dialog, shown in Figure 40.10, or from the **Graphs** menu, shown in Figure 40.46.

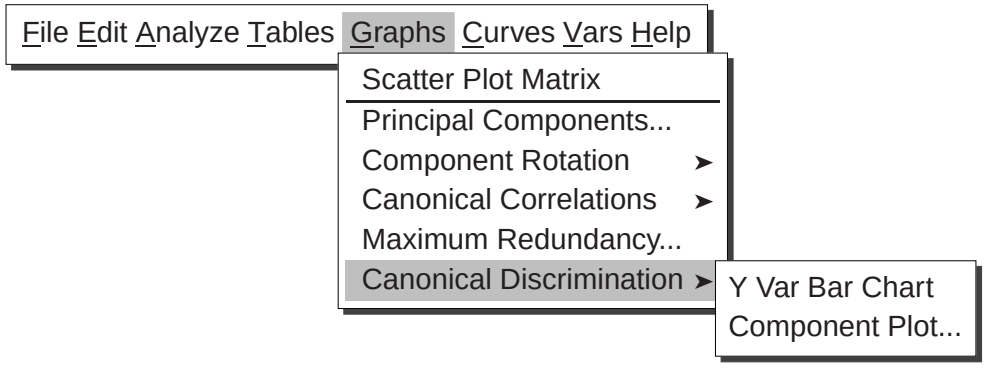


Figure 40.46. Canonical Discrimination Menu

Figure 40.47 shows a bar chart for the variable **SPECIES**.

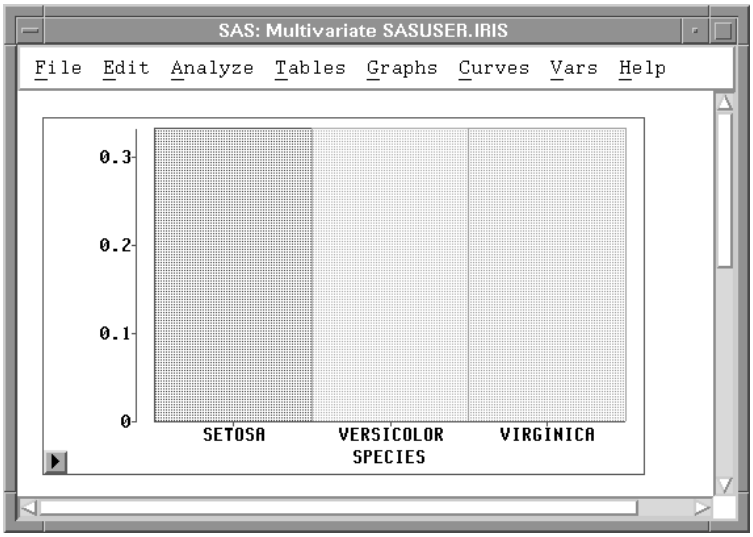


Figure 40.47. Y Var Bar Chart

Select **Component Plot** from the **Canonical Discriminant** menu in Figure 40.48 to display a Canonical Discriminant Component Plots dialog.

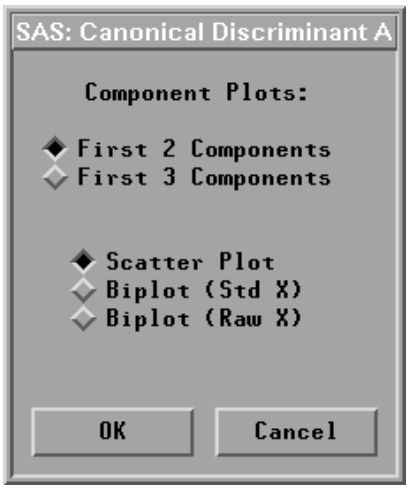


Figure 40.48. Canonical Discriminant Component Plots Dialog

In the dialog, you choose a canonical discriminant component scatter plot (**Scatter Plot**), a component biplot with standardized **X** variables (**Biplot (Std X)**), or a component biplot with centered **X** variables (**Biplot (Raw X)**).

In a canonical discriminant component biplot, the data points are displayed in a scatter plot of canonical discriminant components. With the approximated **X** variable axes also displayed in the scatter plot, the data values of the **X** variables are graphically estimated, as described previously in the "Principal Component Plots" section.

Figure 40.49 shows a biplot of the first two canonical variables from the **X** variable set with centered **X** variables. The biplot shows that the variable SEPALWID (highlighted axis) has a moderate negative correlation with CX1 and the other three variables have high correlation with CX1.

† **Note:** Use caution when evaluating distances in the biplot when the axes do not have comparable scales.

