Statistical Discovery.™ From SAS.

Version 16

Multivariate Methods

"The real voyage of discovery consists not in seeking new landscapes,
but in having new eyes."
Marcel Proust

JMP, A Business Unit of SAS
SAS Campus Drive
Cary, NC 27513

16.0

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2020–2021. *JMP® 16 Multivariate Methods*. Cary, NC: SAS Institute Inc.

JMP® 16 Multivariate Methods

Copyright © 2020–2021, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

March 2021

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to http://support.sas.com/thirdpartylicenses.

Contents

Multivariate Methods

1 Learn about JMP

6 Partial Least Squares Models 121

Develop Models Using Correlations between Ys and Xs

16 Cluster Variables 341

Group Similar Variables into Representative Groups

A Statistical Details 355

Multivariate Methods

Chapter 1

Learn about JMP

Documentation and Additional Resources

Learn about JMP documentation, such as book conventions, descriptions of each JMP document, the Help system, and where to find additional support.

Contents

Formatting Conventions in JMP Documentation

These conventions help you relate written material to information that you see on your screen:

- Sample data table names, column names, pathnames, filenames, file extensions, and folders appear in Helvetica (or sans-serif online) font.
- Code appears in `Lucida Sans Typewriter` (or monospace online) font.
- Code output appears in *`Lucida Sans Typewriter`* italic (or monospace italic online) font and is indented farther than the preceding code.
- **Helvetica bold** formatting (or bold sans-serif online) indicates items that you select to complete a task:
  - buttons
  - check boxes
  - commands
  - list names that are selectable
  - menus
  - options
  - tab names
  - text boxes
- The following items appear in italics:
  - words or phrases that are important or have definitions specific to JMP
  - book titles
  - variables
- Features that are for JMP Pro only are noted with the JMP Pro icon JMP PRO. For an overview of JMP Pro features, visit https://www.jmp.com/software/pro.

Note: Special information and limitations appear within a Note.

Tip: Helpful information appears within a Tip.

JMP Help

JMP Help in the Help menu enables you to search for information about JMP features, statistical methods, and the JMP Scripting Language (or *JSL*). You can open JMP Help in several ways:

- Search and view JMP Help on Windows by selecting **Help > JMP Help**.
- On Windows, press the F1 key to open the Help system in the default browser.
- Get help on a specific part of a data table or report window. Select the Help tool ? from the **Tools** menu and then click anywhere in a data table or report window to see the Help for that area.
- Within a JMP window, click the **Help** button.

Note: The JMP Help is available for users with Internet connections. Users without an Internet connection can search all books in a PDF file by selecting **Help > JMP Documentation Library**. See "JMP Documentation Library" on page 16 for more information.

JMP Documentation Library

The Help system content is also available in one PDF file called *JMP Documentation Library*. Select **Help > JMP Documentation Library** to open the file. If you prefer searching individual PDF files of each document in the JMP library, download the files from https://www.jmp.com/documentation.

The following table describes the purpose and content of each document in the JMP library.

Document Title	**Document Purpose**	**Document Content**
Discovering JMP	If you are not familiar with JMP, start here.	Introduces you to JMP and gets you started creating and analyzing data. Also learn how to share your results.
Using JMP	Learn about JMP data tables and how to perform basic operations.	Covers general JMP concepts and features that span across all of JMP, including importing data, modifying columns properties, sorting data, and connecting to SAS.

Document Title	**Document Purpose**	**Document Content**
Basic Analysis	Perform basic analysis using this document.	Describes these Analyze menu platforms: • Distribution • Fit Y by X • Tabulate • Text Explorer Covers how to perform bivariate, one-way ANOVA, and contingency analyses through Analyze > Fit Y by X. How to approximate sampling distributions using bootstrapping and how to perform parametric resampling with the Simulate platform are also included.
Essential Graphing	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Scatterplot Matrix • Ternary Plot • Treemap • Chart • Overlay Plot The book also covers how to create background and custom maps.
Profilers	Learn how to use interactive profiling tools, which enable you to view cross-sections of any response surface.	Covers all profilers listed in the Graph menu. Analyzing noise factors is included along with running simulations using random inputs.

Document Title	**Document Purpose**	**Document Content**
Design of Experiments Guide	Learn how to design experiments and determine appropriate sample sizes.	Covers all topics in the DOE menu.
Fitting Linear Models	Learn about Fit Model platform and many of its personalities.	Describes these personalities, all available within the Analyze menu Fit Model platform: • Standard Least Squares • Stepwise • Generalized Regression • Mixed Model • MANOVA • Loglinear Variance • Nominal Logistic • Ordinal Logistic • Generalized Linear Model

Document Title	**Document Purpose**	**Document Content**
Predictive and Specialized Modeling	Learn about additional modeling techniques.	Describes these Analyze > Predictive Modeling menu platforms: • Neural • Partition • Bootstrap Forest • Boosted Tree • K Nearest Neighbors • Naive Bayes • Support Vector Machines • Model Comparison • Model Screening • Make Validation Column • Formula Depot Describes these Analyze > Specialized Modeling menu platforms: • Fit Curve • Nonlinear • Functional Data Explorer • Gaussian Process • Time Series • Matched Pairs Describes these Analyze > Screening menu platforms: • Modeling Utilities • Response Screening • Process Screening • Predictor Screening • Association Analysis • Process History Explorer

Document Title	Document Purpose	Document Content
Multivariate Methods	Read about techniques for analyzing several variables simultaneously.	Describes these Analyze > Multivariate Methods menu platforms: • Multivariate • Principal Components • Discriminant • Partial Least Squares • Multiple Correspondence Analysis • Structural Equation Models • Factor Analysis • Multidimensional Scaling • Item Analysis Describes these Analyze > Clustering menu platforms: • Hierarchical Cluster • K Means Cluster • Normal Mixtures • Latent Class Analysis • Cluster Variables
Quality and Process Methods	Read about tools for evaluating and improving processes.	Describes these Analyze > Quality and Process menu platforms: • Control Chart Builder and individual control charts • Measurement Systems Analysis • Variability / Attribute Gauge Charts • Process Capability • Model Driven Multivariate Control Chart • Legacy Control Charts • Pareto Plot • Diagram • Manage Spec Limits • OC Curves

Document Title	**Document Purpose**	**Document Content**
Reliability and Survival Methods	Learn to evaluate and improve reliability in a product or system and analyze survival data for people and products.	Describes these Analyze > Reliability and Survival menu platforms: • Life Distribution • Fit Life by X • Cumulative Damage • Recurrence Analysis • Degradation • Destructive Degradation • Reliability Forecast • Reliability Growth • Reliability Block Diagram • Repairable Systems Simulation • Survival • Fit Parametric Survival • Fit Proportional Hazards
Consumer Research	Learn about methods for studying consumer preferences and using that insight to create better products and services.	Describes these Analyze > Consumer Research menu platforms: • Categorical • Choice • MaxDiff • Uplift • Multiple Factor Analysis
Scripting Guide	Learn about taking advantage of the powerful JMP Scripting Language (JSL).	Covers a variety of topics, such as writing and debugging scripts, manipulating data tables, constructing display boxes, and creating JMP applications.
JSL Syntax Reference	Read about many JSL functions on functions and their arguments, and messages that you send to objects and display boxes.	Includes syntax, examples, and notes for JSL commands.

Additional Resources for Learning JMP

In addition to reading JMP help, you can also learn about JMP using the following resources:

- "JMP Tutorials"
- "Sample Data Tables"
- "Learn about Statistical and JSL Terms"
- "Learn JMP Tips and Tricks"
- "JMP Tooltips"
- "JMP User Community"
- "Free Online Statistical Thinking Course"
- "JMP New User Welcome Kit"
- "Statistics Knowledge Portal"
- "JMP Training"
- "JMP Books by Users"
- "The JMP Starter Window"

JMP Tutorials

You can access JMP tutorials by selecting **Help > Tutorials**. The first item on the **Tutorials** menu is **Tutorials Directory**. This opens a new window with all the tutorials grouped by category.

If you are not familiar with JMP, start with the **Beginners Tutorial**. It steps you through the JMP interface and explains the basics of using JMP.

The rest of the tutorials help you with specific aspects of JMP, such as designing an experiment and comparing a sample mean to a constant.

Sample Data Tables

All of the examples in the JMP documentation suite use sample data. Select **Help > Sample Data Library** to open the sample data directory.

To view an alphabetized list of sample data tables or view sample data within categories, select **Help > Sample Data**.

Sample data tables are installed in the following directory:

On Windows: C:\Program Files\SAS\JMP\16\Samples\Data

On macOS: \Library\Application Support\JMP\16\Samples\Data

In JMP Pro, sample data is installed in the JMPPRO (rather than JMP) directory.

To view examples using sample data, select **Help > Sample Data** and navigate to the Teaching Resources section. To learn more about the teaching resources, visit https://jmp.com/tools.

Learn about Statistical and JSL Terms

For help with statistical terms, select Help > Statistics Index. For help with JSL scripting and examples, select **Help > Scripting Index**.

Statistics Index Provides definitions of statistical terms.

Scripting Index Lets you search for information about JSL functions, objects, and display boxes. You can also edit and run sample scripts from the Scripting Index and get help on the commands.

Learn JMP Tips and Tricks

When you first start JMP, you see the Tip of the Day window. This window provides tips for using JMP.

To turn off the Tip of the Day, clear the **Show tips at startup** check box. To view it again, select **Help > Tip of the Day**. Or, you can turn it off using the Preferences window.

JMP Tooltips

JMP provides descriptive tooltips (or *hover labels*) when you hover over items, such as the following:

- Menu or toolbar options
- Labels in graphs
- Text results in the report window (move your cursor in a circle to reveal)
- Files or windows in the Home Window
- Code in the Script Editor

Tip: On Windows, you can hide tooltips in the JMP Preferences. Select **File > Preferences > General** and then deselect **Show menu tips**. This option is not available on macOS.

JMP User Community

The JMP User Community provides a range of options to help you learn more about JMP and connect with other JMP users. The learning library of one-page guides, tutorials, and demos is a good place to start. And you can continue your education by registering for a variety of JMP training courses.

Other resources include a discussion forum, sample data and script file exchange, webcasts, and social networking groups.

To access JMP resources on the website, select **Help > JMP User Community** or visit https://community.jmp.com.

Free Online Statistical Thinking Course

Learn practical statistical skills in this free online course on topics such as exploratory data analysis, quality methods, and correlation and regression. The course consists of short videos, demonstrations, exercises, and more. Visit https://www.jmp.com/statisticalthinking.

JMP New User Welcome Kit

The JMP New User Welcome Kit is designed to help you quickly get comfortable with the basics of JMP. You'll complete its thirty short demo videos and activities, build your confidence in using the software, and connect with the largest online community of JMP users in the world. Visit https://www.jmp.com/welcome.

Statistics Knowledge Portal

The Statistics Knowledge Portal combines concise statistical explanations with illuminating examples and graphics to help visitors establish a firm foundation upon which to build statistical skills. Visit https://www.jmp.com/skp.

JMP Training

SAS offers training on a variety of topics led by a seasoned team of JMP experts. Public courses, live web courses, and on-site courses are available. You might also choose the online e-learning subscription to learn at your convenience. Visit https://www.jmp.com/training.

JMP Books by Users

Additional books about using JMP that are written by JMP users are available on the JMP website. Visit https://www.jmp.com/books.

The JMP Starter Window

The JMP Starter window is a good place to begin if you are not familiar with JMP or data analysis. Options are categorized and described, and you launch them by clicking a button. The JMP Starter window covers many of the options found in the Analyze, Graph, Tables, and File menus. The window also lists JMP Pro features and platforms.

- To open the JMP Starter window, select **View** (**Window** on macOS) **> JMP Starter**.
- To display the JMP Starter automatically when you open JMP on Windows, select **File > Preferences > General**, and then select **JMP Starter** from the Initial JMP Window list. On macOS, select **JMP > Preferences > Initial JMP Starter Window**.

JMP Technical Support

JMP technical support is provided by statisticians and engineers educated in SAS and JMP, many of whom have graduate degrees in statistics or other technical disciplines.

Many technical support options are provided at https://www.jmp.com/support, including the technical support phone number.

Chapter 2

Introduction to Multivariate Analysis

Overview of Multivariate Techniques

Multivariate Methods describes the following techniques for analyzing several variables simultaneously:

- The Multivariate platform examines multiple variables to see how they relate to each other. See Chapter 3, “Correlations and Multivariate Techniques”.
- The Principal Components platform derives a small number of independent linear combinations (principal components) of a set of measured variables that capture as much of the variability in the original variables as possible. It is a useful exploratory technique and can help you create predictive models. See Chapter 4, “Principal Components”.
- The Discriminant platform looks to find a way to predict a classification (X) variable (nominal or ordinal) based on known continuous responses (Y). It can be regarded as inverse prediction from a multivariate analysis of variance (MANOVA). See Chapter 5, “Discriminant Analysis”.
- The Partial Least Squares platform fits linear models based on factors, namely, linear combinations of the explanatory variables (Xs). PLS exploits the correlations between the Xs and the Ys to reveal underlying latent structures. See Chapter 6, “Partial Least Squares Models”.
- The Multiple Correspondence Analysis (MCA) platform takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA is frequently used in the social sciences particularly in France and Japan. It can be used in survey analysis to identify question agreement. See Chapter 7, “Multiple Correspondence Analysis”.
- JMP PRO The Structural Equation Models platform enables you to fit a variety of models, including confirmatory factor analysis, path models with or without latent variables, measurement error models, and latent growth curve models. See Chapter 8, “Structural Equation Models”.
- The Factor Analysis platform enables you to construct factors from a larger set of observed variables. These factors are expressed as linear combinations of a subset of the observed variables. Factor analysis enables you to explore the number of factors that are explained by a set of measured, observed variables, and the strength of the relationship between factors and variables. See Chapter 9, “Factor Analysis”.
- The Multidimensional Scaling (MDS) platform enables you to create a visual representation of the pattern of proximities (similarities, dissimilarities, or distances) among a set of objects. See Chapter 10, “Multidimensional Scaling”.

- The Item Analysis platform enables you to fit item response theory models. The Item Response Theory (IRT) method is used for the analysis and scoring of measurement instruments such as tests and questionnaires. IRT uses a system of models to relate a trait or ability to an individual’s probability of endorsing or correctly responding to an item. IRT can be used to study standardized tests, cognitive development, and consumer preferences. See Chapter 11, “Item Analysis”.
- The Hierarchical Cluster platform groups rows together that share similar values across a number of variables. It is a useful exploratory technique to help you understand the clumping structure of your data. See Chapter 12, “Hierarchical Cluster”.
- The KMeans Clustering platform groups observations that share similar values across a number of variables. See Chapter 13, “K Means Cluster”.
- The Normal Mixtures platform enables you to cluster observations when your data come from overlapping normal distributions. See Chapter 14, “Normal Mixtures”.
- The Latent Class Analysis platform finds clusters of observations for categorical response variables. The model takes the form of a multinomial mixture model. See Chapter 15, “Latent Class Analysis”.
- The Cluster Variables platform groups similar variables into representative groups. You can use Cluster Variables as a dimension-reduction method. Instead of using a large set of variables in modeling, the cluster components of the most representative variable in the cluster can be used to explain most of the variation in the data. See Chapter 16, “Cluster Variables”.

Chapter 3

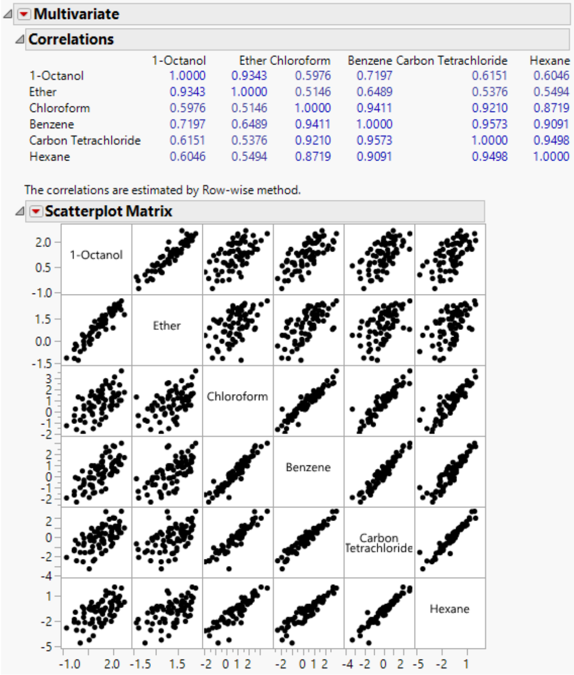
Correlations and Multivariate Techniques

Explore the Multidimensional Behavior of Variables

Multivariate data involve many variables instead of one (univariate) or two (bivariate). Use the Multivariate platform to explore how multiple variables relate to each other. The Multivariate platform provides many techniques to summarize and test the strength of the linear relationship between each pair of response variables. Both parametric and nonparametric correlations tests are available in the platform. You can also use graphical features, such as the Scatterplot Matrix and Color Maps, to identify dependencies, outliers, and clusters among the variables.

There are additional multivariate analysis techniques to further examine the relationship between variables, including principal components analysis, outlier analysis, and item reliability. These techniques are available through the Multivariate report. You can also use the Principal Components Analysis and Outlier Analysis platforms in JMP for more in-depth implementations of these techniques.

Figure 3.1 Example of a Multivariate Report



Multivariate

Correlations

	1-Octanol	Ether	Chloroform	Benzene	Carbon Tetrachloride	Hexane
1-Octanol	1.0000	0.9343	0.5976	0.7197	0.6151	0.6046
Ether	0.9343	1.0000	0.5146	0.6489	0.5376	0.5494
Chloroform	0.5976	0.5146	1.0000	0.9411	0.9210	0.8719
Benzene	0.7197	0.6489	0.9411	1.0000	0.9573	0.9091
Carbon Tetrachloride	0.6151	0.5376	0.9210	0.9573	1.0000	0.9498
Hexane	0.6046	0.5494	0.8719	0.9091	0.9498	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

Contents

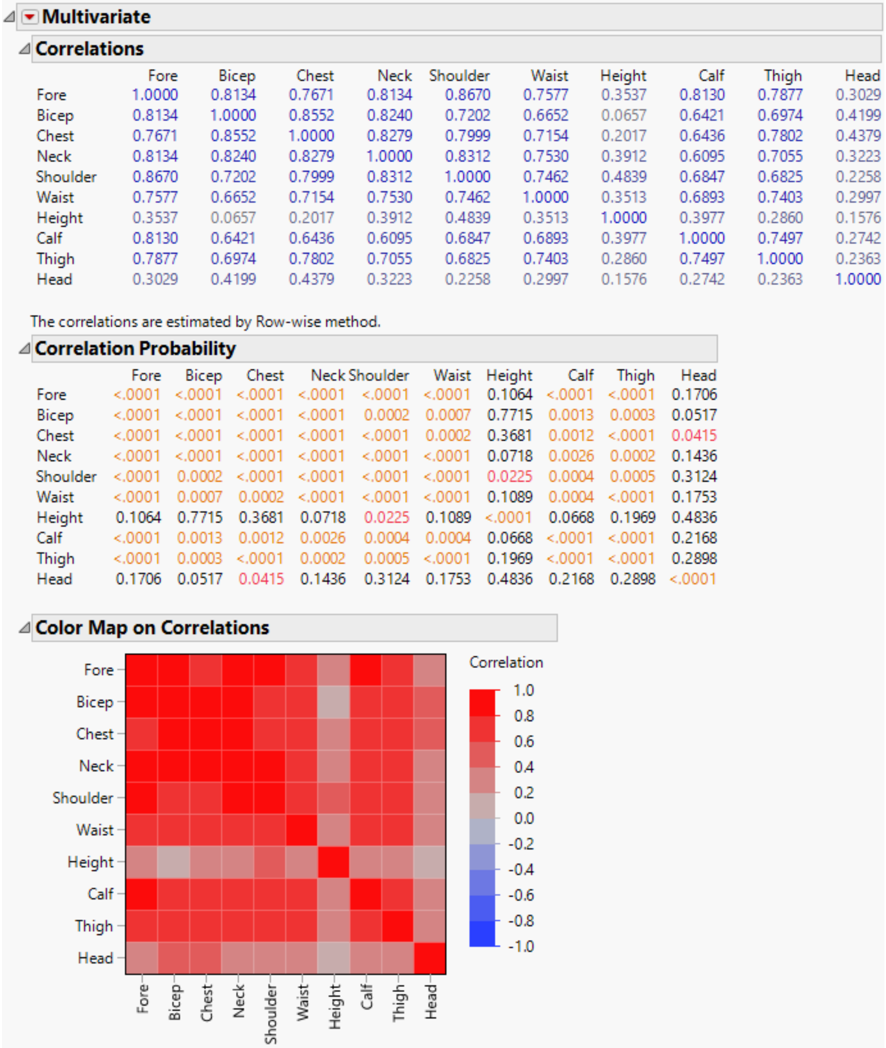
Example of the Multivariate Platform

This example uses the Body Measurements.jmp sample data table to examine the relationship between different measurements on the body.

1. Select **Help > Sample Data Library** and open Body Measurements.jmp.
2. Select **Analyze > Multivariate Methods > Multivariate**.
3. Select all columns except Mass and click **Y, Columns**.
4. Click **OK**.

   The initial multivariate report contains a correlation matrix and scatterplot matrix. There is also a note that tells you that the Row-wise estimation method was used. All of the variables are positively correlated, but at varying strengths.
5. Click the Multivariate red triangle and deselect **Scatterplot Matrix**.
6. Click the Multivariate red triangle and select **Correlation Probability**.
7. Click the Multivariate red triangle and select **Color Maps > Color Map on Correlations**.

Figure 3.2 Multivariate Report for Body Measurements Data



Multivariate

Correlations

	Fore	Bicep	Chest	Neck	Shoulder	Waist	Height	Calf	Thigh	Head
Fore	1.0000	0.8134	0.7671	0.8134	0.8670	0.7577	0.3537	0.8130	0.7877	0.3029
Bicep	0.8134	1.0000	0.8552	0.8240	0.7202	0.6652	0.0657	0.6421	0.6974	0.4199
Chest	0.7671	0.8552	1.0000	0.8279	0.7999	0.7154	0.2017	0.6436	0.7802	0.4379
Neck	0.8134	0.8240	0.8279	1.0000	0.8312	0.7530	0.3912	0.6095	0.7055	0.3223
Shoulder	0.8670	0.7202	0.7999	0.8312	1.0000	0.7462	0.4839	0.6847	0.6825	0.2258
Waist	0.7577	0.6652	0.7154	0.7530	0.7462	1.0000	0.3513	0.6893	0.7403	0.2997
Height	0.3537	0.0657	0.2017	0.3912	0.4839	0.3513	1.0000	0.3977	0.2860	0.1576
Calf	0.8130	0.6421	0.6436	0.6095	0.6847	0.6893	0.3977	1.0000	0.7497	0.2742
Thigh	0.7877	0.6974	0.7802	0.7055	0.6825	0.7403	0.2860	0.7497	1.0000	0.2363
Head	0.3029	0.4199	0.4379	0.3223	0.2258	0.2997	0.1576	0.2742	0.2363	1.0000

The correlations are estimated by Row-wise method.

Correlation Probability

	Fore	Bicep	Chest	Neck	Shoulder	Waist	Height	Calf	Thigh	Head
Fore	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1064	<.0001	<.0001	0.1706
Bicep	<.0001	<.0001	<.0001	<.0001	0.0002	0.0007	0.7715	0.0013	0.0003	0.0517
Chest	<.0001	<.0001	<.0001	<.0001	<.0001	0.0002	0.3681	0.0012	<.0001	0.0415
Neck	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0718	0.0026	0.0002	0.1436
Shoulder	<.0001	0.0002	<.0001	<.0001	<.0001	<.0001	0.0225	0.0004	0.0005	0.3124
Waist	<.0001	0.0007	0.0002	<.0001	<.0001	<.0001	0.1089	0.0004	<.0001	0.1753
Height	0.1064	0.7715	0.3681	0.0718	0.0225	0.1089	<.0001	0.0668	0.1969	0.4836
Calf	<.0001	0.0013	0.0012	0.0026	0.0004	0.0004	0.0668	<.0001	<.0001	0.2168
Thigh	<.0001	0.0003	<.0001	0.0002	0.0005	<.0001	0.1969	<.0001	<.0001	0.2898
Head	0.1706	0.0517	0.0415	0.1436	0.3124	0.1753	0.4836	0.2168	0.2898	<.0001

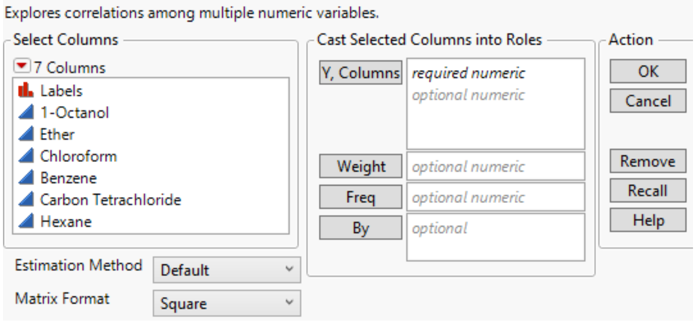
Color Map on Correlations

The Color Map on Correlations report provides a more concise version of the information in the Correlation Matrix. The majority of the plot is dark red colors indicating that most of the variables are highly correlated. The two lighter colored rows and columns indicate that height and head measurements are not highly correlated with other variables. This is further supported by the mostly non-significant *p*-values for Height and Head in the Correlation Probability table.

Launch the Multivariate Platform

Launch the Multivariate platform by selecting **Analyze > Multivariate Methods > Multivariate**.

Figure 3.3 The Multivariate Launch Window



Explores correlations among multiple numeric variables.

Estimation Method: Default

Matrix Format: Square

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns Identifies one or more response columns. The response columns must have a numeric data type, but the modeling type can be continuous or ordinal.

Note: If you specify an ordinal response variable, a JMP alert appears after you click OK in the launch window. The alert indicates which variables are ordinal and confirms that you intended to include the ordinal variables in the analysis.

Weight Identifies one column whose numeric values assign a weight to each row in the analysis.

Freq Identifies one column whose numeric values assign a frequency to each row in the analysis.

By Produces a separate report of each level of By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Estimation Method Specifies the method for calculating the correlations. REML and Pairwise are the methods used most frequently. Several of these methods address the treatment of missing data. You can also estimate missing values by using a method other than Row-wise and then selecting the Impute Missing Data command. See "Impute Missing Data" on page 45.

Default The Default option uses either the Row-wise, Pairwise, or REML methods.

- Row-wise estimation is used for data tables with no missing values.

- Pairwise estimation is used for data tables with missing values and either more than 10 columns, more than 5,000 rows, or more columns than rows.
- REML estimation is used otherwise.

Note: When the Default option would otherwise result in REML estimation, but the fit does not converge properly, the platform reverts to the Pairwise method. This can happen when there are missing values in your data table and at least one of the following situations applies: if your data table has fewer than 10 columns, fewer than 5,000 rows, or fewer columns than rows. If the estimation method shown is Pairwise, this means that the REML fit did not converge.