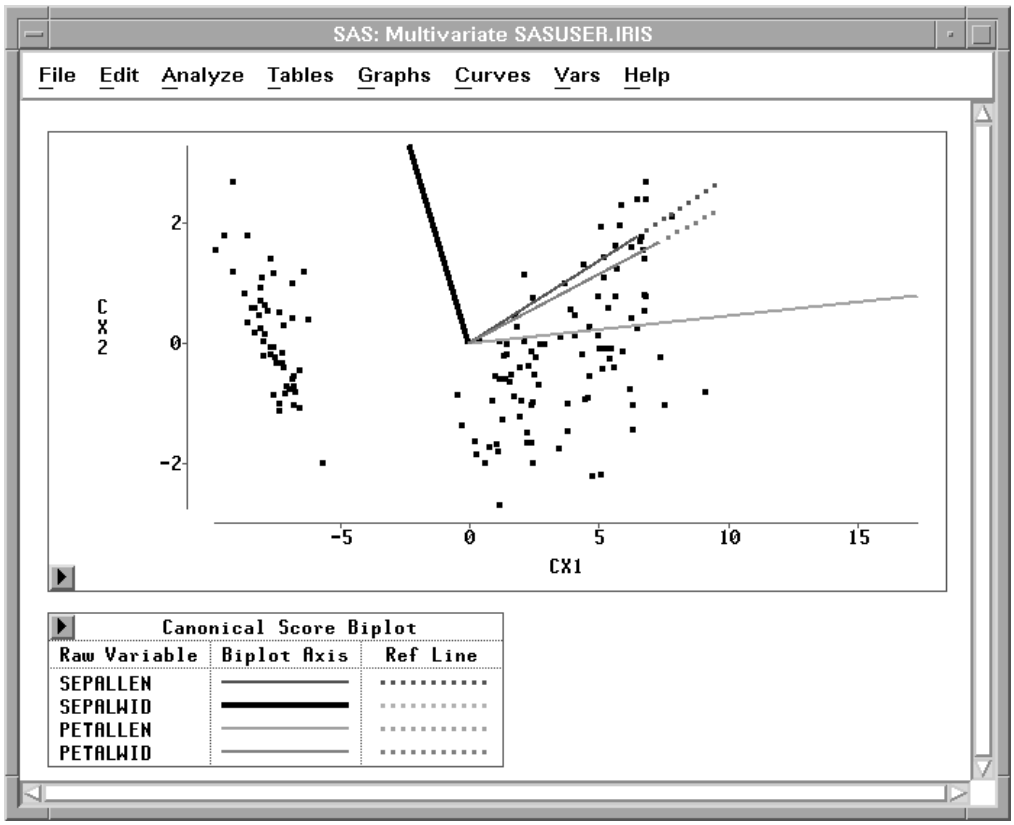


Figure 40.49. Canonical Discrimination Component Plot

Confidence Ellipses

SAS/INSIGHT software provides two types of confidence ellipses for pairs of analysis variables. One is a confidence ellipse for the population mean, and the other is a confidence ellipse for prediction. A confidence ellipse for the population mean is displayed with dashed lines, and a confidence ellipse for prediction is displayed with dotted lines.

Using these confidence ellipses assumes that each pair of variables has a bivariate normal distribution. Let $\overline{\mathbf{Z}}$ and $\mathbf{S}$ be the sample mean and the unbiased estimate of the covariance matrix of a random sample of size n from a bivariate normal distribution with mean μ and covariance matrix Σ.

The variable $\overline{\mathbf{Z}} - \mu$ is distributed as a bivariate normal variate with mean 0 and covariance $n^{-1}\mathbf{\Sigma}$, and it is independent of $\mathbf{S}$. The confidence ellipse for μ is based on Hotelling's T^2 statistic:

$$T^2 = n(\overline{\mathbf{Z}} - \mu)'\mathbf{S}^{-1}(\overline{\mathbf{Z}} - \mu)$$

A $100(1-\alpha)\%$ confidence ellipse for μ is defined by the equation

$$(\overline{\mathbf{Z}} - \mu)'\mathbf{S}^{-1}(\overline{\mathbf{Z}} - \mu) = \frac{2(n-1)}{n(n-2)}F_{2,n-2}(1-\alpha)$$

where $F_{2,n-2}(1-\alpha)$ is the $(1-\alpha)$ critical value of an F variate with degrees of freedom 2 and $n-2$.

A confidence ellipse for prediction is a confidence region for predicting a new observation in the population. It also approximates a region containing a specified percentage of the population.

Consider $\mathbf{Z}$ as a bivariate random variable for a new observation. The variable $\mathbf{Z} - \overline{\mathbf{Z}}$ is distributed as a bivariate normal variate with mean 0 and covariance $(1+1/n)\mathbf{\Sigma}$, and it is independent of $\mathbf{S}$.

A $100(1-\alpha)\%$ confidence ellipse for prediction is then given by the equation

$$(\mathbf{Z} - \overline{\mathbf{Z}})'\mathbf{S}^{-1}(\mathbf{Z} - \overline{\mathbf{Z}}) = \frac{2(n+1)(n-1)}{n(n-2)}F_{2,n-2}(1-\alpha)$$

The family of ellipses generated by different F critical values has a common center (the sample mean) and common major and minor axes.

The ellipses graphically indicate the correlation between two variables. When the variable axes are standardized (by dividing the variables by their respective standard deviations), the ratio of the two axis lengths (in Euclidean distances) reflects the magnitude of the correlation between the two variables. A ratio of 1 between the major and minor axes corresponds to a circular confidence contour and indicates that the variables are uncorrelated. A larger value of the ratio indicates a larger positive or negative correlation between the variables.

Scatter Plot Confidence Ellipses

You can generate confidence ellipses by setting the options in the multivariate output options dialog, shown in Figure 40.5, or by choosing from the **Curves** menu, shown in Figure 40.50.

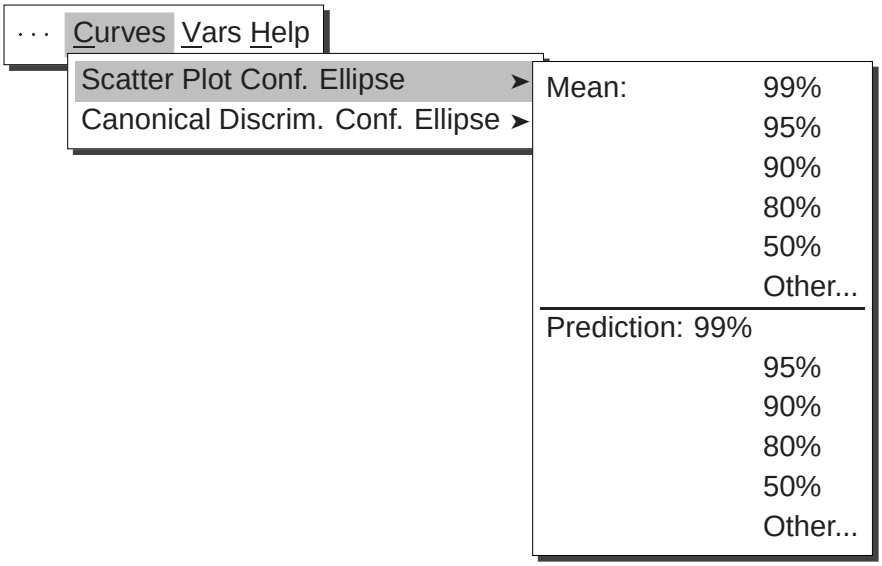


Figure 40.50. Curves Menu

Only 80% prediction confidence ellipses can be selected in the multivariate output options dialog. You must use the **Curves** menu to display mean confidence ellipses. You can use the confidence coefficient slider in the **Confidence Ellipses** table to change the coefficient for these ellipses.

Figure 40.35 displays part of a scatter plot matrix with 80% prediction confidence ellipses and the **Correlation Matrix** table with corresponding correlations highlighted. The ellipses graphically show a small negative correlation (-0.1176) between variables **SEPALLEN** and **SEPALWID**, a moderate negative correlation (-0.4284) between variables **SEPALWID** and **PETALLEN**, and a large positive correlation (0.8718) between variables **SEPALLEN** and **PETALLEN**.

† **Note:** The confidence ellipses displayed in this illustration may not be appropriate since none of the scatter plots suggest bivariate normality.

Canonical Discriminant Confidence Ellipses

You can also generate class-specific confidence ellipses for the first two canonical components in canonical discriminant analysis by setting the options in the Canonical Discriminant Options dialog, shown in Figure 40.10, or by choosing from the predeeding **Curves** menu.

Figure 40.51 displays a scatter plot of the first two canonical components with class-specific 80% prediction confidence ellipses. The figure shows that the first canonical variable **CX1** has most of the discriminntory power between the two canonical variables.