REML Restricted maximum likelihood (REML) estimation uses all of the data, even if missing values are present. Due to a bias-correction factor, this method is slow if the data set is large and there are many missing values. Therefore, REML is most useful for smaller data sets. If there are no missing cells in the data, then the REML and ML estimates are equivalent and equal to the sample covariance matrix. If there are missing cells, REML's variance and covariance estimates are less biased than the estimates from ML estimation. For statistical details, see "REML" on page 47.

ML Maximum likelihood (ML) estimation uses all of the data, even if missing values are present. Because the estimates from ML are generated quickly, this method is most useful for large data tables with missing data.

Robust Robust estimation uses all of the data, even if missing values are present. This method down-weights extreme values and is therefore useful for data tables that might have outliers. For statistical details, see "Robust" on page 47.

Row-wise Row-wise estimation calculates the Pearson correlation for each pair of columns. For statistical details, see "Pearson Product-Moment Correlation" on page 48. Row-wise estimation does not use rows with missing values. This method is useful for excluding observations with missing data.

Pair-wise Pair-wise estimation uses all of the data, even if missing values are present. This estimation method calculates Pearson correlations for each pair of columns using all observations with nonmissing values for those two columns. For statistical details, see "Pearson Product-Moment Correlation" on page 48. Pair-wise estimation is most useful when a data table has missing values and either more columns than rows, more than 10 columns, or more than 5,000 rows.

Note: If you select REML, ML, or Robust and your data table contains more columns than rows and has missing values, JMP switches the Estimation Method to Pairwise.

Matrix Format Select a format option for the Scatterplot Matrix. The Square option displays plots for all ordered combinations of columns. Lower Triangular displays plots below the diagonal, with the first n - 1 columns on the horizontal axis. Upper Triangular displays plots above the diagonal, with the first n - 1 columns on the vertical axis.

The Multivariate Report

The default multivariate report shows the standard correlation matrix, the scatterplot matrix, and a note that indicates which method was used to estimate the correlations. In some cases, information that explains why the given method was used is also displayed. The platform menu lists additional correlation options and other techniques for looking at multiple variables. See "Multivariate Platform Options" on page 35.

Multivariate Platform Options

Note: The correlation values and p-values in the Multivariate report are formatted using the conditional formatting settings in Preferences. See *Using JMP*.

The Multivariate red triangle menu contains the following options:

Correlations Multivariate Shows or hides the Correlations table. The table is a matrix of correlation coefficients that summarizes the strength of the linear relationships between each pair of response (Y) variables. This option is on by default. See "Pearson Product-Moment Correlation" on page 48.

Note: This correlation matrix is calculated by the method that you select in the launch window.

Correlation Probability Shows or hides the Correlation Probability table. The table is a matrix of p-values. Each p-value corresponds to a test of the null hypothesis that the true correlation between the variables is zero. This is a test of no linear relationship between the two response variables.

CI of Correlations Shows or hides the two-tailed confidence intervals of the correlations.

Tip: The default confidence coefficient is 95%. Use the Set α Level option to change the confidence coefficient.

Inverse Correlations Shows or hides the Inverse Corr report, which is the inverse correlation matrix. The diagonal elements of the matrix are a function of how closely the variable is a linear function of the other variables. In the inverse correlation table, the diagonal is the value of $1/(1 - R^2)$. R^2 is calculated from the simple linear model regressing that variable on all the other variables. If the multiple correlation is zero, the diagonal inverse element is 1. If the multiple correlation is 1, then the inverse element becomes infinite and is reported missing. For statistical details about inverse correlations, see the "Inverse Correlation Matrix" on page 50.

Partial Correlations Shows or hides the Partial Corr report, which is the partial correlation matrix. The partial correlation matrix shows the measure of the relationship between a pair of variables after adjusting for the effects of all the other variables. The table is the negative of the inverse correlation matrix, scaled to unit diagonal. This means that the matrix is scaled such that the diagonal elements are equal to one.

Partial Correlation Probability Shows or hides the Partial Correlation Probability report, which is a matrix of *p*-values. Each *p*-value corresponds to a test of the null hypothesis that the true partial correlation between the variables is zero. This is a test of no linear relationship between the two response variables, after adjusting for the effects of the other variables.

Note: The Partial Correlation Probability option is not available when there are not enough degrees of freedom. This can occur when there are more variables than observations.

Covariance Matrix Shows or hides the covariance matrix, which measures the degree to which a pair of variables change together.

Pairwise Correlations Shows or hides the Pairwise Correlations table, which lists the Pearson product-moment correlations for each pair of Y variables. The correlations are calculated by the pairwise deletion method. The count values differ if any pair has a missing value for either variable. The Pairwise Correlations report also shows significance probabilities and compares the correlations in a bar chart. All results are based on the pairwise method.

Note: This option excludes rows that are missing for either of the variables under consideration.

Hotelling's T^2 Test Enables you to conduct a one-sample test for the mean of the multivariate distribution of the variables that you entered as Y. When you select the Hotelling's T^2 Test option, a window appears that enables you to specify the mean vector under the null hypothesis. Enter a hypothesized mean for each variable. The test assumes multivariate normality of the Y variables. The Hotelling's T^2 Test report gives the following information:

Variable The variables entered as Y.

Mean The sample mean for each variable.

Hypothesized Mean The null hypothesis means that you specified.

Test Statistic The value of Hotelling's T^2 statistic.

F Ratio The value of the test statistic. If you have n rows and k variables, the F ratio is defined as follows:

$$\frac{n-k}{k(n-1)}T^2$$

Prob > F The p-value for the test. Under the null hypothesis the F ratio has an F distribution with k and n - k degrees of freedom.

Note: To remove a report, click the red triangle next to Hotelling's T^2 Test and select **Remove Test**.

Simple Statistics This menu contains two options that each show or hide simple statistics (mean, standard deviation, sum, minimum, and maximum) for each column. The univariate and multivariate simple statistics can differ when there are missing values present, or when the Robust method is used.

Univariate Simple Statistics Shows statistics that are calculated on each column, regardless of values or missing values in other columns. These values match those produced by the Distribution platform.

Multivariate Simple Statistics Shows statistics that correspond to the estimation method selected in the launch window and the presence or absence of missing data. If there are no missing observations, this option is available only for the Robust method. If there are missing observations, this option is available for all estimation methods except for Pairwise. For the REML, ML, or Robust methods, the mean vector and covariance matrix are estimated by the selected method. For the Row-wise method, all rows with at least one missing value are excluded from the calculation of means and variances.

Nonparametric Correlations This menu contains three nonparametric measures: Spearman's Rho, Kendall's Tau, and Hoeffding's D. Each option shows or hides a nonparametric report for the corresponding measure. See "Nonparametric Correlations" on page 39.

Set α Level You can specify any alpha value for the correlation confidence intervals. Four alpha values are listed: 0.01, 0.05, 0.10, and 0.50. Select **Other** to enter any other value.

Scatterplot Matrix Shows or hides a scatterplot matrix of each pair of response variables. This option is on by default. See "Scatterplot Matrix" on page 40.

Color Maps The Color Maps menu contains three types of color maps. Each option in the menu shows or hides a color map of the corresponding type. The following types of color maps are available:

Color Map On Correlations Produces a cell plot that shows the correlations among variables on a scale from red (+1) to blue (-1).

Color Map On p-values Produces a cell plot that shows the significance of the correlations on a scale from $p = 0$ (red) to $p = 1$ (blue).

Cluster the Correlations Produces a cell plot that clusters together similar variables. The correlations are the same as for Color Map on Correlations, but the positioning of the variables might be different.

Color Map on Pairwise Correlations Produces a cell plot that shows the pairwise correlations among variables on a scale from red (+1) to blue (-1).

Color Map on Spearman's ρ Produces a cell plot that shows the Spearman's ρ nonparametric correlations among the variables on a scale from red (+1) to blue (-1).

Color Map on Kendall's τ Produces a cell plot that shows the Kendall's τ nonparametric correlations among the variables on a scale from red (+1) to blue (-1).

Color Map on Hoeffding's D Produces a cell plot that shows the Hoeffding's D nonparametric correlations among the variables on a scale from red (+1) to blue (-1).

Parallel Coord Plot Shows or hides a parallel coordinate plot of the variables.

Ellipsoid 3D Plot Shows or hides a three-dimensional scatterplot with a 95% confidence ellipsoid. You are prompted to specify three variables and their corresponding axes when you select this option.

Principal Components This menu contains options to show or hide a principal components report. You can select a principal components report based on correlations, covariances, or unscaled values. Selecting one of these options when a principal components report is already shown changes the report to the new option. Select **None** to remove the report.

Principal components is a technique that calculates linear combinations of the original variables. The first principal component has maximum variation, the second principal component has the next most variation, subject to being orthogonal to the first, and so on. See the "Principal Components" chapter on page 53.

Outlier Analysis This menu contains options that each show or hide plots that measure distance in the multivariate sense using one of the following methods: the Mahalanobis distance, jackknife distances, and the T^2 statistic. See "Outlier Analysis" on page 42.

Item Reliability This menu contains options that each show or hide an item reliability report. The reports indicate how consistently a set of instruments measures an overall response, using either Cronbach's α or Standardized α. See "Item Reliability" on page 45.

Impute Missing Data (Available only if the data table contains missing values.) Produces a new data table that duplicates your data table and replaces all missing values with estimated values. See "Impute Missing Data" on page 45.

Save Imputed Formulas (Available only if the data table contains missing values.) For columns that contain missing values, saves new columns to the data table that contain the formulas used to estimate the missing values. The new columns are called Imputed_<Column Name>.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Nonparametric Correlations

The Nonparametric Correlations menu offers three nonparametric measures for pairwise correlations. Each Nonparametric correlation report gives the significance probability for the chosen measure of association and displays the association value on a bar chart. There are three nonparametric correlation measures.

Spearman's Rho A correlation coefficient computed on the ranks of the data values instead of on the values themselves.

Kendall's Tau Based on the number of concordant and discordant pairs of observations. A pair is *concordant* if the observation with the larger value of X also has the larger value of Y. A pair is *discordant* if the observation with the larger value of X has the smaller value of Y. There is a correction for tied pairs, which are pairs of observations that have equal values of X or equal values of Y.

Hoeffding's D A statistical scale that ranges from –0.5 to 1. Large positive values indicate dependence. The statistic approximates a weighted sum over observations of chi-square statistics for two-by-two classification tables. The two-by-two tables are made by setting each data value as the threshold. This statistic detects more general departures from independence.

Note: The nonparametric correlations are calculated using the Pairwise method, even if you selected a different Estimation Method in the launch window.

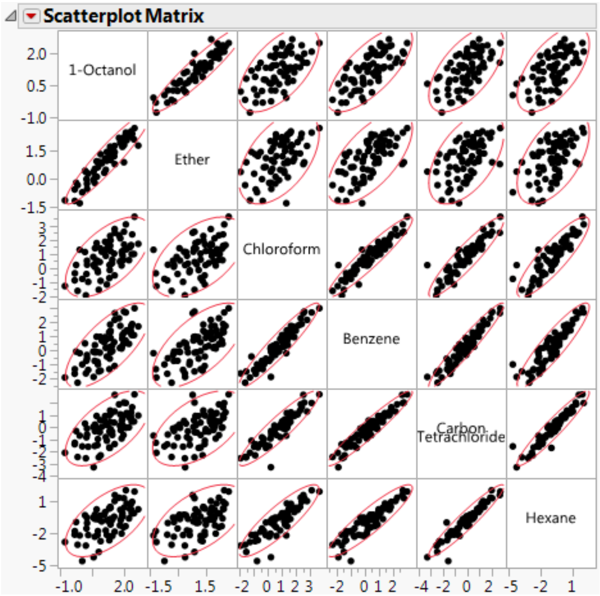
Note: When a Weight variable is specified, missing and zero-valued weights are excluded from the nonparametric correlation calculations. All other weight values are treated as 1.

For statistical details about these three methods, see the "Nonparametric Measures of Association" on page 48.

Scatterplot Matrix

A scatterplot matrix helps you visualize the correlations between each pair of response variables.

Figure 3.4 Scatterplot Matrix



Turning on the Density Ellipses option shows a 95% bivariate normal density ellipse in each scatterplot. Assuming that each pair of variables has a bivariate normal distribution, this ellipse encloses approximately 95% of the points. The narrowness of the ellipse reflects the degree of correlation of the variables. If the ellipse is fairly round and is not diagonally oriented, the variables are uncorrelated. If the ellipse is narrow and diagonally oriented, the variables are correlated.

Tips:

- Re-sizing any cell resizes all the cells.
- Drag a label cell to another label cell to reorder the matrix.
- When you look for patterns in the scatterplot matrix, you can see the variables cluster into groups based on their correlations. Figure 3.4 shows two clusters of correlations: the first two variables (top, left), and the next four (bottom, right).

Scatterplot Matrix Options

The red triangle menu for the Scatterplot Matrix lets you customize the matrix with color and density ellipses and by setting the α-level.

Show Points Shows or hides the points in the scatterplot.

Fit Line Shows or hides the regression line and 95% level confidence curves for the fitted regression line.

Density Ellipses Shows or hides the 95% density ellipses in the scatterplots. Use the Ellipse α menu to change the α-level.

Shaded Ellipses Colors each ellipse. Use the Ellipses Transparency and Ellipse Color menus to change the transparency and color.

Show Correlations Shows or hides the correlation of each pair of variables in the upper left corner of each scatterplot.

Matrix Options (Available only when the Square Matrix Format is selected on the launch window.) Shows a submenu of options to change the appearance of the upper right triangle of the scatterplot matrix. Only one of the following options can be selected at a time.

Significance Circles Shows or hides correlation circles in the upper right triangle of the scatterplot matrix. The color of each circle represents the correlation between each pair of variables on a scale from red (+1) to blue (-1). The size of each circle represents the significance test between the variables. A larger circle indicates a more significant relationship.

Heat Map Shows or hides a correlation heat map in the upper right triangle of the scatterplot matrix. The color of each cell in the heat map represents the correlation between each pair of variables on a scale from red (+1) to blue (-1).

Show Histograms Shows or hides horizontal or vertical histograms in the label cells. Once histograms have been added, select **Show Counts** to label each bar of the histogram with its count. Select **Horizontal** or **Vertical** to either change the orientation of the histograms or remove the histograms.

Ellipse α Sets the α-level used for the ellipses. Select one of the standard α-levels in the menu, or select **Other** to enter a different value.

Ellipses Transparency Sets the transparency of the ellipses if they are colored. Select one of the default levels, or select **Other** to enter a different value. The default value is 0.2.

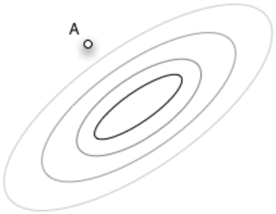
Ellipse Color Sets the color of the ellipses if they are colored. Select one of the colors in the palette, or select **Other** to use another color. The default value is red.

Nonpar Density Shows or hides shaded density contours based on a smooth nonparametric bivariate surface that describes the density of data points. Contours for the 10% and 50% quantiles of the nonparametric surface are shown.

Outlier Analysis

The Outlier Analysis menu contains options that each show or hide plots that measure distance in the multivariate sense, with respect to the correlation structure. For example, in Figure 3.5, Point A is an outlier because it is outside the correlation structure, even though it is not an outlier in any of the coordinate directions.

Figure 3.5 Example of an Outlier



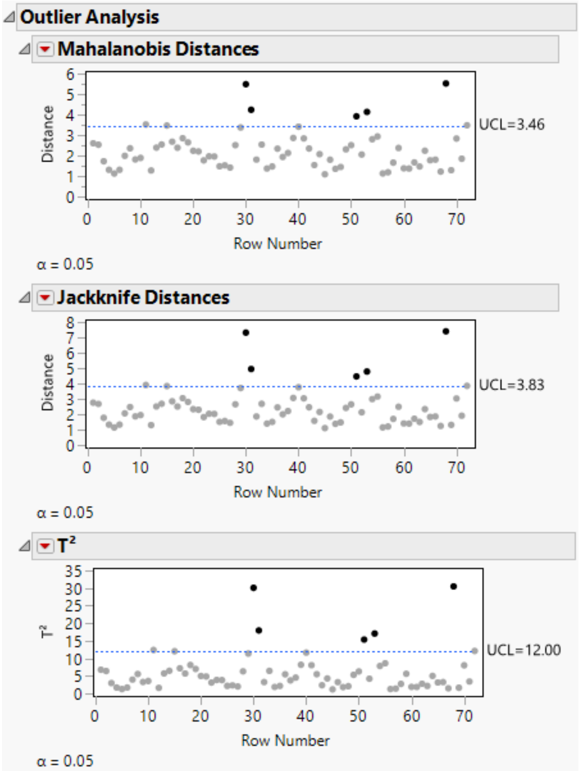
Each option in the Outlier Analysis menu shows or hides a plot of the corresponding distances. Testing is done at the alpha level that appears at the bottom of the plot. The following options are available:

Mahalanobis Distances Shows or hides the Mahalanobis distance of each point from the multivariate mean (centroid). The standard Mahalanobis distance depends on estimates of the mean, standard deviation, and correlation for the data. The distance is plotted for each observation number. Extreme multivariate outliers can be identified by highlighting the points with the largest distance values. See "Mahalanobis Distance Measures" on page 50.

Jackknife Distances Shows or hides distances that are calculated using a jackknife technique. The distance for each observation is calculated with estimates of the mean, standard deviation, and correlation matrix that do not include the observation itself. The jack-knifed distances are useful when there is an outlier. In this case, the Mahalanobis distance is distorted and tends to disguise the outlier or make other points look more outlying than they are. See "Jackknife Distance Measures" on page 51.

T^2 Shows or hides distances that are the square of the Mahalanobis distance. This plot is preferred for multivariate control charts. The plot includes the value of the calculated T^2 statistic, as well as its upper control limit. Values that fall outside this limit might be outliers. See "T^2 Distance Measures" on page 51.

Figure 3.6 Outlier Analysis Plots



Saving Distances and Values

You can save any of the distances to the data table by selecting the **Save** option from the red triangle menu for the plot.

Note: There is no formula saved with the jackknife distance column. This means that the distance is *not* recomputed if you modify the data table. If you add or delete columns, or change values in the data table, select **Analyze > Multivariate Methods > Multivariate** again to compute new jackknife distances.

In addition to saving the distance values for each row, a column property is created that holds the upper control limit (UCL) value for the Outlier Analysis type specified.

Item Reliability

Item reliability indicates how consistently a set of instruments measures an overall response. Cronbach's α (Cronbach 1951) is one measure of reliability. Two primary applications for Cronbach's α are industrial instrument reliability and questionnaire analysis.

Cronbach's α is based on the average correlation of items in a measurement scale. It is equivalent to computing the average of all split-half correlations in the data table. The Standardized α can be requested if the items have variances that vary widely.

Note: Cronbach's α is not related to a significance level α. Also, item reliability is unrelated to survival time reliability analysis.

To look at the influence of an individual item, JMP excludes it from the computations and shows the effect of the Cronbach's α value. If α increases when you exclude a variable (item), that variable is not highly correlated with the other variables. If the α decreases, you can conclude that the variable is correlated with the other items in the scale.

See "Cronbach's α on page 52 for more information about computations.

Impute Missing Data

To impute missing data, click the Multivariate red triangle and select **Impute Missing Data**. A new data table is created. This data table is a duplicate of your data table except that the missing values are replaced with estimated values.

Imputed values are expectations conditional on the nonmissing values for each row. The mean and covariance matrix, estimated by the method chosen in the launch window, are used for the imputation calculation. All multivariate tests and options are then available for the imputed data set.

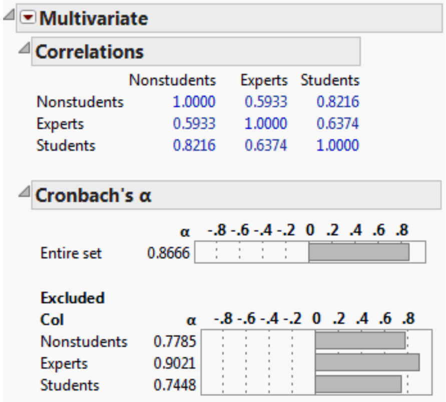
Example of Item Reliability

This example uses the Danger.jmp data in the sample data folder. This table lists 30 items having some level of inherent danger. Three groups of people (students, non-students, and experts) ranked the items according to perceived level of danger. Note that Nuclear power is rated as very dangerous (1) by both students and non-students, but is ranked low (20) by experts. On the other hand, motorcycles are ranked either fifth or sixth by all three judging groups.

You can use Cronbach's α to evaluate the agreement in the perceived way the groups ranked the items. Note that in this type of example, where the values are the same set of ranks for each group, standardizing the data has no effect.

1. Select **Help > Sample Data Library** and open Danger.jmp.
2. Select **Analyze > Multivariate Methods > Multivariate**.
3. Select all the columns except for Activity and click **Y, Columns**.
4. Click **OK**.
5. Click the Multivariate red triangle and select **Item Reliability > Cronbach's** α.
6. (Optional) Click the Multivariate red triangle and select **Scatterplot Matrix** to hide that plot.

Figure 3.7 Cronbach's α Report



Multivariate

Correlations

	Nonstudents	Experts	Students
Nonstudents	1.0000	0.5933	0.8216
Experts	0.5933	1.0000	0.6374
Students	0.8216	0.6374	1.0000

Cronbach's α

	α
Entire set	0.8666

Excluded Col	α
Nonstudents	0.7785
Experts	0.9021
Students	0.7448

The Cronbach's α results show an overall α of 0.8666, which indicates a high correlation of the ranked values among the three groups. Further, when you remove the experts from the analysis, the Nonstudents and Students ranked the dangers nearly the same, with Cronbach's α scores of 0.7785 and 0.7448, respectively.

Computations and Statistical Details

Estimation Methods

REML

REML (restricted maximum likelihood) estimates are less biased than the ML (maximum likelihood) estimation method when the data contains missing values. The REML method maximizes marginal likelihoods based on error contrasts. The REML method is often used for estimating variances and covariances. The REML method in the Principal Components platform is the same as the REML estimation of mixed models for repeated measures data with an unstructured covariance matrix. See the MIXED Procedure chapter in SAS Institute Inc. (2020e) about REML estimation of mixed models.

Robust

This method essentially ignores any outlying values by substantially down-weighting them. A sequence of iteratively reweighted fits of the data is done using the following weight:

$w_i = 1.0$ if $Q < K$ and $w_i = K/Q$ otherwise

Here, K is a constant equal to the 0.75 quantile of a chi-square distribution with the degrees of freedom equal to the number of columns in the data table. Q is defined as follows:

$$Q = (y_i - \mu)^T (S^2)^{-1} (y_i - \mu)$$

In this equation, y_i = the response for the i^{th} observation, μ = the current estimate of the mean vector, S^2 = current estimate of the covariance matrix, and T = the transpose matrix operation. The final step is a bias reduction of the variance matrix.

The trade off of this method is that you can have higher variance estimates when the data do not have many outliers, but can have a much more precise estimate of the variances when the data do have outliers.

Pearson Product-Moment Correlation

The Pearson product-moment correlation coefficient measures the strength of the linear relationship between two variables. For response variables *X* and *Y*, it is denoted as *r* and computed as follows:

$$r = \frac{\sum(x - x)(y - y)}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(y - \bar{y})^2}}$$

If there is an exact linear relationship between two variables, the correlation is 1 or –1, depending on whether the variables are positively or negatively related. If there is no linear relationship, the correlation tends toward zero.

Nonparametric Measures of Association

For the Spearman, Kendall, or Hoeffding correlations, the data are first ranked. Computations are then performed on the ranks of the data values. Average ranks are used in case of ties.

Note: When a Weight variable is specified, missing and zero-valued weights are excluded from the nonparametric correlation calculations. All other weight values are treated as 1.

Spearman’s ρ (rho) Coefficients

Spearman’s ρ correlation coefficient is computed on the ranks of the data using the formula for the Pearson’s correlation previously described.

Kendall’s τ_b Coefficients

Kendall’s τ_b coefficients are based on the number of concordant and discordant pairs. A pair of rows for two variables is *concordant* if they agree in which variable is greater. Otherwise, they are discordant, or tied.

The formula

$$\tau_b = \frac{\sum_{i<j} \operatorname{sgn}(x_i - x_j)\operatorname{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

computes Kendall’s τ_b where:

$$T_0 = (n(n-1))/2$$

$$T_1 = \sum((t_i)(t_i-1))/2$$

$$T_2 = \sum((u_i)(u_i-1))/2$$

Note the following:

- The sgn(z) is equal to 1 if z>0, 0 if z=0, and –1 if z<0.
- The t_i (the u_i) are the number of tied x (respectively y) values in the ith group of tied x (respectively y) values.
- The n is the number of observations.
- Kendall’s τ_b ranges from –1 to 1. If a weight variable is specified, it is ignored.

Computations proceed in the following way:

- Observations are ranked in order according to the value of the first variable.
- The observations are then re-ranked according to the values of the second variable.
- The number of interchanges of the first variable is used to compute Kendall’s τ_b.

Hoeffding’s D Statistic

The formula for Hoeffding’s *D* (1948) is

$$D = 30\left(\frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}\right)$$

where:

$$D_1 = \Sigma_i(Q_i-1)(Q_i-2)$$

$$D_2 = \Sigma_i(R_i-1)(R_i-2)(S_i-1)(S_i-2)$$

$$D_3 = \Sigma_i(R_i-2)(S_i-2)(Q_i-1)$$

Note the following:

- The R_i and S_i are ranks of the x and y values.
- The Q_i (sometimes called bivariate ranks) are one plus the number of points that have both x and y values less than the ith points.
- A point that is tied on its x value or y value, but not on both, contributes 1/2 to Q_i if the other value is less than the corresponding value for the ith point. A point tied on both x and y contributes 1/4 to Q_i.

When there are no ties among observations, the *D* statistic has values between –0.5 and 1, where 1 indicates complete dependence. If a weight variable is specified, it is ignored.

Inverse Correlation Matrix

The inverse correlation matrix provides useful multivariate information. The diagonal elements of the inverse correlation matrix, sometimes called the variance inflation factors (VIF), are a function of how closely the variable is a linear function of the other variables. Specifically, if the correlation matrix is denoted **R** and the inverse correlation matrix is denoted $\mathbf{R}^{-1}$, the diagonal element is denoted r^{ii} and is computed as follows:

$$r^{ii} = \mathrm{VIF}_i = \frac{1}{1 - \mathrm{R}_i^2}$$

where R_i^2 is the coefficient of variation from the model regressing the i^{th} explanatory variable on the other explanatory variables. Thus, a large r^{ii} indicates that the ith variable is highly correlated with any number of the other variables.

Distance Measures

The Outlier Analysis plots show the specified distance measure for each point in the data table.

Mahalanobis Distance Measures

The Mahalanobis distance takes into account the correlation structure of the data and the individual scales. For each value, the Mahalanobis distance is denoted M_i and is computed as follows:

$$M_i = \sqrt{(\mathrm{Y}_i - \bar{\mathrm{Y}})'\mathrm{S}^{-1}(\mathrm{Y}_i - \bar{\mathrm{Y}})}$$

where:

Y_i is the data for the i^{th} row

$\bar{Y}$ is the row of means

S is the estimated covariance matrix for the data

The UCL reference line (Mason and Young 2002) drawn on the Mahalanobis Distances plot is computed as follows:

$$UCL_{Mahalanobis} = \sqrt{\frac{(n-1)^2}{n}\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}}$$

where:

n = number of observations

p = number of variables (columns)

$\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}$ = $(1-\alpha)^{\text{th}}$ quantile of a Beta $\left(\frac{p}{2},\frac{n-p-1}{2}\right)$ distribution

If a variable is an exact linear combination of other variables, then the correlation matrix is singular and the row and the column for that variable are zeroed out. The generalized inverse that results is still valid for forming the distances.

Jackknife Distance Measures

The jackknife distance is calculated with estimates of the mean, standard deviation, and correlation matrix that do not include the observation itself. For each value, the jackknife distance is computed as follows:

$$J_i = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{M_i^2}{1 - \frac{nM_i^2}{(n-1)^2}}}$$

where:

n = number of observations

p = number of variables (columns)

M_i = Mahalanobis distance for the i^{th} observation

The UCL reference line (Penny 1996) drawn on the Jackknife Distances plot is calculated as follows:

$$UCL_{Jackknife} = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{UCL_{Mahalanobis}^{\;2}}{1 - \frac{n \bullet UCL_{Mahalanobis}^{\;2}}{(n-1)^2}}}$$

T^2 Distance Measures

The T^2 distance is the square of the Mahalanobis distance, so $T_i^2 = M_i^2$.

The UCL on the T^2 distance is:

$$UCL_{T^2} = \frac{(n-1)^2}{n}\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]} = (UCL_{Mahalanobis})^2$$

where

n = number of observations

p = number of variables (columns)

$\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}$ = $(1-\alpha)^{\text{th}}$ quantile of a Beta $\left(\frac{p}{2},\frac{n-p-1}{2}\right)$ distribution

Multivariate distances are useful for spotting outliers in many dimensions. However, if the variables are highly correlated in a multivariate sense, then a point can be seen as an outlier in multivariate space without looking unusual along any subset of dimensions. In other words, when the values are correlated, it is possible for a point to be unremarkable when seen along one or two axes but still be an outlier by violating the correlation.

Cronbach's α

Cronbach's α is defined as follows:

$$\alpha = \frac{kc}{v+(k-1)c}$$

where

k = the number of items in the scale

c = the average covariance between items

v = the average variance between items

If the items are standardized to have a constant variance, the formula becomes

$$\alpha = \frac{k(r)}{1+(k-1)r}$$

where

r = the average correlation between items

The larger the overall α coefficient, the more confident you can feel that your items contribute to a reliable scale or test. The coefficient can approach 1.0 if you have many highly correlated items.

Chapter 4

Principal Components

Reduce the Dimensionality of Your Data

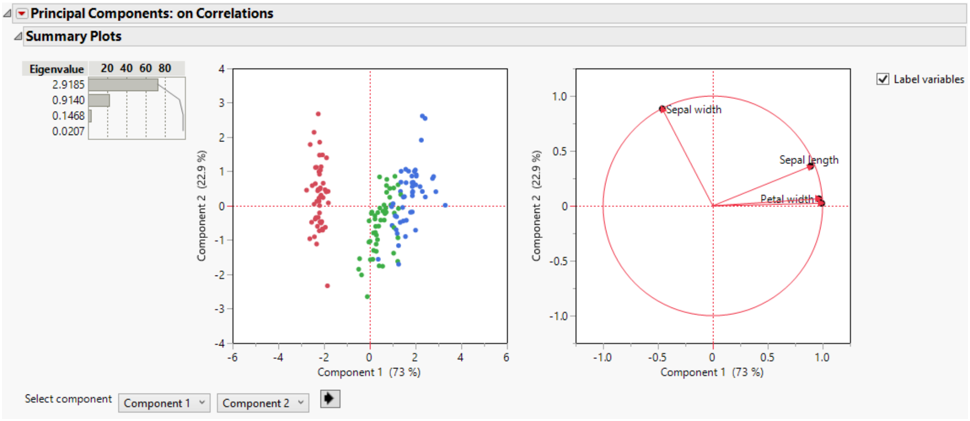
The purpose of principal component analysis is to derive a small number of independent linear combinations (*principal components*) of a set of measured variables that capture as much of the variability in the original variables as possible. Principal component analysis is a dimension-reduction technique, as well as an exploratory data analysis tool. Principal component analysis is also useful for constructing predictive models, as in *principal components analysis regression* (also known as PCA regression or PCR).

For data with a very large number of variables, the Principal Components platform provides an estimation method called the Wide method. The Wide method enables you to calculate principal components in short computing times. These principal components can then be used in PCA regression.

For data that contain mostly zeros, also called *sparse data*, the Principal Components platform provides the Sparse estimation method. Similar to the Wide method, the Sparse method calculates principal components in short computing times. Unlike the Wide method, the Sparse method calculates a fixed, user-defined number of principal components rather than the full set.

The Principal Components platform also supports factor analysis. JMP offers several types of orthogonal and oblique factor analysis-style rotations to help interpret the extracted components. For factor analysis, see the "Factor Analysis" chapter on page 219.

Figure 4.1 Example of Principal Components



Contents

Overview of Principal Component Analysis Platform

A principal component analysis models the variation in a set of variables in terms of a smaller number of independent linear combinations (*principal components*) of those variables.

If you want to see the arrangement of points across many correlated variables, you can use principal component analysis to show the most prominent directions of the high-dimensional data. Using principal component analysis reduces the dimensionality of a set of data. Principal components is a way to picture the structure of the data as completely as possible by using as few variables as possible.

For p variables, this is how the p principal components are formed:

- The first principal component is the linear combination of the standardized original variables that has the greatest possible variance.
- Each subsequent principal component is the linear combination of the variables that has the greatest possible variance and is uncorrelated with all previously defined components.

Each principal component is calculated by taking a linear combination of an eigenvector of the correlation matrix (or covariance matrix or sum of squares and cross products matrix) with the variables. The eigenvalues represent the variance of each component.

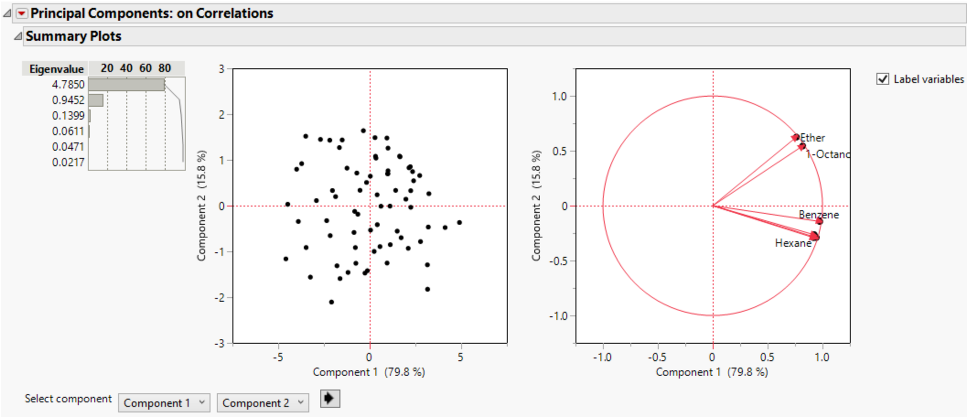
The Principal Components platform enables you to conduct your analysis on the correlation matrix, the covariance matrix, or the unscaled data. You can also conduct Factor Analysis within the Principal Components platform. See the “Factor Analysis” chapter on page 219.

Example of Principal Component Analysis

To view an example Principal Component Analysis report for a data table for two factors:

1. Select **Help > Sample Data Library** and open Solubility.jmp.
2. Select **Analyze > Multivariate Methods > Principal Components**.
3. Select all of the continuous columns and click **Y, Columns**.
4. Keep the default Estimation Method and then click **OK**.

Figure 4.2 Principal Components on Correlations Report



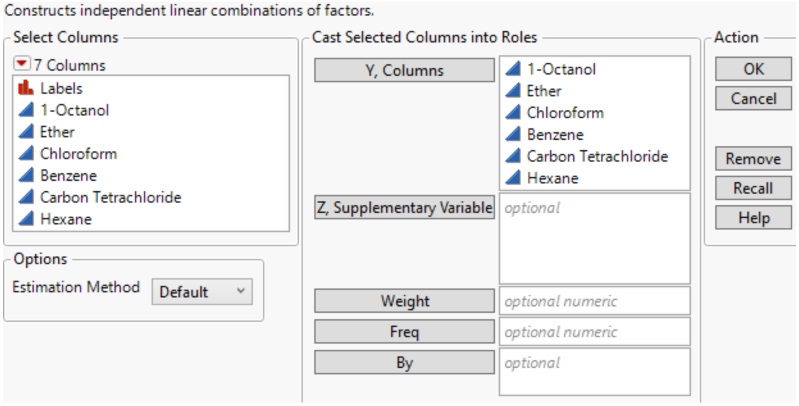
The report gives the eigenvalues and a bar chart of the percent of the variation accounted for by each principal component. In this example, the first principal component accounts for almost 80% of the variation in the data. Together, the first two principal components account for almost all of the variation in the data (95.5%). There is a Score Plot and a Loadings Plot as well. See "Principal Components Report" on page 60.

Launch the Principal Components Platform

Launch the Principal Components platform by selecting **Analyze > Multivariate Methods > Principal Components**. Principal Component analysis is also available using the Multivariate and the Scatterplot 3D platforms.

The example described in "Example of Principal Component Analysis" on page 55 uses all of the continuous variables from the Solubility.jmp sample data table.

Figure 4.3 Principal Components Launch Window



Constructs independent linear combinations of factors.

Y, Columns: 1-Octanol, Ether, Chloroform, Benzene, Carbon Tetrachloride, Hexane

Estimation Method: Default

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns The variables to analyze for components.

Z, Supplementary Variable The supplementary variables to be displayed. Supplementary variables are not included in the calculation of principal components and including them does not affect the results. Supplementary variables that are continuous can be projected on to the loading plot and used to enhance interpretation.

Weight Identifies one column whose numeric values assign a weight to each row in the analysis.

Note: The Weight role is ignored for the Wide and Sparse estimation methods.

Freq Identifies one column whose numeric values assign a frequency to each row in the analysis.

Note: The Freq role is ignored for the Wide and Sparse estimation methods.

By Creates a Principal Component report for each value specified by the By column so that you can perform separate analyses for each group.

Estimation Method Specifies the method for calculating the correlations. Several of these methods address the treatment of missing data.

Default The **Default** option uses either the Row-wise, Pairwise, or REML methods. A JMP Alert also recommends switching to the Wide method when appropriate.

- **Row-wise** estimation is used for data tables with no missing values.
- **Pairwise** estimation is used for data tables with missing values and either more than 10 columns, more than 5,000 rows, or more columns than rows.
- **REML** estimation is used otherwise.
- **Wide** estimation is recommended by a JMP Alert window for data tables with more than 500 columns and more columns than rows. This is because computation time can be considerable when you use the other methods with a large number of columns. Click **Wide** to switch to the Wide method or click **Continue** to use the method you originally selected.

REML Restricted maximum likelihood (REML) estimation uses all of the data, even if missing values are present. Due to a bias-correction factor, this method is slow if the dataset is large and there are many missing values. Therefore, REML is most useful for smaller datasets. If there are no missing cells in the data, then the REML and ML estimates are equivalent and equal to the sample covariance matrix. If there are missing cells, REML's variance and covariance estimates are less biased than the estimates from ML estimation. For statistical details, see "REML" on page 73.

ML Maximum likelihood (ML) estimation uses all of the data, even if missing values are present. Because the estimates from ML are generated quickly, this method is most useful for large data tables with missing data.

Robust Robust estimation uses all of the data, even if missing values are present. This method down-weights extreme values and is therefore useful for data tables that might have outliers. For statistical details, see "Robust" on page 47 in the "Correlations and Multivariate Techniques" chapter.

Row-wise Row-wise estimation calculates the Pearson correlation for each pair of columns. For statistical details, see "Pearson Product-Moment Correlation" on page 48 in the "Correlations and Multivariate Techniques" chapter. Row-wise estimation does not use observations with missing values. This method is useful for excluding any observations that have missing data.

Pair-wise Pair-wise estimation uses all of the data, even if missing values are present. This estimation method calculates Pearson correlations for each pair of columns using all observations with nonmissing values for those two columns. For statistical details, see “Pearson Product-Moment Correlation” on page 48 in the “Correlations and Multivariate Techniques” chapter. Pair-wise estimation is most useful when a data table has missing values and either more columns than rows, more than 10 columns, or more than 5,000 rows.

Wide Wide estimation does not use observations with missing values, so rows that contain missing cells are deleted before the method is applied. This estimation method uses an algorithm based on the full singular value decomposition. The algorithm avoids calculating the covariance matrix and is therefore computationally efficient. It is useful when you have a very large number of columns in your data. For statistical details, see “Wide” on page 73.

JMP PRO **Sparse** Sparse estimation uses all of the data, even if missing values are present. This estimation method uses an algorithm based on the partial singular value decomposition, which computes only the first specified number of singular values and singular value vectors. The algorithm avoids calculating the covariance matrix, as well as unnecessary principal components and is therefore computationally efficient. It is useful when your data are sparse, meaning they contain many zeros, or when there are a large number of columns in the data. For statistical details, see “Sparse” on page 74.

Note: If you select REML, ML, or Robust and your data table contains more columns than rows and has missing values, JMP switches the Estimation Method to Pairwise.

Number of Components (Available only when Sparse is specified as the Estimation Method.) Specifies the number of components to be estimated. Typically, the Number of Components is much smaller than the dimension of your data.

Missing Data

The different estimation methods are equipped to handle missing data in a variety of ways. You can also estimate missing values in the following ways:

- Use the Impute Missing Data option found under **Multivariate Methods > Multivariate**. See “Impute Missing Data” on page 45 in the “Correlations and Multivariate Techniques” chapter.
- Use the Multivariate Normal Imputation or Multivariate SVD Imputation utilities found in **Analyze > Screening > Explore Missing Values**. See *Predictive and Specialized Modeling*.

Principal Components Report

If you selected any estimation method other than Wide or Sparse, the Principal Components: on Correlations report initially appears. (The title of this report changes if you select on Covariances or on Unscaled for the Principal Components option in the Principal Components red triangle menu.)

If you select the Wide method, the Wide Principal Components report appears. If you select the Sparse method, the Sparse Principal Components report appears.

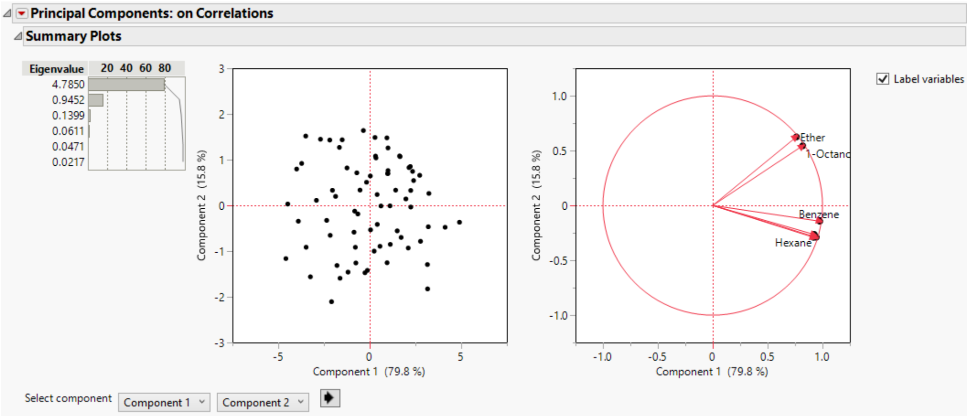
The initial Principal Components report is for an analysis on Correlations. It summarizes the variation of the specified *Y* variables with principal components (Figure 4.4). You can switch to an analysis based on the covariance matrix or unscaled data by selecting the Principal Components option from the red triangle menu.

Based on your selection, the principal components are derived from an eigenvalue decomposition of one of the following:

- the correlation matrix
- the covariance matrix
- the sum of squares and cross products matrix for the unscaled and uncentered data

The details in the report show how the principal components absorb the variation in the data. The principal component points are derived from the eigenvector linear combination of the variables.

Figure 4.4 Principal Components on Correlations Report



The report gives the eigenvalues and a bar chart of the percent of the variation accounted for by each principal component. There is a Score Plot and a Loadings Plot as well. The eigenvalues indicate the total number of components extracted based on the amount of variance contributed by each component.

The Score Plot graphs each component's calculated values in relation to the other, adjusting each value for the mean and standard deviation.

The Loadings Plot graphs the unrotated loading matrix between the variables and the components. The closer the value is to 1, the greater the effect of the component on the variable.

By default, the report shows the Score Plot and the Loadings Plot for the first two principal components. Use the list next to Select component to specify the principal components that are graphed on the Score Plot and the Loadings Plot.

Principal Components Report Options

The Principal Components red triangle menu contains the following options:

Note: Some of the options are not available for the Wide or Sparse estimation methods.

Principal Components (Not available for the Wide or Sparse estimation methods.) Enables you to create the principal components based on **Correlations**, **Covariances**, or **Unscaled**.

Correlations (Not available for the Wide or Sparse estimation methods.) The matrix of correlations between the variables.

Note: The values on the diagonals are 1.0.

Covariance Matrix (Not available for the Wide or Sparse estimation methods.) Shows or hides the covariances of the variables.

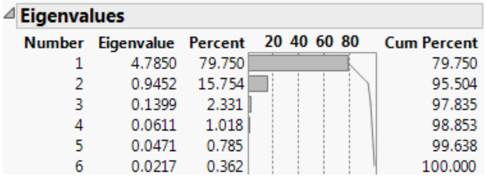
Eigenvalues Lists the eigenvalue that corresponds to each principal component in order from largest to smallest. The eigenvalues represent a partition of the total variation in the multivariate sample.

The scaling of the eigenvalues depends on which matrix you select for extraction of principal components:

- For the on Correlations option, the eigenvalues are scaled to sum to the number of variables.
- For the on Covariances options, the eigenvalues are not scaled.
- For the on Unscaled option, the eigenvalues are divided by the total number of observations.

If you select the **Bartlett Test** option from the red triangle menu, hypothesis tests (Figure 4.6) are given for each eigenvalue (Jackson 2003).

Figure 4.5 Eigenvalues



Eigenvalues

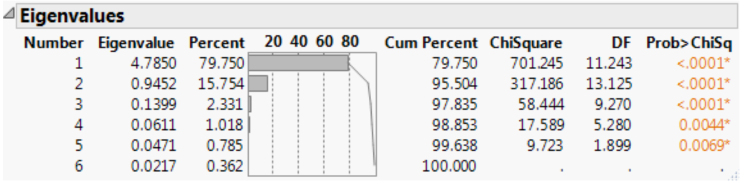
Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.7850	79.750		79.750
2	0.9452	15.754		95.504
3	0.1399	2.331		97.835
4	0.0611	1.018		98.853
5	0.0471	0.785		99.638
6	0.0217	0.362		100.000

Eigenvectors Shows or hides a table of the eigenvectors for each of the principal components, in order, from left to right. Using these coefficients to form a linear combination of the original variables produces the principal component variables. Following the standard convention, eigenvectors have norm 1.

Note: The number of eigenvectors shown is equal to the rank of the correlation matrix, or, if the Sparse method is selected, the number of components specified on the launch window.

Bartlett Test (Not available for the Wide or Sparse estimation methods.) Shows or hides the results of the homogeneity test (appended to the Eigenvalues table). The test determines whether the eigenvalues have the same variance by calculating the Chi-square, degrees of freedom (DF), and the *p*-value (prob > ChiSq) for the test. See Bartlett (1937, 1954).

Figure 4.6 Bartlett Test



Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent	ChiSquare	DF	Prob>ChiSq
1	4.7850	79.750		79.750	701.245	11.243	<.0001*
2	0.9452	15.754		95.504	317.186	13.125	<.0001*
3	0.1399	2.331		97.835	58.444	9.270	<.0001*
4	0.0611	1.018		98.853	17.589	5.280	0.0044*
5	0.0471	0.785		99.638	9.723	1.899	0.0069*
6	0.0217	0.362		100.000	.	.	.

Loading Matrix Shows or hides a table of the loadings for each component. These values are graphed in the loading plot. The degree of transparency for the table values indicates the distance of the absolute loading value from zero. Absolute loading values that are closer to zero are more transparent than absolute loading values that are farther from zero.

If you specify a supplementary variable, an additional table of coordinates is shown for each supplementary continuous variable and each level of supplementary categorical variable. These values are graphed in the loading plot for continuous supplementary variables.

The scaling of the loadings and coordinates depends on which matrix you select for extraction of principal components:

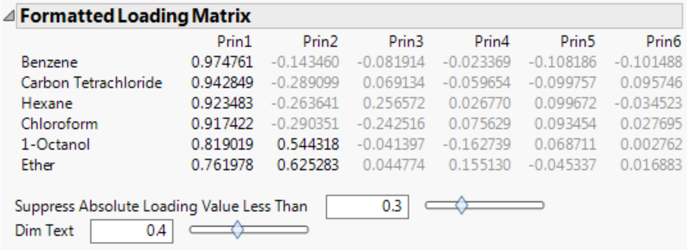
- For the on Correlations option, the i^{th} column of loadings is the i^{th} eigenvector multiplied by the square root of the i^{th} eigenvalue. The i,j^{th} loading is the correlation between the i^{th} variable and the j^{th} principal component.

- For the on Covariances option, the j^{th} entry in the i^{th} column of loadings is the i^{th} eigenvector multiplied by the square root of the i^{th} eigenvalue and divided by the standard deviation of the j^{th} variable. The i,j^{th} loading is the correlation between the i^{th} variable and the j^{th} principal component.
- For the on Unscaled option, the j^{th} entry in the i^{th} column of loadings is the i^{th} eigenvector multiplied by the square root of the i^{th} eigenvalue and divided by the standard error of the j^{th} variable. The standard error of the j^{th} variable is the j^{th} diagonal entry of the sum of squares and cross products matrix divided by the number of rows ($\mathbf{X'X}$/n).

Note: When you are analyzing the unscaled data, the i,j^{th} loading is *not* the correlation between the i^{th} variable and the j^{th} principal component.

Formatted Loading Matrix Shows or hides a table of the loadings for each component. The table is sorted in order of decreasing loadings on the first principal component. Therefore, the variables are listed in the order of decreasing loadings on the first component.

Figure 4.7 Formatted Loading Matrix



Formatted Loading Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
Benzene	0.974761	-0.143460	-0.081914	-0.023369	-0.108186	-0.101488
Carbon Tetrachloride	0.942849	-0.289099	0.069134	-0.059654	-0.099757	0.095746
Hexane	0.923483	-0.263641	0.256572	0.026770	0.099672	-0.034523
Chloroform	0.917422	-0.290351	-0.242516	0.075629	0.093454	0.027695
1-Octanol	0.819019	0.544318	-0.041397	-0.162739	0.068711	0.002762
Ether	0.761978	0.625283	0.044774	0.155130	-0.045337	0.016883

Suppress Absolute Loading Value Less Than 0.3
Dim Text 0.4

Suppress Absolute Loading Value Less Than The value that determines which loadings are unavailable in the Formatted Loading Matrix report. You can use the text box or the slider to dim the loadings whose absolute values fall below the selected value.

Dim Text The transparency of the dimmed values in the Formatted Loading Matrix report. You can use the text box or the slider to set the degree of transparency for the dimmed loadings. The degree of transparency ranges from 0 to 1, where lower values are more transparent than higher values. For example, setting the transparency to 0 completely removes the unavailable loadings from the matrix, while the loadings are still available when you set the transparency to 1.

Squared Cosines of Variables Shows or hides a table that contains the squared cosines of variables. If you specify a supplementary variable, an additional table of squared cosines of supplementary variables is shown. The sum of the squared cosine values across principal components is equal to one for each variable. The squared cosines enable you to see how well the variables are represented by the principal components. You can also determine how many principal components are necessary to represent certain variables.

This option also shows a plot of the squared cosines for the first three principal components.

Note: If the Sparse estimation method is used and the number of components selected is less than three, only the specified number of components are displayed in the plot.

Partial Contribution of Variables Shows or hides a table that contains the partial contributions of variables. The partial contributions enable you to see the percentage that each variable contributes to each principal component. This option also shows a plot of the partial contributions for the first three principal components.

Note: If the Sparse estimation method is used and the number of components selected is less than three, only the specified number of components are displayed in the plot.

Summary Plots Shows or hides the summary information produced in the default report. This summary information includes a plot of the eigenvalues, a score plot, and a loading plot. By default, the report shows the score and loading plots for the first two principal components. There are options in the report to specify which principal components to plot. See "Principal Components Report" on page 60.

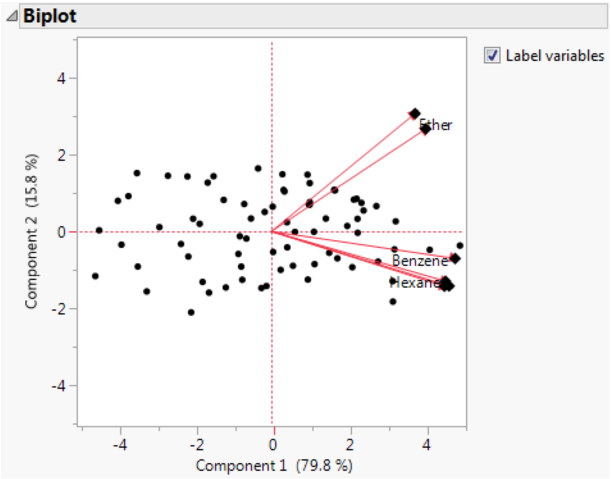
Note: If your data contain missing values, the imputed scores are plotted on the score plot.

Tip: Select the tips of arrows in the loading plot to select the corresponding columns in the data table. Press Ctrl and click an arrow tip to deselect the column.

Biplot Shows or hides a plot that overlays the score plot and the loading plot for the specified number of components.

Note: If your data contain missing values, the imputed scores are plotted on the Biplot.