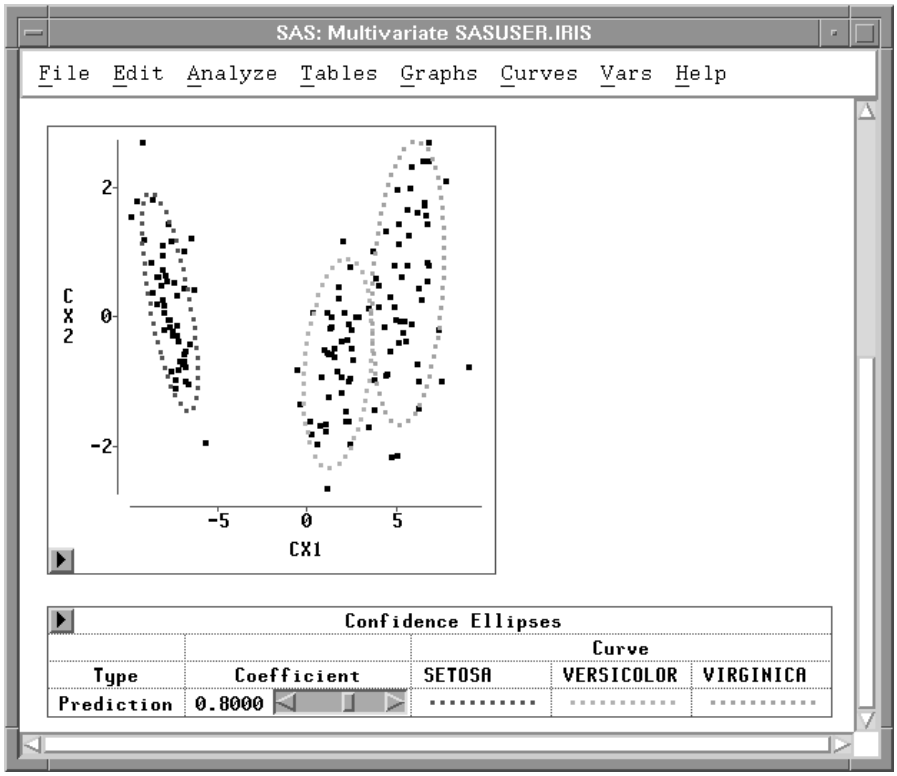


Figure 40.51. Canonical Discriminant Confidence Ellipses

Output Variables

You can save component scores from principal component analysis, component rotation, canonical correlation analysis, maximum redundancy analysis, and canonical discriminant analysis in the data window for use in subsequent analyses. For component rotation, the number of component output variables is the number of components rotated, as specified in Figure 40.4. For other analyses, you specify the number of component output variables in the Output Options dialogs, shown in Figure 40.6 to Figure 40.10, or from the **Vars** menu, shown in Figure 40.52. For component rotation, you specify the number of output rotated components in the Rotation Options dialog, shown in Figure 40.4.

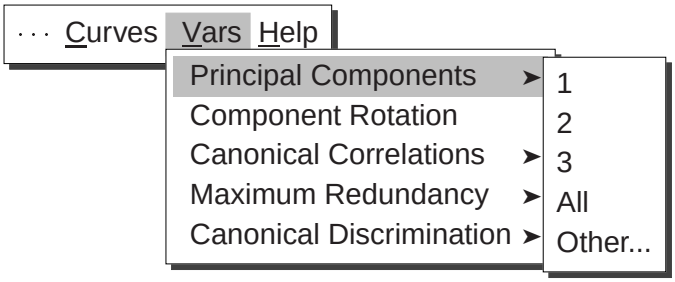


Figure 40.52. Vars Menu

Selecting **1**, **2**, or **3** gives you 1, 2, or 3 components. **All** gives you all components. Selecting **0** in the component options dialogs suppresses the output variables in the corresponding analysis. Selecting **Other** in the **Vars** menu displays the dialog shown in Figure 40.53. You specify the number of components you want to save in the dialog.

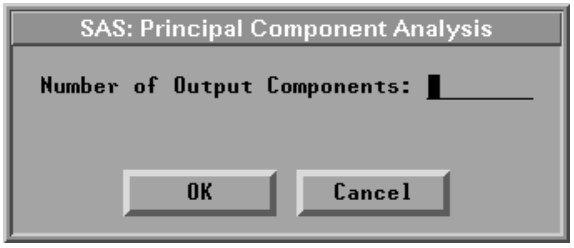


Figure 40.53. Output Components Dialog

Principal Components

For principal components from a covariance matrix, the names of the variables containing principal component scores are **PCV1**, **PCV2**, **PCV3**, and so on. The output component scores are a linear combination of the centered **Y** variables with coefficients equal to the eigenvectors of the covariance matrix.