Figure 4.8 Biplot

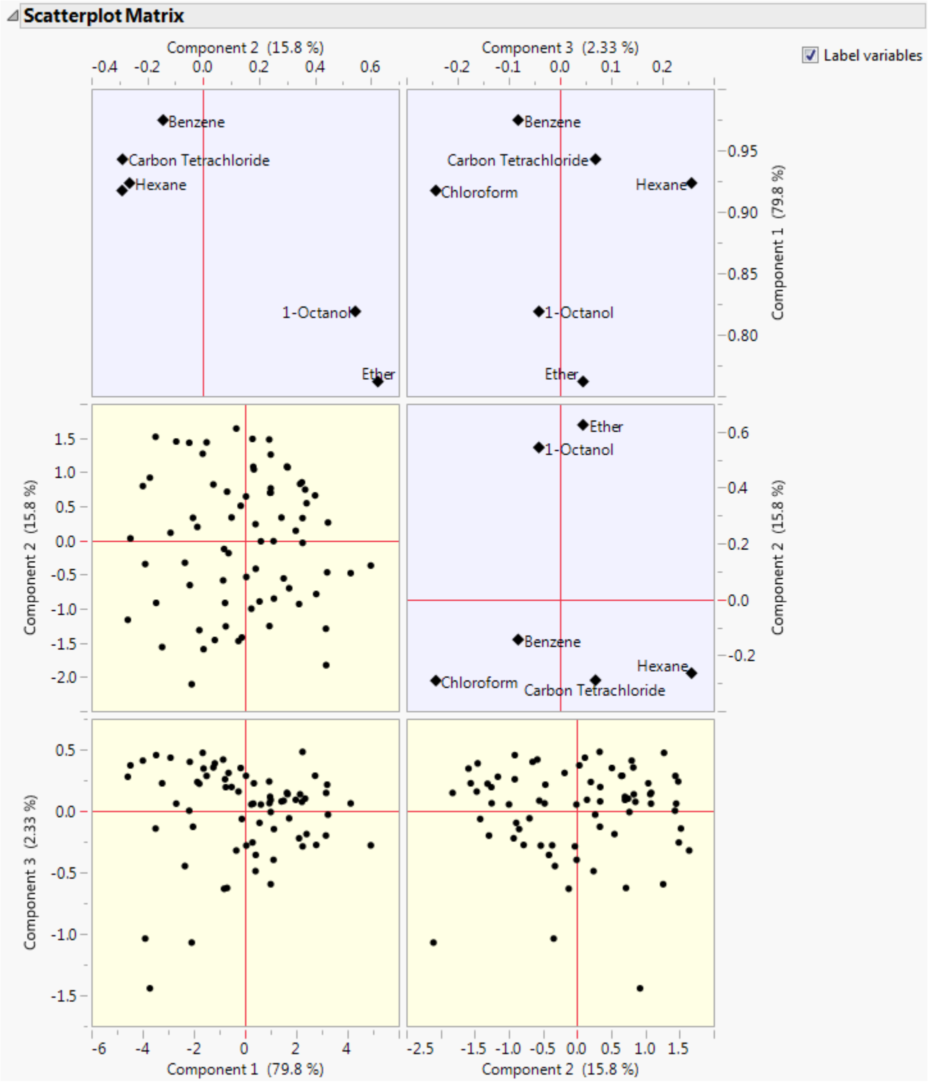


Note: The score plot markers are dots and the loading plot markers are diamonds.

Scatterplot Matrix Shows or hides a matrix of score and loading plots for a specified number of principal components. The scatterplot matrix arranges both the score plots and the loading plots in one space. The score plots have a yellow shaded background. The loading plots have a blue shaded background.

Note: If your data contain missing values, the imputed scores are plotted in the Scatterplot Matrix.

Figure 4.9 Scatterplot Matrix



Note: The loading plot matrix displayed in the Scatterplot Matrix is the transpose of the loading plot matrix that you obtain when you select the Loading Plot option.

Scree Plot Shows or hides a graph of the eigenvalue for each component. This scree plot helps in visualizing the dimensionality of the data space.

Score Plot Shows or hides a matrix of scatterplots of the scores for pairs of principal components for the specified number of components. This plot is shown in Figure 4.4 (left-most plot).

Loading Plot Shows or hides a matrix of two-dimensional representations of factor loadings for the specified number of components. The loading plot labels variables if the number of variables is 30 or fewer. If there are more than 30 variables, the labels are off by default. This information is shown in Figure 4.4 (right-most plot).

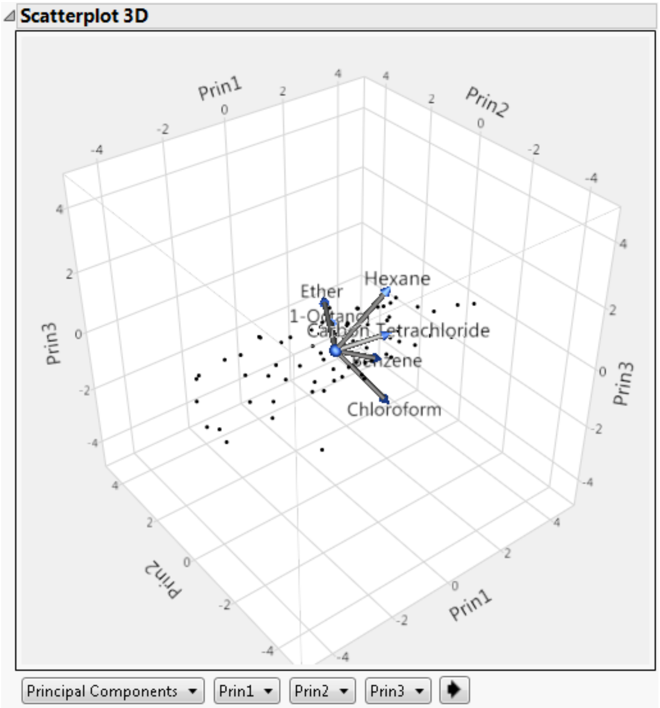
Tip: Select the tips of arrows in the loading plot to select the corresponding columns in the data table. Press Ctrl and click an arrow tip to deselect the column.

Score Plot with Imputation (Not available for the Wide or Sparse estimation methods.) Imputes any missing values and creates a score plot. This option is available only if there are missing values.

3D Score Plot (Not available for the Wide or Sparse estimation methods.) Shows or hides a 3D scatterplot of any three principal component scores. When you first invoke the command, the first three principal components are presented.

Note: If your data contain missing values, the imputed scores are plotted on the 3D Score Plot.

Figure 4.10 Scatterplot 3D Score Plot



Plot Source The source of the data points in the plot. The available options are Principal Components, Rotated Principal Components, and Data Columns.

Axis Controls The contents of each axis. If the Principal Components option or the Rotated Components option is selected, the options for the Axis Controls are principal components. If the Data Columns option is selected, the options are variables from the analysis.

Cycle Button ◆ Cycles through all axis content possibilities.

The variables show as rays in the plot. These rays, called *biplot rays*, approximate the variables as a function of the principal components on the axes. If there are only two or three variables, the rays represent the variables exactly. The ray corresponds to the principal component loadings.

Score Ellipses Shows or hides ellipses on the summary score plot for each pair of principal components. The ellipses are constructed as either confidence ellipses based on the alpha level or a control limit ellipses based on how far the observations are from the center. By default, the ellipses are 95% confidence ellipses.

Score Ellipse Coverage Displays a submenu that enables you to change how the score ellipses are constructed. Specify the score ellipses by confidence level or the distance from the center in terms of *k*-sigma. The relationship between the confidence level, p, and k-sigma is $p = 1 - \exp(-k^2/2)$.

Display Options

Arrow Lines Enables you to show or hide arrows on all plots that can display arrows. Arrows are shown if the number of variables is 1000 or fewer. If there are more than 1000 variables, the arrows are off by default.

Show Supplementary Variable (Available only if you specify a supplementary variable.) Shows or hides the arrow lines for continuous supplementary variables or label markers for categorical supplementary variables in the biplot, score plot, and loading plots.

Outlier Analysis Shows or hides the Outlier Analysis report, which enables you to detect outliers in the data through T^2 and contribution statistics. See "Outlier Analysis" on page 71.

Factor Analysis (Not available for the Wide or Sparse estimation methods.) Performs factor analysis-style rotations of the principal components, or factor analysis. See the "Factor Analysis" chapter on page 219.

JMP PRO **Cluster Variables** (Not available for the Wide or Sparse estimation methods.) Performs a cluster analysis on the variables by dividing the variables into non-overlapping clusters. Variable clustering provides a method for grouping similar variables into representative groups. Each cluster can then be represented by a single component or variable. The component is a linear combination of all variables in the cluster. Alternatively, the cluster can be represented by the variable identified to be the most representative member in the cluster. See the "Cluster Variables" chapter on page 341.

Note: Cluster Variables uses correlation matrices for all calculations, even when you select the on Covariance or on Unscaled options.

Save Principal Components Saves the number of principal components that you specify to the data table with a formula for computing each component. The formula cannot evaluate rows with missing values.

The calculation for the principal components depends on which matrix you select for extraction of principal components:

- For the on Correlations option, the i^{th} principal component is a linear combination of the centered and scaled observations using the entries of the i^{th} eigenvector as coefficients.
- For the on Covariances options, the i^{th} principal component is a linear combination of the centered observations using the entries of the i^{th} eigenvector as coefficients.
- For the on Unscaled option, the i^{th} principal component is a linear combination of the raw observations using the entries of the i^{th} eigenvector as coefficients.

Note: If the specified number of components exceeds the rank of the correlation matrix, then the number of components saved is set to the rank of the correlation matrix.

Save Predicteds Saves the predicted variables, given the specified number of principal components, to new columns in the data table.

Save DModX Saves the observation distance to the principal components model (DModX), given the specified number of principal components, to a new column in the data table. Larger DModX values indicate mild to moderate outliers in the data. See "DModX Calculation" on page 74.

Save Individual Squared Cosines Saves the individual squared cosines, given the specified number of principal components, to new columns in the data table.

Save Individual Partial Contributions Saves the individual partial contributions, given the specified number of principal components, to new columns in the data table.

Save Rotated Components (Not available for the Wide or Sparse estimation methods.) Saves the rotated components to the data table, with a formula for computing the components. This option is available only after the Factor Analysis option is used. The formula cannot evaluate rows with missing values.

Save Principal Components with Imputation (Not available for the Wide or Sparse estimation methods.) Imputes missing values, and saves the principal components to the data table. The column contains a formula for doing the imputation and computing the principal components. This option is available only if there are missing values.

Save Rotated Components with Imputation (Not available for the Wide or Sparse estimation methods.) Imputes missing values and saves the rotated components to the data table. The column contains a formula for doing the imputation and computing the rotated components. This option is available only after the Factor Analysis option is used and if there are missing values.

Save Imputed Formulas (Available only if the data table contains missing values.) For columns that contain missing values, saves new columns to the data table that contain the formulas used to estimate the missing values. The new columns are called Imputed_<Column Name>.

JMP PRO **Publish Components Formulas** Creates a specified number of principal component formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See *Predictive and Specialized Modeling*.

JMP PRO **Publish DModX Formula** Saves the DModX formula based on a specified number of principal components as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See *Predictive and Specialized Modeling*.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

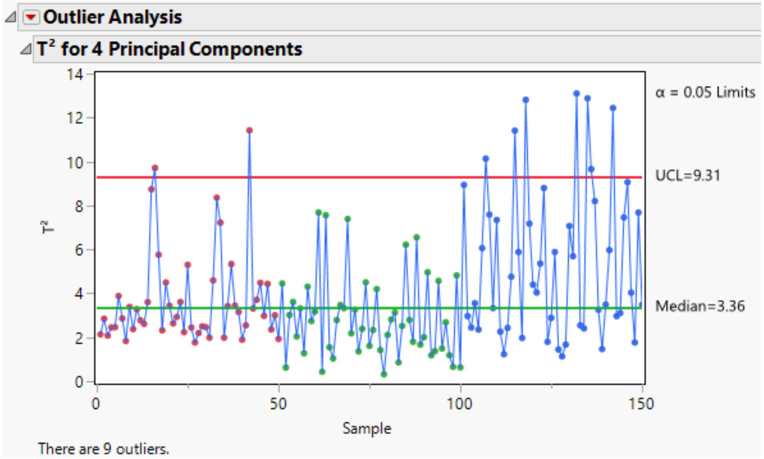
Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Outlier Analysis

Figure 4.11 Outlier Analysis Report



There are 9 outliers.

By default, the Outlier Analysis report displays the T^2 for <*A*> Principal Components plot. The plot shows the T^2 value for each observation and horizontal lines at the Median and Upper Control Limit (UCL). For more information about how the T^2 values, median, and UCL are calculated, see "Calculations for Outlier Analysis" on page 75.

The α level used to calculate the UCL is displayed next to the plot. The number of outliers detected is shown below the plot. This number is the number of observations with T^2 values that are greater than the UCL.

Tip: Hover over a point in the T^2 Plot to view the T^2 Contribution Proportion Plot for that observation. Click the T^2 Contribution Proportion Plot to add it to the report window.

Outlier Analysis Report Options

T^2 Plot Shows or hides the T^2 plot. On by default.

Contribution Heat Map Shows or hides a heat map of the T^2 contribution values for all observations.

Contribution Proportions Heat Map Shows or hides a heat map of the T^2 contribution values for all observations expressed as a proportion of the individual row's T^2. The proportions for an individual row are obtained by computing the square of the contribution and dividing by the sum of the squares of all contributions for that individual row.

Contribution Plots for Selected Samples (Available only if one or more points is selected in the T^2 plot.) Shows a report with a T^2 contribution plot for each selected sample. A T^2

contribution plot shows the contribution of each variable to the sample's T^2 statistic. For more information about how the contributions are calculated, see "Calculations for Outlier Analysis" on page 75. Use contribution plots to investigate outliers. The variables with the largest positive or negative contributions are those that contribute most to a sample having a large T^2 value. See "Contribution Plots Report Options" on page 72 for information on red triangle menu options.

Contribution Proportion Plots for Selected Samples (Available only if one or more points is selected in the T^2 plot.) Shows a report with a T^2 contribution proportion plot for each selected sample. A T^2 contribution proportion plot shows the contribution values for the selected observations expressed as a proportion of the individual row's T^2. This is a different presentation of the information found in the Contribution Proportions Heat Map. See "Contribution Plots Report Options" on page 72 for information on red triangle menu options.

Normalized DModX Plot (Available only when the number of components is less than the number of variables.) Shows or hides a plot of the Normalized DModX values. DModX values are useful for detecting moderate outliers in the data.

Number of Components Enables you to specify the number of principal components used in the T^2 and T^2 contribution statistics. When you change the number of components, the T^2 plot, heap map, and normalized DModX plot automatically update.

Set α level Enables you to specify the α level.

Save T^2 Saves the T^2 values to a new column in the data table.

Save Contributions Saves the T^2 Contributions to new columns in the data table. There is one column for each Y variable.

Save Normalized DModX (Available only when the number of components is less than the number of variables.) Saves the normalized DModX values to a new column in the data table.

Contribution Plots Report Options

The T^2 Contribution Plots for Selected Samples, T^2 Contribution Proportion Plots for Selected Samples, and T^2 Mean Contribution Proportion Plots for Selected Samples contain the following red triangle menu options.

Bar Label Displays a submenu of options to label the bars in the contribution plot. The label options are No Label, Label by Value, and Label by Column.

Remove Plot Removes the contribution plot from the report.

Control Charts for Selected Items (Available only if one or more bar segments are selected in the contribution plot.) Opens a Control Chart Builder window, with control chart results for each of the selected processes and groups.

Tip: You can also view the control chart by hovering over a bar segment in a contribution plot. Click the control chart to open the Control Chart Builder window.

Statistical Details for the Principal Components Analysis Platform

- "Estimation Methods"
- "DModX Calculation"
- "Calculations for Outlier Analysis"

Estimation Methods

REML

REML (restricted maximum likelihood) estimates are less biased than the ML (maximum likelihood) estimation method when the data contains missing values. The REML method maximizes marginal likelihoods based on error contrasts. The REML method is often used for estimating variances and covariances. The REML method in the Principal Components platform is the same as the REML estimation of mixed models for repeated measures data with an unstructured covariance matrix. See the documentation for SAS PROC MIXED about REML estimation of mixed models.

Wide

The Wide method uses a computationally efficient algorithm that avoids calculating the covariance matrix. The algorithm is based on the singular value decomposition. Consider the following notation:

- n = number of rows
- p = number of variables
- $\mathbf{X}$ = n by p matrix of data values

The number of nonzero eigenvalues, and consequently the number of principal components, equals the rank of the correlation matrix of $\mathbf{X}$. The number of nonzero eigenvalues cannot exceed the smaller of n and p.

When you select the Wide method, the data are standardized. To standardize a value, subtract its mean and divide by its standard deviation. Denote the n by p matrix of standardized data values by $\mathbf{X}_s$. Then the covariance matrix of the standardized data is the correlation matrix of $\mathbf{X}$ and it is defined as follows:

$$Cov = \mathbf{X}_s{}'\mathbf{X}_s/(n-1)$$

Using the singular value decomposition, $\mathbf{X}_s$ is written as $\mathbf{U}Diag(\Lambda)\mathbf{V}'$. This representation is used to obtain the eigenvectors and eigenvalues of $\mathbf{X}_s{}'\mathbf{X}_s$. The principal components, or scores, are given by $\mathbf{X}_s\mathbf{V}$. For additional background information, see “Wide Linear Methods and the Singular Value Decomposition” on page 357 in the “Statistical Details” appendix.

JMP PRO

Sparse

Similar to the Wide method, the Sparse method is based on singular value decomposition. Therefore, the algorithm for the Sparse method avoids computing the covariance matrix and is computationally efficient.

Consider the same notation and standardization of $\mathbf{X}$ that is described in “Wide” on page 73. The correlation matrix of $\mathbf{X}$ is represented by the covariance matrix of $\mathbf{X}_s$:

$$Cov(\mathbf{X}_s) = \mathbf{X}_s{}'\mathbf{X}_s/(n-1)$$

The Sparse method differs from the Wide method in the calculation of the singular value decomposition. The Wide method performs a full singular value decomposition. However, the Sparse method uses an algorithm that computes only the first specified number of singular values and singular vectors in the singular value decomposition. Therefore, only the first specified number of eigenvalues and principal components are returned. For more information about the algorithm, see Baglama and Reichel (2005).

DModX Calculation

DModX is the observed distance to the principal components model and is defined as follows:

$$\text{DModX} = \sqrt{\frac{\sum_k e_{ik}^2}{K-A}}$$

where

e_{ik} = the residuals from the model

K = the number of variables

A = the number of principal components

Larger values of DModX indicate mild to moderate outliers in the data.

Calculations for Outlier Analysis

The calculations in the Outlier Analysis report use the following notation:

n = number of observations

A = number of principal components

X_{ci} = the standardized data for the i^{th} observation

T^2 Statistic

The T^2 statistic for the i^{th} observation is calculated as follows:

$$T_i^2 = X_{ci} P_A L^{-1} P_A^{\mathrm{T}} X_{ci}^{\mathrm{T}}$$

where P_A is a matrix containing the first A eigenvectors and L is a diagonal matrix containing the first A eigenvalues.

The median and UCL for the T^2 plot are calculated as follows:

$$CL_{T^2,\,q} = \frac{(n-1)^2}{n}\beta\left[q, \frac{A}{2}, \frac{n-A-1}{2}\right]$$

where

$$\beta\left[q, \frac{A}{2}, \frac{n-A-1}{2}\right] = \text{ the } q^{\text{th}} \text{ quantile of the Beta}\left(\frac{A}{2}, \frac{n-A-1}{2}\right) \text{ distribution.}$$

To calculate the median, use $q = 0.5$. To calculate the UCL, use $q = (1 - \alpha)$.

Contribution Statistic

The vector of T^2 contribution statistics for the ith observation is calculated as follows:

$$con_i = X_{ci} P_A L^{-1/2} P_A^{\mathrm{T}}$$

Note: The sum of the squared contributions for an individual is equal to T_i^2 for that individual.

Chapter 5

Discriminant Analysis

Predict Classifications Based on Continuous Variables

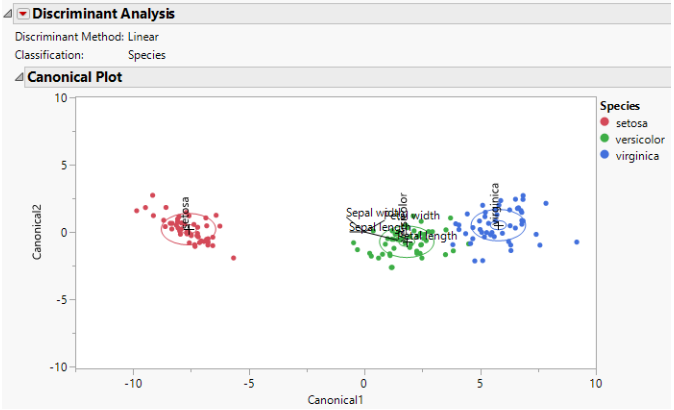
Discriminant analysis predicts membership in a group or category based on observed values of several continuous variables. Specifically, discriminant analysis predicts a classification (*X*) variable (categorical) based on known continuous responses (*Y*). The data for a discriminant analysis consist of a sample of observations with known group membership together with their values on the continuous variables.

For example, you might attempt to classify loan applicants into three loan categories (X) based on expected profitability: low interest rate loan, long term loan, or no loan. You might use continuous variables such as current salary, years in current job, age, and debt burden, (Ys) to predict an individual's most profitable loan category. You could build a predictive model to classify an individual into a loan category using discriminant analysis.

Features of the Discriminant platform include the following:

- A stepwise selection option to help choose variables that discriminate well.
- A choice of fitting methods: Linear, Quadratic, Regularized, and Wide Linear.
- A canonical plot and a misclassification summary.
- Discriminant scores and squared distances to each group.
- Options to save prediction distances and probabilities to the data table.

Figure 5.1 Canonical Plot



Discriminant Analysis

Discriminant Method: Linear

Classification: Species

Canonical Plot

Contents

Overview of the Discriminant Analysis Platform

Discriminant analysis attempts to classify observations described by values on continuous variables into groups. Group membership, defined by a categorical variable *X*, is predicted by the continuous variables. These variables are called *covariates* and are denoted by *Y*.

Discriminant analysis differs from logistic regression. In logistic regression, the classification variable is random and predicted by the continuous variables. In discriminant analysis, the classifications are fixed, and the covariates (*Y*) are realizations of random variables. However, in both techniques, the categorical value is predicted by the continuous variables.

The Discriminant platform provides four methods for fitting models. All methods estimate the distance from each observation to each group's multivariate mean (*centroid*) using Mahalanobis distance. You can specify prior probabilities of group membership and these are accounted for in the distance calculation. Observations are classified into the closest group.

Fitting methods include the following:

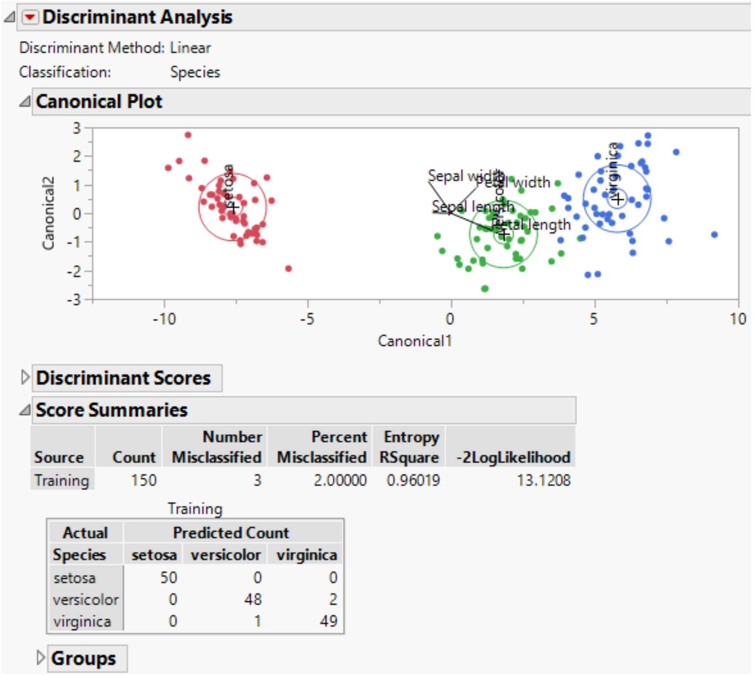
- Linear—Assumes that the within-group covariance matrices are equal. The covariate means for the groups defined by *X* are assumed to differ.
- Quadratic—Assumes that the within-group covariance matrices differ. This requires estimating more parameters than does the Linear method. If group sample sizes are small, you risk obtaining unstable estimates.
- Regularized—Provides two ways to impose stability on estimates when the within-group covariance matrices differ. This is a useful option if group sample sizes are small.
- Wide Linear—Useful in fitting models based on a large number of covariates, where other methods can have computational difficulties. It assumes that all covariance matrices are equal.

Example of Discriminant Analysis

In Fisher's Iris data set, four measurements are taken from a sample of Iris flowers consisting of three different species. The goal is to identify the species accurately using the values of the four measurements.

1. Select **Help > Sample Data Library** and open Iris.jmp.
2. Select **Analyze > Multivariate Methods > Discriminant**.
3. Select Sepal length, Sepal width, Petal length, and Petal width and click **Y, Covariates**.
4. Select Species and click **X, Categories**.
5. Click **OK**.

Figure 5.2 Discriminant Analysis Report Window



Discriminant Method: Linear
Classification: Species

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	150	3	2.00000	0.96019	13.1208

Training

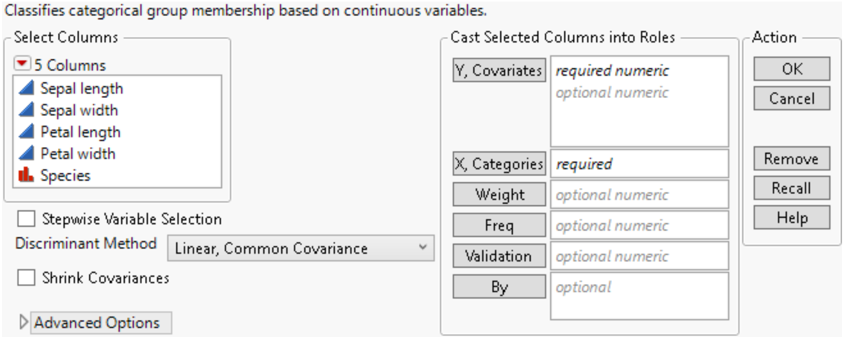
Actual Species	Predicted Count setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

Because there are three classes for Species, there are two canonical variables. In the Canonical Plot, each observation is plotted against the two canonical coordinates. The plot shows that these two coordinates separate the three species. Since there was no validation set, the Score Summaries report shows a panel for the Training set only. When there is no validation set, the entire data set is considered the Training set. Of the 150 observations, only three are misclassified.

Launch the Discriminant Analysis Platform

Launch the Discriminant platform by selecting **Analyze > Multivariate Methods > Discriminant**.

Figure 5.3 Discriminant Launch Window for Iris.jmp



Classifies categorical group membership based on continuous variables.

☐ Stepwise Variable Selection

Discriminant Method: Linear, Common Covariance

☐ Shrink Covariances

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Note: The Validation button appears in JMP Pro only. In JMP, you can define a validation set using excluded rows. See "Validation in JMP and JMP Pro" on page 109.

Y, Covariates Columns that contain the continuous variables used to classify observations into categories.

X, Categories A column that contains the categories or groups into which observations are to be classified.

Weight A column whose values assign a weight to each row for the analysis.

Freq A column whose values assign a frequency to each row for the analysis. In general terms, the effect of a frequency column is to expand the data table, so that any row with integer frequency *k* is expanded to *k* rows. You can specify fractional frequencies.

Validation A numeric column that defines the validation sets. This column should contain at most three distinct values:

- If there are two values, the smaller value defines the training set and the larger value defines the validation set.
- If there are three values, these values define the training, validation, and test sets in order of increasing size.

- – If the validation column has more than three levels, the rows that contain the smallest three values define the validation sets. All other rows are excluded from the analysis.

The Discriminant platform uses the validation column to train and evaluation the model, unless Stepwise Variable Selection is used. If the Stepwise Variable Selection option is selected in the launch, the Discriminant platform uses the validation column to train and tune the model or to train, tune, and evaluate the model. For more information about validation, see "Validation in JMP Modeling" on page 541 in the "Statistical Details" chapter.

Tip: If Stepwise Variable Selection is not used, the validation column should contain only two distinct values.

If you click the Validation button with no columns selected in the Select Columns list, you can add a validation column to your data table. For more information about the Make Validation Column utility, see the *Predictive and Specialized Modeling*.

By Performs a separate analysis for each level of the specified column.

Stepwise Variable Selection Performs stepwise variable selection using covariance analysis and *p*-values. See "Stepwise Variable Selection" on page 83.

If you have specified a validation set, statistics for the validation set also appear. The validation set statistics are used to determine how many steps to take if you use the Go button.

Note: This option is not provided for the Wide Linear discriminant method.

Discriminant Method Provides four methods for conducting discriminant analysis. See "Discriminant Methods" on page 86.

Shrink Covariances Shrinks the off-diagonal elements of the pooled within-group covariance matrix and the within-group covariance matrices. This can improve stability and reduce the variance of prediction. See "Shrink Covariances" on page 89.

Advanced Options Contains the following options:

Uncentered Canonical Suppresses centering of canonical scores for compatibility with older versions of JMP.

Use Pseudoinverses Uses Moore-Penrose pseudoinverses in the analysis when the covariance matrix is singular. The resulting scores involve all covariates. If left unchecked, the analysis drops covariates that are linear combinations of covariates that precede them in the list of **Y, Covariates**.

Cross Validate by Excluded Rows Specifies that the excluded rows form a validation set for which statistics of fit are calculated.

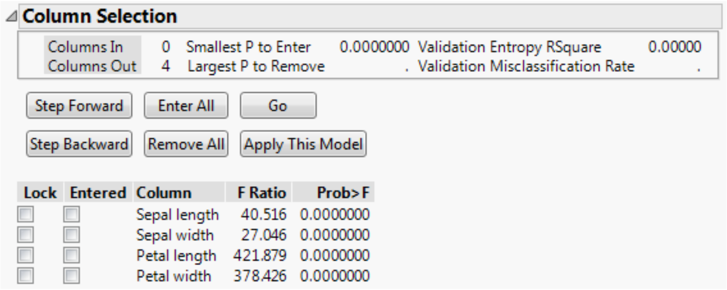
Stepwise Variable Selection

Note: Stepwise Variable Selection is not available for the Wide Linear method.

If you select the Stepwise Variable Selection option in the launch window, the Discriminant Analysis report opens, showing the Column Selection panel. You can perform stepwise analysis using the buttons to select variables or selecting them manually with the Lock and Entered check boxes. Based on your selection *F* ratios and *p*-values are updated. For more information about how these are updated, see "Updating the F Ratio and Prob>F" on page 83.

If you specify any type of validation set, a Go button appears. When you click Go, JMP uses the validation set statistics to determine how many steps to take.

Figure 5.4 Column Selection Panel for Iris.jmp with a Validation Set



Columns In 0
Columns Out 4
Smallest P to Enter 0.0000000
Largest P to Remove .
Validation Entropy RSquare 0.00000
Validation Misclassification Rate .

Lock	Entered	Column	F Ratio	Prob>F
☐	☐	Sepal length	40.516	0.0000000
☐	☐	Sepal width	27.046	0.0000000
☐	☐	Petal length	421.879	0.0000000
☐	☐	Petal width	378.426	0.0000000

Updating the F Ratio and Prob>F

When you enter or remove variables from the model, the F Ratio and Prob>F values are updated based on an analysis of covariance model with the following structure:

- The covariate under consideration is the response.
- The covariates already entered into the model are predictors.
- The group variable is a predictor.

The values for F Ratio and Prob>F given in the Stepwise report are the *F* ratio and *p*-value for the analysis of covariance test for the group variable. The analysis of covariance test for the group variable is an indicator of its discriminatory power relative to the covariate under consideration.

Statistics

Columns In The number of columns currently selected for entry into the discriminant model.

Columns Out The number of columns currently available for entry into the discriminant model.

Smallest P to Enter The smallest *p*-value among the *p*-values for all covariates available to enter the model.

Largest P to Remove The largest *p*-value among the *p*-values for all covariates currently selected for entry into the model.

Validation Entropy RSquare Entropy RSquare for the validation set. Larger values indicate better fit. An Entropy RSquare value of 1 indicates that the classifications are perfectly predicted. Because uncertainty in the predicted probabilities is typical for discriminant models, Entropy RSquare values tend to be small.

See "Entropy RSquare" on page 97. Available only if a validation set is used.

Note: It is possible for the Validation Entropy RSquare to be negative.

Validation Misclassification Rate Misclassification rate for the validation set. Smaller values indicate better classification. Available only if a validation set is used.

Buttons

Step Forward Enters the most significant covariate from the covariates not yet entered. If a validation set is used, the Prob>F values are based on the training set.

Step Backward Removes the least significant covariate from the covariates entered but not locked. If a validation set is used, Prob>F values are based on the training set.

Enter All Enters all covariates by checking all covariates that are not locked in the Entered column.

Remove All Removes all covariates that are not locked by deselecting them in the Entered column.

Apply this Model Produces a discriminant analysis report based on the covariates that are checked in the Entered columns. The Select Columns outline is closed and the Discriminant Analysis window is updated to show analysis results based on your selected Discriminant Method.

Tip: After you click **Apply this Model**, the columns that you select appear at the top of the Score Summaries report.

Go Enters covariates in forward steps until the Validation Entropy RSquare begins to decrease. Entry terminates when two forward steps are taken without improving the Validation Entropy RSquare. Available only with excluded rows in JMP or a validation column in JMP Pro.

Columns

Lock Forces a covariate to stay in its current state regardless of any stepping using the buttons.

Note the following:

- If you enter a covariate and then select **Lock** for that covariate, it remains in the model regardless of selections made using the control buttons. The **Entered** box for the locked covariate shows a dimmed check mark to indicate that it is in the model.
- If you select **Lock** for a covariate that is not Entered, it is not entered into the model regardless of selections made using the control buttons.

Entered Indicates which columns are currently in the model. You can manually select columns in or out of the model. A dimmed check mark indicates a locked covariate that has been entered into the model.

Column The covariate of interest.

F Ratio The *F* ratio for a test for the group variable obtained using an analysis of covariance model. See "Updating the F Ratio and Prob>F" on page 83.

Prob > F The *p*-value for a test for the group variable obtained using an analysis of covariance model. See "Updating the F Ratio and Prob>F" on page 83.

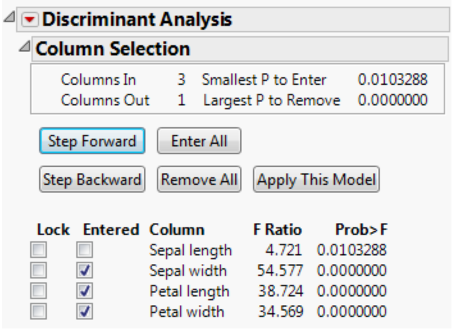
Stepwise Example

For an illustration of how to use Stepwise, consider the Iris.jmp sample data table.

1. Select **Help > Sample Data Library** and open Iris.jmp.
2. Select **Analyze > Multivariate Methods > Discriminant**.
3. Select Sepal length, Sepal width, Petal length, and Petal width and click **Y, Covariates**.
4. Select Species and click **X, Categories**.
5. Select **Stepwise Variable Selection**.
6. Click **OK**.
7. Click **Step Forward** three times.

   Three covariates are entered into the model. The Smallest P to Enter appears in the top panel. It is 0.0103288, indicating that the remaining covariate, Sepal length, might also be valuable in a discriminant analysis model for Species.

Figure 5.5 Stepped Model for Iris.jmp



Columns In 3 Smallest P to Enter 0.0103288
Columns Out 1 Largest P to Remove 0.0000000

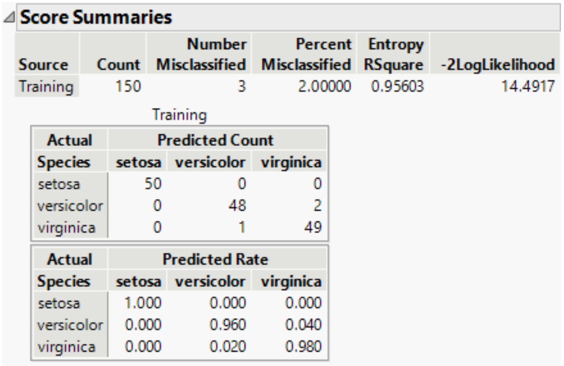
Lock	Entered	Column	F Ratio	Prob>F
☐	☐	Sepal length	4.721	0.0103288
☐	☑	Sepal width	54.577	0.0000000
☐	☑	Petal length	38.724	0.0000000
☐	☑	Petal width	34.569	0.0000000

8. Click **Apply This Model**.

   The Column Selection outline is closed. The window is updated to show reports for a fit based on the entered covariates and your selected discriminant method.

   Note that the covariates that you selected for your model are listed at the top of the Score Summaries report.

Figure 5.6 Score Summaries Report Showing Selected Covariates



Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	150	3	2.00000	0.95603	14.4917

Training

Actual	Predicted Count		
Species	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

Actual	Predicted Rate		
Species	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	0.960	0.040
virginica	0.000	0.020	0.980

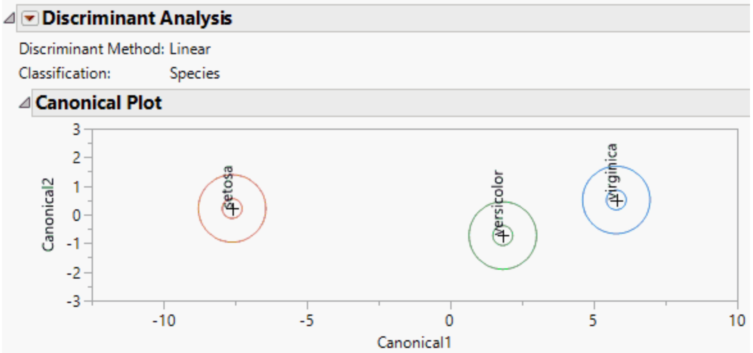
Discriminant Methods

JMP offers these methods for conducting Discriminant Analysis: Linear, Quadratic, Regularized, and Wide Linear. The first three methods differ in terms of the underlying model. The Wide Linear method is an efficient way to fit a Linear model when the number of covariates is large.

Note: When you enter more than 500 covariates, a JMP Alert recommends that you switch to the Wide Linear method. This is because computation time can be considerable when you use the other methods with a large number of columns. Click **Wide Linear, Many Columns** to switch to the Wide Linear method. Click **Continue** to use the method you originally selected.

Figure 5.7 Linear, Quadratic, and Regularized Discriminant Analysis

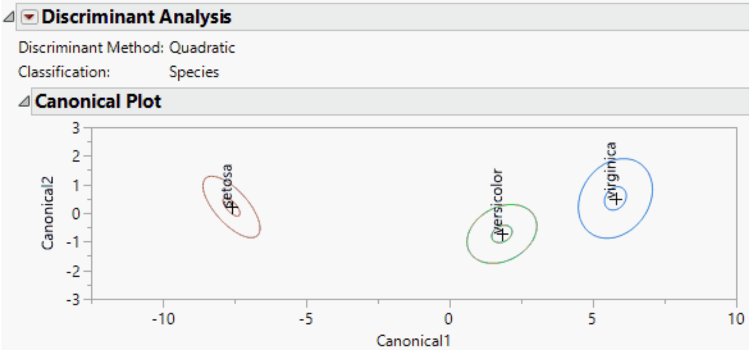
Linear



Discriminant Analysis

Discriminant Method: Linear
Classification: Species

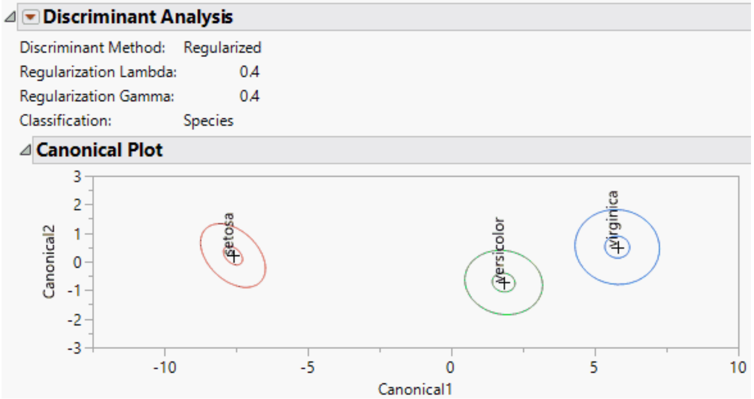
Quadratic



Discriminant Analysis

Discriminant Method: Quadratic
Classification: Species

Regularized
(λ=0.4, γ=0.4)



Discriminant Analysis

Discriminant Method: Regularized
Regularization Lambda: 0.4
Regularization Gamma: 0.4
Classification: Species

The Linear, Quadratic, and Regularized methods are illustrated in Figure 5.7. The methods are described here briefly. See “Saved Formulas” on page 110.

Linear, Common Covariance Performs linear discriminant analysis. This method assumes that the within-group covariance matrices are equal. See “Linear Discriminant Method” on page 111.

Quadratic, Different Covariances Performs quadratic discriminant analysis. This method assumes that the within-group covariance matrices differ. This method requires estimating more parameters than the Linear method requires. If group sample sizes are small, you risk obtaining unstable estimates. See "Quadratic Discriminant Method" on page 112.

If a covariate is constant across a level of the X variable, then its related entries in the within-group covariance matrix have zero covariances. To enable matrix inversion, the zero covariances are replaced with the corresponding pooled within covariances. When this is done, a note appears in the report window identifying the problematic covariate and level of X.

Tip: A shortcoming of the quadratic method surfaces in small data sets. It can be difficult to construct invertible and stable covariance matrices. The Regularized method ameliorates these problems, still allowing for differences among groups.

Regularized, Compromise Method Provides two ways to impose stability on estimates when the within-group covariance matrices differ. This is a useful option when group sample sizes are small. See "Regularized, Compromise Method" on page 88 and "Regularized Discriminant Method" on page 113.

Wide Linear, Many Columns Useful in fitting models based on a large number of covariates, where other methods can have computational difficulties. This method assumes that all within-group covariance matrices are equal. This method uses a singular value decomposition approach to compute the inverse of the pooled within-group covariance matrix. See "Description of the Wide Linear Algorithm" on page 110.

Note: When you use the Wide Linear option, a few of the features that normally appear for other discriminant methods are not available. This is because the algorithm does not explicitly calculate the very large pooled within-group covariance matrix.

Regularized, Compromise Method

Regularized discriminant analysis is governed by two nonnegative parameters.

- The first parameter (**Lambda, Shrinkage to Common Covariance**) specifies how to mix the individual and group covariance matrices. For this parameter, 1 corresponds to Linear Discriminant Analysis and 0 corresponds to Quadratic Discriminant Analysis.
- The second parameter (**Gamma, Shrinkage to Diagonal**) is a multiplier that specifies how much deflation to apply to the non-diagonal elements (the covariances across variables). If you choose 1, then the covariance matrix is forced to be diagonal.

Assigning 0 to each of these two parameters is identical to requesting quadratic discriminant analysis. Similarly, assigning 1 to Lambda and 0 to Gamma requests linear discriminant analysis. Use Table 5.1 to help you decide on the regularization. See Figure 5.7 for examples of linear, quadratic, and regularized discriminant analysis.

Table 5.1 Regularized Discriminant Analysis

Use Smaller Lambda	Use Larger Lambda	Use Smaller Gamma	Use Larger Gamma
Covariance matrices differ	Covariance matrices are identical	Variables are correlated	Variables are uncorrelated
Many rows	Few rows		
Few variables	Many variables		

Shrink Covariances

In the Discriminant launch window, you can select the option to Shrink Covariances. This option is recommended when some groups have a small number of observations. Discriminant analysis requires inversion of the covariance matrices. Shrinking off-diagonal entries improves their stability and reduces prediction variance. The Shrink Covariances option shrinks the off-diagonal entries by a factor that is determined using the method described in Schafer and Strimmer (2005).

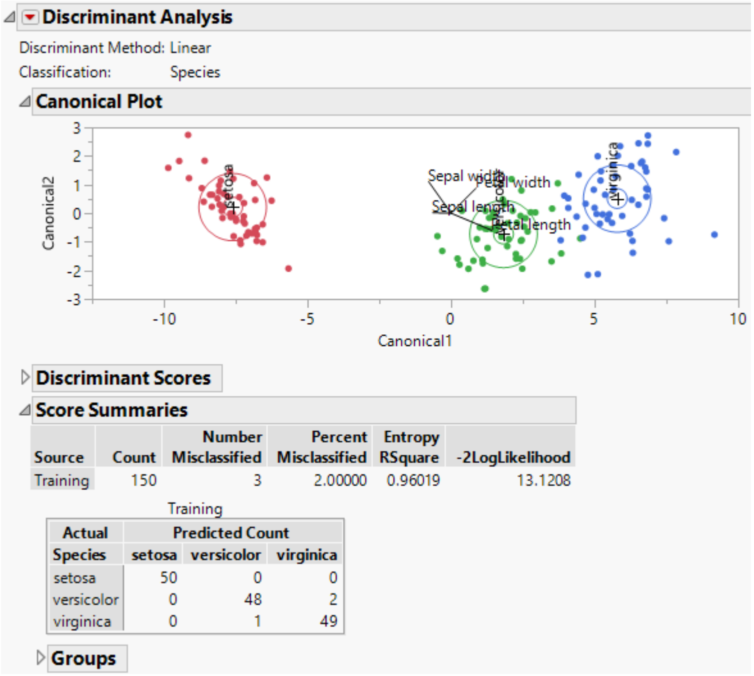
If you select the Shrink Covariances option with the Linear discriminant method in the launch window, this provides a shrinkage of the covariance matrices that is equivalent to the shrinkage provided by the Regularized discriminant method with appropriate Lambda and Gamma values. When you select the Shrink Covariances option and run your analysis, the Shrinkage report gives you an Overall Shrinkage value and an Overall Lambda value. To obtain the same analysis using the Regularized method, enter 1 as Lambda and the Overall Lambda from the Shrinkage report as Gamma in the Regularization Parameters window.

Discriminant Analysis Report

The Discriminant Analysis report provides discriminant results based on your selected Discriminant Method. The Discriminant Method and the Classification variable are shown at the top of the report. If you selected the Regularized method, its associated parameters are also shown.

You can change Discriminant Method by selecting the option from the Discriminant Analysis red triangle menu. The results in the report update to reflect the selected method.

Figure 5.8 Example of a Discriminant Analysis Report



Discriminant Method: Linear
Classification: Species

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	150	3	2.00000	0.96019	13.1208

Training

Actual Species	Predicted Count setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

The default Discriminant Analysis report contains the following sections:

- When you select the Wide Linear discriminant method, a Principal Components report appears. See "Principal Components" on page 90.
- The Canonical Plot shows the points and multivariate means in the two dimensions that best separate the groups. See "Canonical Plot and Canonical Structure" on page 91.
- The Discriminant Scores report provides details about how each observation is classified. See "Discriminant Scores" on page 95.
- The Score Summaries report provides an overview of how well observations are classified. See "Discriminant Scores" on page 95.

Principal Components

This report appears only for the Wide Linear method. Consider the following notation:

- Denote the n by p matrix of covariates by $\mathbf{Y}$, where n is the number of observations and p is the number of covariates.
- For each observation in $\mathbf{Y}$, subtract the covariate mean and divide the difference by the pooled standard deviation for the covariate. Denote the resulting matrix by $\mathbf{Y_s}$.

The report gives the following:

Number The number of eigenvalues extracted. Eigenvalues are extracted until Cum Percent is at least 99.99%, indicating that 99.99% of the variation has been explained.

Eigenvalue The eigenvalues of the covariance matrix for $\mathbf{Y_s}$, namely $(\mathbf{Y_s}'\mathbf{Y_s})/(n - p)$, arranged in decreasing order.

Cum Percent The cumulative sum of the eigenvalues as a percentage of the sum of all eigenvalues. The eigenvalues sum to the rank of $\mathbf{Y_s}'\mathbf{Y_s}$.

Singular Value The singular values of $\mathbf{Y_s}$ arranged in decreasing order.

Canonical Plot and Canonical Structure

The Canonical Plot is a biplot that describes the canonical correlation structure of the variables.

Canonical Structure

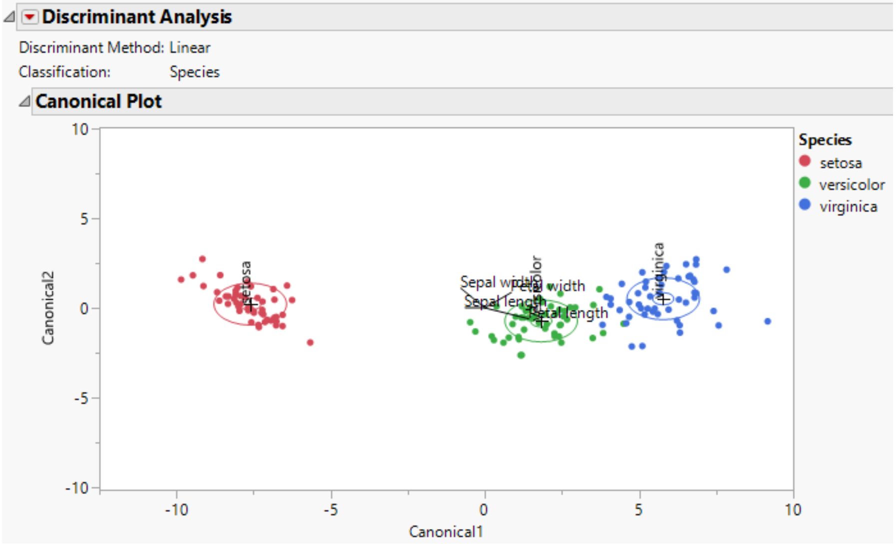
Each of the levels of the X, Categories column defines an indicator variable. A canonical correlation is performed between the set of indicator variables representing the categories and the covariates. Linear combinations of the covariates, called *canonical variables*, are derived. These canonical variables attempt to summarize the between-category variation.

The first canonical variable is the linear combination of the covariates that maximizes the multiple correlation between the category indicator variables and the covariates. The second canonical variable is a linear combination uncorrelated with the first canonical variable that maximizes the multiples correlation with the categories. If the X, Categories column has k levels, then $k - 1$ canonical variables are obtained.

Canonical Plot

Figure 5.9 shows the Canonical Plot for a linear discriminant analysis of the data table Iris.jmp. The points have been colored by Species.

Figure 5.9 Canonical Plot for Iris.jmp



Discriminant Analysis

Discriminant Method: Linear

Classification: Species

Canonical Plot

The biplot axes are the first two canonical variables. These define the two dimensions that provide maximum separation among the groups. Each canonical variable is a linear combination of the covariates. (See “Canonical Structure” on page 91.) The biplot shows how each observation is represented in terms of canonical variables and how each covariate contributes to the canonical variables.

- The observations and the multivariate means of each group are represented as points on the biplot. They are expressed in terms of the first two canonical variables.
  - The point corresponding to each multivariate mean is denoted by a plus (“+”) marker.
  - A 95% confidence level ellipse is plotted for each mean. If two groups differ significantly, the confidence ellipses tend not to intersect.
  - An ellipse denoting a 50% contour is plotted for each group. This depicts a region in the space of the first two canonical variables that contains approximately 50% of the observations, assuming normality.
- The set of rays that appears in the plot represents the covariates.
  - For each canonical variable, the coefficients of the covariates in the linear combination can be interpreted as *weights*.
  - To facilitate comparisons among the weights, the covariates are standardized so that each has mean 0 and standard deviation 1. The coefficients for the standardized covariates are called the *canonical weights*. The larger the canonical weight of a covariate, the greater its association with the canonical variable.

- The length and direction of each ray in the biplot indicates the degree of association of the corresponding covariate with the first two canonical variables. The length of the rays is a multiple of the canonical weights.
- The rays emanate from the point (0,0), which represents the grand mean of the data in terms of the canonical variables.
- You can obtain the values of the weight coefficients by selecting **Canonical Options > Show Canonical Details** from the Discriminant Analysis red triangle menu. At the bottom of the Canonical Details report, click Standardized Scoring Coefficients. See "Standardized Scoring Coefficients" on page 105.

Modifying the Canonical Plot

Additional options enable you to modify the biplot:

- Show or hide the 95% confidence ellipses by selecting **Canonical Options > Show Means CL Ellipses** from the Discriminant Analysis red triangle menu.
- Show or hide the rays by selecting **Canonical Options > Show Biplot Rays** from the Discriminant Analysis red triangle menu.
- Drag the center of the biplot rays to other places in the graph. Specify their position and scaling by selecting **Canonical Options > Biplot Ray Position** from the Discriminant Analysis red triangle menu. The default Radius Scaling shown in the Canonical Plot is 1.5, unless an adjustment is needed to make the rays visible.
- Show or hide the 50% contours by selecting **Canonical Options > Show Normal 50% Contours** from the Discriminant Analysis red triangle menu.
- Color code the points to match the ellipses by selecting **Canonical Options > Color Points** from the Discriminant Analysis red triangle menu.

Classification into Three or More Categories

For the Iris.jmp data, there are three Species, so there are only two canonical variables. The plot in Figure 5.9 shows good separation of the three groups using the two canonical variables.

The rays in the plot indicate the following:

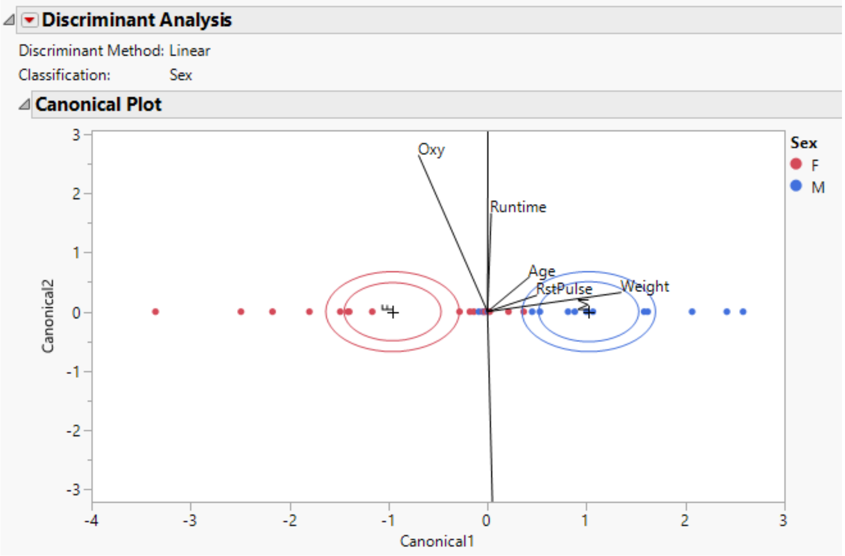
- Petal length is positively associated with Canonical1 and negatively associated with Canonical2. It carries more weight in defining Canonical1 than Canonical2.
- Petal width is positively associated with both Canonical1 and Canonical2. It carries about the same weight in defining both canonical variates.
- Sepal width is negatively associated with Canonical1 and positively associated with Canonical2. It carries more weight in defining Canonical2 than Canonical1.
- Sepal length is negatively weighted in terms of defining Canonical1 and very weakly associated in defining Canonical2.