For principal components from a correlation matrix, the names of the variables containing principal component scores are **PCR1**, **PCR2**, **PCR3**, and so on. The output component scores are a linear combination of the standardized **Y** variables with coefficients equal to the eigenvectors of the correlation matrix.

If you specify **Variance=Eigenvalues** in the multivariate method options dialog, the new variables of principal component scores have mean zero and variance equal to the associated eigenvalues. If you specify **Variance=1**, the new variables have variance equal to one.

Principal Component Rotation

The names of the variables containing rotated principal component scores are **RT1**, **RT2**, **RT3**, and so on. The new variables of rotated principal component scores have mean zero and variance equal to one.

Canonical Variables

The names of the variables containing canonical component scores are **CY1**, **CY2**, **CY3**, and so on, from the **Y** variable list, and **CX1**, **CX2**, **CX3**, from the **X** variable list. The new variables of canonical component scores have mean zero and variance equal to one.

Maximum Redundancy

The names of the variables containing maximum redundancy scores are **RY1**, **RY2**, **RY3**, and so on, from the **Y** variable list, and **RX1**, **RX2**, **RX3**, from the **X** variable list. The new variables of maximum redundancy scores have mean zero and variance equal to one.

Canonical Discriminant

The names of the variables containing canonical component scores are **CX1**, **CX2**, **CX3**, and so on. If you specify **Std Pooled Variance** in the multivariate method options dialog, the new variables of canonical component scores have mean zero and pooled within-class variance equal to one. If you specify **Std Total Variance**, the new variables have total-sample variance equal to one.

Weighted Analyses

When the observations are independently distributed with a common mean and unequal variances, a weighted analysis may be appropriate. The individual weights are the values of the **Weight** variable you specify.

The following statistics are modified to incorporate the observation weights:

- Mean $\overline{\mathbf{y}}_w, \overline{\mathbf{x}}_w$
- SSCP $\mathbf{U}_{yy}, \mathbf{U}_{yx}, \mathbf{U}_{xx}$
- CSSCP $\mathbf{C}_{yy}, \mathbf{C}_{yx}, \mathbf{C}_{xx}$
- COV $\mathbf{S}_{yy}, \mathbf{S}_{yx}, \mathbf{S}_{xx}$
- CORR $\mathbf{R}_{yy}, \mathbf{R}_{yx}, \mathbf{R}_{xx}$

The formulas for these weighted statistics are given in the "Method" section earlier in this chapter. The resulting weighted statistics are used in the multivariate analyses.

References

Cooley, W.W. and Lohnes, P.R. (1971), *Multivariate Data Analysis*, New York: John Wiley & Sons, Inc.

Dillon, W.R. and Goldstein, M. (1984), *Multivariate Analysis*, New York: John Wiley & Sons, Inc.

Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

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

Gnanadesikan, R. (1997), *Methods for Statistical Data Analysis of Multivariate Observations*, Second Edition, New York: John Wiley & Sons, Inc.

Gower, J.C. and Hand, D.J. (1996), *Biplots*, New York: Chapman and Hall.

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

Hotelling, H. (1935), "The Most Predictable Criterion," *Journal of Educational Psychology*, 26, 139–142.

Hotelling, H. (1936), "Relations Between Two Sets of Variables," *Biometrika*, 28, 321–377.

Jobson, J.D. (1992), *Applied Multivariate Data Analysis, Vol 2: Categorical and Multivariate Methods*, New York: Springer-Verlag.

Kaiser, H.F. (1958), "The Varimax Criterion of Analytic Rotation in Factor Analysis," *Psychometrika*, 23, 187–200.

Krzanowski, W.J. (1988), *Principles of Multivariate Analysis: A User's Perspective*, New York: Oxford University Press.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, New York: Marcel Dekker, Inc.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, New York: Academic Press.

Morrison, D.F. (1976), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill Book Co.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, 6(2), 559–572.

Pringle, R.M. and Raynor, A.A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing Co.

Rao, C.R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya A*, 26, 329–358.

Rao, C.R. (1973), *Linear Statistical Inference*, New York: John Wiley & Sons, Inc.

Stewart, D.K. and Love, W.A. (1968), "A General Canonical Correlation Index," *Psychological Bulletin*, 70, 160–163.

van den Wollenberg, A.L. (1977), "Redundancy Analysis—An Alternative to Canonical Correlation Analysis," *Psychometrika*, 42, 207–219.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/INSIGHT User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 752 pp.

SAS/INSIGHT User's Guide, Version 8

ISBN 1–58025–490–X

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999