Classification into Two Categories

When the classification variable has only two levels, the points are plotted against the single canonical variable, denoted by Canonical1 in the plot. The canonical weights for each covariate relate to Canonical1 only. The rays are shown with a vertical component only in order to separate them. Project the rays onto the Canonical1 axis to compare their relative association with the single canonical variable.

Figure 5.10 shows a Canonical Plot for the sample data table Fitness.jmp. The seven continuous variates are used to classify an individual into the categories M (male) or F (female). Since the classification variable has only two categories, there is only one canonical variable.

Figure 5.10 Canonical Plot for Fitness.jmp



The points in the Canonical Plot have been colored by Sex. Note that the two groups are well separated by their values on Canonical1.

Although the rays corresponding to the seven covariates have a vertical component, in this case you must interpret the rays only in terms of their projection onto the Canonical1 axis. You note the following:

- MaxPulse, Runtime, and RunPulse have little association with Canonical1.
- Weight, RstPulse, and Age are positively associated with Canonical1. Weight has the highest degree of association. The covariates RstPulse and Age have a similar, but smaller, degree of association.
- Oxy is negatively associated with Canonical1.

Discriminant Scores

The Discriminant Scores report provides the predicted classification of each observation and supporting information.

Row Row of the observation in the data table.

Actual Classification of the observation as given in the data table.

SqDist(Actual) Value of the saved formula SqDist[<level>] for the classification of the observation given in the data table. See "Score Options" on page 100.

Note: Due to an offset term in the formula, SqDist(Actual) can be negative.

Prob(Actual) Estimated probability of the observation's actual classification.

-Log(Prob) Negative of the log of Prob(Actual). Large values of this negative log-likelihood identify observations that are poorly predicted in terms of membership in their actual categories.

A plot of -Log(Prob) appears to the right of the -Log(Prob) values. A large bar indicates a poor prediction. An asterisk(*) indicates observations that are misclassified.

If you are using a validation or a test set, observations in the validation set are marked with a "v" and those in the test set are marked with a "t".

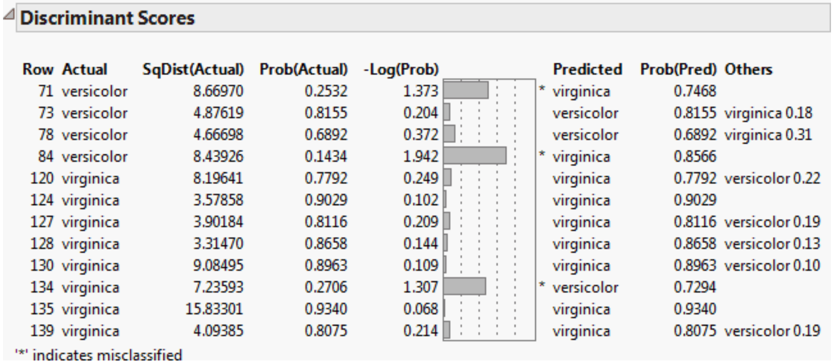
Predicted Predicted classification of the observation. The predicted classification is the category with the highest predicted probability of membership.

Prob(Pred) Estimated probability of the observation's predicted classification.

Others Lists other categories, if they exist, that have a predicted probability that exceeds 0.1.

Figure 5.11 shows the Discriminant Scores report for the Iris.jmp sample data table using the Linear discriminant method. The option **Score Options > Show Interesting Rows Only** option is selected, showing only misclassified rows or rows with predicted probabilities between 0.05 and 0.95.

Figure 5.11 Show Interesting Rows Only



Discriminant Scores

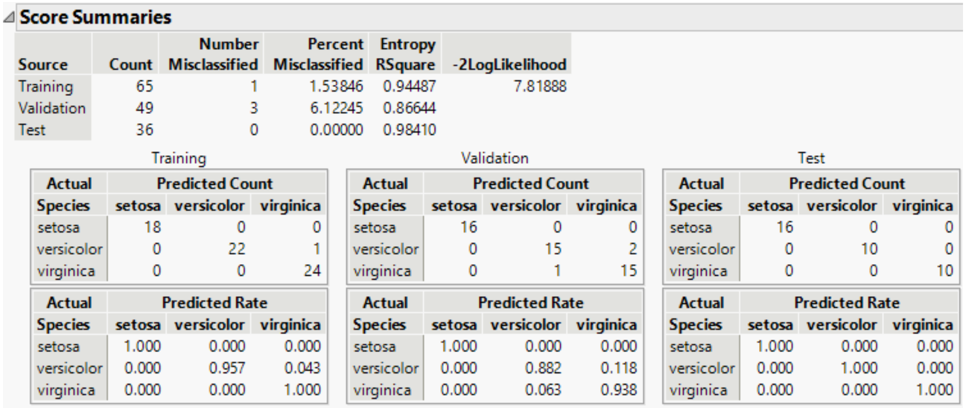
Row	Actual	SqDist(Actual)	Prob(Actual)	-Log(Prob)		Predicted	Prob(Pred)	Others
71	versicolor	8.66970	0.2532	1.373	*	virginica	0.7468	
73	versicolor	4.87619	0.8155	0.204		versicolor	0.8155	virginica 0.18
78	versicolor	4.66698	0.6892	0.372		versicolor	0.6892	virginica 0.31
84	versicolor	8.43926	0.1434	1.942	*	virginica	0.8566	
120	virginica	8.19641	0.7792	0.249		virginica	0.7792	versicolor 0.22
124	virginica	3.57858	0.9029	0.102		virginica	0.9029	
127	virginica	3.90184	0.8116	0.209		virginica	0.8116	versicolor 0.19
128	virginica	3.31470	0.8658	0.144		virginica	0.8658	versicolor 0.13
130	virginica	9.08495	0.8963	0.109		virginica	0.8963	versicolor 0.10
134	virginica	7.23593	0.2706	1.307	*	versicolor	0.7294	
135	virginica	15.83301	0.9340	0.068		virginica	0.9340	
139	virginica	4.09385	0.8075	0.214		virginica	0.8075	versicolor 0.19

'*' indicates misclassified

Score Summaries

The Score Summaries report provides an overview of the discriminant scores. The table in Figure 5.12 shows Actual and Predicted classifications. If all observations are correctly classified, the off-diagonal counts are zero.

Figure 5.12 Score Summaries for Iris.jmp



Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	65	1	1.53846	0.94487	7.81888
Validation	49	3	6.12245	0.86644	
Test	36	0	0.00000	0.98410	

Training

Actual	Predicted Count		
Species	setosa	versicolor	virginica
setosa	18	0	0
versicolor	0	22	1
virginica	0	0	24

Actual	Predicted Rate		
Species	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	0.957	0.043
virginica	0.000	0.000	1.000

Validation

Actual	Predicted Count		
Species	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	15	2
virginica	0	1	15

Actual	Predicted Rate		
Species	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	0.882	0.118
virginica	0.000	0.063	0.938

Test

Actual	Predicted Count		
Species	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	10	0
virginica	0	0	10

Actual	Predicted Rate		
Species	setosa	versicolor	virginica
setosa	1.000	0.000	0.000
versicolor	0.000	1.000	0.000
virginica	0.000	0.000	1.000

The Score Summaries report provides the following information:

Columns If you used Stepwise Variable Selection to construct the model, the columns entered into the model are listed (Figure 5.6).

Source If no validation is used, all observations comprise the Training set. If validation is used, a row is shown for the Training and Validation sets, or for the Training, Validation, and Test sets.

Number Misclassified Provides the number of observations in the specified set that are incorrectly classified.

Percent Misclassified Provides the percent of observations in the specified set that are incorrectly classified.

Entropy RSquare A measure of fit. Larger values indicate better fit. An Entropy RSquare value of 1 indicates that the classifications are perfectly predicted. Because uncertainty in the predicted probabilities is typical for discriminant models, Entropy RSquare values tend to be small.

See “Entropy RSquare” on page 97.

Note: It is possible for Entropy RSquare to be negative.

-2LogLikelihood Twice the negative log-likelihood of the observations in the training set, based on the model. Smaller values indicate better fit. Provided for the training set only. See *Fitting Linear Models*.

Confusion Matrices Shows matrices of actual by predicted counts for each level of the categorical X. If you are using JMP Pro with validation, a matrix is given for each set of observations. If you are using JMP with excluded rows, the excluded rows are considered the validation set and a separate Validation matrix is given. See “Validation in JMP and JMP Pro” on page 109.

Entropy RSquare

The Entropy RSquare is a measure of fit. It is computed for the training set and for the validation and test sets if validation is used.

Entropy RSquare for the Training Set

For the training set, Entropy RSquare is computed as follows:

- A discriminant model is fit using the training set.
- Predicted probabilities based on the model are obtained.
- Using these predicted probabilities, the likelihood is computed for observations in the training set. Call this $Likelihood\_Full_{Training}$.
- The reduced model (no predictors) is fit using the training set.
- The predicted probabilities for the levels of X from the reduced model are used to compute the likelihood for observations in the training set. Call this quantity $Likelihood\_Reduced_{Training}$.
- The Entropy RSquare for the training set is:

$$\text{Entropy RSquare}_{Training} = 1 - \frac{\log(Likelihood\_Full_{Training})}{\log(Likelihood\_Reduced_{Training})}$$

Entropy RSquare for Validation and Test Sets

For the validation set, Entropy RSquare is computed as follows:

- A discriminant model is fit using only the training set.
- Predicted probabilities based on the training set model are obtained for all observations.
- Using these predicted probabilities, the likelihood is computed for observations in the validation set. Call this $Likelihood\_Full_{Validation}$.
- The reduced model (no predictors) is fit using only the training set.
- The predicted probabilities for the levels of X from the reduced model are used to compute the likelihood for observations in the validation set. Call this quantity $Likelihood\_Reduced_{Validation}$.
- The Validation Entropy RSquare is:

$$\text{Validation Entropy RSquare} = 1 - \frac{\log(Likelihood\_Full_{Validation})}{\log(Likelihood\_Reduced_{Validation})}$$

The Entropy RSquare for the test set is computed in a manner analogous to the Entropy RSquare for the Validation set.

Discriminant Analysis Options

The Discriminant Analysis red triangle menu contains the following options:

Stepwise Variable Selection (Not available for the Wide Linear method.) Shows or hides the Column Selection control panel. This control panel contains options that enable you to perform stepwise variable selection using covariance analysis and p-values. See "Stepwise Variable Selection" on page 83.

Discriminant Method Specifies the discriminant method. Choose from Linear, Quadratic, Regularized, and Wide Linear. See "Discriminant Methods" on page 86.

Discriminant Scores Shows or hides a table of the discriminant scores for each row.

Score Options Provides options for the scoring of the observations. In particular, you can save the scoring formulas. See "Score Options" on page 100.

Canonical Plot Shows or hides the Canonical Plot. See "Canonical Plot and Canonical Structure" on page 91.

Canonical Options Provides options that affect the Canonical Plot. See "Canonical Options" on page 102.

Canonical 3D Plot Shows a three-dimensional canonical plot. This option is available only when there are four or more levels of the categorical X. See "Example of a Canonical 3D Plot" on page 105.

Specify Priors Enables you to specify prior probabilities for each level of the *X* variable. See "Specify Priors" on page 107.

Consider New Levels Specifies that some points might not fit into any known group and should be considered to be from an unscored new group. See "Consider New Levels" on page 107.

Show Within Covariances Shows or hides these reports:

- A Covariance Matrices report that gives the pooled-within covariance and correlation matrices.
- For the Quadratic and Regularized methods, a Correlations for Each Group report that shows the within-group correlation matrices.

  For each group, the log of the determinant of the within-group covariance matrix is also shown.
- For the Quadratic discriminant method, adds a Group Covariances outline to the Covariance Matrices report that shows the within-group covariance matrices.

Show Within Covariances is not available for the Wide Linear discriminant method.

Show Group Means Shows or hides the Group Means report that provides the mean of each covariate. Means for each level of the *X* variable and overall means appear.

Save Discrim Matrices Saves a script called `Discrim Results` to the data table. The script is a list of the following objects for use in JSL:

- a list of the covariates (Ys)
- the categorical variable X
- a list of the levels of X
- a matrix of the means of the covariates by the levels of X
- the pooled-within covariance matrix

Save Discrim Matrices is not available for the Wide Linear discriminant method. See "Save Discrim Matrices" on page 107.

Scatterplot Matrix Opens a Scatterplot Matrix report that shows a matrix with a scatterplot for each pair of covariates. The option invokes the Scatterplot Matrix platform with shaded density ellipses for each group. The scatterplots include all observations in the data table, even if validation is used. See "Scatterplot Matrix" on page 108.

Not available for the Wide Linear discriminant method.

Profiler Shows or hides an interactive profiler report where the categorical probabilities are combined into a single profiler row. Changes in the factor values are reflected in the estimated classification probabilities. For more information about the options in the red triangle menu, see *Profilers*.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Score Options

Score Options provides the following selections that deal with scores:

Show Interesting Rows Only In the Discriminant Scores report, shows only rows that are misclassified and those with predicted probability between 0.05 and 0.95.

Show Classification Counts Shows or hides the confusion matrices and confusion rates matrices in the Score Summaries report. A confusion matrix is a two-way classification of actual and predicted responses. A confusion rates matrix is equal to the confusion matrix, with the numbers divided by the row totals. By default, the Score Summaries report shows a confusion matrix and a confusion rates matrix for each level of the categorical X. If you are using JMP Pro with validation, a matrix is given for each set of observations. If you are using JMP with excluded rows, these rows are considered the validation set and a separate Validation matrix is given. See “Validation in JMP and JMP Pro” on page 109.

Show Distances to Each Group Shows or hides a report that contains each observation’s squared Mahalanobis distance to each group mean.

Show Probabilities to Each Group Shows or hides a report that contains the probability that an observation belongs to each of the groups defined by the categorical X.

ROC Curve Shows or hides a Receiver Operating Characteristic (ROC) curve to the Score Summaries report. For more information about the ROC Curve, see *Predictive and Specialized Modeling*.

Select Misclassified Rows Selects the misclassified rows in the data table and in report windows that display a listing by row.

Select Uncertain Rows Selects rows with uncertain classifications in the data table and in report windows that display a listing by row. An uncertain row is one whose probability of group membership for *any* group is neither close to 0 nor close to 1.

When you select this option, a window opens where you can specify the range of predicted probabilities that reflect uncertainty. By default, any row whose probability differs from 0 or 1 by more than 0.1 is defined to be uncertain. Therefore, the default selects rows with probabilities between 0.1 and 0.9.

Save Formulas Saves distance, probability, and predicted membership formulas to the data table. See “Saved Formulas” on page 110.

- The distance formulas are SqDist0 and SqDist[<level>], where <level> represents a level of X. The distance formulas produce intermediate values connected with the Mahalanobis distance calculations.
- The probability formulas are Prob[<level>], where <level> represents a level of X. Each probability column gives the posterior probability of an observation’s membership in that level of X. The Response Probability column property is saved to each probability column. For more information about the Response Probability column property, see *Using JMP*.
- The predicted membership formula is Pred <X> and contains the “most likely level” classification rule.
- The Wide Linear method also saves a Discrim Data Matrix column containing the vector of covariates and a Discrim Prin Comp formula. See “Wide Linear Discriminant Method” on page 115.

Note: For any method other than Wide Linear, when you Save Formulas, a RowEdit Prob script is saved to the data table. This script selects uncertain rows in the data table. The script defines any row whose probability differs from 0 or 1 by more than 0.1 as uncertain. The script also opens a Row Editor window that enables you to examine the uncertain rows. If you fit a new model (other than Wide Linear) and select Save Formulas, any existing RowEdit Prob script is replaced with a script that applies to the new fit.

Make Scoring Script (Available only in JMP Standard.) Creates a script that constructs the formula columns saved by the Save Formulas option. You can save this script and use it, perhaps with other data tables, to create the formula columns that calculate membership probabilities and predict group membership.

JMP PRO **Publish Probability Formulas** (Available only in JMP Pro.) Creates probability formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula

Depot report is not open, this option creates a Formula Depot report. See *Predictive and Specialized Modeling*.

Canonical Options

The first options listed below relate to the appearance of the Canonical Plot or the Canonical 3D Plot. The remaining options provide detail on the calculations related to the plot.

Note: The Canonical 3D Plot is available only when there are three or more covariates and when the grouping variable has four or more categories.

Options Relating to Plot Appearance

Show Points Shows or hides the points in the Canonical Plot and the Canonical 3D Plot.

Show Means CL Ellipses Shows or hides 95% confidence ellipses for the mean of each group in the Canonical Plot and the Canonical 3D Plot, assuming normality.

Show Normal 50% Contours Shows or hides a 50% prediction ellipse or ellipsoid for each group. In the Canonical Plot, each ellipse depicts a region in the space of the first two canonical variables that estimates where 50% of the observations for each group should fall, assuming multivariate normality. In the Canonical 3D Plot, each ellipsoid depicts a region in the space of the first three canonical variables that estimates where 50% of the observations should fall, assuming multivariate normality.

Show Biplot Rays Shows or hides the biplot rays in the Canonical Plot and the Canonical 3D Plot. The labeled rays show the directions of the covariates in the canonical space. They represent the degree of association of each covariate with each canonical variable.

Biplot Ray Position Enables you to specify the position and radius scaling of the biplot rays in the Canonical Plot and in the Canonical 3D Plot.

- By default, the rays emanate from the point (0,0), which represents the grand mean of the data in terms of the canonical variables. In the Canonical Plot, you can drag the rays or use this option to specify coordinates.
- The default Radius Scaling in the canonical plots is 1.5, unless an adjustment is needed to make the rays visible. Radius Scaling is done relative to the Standardized Scoring Coefficients.

Color Points Colors the points in the Canonical Plot and the Canonical 3D Plot based on the levels of the *X* variable. Color markers are added to the rows in the data table. This option is equivalent to selecting **Rows > Color or Mark by Column** and selecting the *X* variable. It is also equivalent to right-clicking the graph and selecting **Row Legend**, and then coloring by the classification column.

Options Relating to Calculations

Show Canonical Details Shows or hides the Canonical Details report. See "Show Canonical Details" on page 103.

Show Canonical Structure Shows or hides the Canonical Structures report. See "Show Canonical Structure" on page 105. Not available for the Wide Linear discriminant method.

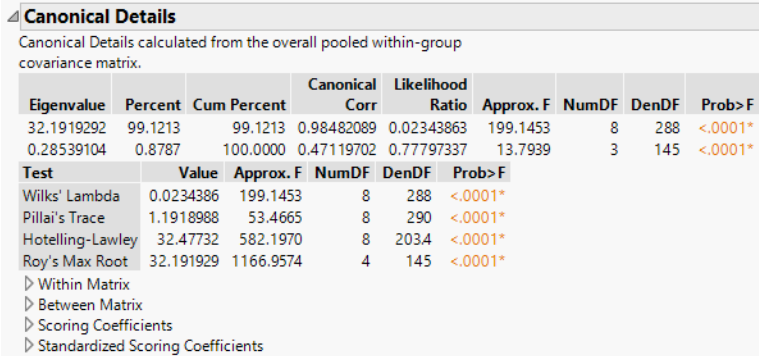
Save Canonical Scores Creates columns in the data table that contain canonical score formulas for each observation. The column for the k^{th} canonical score is named Canon[<k>].

Tip: In a script, sending the scripting command **Save to New Data Table** to the Discriminant object saves the following to a new data table: group means on the canonical variables; the biplot rays with 1.5 Radius Scaling of the Standardized Scoring Coefficients; and the canonical scores. Not available for the Wide Linear discriminant method.

Show Canonical Details

The Canonical Details report shows tests that address the relationship between the covariates and the grouping variable *X*. Relevant matrices are presented at the bottom of the report.

Figure 5.13 Canonical Details for Iris.jmp



Canonical Details

Canonical Details calculated from the overall pooled within-group covariance matrix.

Eigenvalue	Percent	Cum Percent	Canonical Corr	Likelihood Ratio	Approx. F	NumDF	DenDF	Prob>F
32.1919292	99.1213	99.1213	0.98482089	0.02343863	199.1453	8	288	<.0001*
0.28539104	0.8787	100.0000	0.47119702	0.77797337	13.7939	3	145	<.0001*

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.0234386	199.1453	8	288	<.0001*
Pillai's Trace	1.1918988	53.4665	8	290	<.0001*
Hotelling-Lawley	32.47732	582.1970	8	203.4	<.0001*
Roy's Max Root	32.191929	1166.9574	4	145	<.0001*

Note: The matrix used in computing the results in the report is the pooled within-covariance matrix (given as the Within Matrix). This matrix is used as a basis for the Canonical Details report for all discriminant methods. The statistics and tests in the Canonical Details report are the same for all discriminant methods.

Statistics and Tests

The Canonical Details report lists eigenvalues and gives a likelihood ratio test for zero eigenvalues. Four tests are provided for the null hypothesis that the canonical correlations are zero.

Eigenvalue Eigenvalues of the product of the Between Matrix and the inverse of the Within Matrix. These are listed from largest to smallest. The size of an eigenvalue reflects the amount of variance explained by its associated discriminant function.

Percent Proportion of the sum of the eigenvalues represented by the given eigenvalue.

Cum Percent Cumulative sum of the proportions.

Canonical Corr Canonical correlations between the covariates and the groups defined by the categorical *X*. Suppose that you define numeric indicator variables to represent the groups defined by *X*. Then perform a canonical correlation analysis using the covariates as one set of variables and the indicator variables representing the groups in *X* as the other. The Canonical Corr values are the canonical correlation values that result from this analysis.

Likelihood Ratio Likelihood ratio statistic for a test of whether the population values of the corresponding canonical correlation and all smaller correlations are zero. The ratio equals the product of the values (1 - Canonical Corr^2) for the given and all smaller canonical correlations.

Test Lists four standard tests for the null hypothesis that the means of the covariates are equal across groups: Wilk's Lambda, Pillai's Trace, Hotelling-Lawley, and Roy's Max Root. See "Multivariate Tests" on page 117 and "Approximate F-Tests" on page 118 in the "Discriminant Analysis" appendix.

Approx. F The *F* value associated with the corresponding test. For certain tests, the *F* value is approximate or an upper bound. See "Approximate F-Tests" on page 118 in the "Discriminant Analysis" appendix.

NumDF The numerator degrees of freedom for the corresponding test.

DenDF The denominator degrees of freedom for the corresponding test.

Prob>F The *p*-value for the corresponding test.

Matrices

Four matrices that relate to the canonical structure are presented at the bottom of the report. To view a matrix, click the disclosure icon beside its names. To hide it, click the name of the matrix.

Within Matrix Pooled within-covariance matrix.

Between Matrix Between groups covariance matrix, $\mathbf{S}_B$. See "Between Groups Covariance Matrix" on page 119.

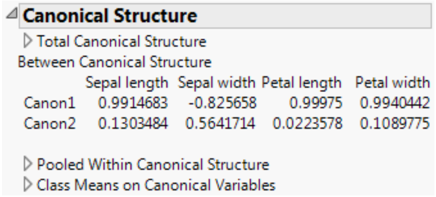
Scoring Coefficients Coefficients used to compute canonical scores in terms of the raw data. These are the coefficients used for the option **Canonical Options > Save Canonical Scores**. For more information about how these are computed, see the CANDISC Procedure chapter in SAS Institute Inc. (2020b).

Standardized Scoring Coefficients Coefficients used to compute canonical scores in terms of the standardized data. Often called *canonical weights*. For more information about how these are computed, see the CANDISC Procedure chapter in SAS Institute Inc. (2020b).

Show Canonical Structure

The Canonical Structure report gives three matrices that provide correlations between the canonical variables and the covariates. Another matrix shows means across the levels of the group variable. To view a matrix, click the disclosure icon beside its names. To hide it, click the name of the matrix.

Figure 5.14 Canonical Structure for Iris.jmp Showing between Canonical Structure



Between Canonical Structure

	Sepal length	Sepal width	Petal length	Petal width
Canon1	0.9914683	-0.825658	0.99975	0.9940442
Canon2	0.1303484	0.5641714	0.0223578	0.1089775

Total Canonical Structure Correlations between the canonical variables and the covariates. Often called *loadings*.

Between Canonical Structure Correlations between the group means on the canonical variables and the group means on the covariates.

Pooled Within Canonical Structure Partial correlations between the canonical variables and the covariates, adjusted for the group variable.

Class Means on Canonical Variables Provides means across the levels of the group variable for each canonical variable.

Example of a Canonical 3D Plot

1. Select **Help > Sample Data Library** and open Owl Diet.jmp.
2. Select rows 180 through 294.

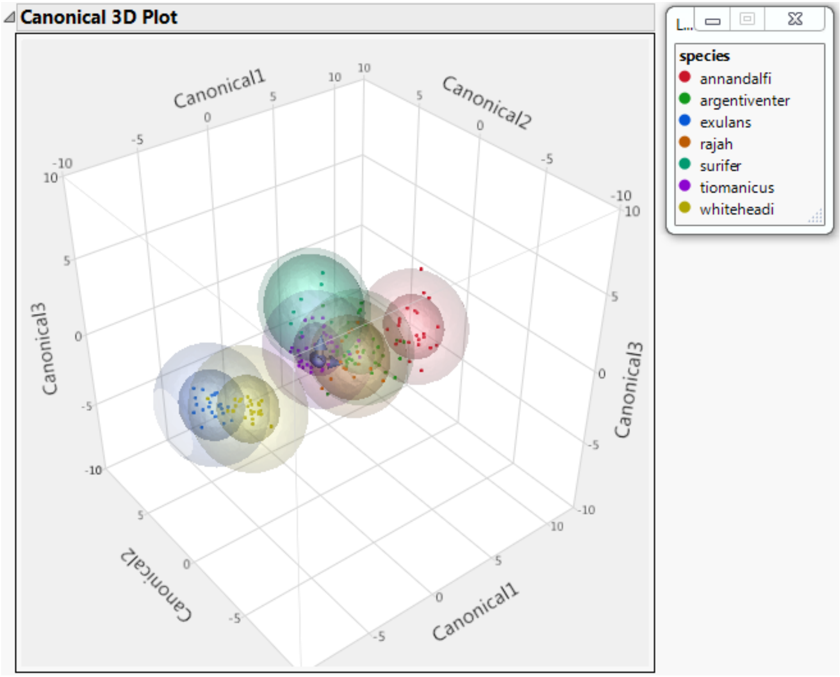
   These are the rows for which species is missing. Hide and exclude these rows.

3. Select **Rows > Hide and Exclude**.
4. Select **Rows > Color or Mark by Column**.
5. Select species.
6. From the Colors menu, select **JMP Dark**.
7. Check **Make Window with Legend**.
8. Click **OK**.

   A small Legend window appears. The rows in the data table are assigned colors by species.
9. **Select Analyze > Multivariate Methods > Discriminant**.
10. Specify skull length, teeth row, palatine foramen, and jaw length as **Y, Covariates**.
11. Specify species as **X, Categories**.
12. Click **OK**.
13. Click the Discriminant Analysis red triangle and click **Canonical 3D Plot**.

Tip: Click categories in the Legend to highlight those points in the Canonical 3D plot. Click and drag inside the 3D plot to rotate it.

Figure 5.15 Canonical 3D Plot with Legend Window



Specify Priors

The following options are available for specifying priors:

Equal Probabilities Assigns equal prior probabilities to all groups. This is the default.

Proportional to Occurrence Assigns prior probabilities to the groups that are proportional to their frequency in the observed data.

Other Enables you to specify custom prior probabilities.

Consider New Levels

Use the Consider New Levels option if you suspect that some of your observations are outliers with respect to the specified levels of the categorical variable. When you select the option, a menu asks you to specify the prior probability of the new level.

Observations that would be better fit using a new group are assigned to the new level, called “Other”. Probability of membership in the Other group assumes that these observations have the distribution of the entire set of observations where *no group structure* is assumed. This leads to correspondingly wide normal contours associated with the covariance structure. Distance calculations are adjusted by the specified prior probability.

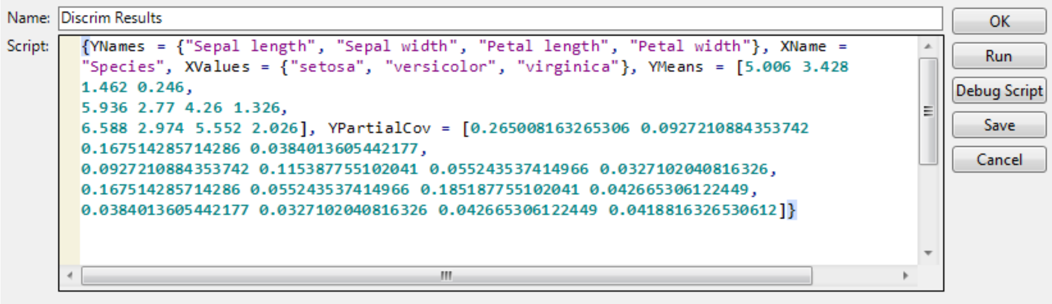
Save Discrim Matrices

Save Discrim Matrices creates a global list (`DiscrimResults`) for use in the JMP scripting language. The list contains the following, calculated for the training set:

- YNames, a list of the covariates (Ys)
- XName, the categorical variable
- XValues, a list of the levels of X
- YMeans, a matrix of the means of the covariates by the levels of X
- YPartialCov, the within covariance matrix

Consider the analysis obtained using the **Discriminant** script in the Iris.jmp sample data table. If you select **Save Discrim Matrices** from the Discriminant Analysis red triangle menu, the script **Discrim Results** is saved to the data table.

Figure 5.16 Discrim Results Table Script for Iris.jmp



Name: Discrim Results

Script:

```
{YNames = {"Sepal length", "Sepal width", "Petal length", "Petal width"}, XName =
"Species", XValues = {"setosa", "versicolor", "virginica"}, YMeans = [5.006 3.428
1.462 0.246,
5.936 2.77 4.26 1.326,
6.588 2.974 5.552 2.026], YPartialCov = [0.265008163265306 0.0927210884353742
0.167514285714286 0.0384013605442177,
0.0927210884353742 0.115387755102041 0.0552435374149660 0.0327102040816326,
0.167514285714286 0.0552435374149660 0.185187755102041 0.042665306122449,
0.0384013605442177 0.0327102040816326 0.042665306122449 0.0418816326530612]}
```

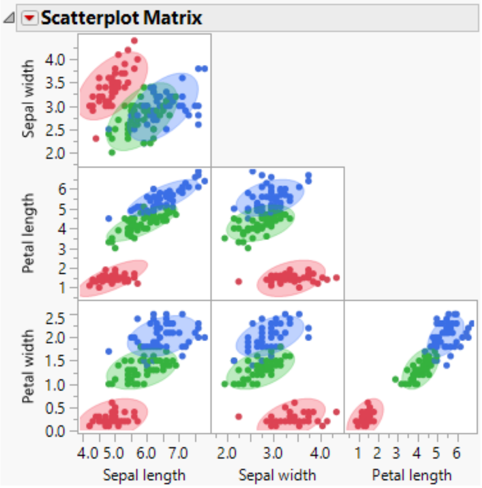
Note: In a script, you can send the scripting command Get Discrim Matrices to the Discriminant platform object. This obtains the same values as Save Discrim Matrices, but does not store them in the data table.

Scatterplot Matrix

The Scatterplot Matrix command invokes the Scatterplot Matrix platform in a separate window containing a lower triangular scatterplot matrix for the covariates. Points are plotted for all observations in the data table.

Ellipses with 90% coverage are shown for each level of the categorical variable X. For the Linear discriminant method, these are based on the pooled within covariance matrix. Figure 5.17 shows the Scatterplot Matrix window for the Iris.jmp sample data table.

Figure 5.17 Scatterplot Matrix for Iris.jmp



The options in the Scatterplot Matrix red triangle menu are described in *Essential Graphing*.

Validation in JMP and JMP Pro

In JMP, you can specify a validation set by excluding the rows that form the validation set. Select the rows that you want to use as your validation set and then select **Rows > Exclude/Unexclude**. The unexcluded rows are treated as the training set.

Note: In JMP Pro, you can specify a Validation column in the Discriminant launch window. A validation column must have a numeric data type and should contain at least two distinct values.

Notice the following:

- If the column contains two values, the smaller value defines the training set and the larger value defines the validation set.
- If the column contains three values, the values define the training, validation, and test sets in order of increasing size.
- If the column contains four or more distinct values, only the smallest three values and their associated observations are used to define the training, validation, and test sets, in that order.

When a validation set is specified, the Discriminant platform does the following:

- Models are fit using the training data.
- The Stepwise Variable Selection option gives the Validation Entropy RSquare and Validation Misclassification Rate statistics for the model. See “Statistics” on page 83 and “Entropy RSquare for Validation and Test Sets” on page 98.
- The Discriminant Scores report shows an indicator identifying rows in the validation and test sets.
- The Score Summaries report shows actual by predicted classifications for the training, validation, and test sets.

Statistical Details for the Discriminant Analysis Platform

- "Description of the Wide Linear Algorithm"
- "Saved Formulas"
- "Multivariate Tests"
- "Approximate F-Tests"
- "Between Groups Covariance Matrix"

Description of the Wide Linear Algorithm

Wide Linear discriminant analysis is performed as follows:

- The data are standardized by subtracting group means and dividing by pooled standard deviations.
- The singular value decomposition is used to obtain a principal component transformation matrix from the set of singular vectors.
- The number of components retained represents a minimum of 0.9999 of the sum of the squared singular values.
- A linear discriminant analysis is performed on the transformed data, where the data are not shifted by group means. This is a fast calculation because the pooled-within covariance matrix is diagonal.

Saved Formulas

This section gives the derivation of formulas saved by **Score Options > Save Formulas**. The formulas depend on the Discriminant Method.

For each group defined by the categorical variable X, observations on the covariates are assumed to have a p-dimensional multivariate normal distribution, where p is the number of covariates. The notation used in the formulas is given in Table 5.2.

Table 5.2 Notation for Formulas Given by Save Formulas Options

p	number of covariates
T	total number of groups (levels of X)

Table 5.2 Notation for Formulas Given by Save Formulas Options *(Continued)*

$t = 1,\ldots, T$	subscript to distinguish groups defined by X
n_t	number of observations in group t
$n = n_1 + n_2 + \ldots + n_T$	total number of observations
$\mathbf{y}$	p by 1 vector of covariates for an observation
$\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \ldots, y_{ipt})$	i^{th} observation in group t, consisting of a vector of p covariates
$\bar{\mathbf{y}}_t$	p by 1 vector of means of the covariates $\mathbf{y}$ for observations in group t
$\mathbf{y}_{bar}$	p by 1 vector of means for the covariates across all observations
$\mathbf{S}_t = \frac{1}{n_t - 1}\sum_{i=1}^{n_t}(\mathbf{y}_{it} - \bar{\mathbf{y}}_t)(\mathbf{y}_{it} - \bar{\mathbf{y}}_t)'$	estimated (p by p) within-group covariance matrix for group t
$\mathbf{S}_p = \frac{1}{n - T}\sum_{t=1}^{T}(n_t - 1)\mathbf{S}_t$	estimated (p by p) pooled within covariance matrix
q_t	prior probability of membership for group t
$p(t \mid \mathbf{y})$	posterior probability that $\mathbf{y}$ belongs to group t
$\lvert \mathbf{A} \rvert$	determinant of a matrix A

Linear Discriminant Method

In linear discriminant analysis, all within-group covariance matrices are assumed equal. The common covariance matrix is estimated by $\mathbf{S}_p$. See Table 5.2 for notation.

The Mahalanobis distance from an observation $\mathbf{y}$ to group t is defined as follows:

$$d_t^2 = (\mathbf{y} - \bar{\mathbf{y}}_t)'\mathbf{S}_p^{-1}(\mathbf{y} - \bar{\mathbf{y}}_t)$$

The likelihood for an observation **y** in group *t* is estimated as follows:

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2}\left|\mathbf{S}_p\right|^{-1/2}\exp(-(\mathbf{y}-\bar{\mathbf{y}}_t)'\mathbf{S}_p^{-1}(\mathbf{y}-\bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2}\left|\mathbf{S}_p\right|^{-1/2}\exp(-d_t^2/2) \end{aligned}$$

Note that the number of parameters that must be estimated for the pooled covariance matrix is *p(p+1)/2* and for the means is *Tp*. The total number of parameters that must be estimated is *p(p+1)/2* + *Tp*.

The posterior probability of membership in group *t* is defined as follows:

$$p(t|\mathbf{y}) = \frac{q_t l_t(\mathbf{y})}{\sum_{u=1}^{T} q_u l_u(\mathbf{y})} = \frac{1}{1 + \sum_{u \neq t} \exp(-[(d_u^2 - 2\log(q_u)) - (d_t^2 - 2\log(q_t))]/2)}$$

An observation **y** is assigned to the group for which its posterior probability is the largest.

The formulas saved by the Linear discriminant method are defined as follows:

SqDist[0]	$\mathbf{y}'\mathbf{S}_p^{-1}\mathbf{y}$
SqDist[<group *t*>]	$d_t^2 - 2\log(q_t)$
Prob[<group *t*>]	$p(t\|\mathbf{y})$
Pred <*X*>	*t* for which $p(t\|\mathbf{y})$ is maximum, $t = 1, \ldots, T$

Quadratic Discriminant Method

In quadratic discriminant analysis, the within-group covariance matrices are not assumed equal. The within-group covariance matrix for group *t* is estimated by $\mathbf{S}_t$. This means that the number of parameters that must be estimated for the within-group covariance matrices is *Tp(p+1)/2* and for the means is *Tp*. The total number of parameters that must be estimated is *Tp(p+3)/2*.

When group sample sizes are small relative to *p*, the estimates of the within-group covariance matrices tend to be highly variable. The discriminant score is heavily influenced by the smallest eigenvalues of the inverse of the within-group covariance matrices. See Friedman (1989). For this reason, if your group sample sizes are small compared to *p*, you might want to consider the Regularized method, described in "Regularized Discriminant Method" on page 113.

See Table 5.2 for notation. The Mahalanobis distance from an observation **y** to group *t* is defined as follows:

$$d_t^2 = (\mathbf{y} - \bar{\mathbf{y}}_t)'\mathbf{S}_t^{-1}(\mathbf{y} - \bar{\mathbf{y}}_t)$$

The likelihood for an observation **y** in group *t* is estimated as follows:

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2}\left|\mathbf{S}_t\right|^{-1/2}\exp(-(\mathbf{y} - \bar{\mathbf{y}}_t)'\mathbf{S}_t^{-1}(\mathbf{y} - \bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2}\left|\mathbf{S}_t\right|^{-1/2}\exp(-d_t^2/2) \end{aligned}$$

The posterior probability of membership in group *t* is the following:

$$\begin{aligned} p(t|\mathbf{y}) &= (q_t l_t(\mathbf{y}))/\left(\sum_{u=1}^{T} q_u l_u(\mathbf{x})\right) \\ &= \frac{1}{1 + \sum\limits_{u \neq t} \exp(-[(d_u^2 + \log\left|\mathbf{S}_u\right| - 2\log(q_u)) - (d_t^2 + \log\left|\mathbf{S}_t\right| - 2\log(q_t))]/2)} \end{aligned}$$

An observation **y** is assigned to the group for which its posterior probability is the largest.

The formulas saved by the Quadratic discriminant method are defined as follows:

SqDist[<group *t*>]	$d_t^2 + \log\left\|\mathbf{S}_t\right\| - 2\log(q_t)$
Prob[<group *t*>]	$p(t\|\mathbf{y})$
Pred <*X*>	*t* for which $p(t\|\mathbf{y})$ is maximum, $t = 1, \ldots, T$

Note: SqDist[<group *t*>] can be negative.

Regularized Discriminant Method

Regularized discriminant analysis allows for two parameters: λ and γ.

- The parameter λ balances weights assigned to the pooled covariance matrix and the within-group covariance matrices, which are not assumed equal.
- The parameter γ determines the amount of shrinkage toward a diagonal matrix.

This method enables you to leverage two aspects of regularization to bring stability to estimates for quadratic discriminant analysis. See Friedman (1989). See Table 5.2 for notation.

For the regularized method, the covariance matrix for group *t* is:

$$\Sigma_t = (1-\gamma)(\lambda \mathbf{S}_p + (1-\lambda)\mathbf{S}_t) + \gamma Diag((\lambda \mathbf{S}_p + (1-\lambda)\mathbf{S}_t))$$

The Mahalanobis distance from an observation **y** to group *t* is defined as follows:

$$d_t^2 = (\mathbf{y} - \bar{\mathbf{y}}_t)'\Sigma_t^{-1}(\mathbf{y} - \bar{\mathbf{y}}_t)$$

The likelihood for an observation **y** in group *t* is estimated as follows:

$$\begin{aligned} l_t(\mathbf{y}) &= (2\pi)^{-T/2}|\Sigma_t|^{-1/2}\exp(-(\mathbf{y} - \bar{\mathbf{y}}_t)'\Sigma_t^{-1}(\mathbf{y} - \bar{\mathbf{y}}_t)/2) \\ &= (2\pi)^{-T/2}|\Sigma_t|^{-1/2}\exp(-d_t^2/2) \end{aligned}$$

The posterior probability of membership in group *t* given by the following:

$$\begin{aligned} p(t|\mathbf{y}) &= (q_t l_t(\mathbf{y}))\Big/\left(\sum_{u=1}^{T} q_u l_u(\mathbf{x})\right) \\ &= \frac{1}{1 + \sum_{u \neq t} \exp(-[(d_u^2 + \log|\Sigma_u| - 2\log(q_u)) - (d_t^2 + \log|\Sigma_t| - 2\log(q_t))]/2)} \end{aligned}$$

An observation **y** is assigned to the group for which its posterior probability is the largest.

The formulas saved by the Regularized discriminant method are defined below:

SqDist[<group *t*>]	$d_t^2 + \log\lvert\Sigma_t\rvert - 2\log(q_t)$
Prob[<group *t*>]	$p(t\lvert\mathbf{y})$
Pred <*X*>	*t* for which $p(t\lvert\mathbf{y})$ is maximum, $t = 1, \ldots, T$

Note: SqDist[<group *t*>] can be negative.

Wide Linear Discriminant Method

The Wide Linear method is useful when you have a large number of covariates and, in particular, when the number of covariates exceeds the number of observations ($p > n$). This approach centers around an efficient calculation of the inverse of the pooled within-covariance matrix $\mathbf{S}_p$ or of its transpose, if $p > n$. It uses a singular value decomposition approach to avoid inverting and allocating space for large covariance matrices.

The Wide Linear method assumes equal within-group covariance matrices and is equivalent to the Linear method if the number of observations equals or exceeds the number of covariates.

Wide Linear Calculation

See Table 5.2 for notation. The following steps are used in the Wide Linear calculation:

1. Compute the T by p matrix **M** of within-group sample means. The $(t,j)^{\text{th}}$ entry of **M**, m_{tj}, is the sample mean for members of group t on the j^{th} covariate.
2. For each covariate j, calculate the pooled standard deviation across groups. Call this s_{jj}.
3. Denote the diagonal matrix with diagonal entries s_{jj} by $\mathbf{S}_{diag}$.
4. Center and scale values for each covariate:
   - Subtract the mean for the group to which the observation belongs.
   - Divide the difference by the pooled standard deviation.

   Using notation, for an observation i in group t, the group-centered and scaled value for the j^{th} covariate is:

   $$y_{ij}^{*} = \frac{y_{ij} - m_{t(i)j}}{s_{jj}}$$

   The notation *t(i)* indicates the group t to which observation i belongs.
5. Denote the matrix of y_{ij}^{*} values by $\mathbf{Y}_s$.
6. Denote the pooled within-covariance matrix for the group-centered and scaled covariates by **R**. The matrix **R** is given by the following:

   $$\mathbf{R} = (\mathbf{Y}_s{'}\mathbf{Y}_s)/(n - T)$$

7. Apply the singular value decomposition to $\mathrm{Y}_{\mathbf{s}}$:

   $$\mathbf{Y}_s = \mathbf{UDV'}$$

   where **U** and **V** are orthonormal and **D** is a diagonal matrix with positive entries (the singular values) on the diagonal. See “The Singular Value Decomposition” on page 357 in the “Statistical Details” appendix.

   Then **R** can be written as follows:

$$\mathbf{R} = (\mathbf{Y}_s'\mathbf{Y}_s)/(n-T) = (\mathbf{V}\mathbf{D}^2\mathbf{V}')/(n-T)$$

8. If **R** is of full rank, obtain $\mathbf{R}^{-1/2}$ as follows:

$$\mathbf{R}^{-1/2} = (\mathbf{V}\mathbf{D}^{-1}\mathbf{V}')/\sqrt{n-T}$$

where $\mathbf{D}^{-1}$ is the diagonal matrix whose diagonal entries are the inverses of the diagonal entries of **D**.

If **R** is not of full rank, define a pseudo-inverse for **R** as follows:

$$\mathbf{R}^{-} = (\mathbf{V}\mathbf{D}^{-2}\mathbf{V}')/(n-T)$$

Then define the inverse square root of **R** as follows:

$$(\mathbf{R}^{-})^{1/2} = (\mathbf{V}\mathbf{D}^{-1}\mathbf{V}')/\sqrt{n-T}$$

9. If **R** is of full rank, it follows that $\mathbf{R}^{-} = \mathbf{R}^{-1}$. So, for completeness, the discussion continues using pseudo-inverses.

Define a p by p matrix $\mathbf{T}_s$ as follows:

$$\mathbf{T}_s = (\mathbf{S}_{diag}^{-1}\mathbf{V}\mathbf{D}^{-})/(\sqrt{n-T})$$

Then:

$$(\mathbf{T}_s\mathbf{T}_s') = \left(\mathbf{S}_{diag}^{-1}\mathbf{V}(\mathbf{D}^{-})^2\mathbf{V}'\mathbf{S}_{diag}^{-1}\right)/(n-T) = \mathbf{S}_{diag}^{-1}\mathbf{R}^{-}\mathbf{S}_{diag}^{-1} = \mathbf{S}_p^{-}$$

where $\mathbf{S}^{-}_p$ is a generalized inverse of the pooled within-covariance matrix for the original data that is calculated using the SVD.

Mahalanobis Distance

The formulas for the Mahalanobis distance, the likelihood, and the posterior probabilities are identical to those in "Linear Discriminant Method" on page 111. However, the inverse of $\mathbf{S}_p$ is replaced by a generalized inverse computed using the singular value decomposition.

When you save the formulas, the Mahalanobis distance is given in terms of the decomposition. For an observation **y**, the squared distance to group *t* is the following, where *SqDist*[0] and *Discrim Prin Comp* in the last equality are defined in “Saved Formulas” on page 117:

$$
\begin{aligned}
d_t^2 &= (\mathbf{y}-\bar{\mathbf{y}}_t)'\mathbf{S}_p^{-}(\mathbf{y}-\bar{\mathbf{y}}_t) \\
&= (\mathbf{y}-\bar{\mathbf{y}}_t)'\mathbf{T}_s\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}_t) \\
&= ((\mathbf{y}-\bar{\mathbf{y}})-(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'\mathbf{T}_s\mathbf{T}_s'((\mathbf{y}-\bar{\mathbf{y}})-(\bar{\mathbf{y}}_t-\bar{\mathbf{y}})) \\
&= (\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}))-2(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}}))+(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}})) \\
&= SqDist[0]-2(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'\mathit{Discrim\ Prin\ Comp}+(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))'(\mathbf{T}_s'(\bar{\mathbf{y}}_t-\bar{\mathbf{y}}))
\end{aligned}
$$

Saved Formulas

These are the formulas saved by the Wide Linear discriminant method:

Discrim Data Matrix	Vector of observations on the covariates
Discrim Prin Comp	The data transformed by the principal component scoring matrix, which renders the data uncorrelated within groups. Given by $\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}})$, where $\bar{\mathbf{y}}$ is a *p* by *1* vector containing the overall means.
SqDist[0]	$(\mathbf{y}-\bar{\mathbf{y}})'\mathbf{T'}_s\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}})$
SqDist[<group *t*>]	The Mahalanobis distance from the from observation to the group centroid. See “Mahalanobis Distance” on page 116.
Prob[<group *t*>]	$p(t\|\mathbf{y})$, given in “Linear Discriminant Method” on page 111
Pred <*X*>	*t* for which $p(t\|\mathbf{y})$ is maximum, $t = 1, \ldots, T$

Multivariate Tests

In the following, **E** is the residual cross product matrix and **H** is the model cross product matrix. Diagonal elements of **E** are the residual sums of squares for each variable. Diagonal elements of **H** are the sums of squares for the model for each variable. In the discriminant analysis literature, **E** is often called **W**, where **W** stands for *within*.

Test statistics in the multivariate results tables are functions of the eigenvalues λ of $\mathbf{E}^{-1}\mathbf{H}$. The following list describes the computation of each test statistic.

Note: After specification of a response design, the initial **E** and **H** matrices are premultiplied by **M′** and postmultiplied by **M**.

- Wilks' Lambda

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{H}+\mathbf{E})} = \prod_{i=1}^{n}\left(\frac{1}{1+\lambda_i}\right)$$

- Pillai's Trace

$$V = \mathrm{Trace}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}] = \sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_i}$$

- Hotelling-Lawley Trace

$$U = \mathrm{Trace}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^{n}\lambda_i$$

- Roy's Max Root

$\Theta = \lambda_1$, the maximum eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$.

E and **H** are defined as follows:

$$\mathbf{E} = \mathbf{Y'Y} - \mathbf{b'}(\mathbf{X'X})\mathbf{b}$$

$$\mathbf{H} = (\mathbf{Lb})'(\mathbf{L}(\mathbf{X'X})^{-}\mathbf{L}')^{-1}(\mathbf{Lb})$$

where **b** is the estimated vector for the model coefficients and $\mathbf{A}^{-}$ denotes the generalized inverse of a matrix **A**.

The whole model **L** is a column of zeros (for the intercept) concatenated with an identity matrix having the number of rows and columns equal to the number of parameters in the model. **L** matrices for effects are subsets of rows from the whole model **L** matrix.

Approximate F-Tests

To compute *F*-values and degrees of freedom, let p be the rank of $\mathbf{H}+\mathbf{E}$. Let q be the rank of $\mathbf{L}(\mathbf{X'X})^{-1}\mathbf{L}'$, where the **L** matrix identifies elements of $\mathbf{X'X}$ associated with the effect being tested. Let v be the error degrees of freedom and s be the minimum of p and q. Also let $m = 0.5(|p-q|-1)$ and $n = 0.5(v-p-1)$.

Table 5.3 on page 119, gives the computation of each approximate F from the corresponding test statistic.

Table 5.3 Approximate *F*-statistics

Test	Approximate F	Numerator DF	Denominator DF
Wilks' Lambda	$F = \left(\frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\right)\left(\frac{rt-2u}{pq}\right)$	pq	$rt-2u$
Pillai's Trace	$F = \left(\frac{V}{s-V}\right)\left(\frac{2n+s+1}{2m+s+1}\right)$	$s(2m+s+1)$	$s(2n+s+1)$
Hotelling-Lawley Trace	$F = \frac{2(sn+1)U}{s^2(2m+s+1)}$	$s(2m+s+1)$	$2(sn+1)$
Roy's Max Root	$F = \frac{\Theta(v-\max(p,q)+q)}{\max(p,q)}$	$\max(p, q)$	$v-\max(p,q)+q$

Between Groups Covariance Matrix

Using the notation in Table 5.2, this matrix is defined as follows:

$$\mathbf{S}_B = \frac{1}{T-1}\sum_{t=1}^{T} T\left(\frac{n_t}{n}\right)(\bar{\mathbf{y}}_t - \mathbf{y}_{bar})(\bar{\mathbf{y}}_t - \mathbf{y}_{bar})'$$

Chapter 6

Partial Least Squares Models

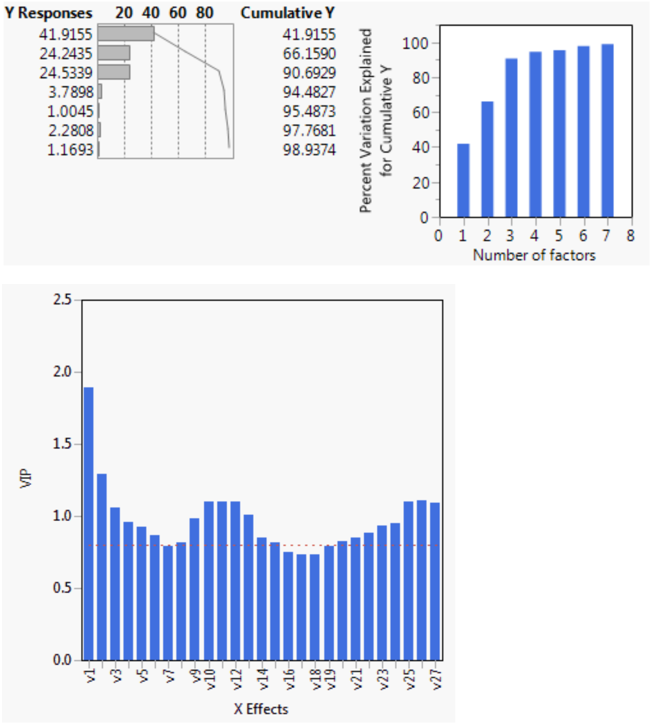
Develop Models Using Correlations between Ys and Xs

The Partial Least Squares (PLS) platform fits linear models based on factors, namely, linear combinations of the explanatory variables (*X*s). These factors are obtained in a way that attempts to maximize the covariance between the *X*s and the response or responses (*Y*s). PLS exploits the correlations between the *X*s and the *Y*s to reveal underlying latent structures.

JMP Pro provides additional functionality, enabling you to conduct PLS Discriminant Analysis (PLS-DA), include a variety of model effects, use several validation methods, impute missing data, and obtain bootstrap estimates of the distributions of various statistics.

Partial least squares performs well in situations such as the following, where the use of ordinary least squares does not produce satisfactory results: More X variables than observations; highly correlated X variables; a large number of X variables; several Y variables and many X variables.

Figure 6.1 A Portion of a Partial Least Squares Report



Y Responses	Cumulative Y
41.9155	41.9155
24.2435	66.1590
24.5339	90.6929
3.7898	94.4827
1.0045	95.4873
2.2808	97.7681
1.1693	98.9374

Contents

Overview of the Partial Least Squares Platform

In contrast to ordinary least squares, PLS can be used when the predictors outnumber the observations. PLS is used widely in modeling high-dimensional data in areas such as spectroscopy, chemometrics, genomics, psychology, education, economics, political science, and environmental science.

The PLS approach to model fitting is particularly useful when there are more explanatory variables than observations or when the explanatory variables are highly correlated. You can use PLS to fit a single model to several responses simultaneously. See Garthwaite (1994), Wold (1994), Wold et al. (2001), Eriksson et al. (2006), and Cox and Gaudard (2013).

Two model fitting algorithms are available: nonlinear iterative partial least squares (NIPALS) and a "statistically inspired modification of PLS" (SIMPLS). For more information about NIPALS, see Wold (1980). For more information about SIMPLS, see De Jong (1993). For a description of both methods, see Boulesteix and Strimmer (2007). The SIMPLS algorithm was developed with the goal of solving a specific optimality problem. For a single response, both methods give the same model. For multiple responses, there are slight differences.

In JMP, the PLS platform is accessible only through Analyze > Multivariate Methods > Partial Least Squares. In JMP Pro, you can also access the Partial Least Squares personality through Analyze > Fit Model.

JMP PRO In JMP Pro, you can do the following:

- Conduct PLS-DA (PLS discriminant analysis) by fitting responses with a nominal modeling type, using the Partial Least Squares personality in Fit Model.
- Fit polynomial, interaction, and categorical effects, using the Partial Least Squares personality in Fit Model.
- Select among several methods for validation and cross validation.
- Impute missing data.
- Obtain bootstrap estimates of the distributions of various statistics. Right-click in the report of interest. For more information about bootstrap estimates, see *Basic Analysis*.

Partial Least Squares uses the van der Voet T^2 test and cross validation to help you choose the optimal number of factors to extract.

- In JMP, the platform uses the leave-one-out method of cross validation. You can also choose not to use validation.
- JMP PRO In JMP Pro, you can choose KFold, Leave-One-Out, or random holdback cross validation, or you can specify a validation column. You can also choose not to use validation.

Example of Partial Least Squares

This example is from spectrometric calibration, which is an area where partial least squares is very effective. Suppose you are researching pollution in the Baltic Sea. You would like to use the spectra of samples of sea water to determine the amounts of three compounds that are present in these samples.

The three compounds of interest are:

- lignin sulfonate (ls), which is pulp industry pollution
- humic acid (ha), which is a natural forest product
- an optical whitener from detergent (dt)

The amounts of these compounds in each of the samples are the responses. The predictors are spectral emission intensities measured at a range of wavelengths (v1–v27).

For the purposes of calibrating the model, samples with known compositions are used. The calibration data consist of 16 samples of known concentrations of lignin sulfonate, humic acid, and detergent. Emission intensities are recorded at 27 equidistant wavelengths. Use the Partial Least Squares platform to build a model for predicting the amount of the compounds from the spectral emission intensities.

1. Select **Help > Sample Data Library** and open Baltic.jmp.

   Note: The data in the Baltic.jmp data table are reported in Umetrics (1995). The original source is Lindberg, Persson, and Wold (1983).

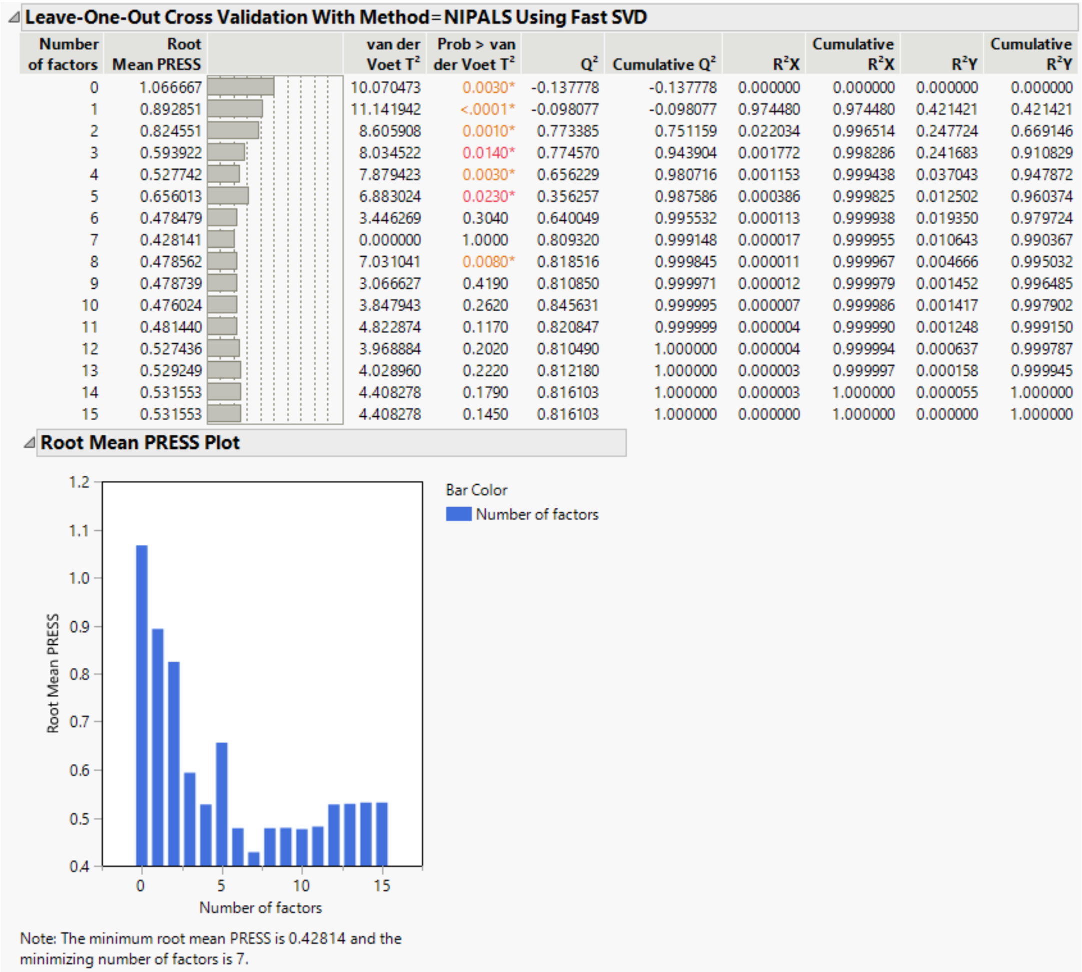
2. Select **Analyze > Multivariate Methods > Partial Least Squares**.
3. Assign ls, ha, and dt to the **Y, Response** role.
4. Assign Intensities, which contains the 27 intensity variables v1 through v27, to the **X, Factor** role.
5. Click **OK**.

   The Partial Least Squares Model Launch control panel appears.

6. Select **Leave-One-Out** as the Validation Method.
7. Click **Go**.

   Since the van der Voet test is a randomization test, your Prob > van der Voet T^2 values may differ slightly.

Figure 6.2 Partial Least Squares Report



Leave-One-Out Cross Validation With Method=NIPALS Using Fast SVD

Number of factors	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2	Q^2	Cumulative Q^2	R^2X	Cumulative R^2X	R^2Y	Cumulative R^2Y
0	1.066667	10.070473	0.0030*	-0.137778	-0.137778	0.000000	0.000000	0.000000	0.000000
1	0.892851	11.141942	<.0001*	-0.098077	-0.098077	0.974480	0.974480	0.421421	0.421421
2	0.824551	8.605908	0.0010*	0.773385	0.751159	0.022034	0.996514	0.247724	0.669146
3	0.593922	8.034522	0.0140*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0030*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0230*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3040	0.640049	0.995532	0.000113	0.999938	0.019350	0.979724
7	0.428141	0.000000	1.0000	0.809320	0.999148	0.000017	0.999955	0.010643	0.990367
8	0.478562	7.031041	0.0080*	0.818516	0.999845	0.000011	0.999967	0.004666	0.995032
9	0.478739	3.066627	0.4190	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2620	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822874	0.1170	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.2020	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2220	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1790	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1450	0.816103	1.000000	0.000000	1.000000	0.000000	1.000000

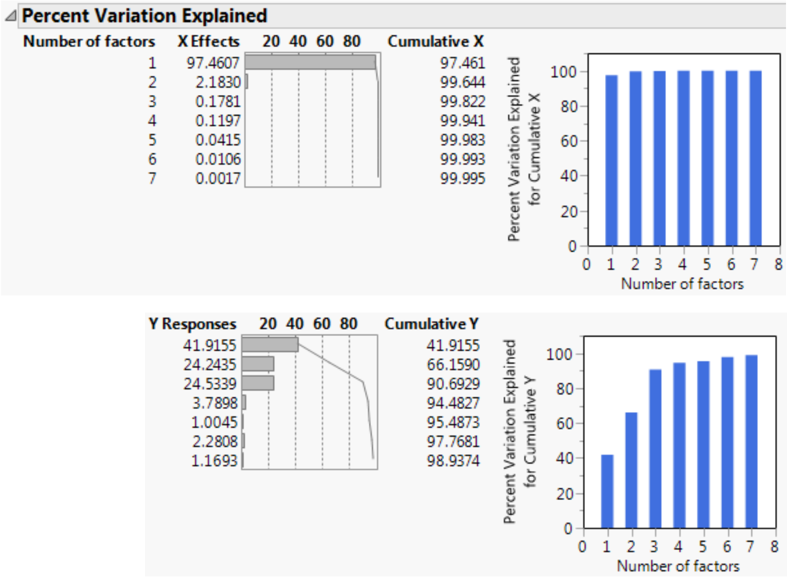
Root Mean PRESS Plot

Note: The minimum root mean PRESS is 0.42814 and the minimizing number of factors is 7.

The Root Mean PRESS (predicted residual sum of squares) Plot shows that Root Mean PRESS is minimized when the number of factors is 7. This is stated in the note beneath the Root Mean PRESS Plot. A report called **NIPALS Fit with 7 Factors Using Fast SVD** is produced. A portion of that report is shown in Figure 6.3.

The van der Voet T^2 statistic tests to determine whether a model with a different number of factors differs significantly from the model with the minimum PRESS value. A common practice is to extract the smallest number of factors for which the van der Voet significance level exceeds 0.10 (SAS Institute Inc 2020f; Tobias 1995). If you were to apply this thinking here, you would fit a new model by entering 6 as the **Number of Factors** in the **Model Launch** panel.

Figure 6.3 Seven Extracted Factors



Percent Variation Explained

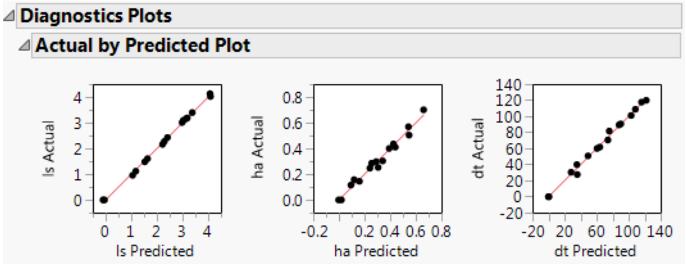
Number of factors	X Effects	Cumulative X
1	97.4607	97.461
2	2.1830	99.644
3	0.1781	99.822
4	0.1197	99.941
5	0.0415	99.983
6	0.0106	99.993
7	0.0017	99.995

Y Responses	Cumulative Y
41.9155	41.9155
24.2435	66.1590
24.5339	90.6929
3.7898	94.4827
1.0045	95.4873
2.2808	97.7681
1.1693	98.9374

8. Click the NIPALS Fit with 7 Factors Using Fast SVD red triangle and select **Diagnostics Plots**.

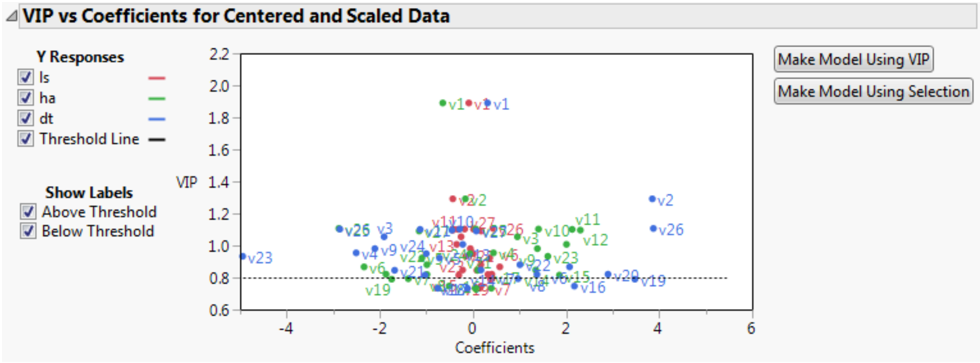
   This gives a report showing actual by predicted plots and three reports showing various residual plots. The Actual by Predicted Plot shows the degree to which predicted compound amounts agree with actual amounts.

Figure 6.4 Diagnostics Plots



9. Click the NIPALS Fit with 7 Factors Using Fast SVD red triangle and select **VIP vs Coefficients Plot**.

Figure 6.5 VIP vs Coefficients Plot



◿ VIP vs Coefficients for Centered and Scaled Data

Y Responses
☑ ls
☑ ha
☑ dt
☑ Threshold Line

Show Labels
☑ Above Threshold
☑ Below Threshold

The VIP vs Coefficients plot helps identify variables that are influential relative to the fit for the various responses. For example, v23, v2, and v26 have both VIP values that exceed 0.8 and relatively large coefficients.

Launch the Partial Least Squares Platform

There are two ways to launch the Partial Least Squares platform:

- Select **Analyze > Multivariate Methods > Partial Least Squares**.
- Select **Analyze > Fit Model** and select **Partial Least Squares** from the Personality menu. This approach enables you to do the following:
  - Enter categorical variables as *Y*s or *X*s. Conduct PLS-DA by entering categorical *Y*s.
  - Add interaction and polynomial terms to your model.
  - Use the Standardize X option to construct higher-order terms using centered and scaled columns.
  - Save your model specification script.

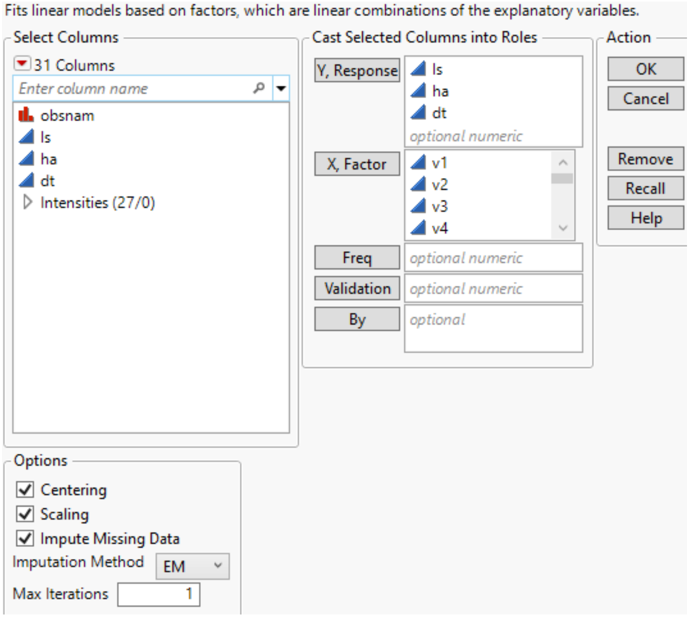
Some features on the Fit Model launch window are not applicable for the Partial Least Squares personality:

- Weight, Nest, Attributes, Transform, and No Intercept.

Tip: You can transform a variable by right-clicking it in the Select Columns box and selecting a Transform option.

- The following Macros: Mixture Response Surface, and Scheffé Cubic.

Figure 6.6 JMP Pro Partial Least Squares Launch Window (Imputation Method EM Selected)

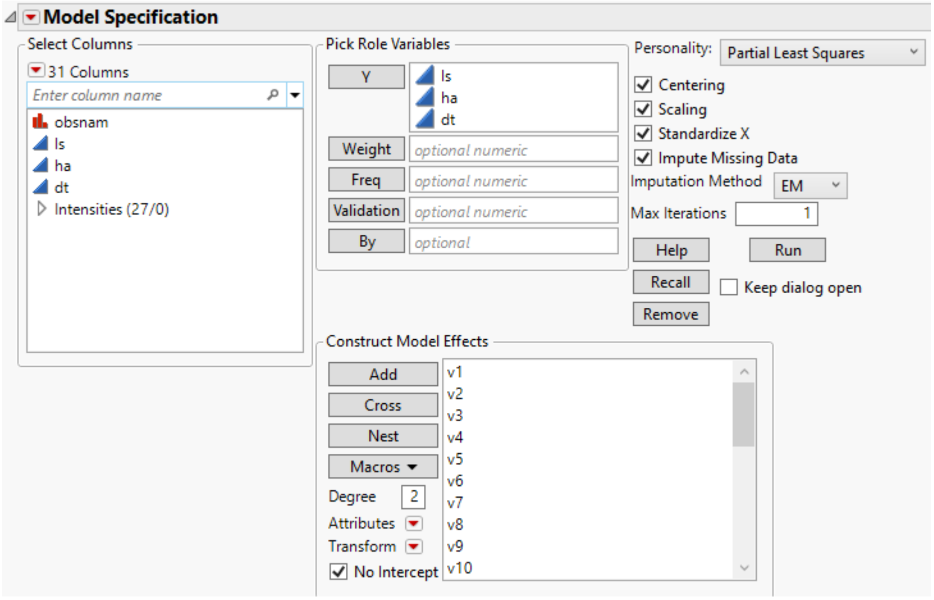

Fits linear models based on factors, which are linear combinations of the explanatory variables.

Y, Response: ls, ha, dt

X, Factor: v1, v2, v3, v4

☑ Centering
☑ Scaling
☑ Impute Missing Data
Imputation Method: EM
Max Iterations: 1

Figure 6.7 JMP Pro Fit Model Partial Least Squares Launch Window


Personality: Partial Least Squares

Y: ls, ha, dt

☑ Centering
☑ Scaling
☑ Standardize X
☑ Impute Missing Data
Imputation Method: EM
Max Iterations: 1
☐ Keep dialog open

Construct Model Effects: v1, v2, v3, v4, v5, v6, v7, v8, v9, v10

Degree: 2
☑ No Intercept

For more information about the Fit Model launch options, see *Fitting Linear Models*.

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Partial Least Squares launch window contains the following options:

Y, Response Enter numeric response columns. If you enter multiple columns, they are modeled jointly.

In JMP Pro, you can enter nominal response columns in the Fit Model launch window to conduct PLS-DA. See “PLS Discriminant Analysis (PLS-DA)” on page 153.

X, Factor Enter the predictor columns. The Partial Least Squares launch window allows only numeric predictors.

In JMP Pro, you can enter nominal and ordinal model effects in the Fit Model launch window. Ordinal effects are treated as nominal.

Freq If your data are summarized, enter the column whose values contain counts for each row.

Validation A numeric column that defines the validation sets. A validation column must contain only consecutive integer values:

- If the validation column has two levels, the smaller value defines the training set and the larger value defines the validation set.
- If the validation column has three levels, the values define the training, validation, and test sets in order of increasing size.
- If the validation column has more than three levels, then KFold Cross Validation is used. For information about other validation options, see “Validation Method” on page 132.

The PLS platform uses the validation column to train and tune the model or to train, tune, and evaluate the model. For more information about validation, see “Validation in JMP Modeling” on page 541 in the “Statistical Details” chapter.

Note: If you click the Validation button with no columns selected in the Select Columns list, you can add a validation column to your data table. For more information about the Make Validation Column utility, the *Predictive and Specialized Modeling*.

By Enter a column that creates a report consisting of separate analyses for each level of the variable. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

Centering Centers all Y variables and model effects by subtracting the mean from each column. See “Centering and Scaling” on page 131.

Scaling Scales all Y variables and model effects by dividing each column by its standard deviation. See “Centering and Scaling” on page 131.

JMP PRO **Standardize X** (Available only in the Fit Model launch window.) Centers and scales all columns that are used in the construction of model effects. If this option is not selected, higher-order effects are constructed using the original data table columns. Then each higher-order effect is centered or scaled, based on the selected Centering and Scaling options. Note that Standardize X does not center or scale Y variables. See "Standardize X" on page 131.

JMP PRO **Impute Missing Data** Replaces missing data values in Ys or Xs with nonmissing values. Select the appropriate method from the **Imputation Method** list.

If **Impute Missing Data** is not selected, rows that are missing observations on any X variable are excluded from the analysis and no predictions are computed for these rows. Rows with no missing observations on X variables but with missing observations on Y variables are also excluded from the analysis, but predictions are computed.

JMP PRO **Imputation Method** (Appears only when **Impute Missing Data** is selected.) Select from the following imputation methods:

Mean For each model effect or response column, replaces the missing value with the mean of the nonmissing values.

EM Uses an iterative Expectation-Maximization (EM) approach to impute missing values. On the first iteration, the specified model is fit to the data with missing values for an effect or response replaced by their means. Predicted values from the model for Y and the model for X are used to impute the missing values. For subsequent iterations, the missing values are replaced by their predicted values, given the conditional distribution using the current estimates.

For the purpose of imputation, polynomial terms are treated as separate predictors. When a polynomial term is specified, that term is calculated from the original data, or, if Standardize X is checked, from the standardized column values. If a row has a missing value for a column involved in the definition of the polynomial term, then that entry is missing for the polynomial term. Imputation is conducted using polynomial terms defined in this way.

For more information about the EM approach, see Nelson, Taylor, and MacGregor (1996).

JMP PRO **Max Iterations** (Appears only when EM is selected as the Imputation Method.) Enables you to set the maximum number of iterations used by the algorithm. The algorithm terminates if the maximum difference between the current and previous estimates of missing values is bounded by 10^{-8}.

After completing the launch window and clicking **OK**, the Model Launch control panel appears. See "Model Launch Control Panel" on page 132.

Centering and Scaling

The Centering and Scaling options are selected by default. This means that predictors and responses are centered and scaled to have mean 0 and standard deviation 1. Centering the predictors and the responses places them on an equal footing relative to their variation. Without centering, both the variable's mean and its variation around that mean are involved in constructing successive factors. To illustrate, suppose that Time and Temp are two of the predictors. Scaling them indicates that a change of one standard deviation in Time is approximately equivalent to a change of one standard deviation in Temp.

Standardize X

When the Partial Least Square personality is selected in the Fit Model window, the Standardize X option is selected by default. This ensures that all columns entered as model effects and that all columns that are involved in an interaction or polynomial term are standardized.

Suppose that you have two columns, X1 and X2, and you enter the interaction term X1*X2 as a model effect in the Fit Model window. When the Standardize X option is selected, both X1 and X2 are centered and scaled before forming the interaction term. The interaction term that is formed is calculated as follows:

$$\left(\frac{X1 - mean(X1)}{std(X1)}\right) \times \left(\frac{X2 - mean(X2)}{std(X2)}\right)$$

All model effects are then centered or scaled, in accordance with your selections of the **Centering** and **Scaling** options, prior to inclusion in the model.

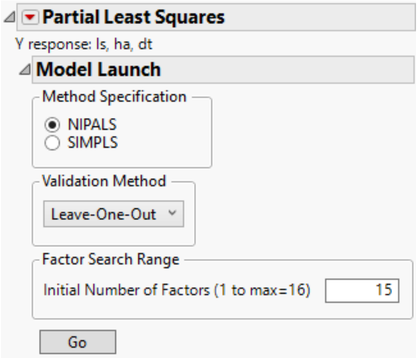
If the Standardize X option is not selected, and **Centering** and **Scaling** are both selected, then the term that is entered into the model is calculated as follows:

$$\frac{X1 \times X2 - mean(X1 \times X2)}{std(X1 \times X2)}$$

Model Launch Control Panel

After you click **OK** in the platform launch window (or **Run** in the Fit Model window), the Model Launch control panel appears.

Figure 6.8 Partial Least Squares Model Launch Control Panel



Partial Least Squares
Y response: ls, ha, dt
Model Launch
Method Specification
● NIPALS
○ SIMPLS
Validation Method
Leave-One-Out
Factor Search Range
Initial Number of Factors (1 to max=16) 15

Note: The Validation Method portion of the Model Launch control panel appears differently in JMP Pro.

The Model Launch control panel contains the following selections:

Method Specification Select the type of model fitting algorithm. There are two algorithm choices: **NIPALS** and **SIMPLS**. The two methods produce the same coefficient estimates when there is only one response variable. See "Statistical Details for the Partial Least Squares Platform" on page 148 for more information about differences between the two algorithms.

Validation Method Select the validation method. Validation is used to determine the optimum number of factors to extract. For JMP Pro, if a validation column is specified on the platform launch window, these options do not appear.

Holdback Randomly selects the specified proportion of the data for a validation set, and uses the other portion of the data to fit the model. The random selection is based on stratified sampling across the model factors to attempt to create training and validation sets that are more balanced than ones based on simple random sampling.

KFold Partitions the data into *K* subsets, or *folds*. In turn, each fold is used to validate the model that is fit to the rest of the data, fitting a total of *K* models. This method is best for small data sets because it makes efficient use of limited amounts of data.

Leave-One-Out Performs leave-one-out cross validation.

None Does not use validation to choose the number of factors to extract. The number of factors is specified in the Factor Search Range.

Factor Search Range Specifies how many latent factors to extract if not using validation. If validation is being used, this is the maximum number of factors the platform attempts to fit before choosing the optimum number of factors. The maximum number of factors is the minimum of the number of non-missing rows and the number of factors. Then, the initial number of factors is the minimum of 15 and the maximum number of factors.

Factor Specification Appears once you click **Go** to fit an initial model. Specify a number of factors to be used in fitting a new model.

Go Launches the model fit with the given specifications.

Partial Least Squares Options

The Partial Least Squares red triangle menu contains some options that are useful before clicking Go in the Model Launch Control Panel.

JMP® PRO **Set Random Seed** Sets the seed for the randomization process used for KFold and Holdback validation. This is useful if you want to reproduce an analysis. Set the seed to a positive value, save the script, and the seed is automatically saved in the script. Running the script always produces the same cross validation analysis. This option does not appear when Validation Method is set to None, or when a validation column is used.

JMP® PRO **SVD** Shows a submenu that enables you to choose between a fast or classical implementation of the SVD algorithm in the model fitting algorithm. The default is Fast SVD.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Partial Least Squares Report

The first time you click **Go** in the Model Launch control panel (Figure 6.8), the Validation Method panel is removed from the Model Launch window. If you specified a Validation column or if you selected Holdback in the Validation Method panel, all model fits in the report are based on the training data. Otherwise, all model fits are based on the entire data set.

If you used validation, three reports appear:

- Model Comparison Summary
- Cross Validation Report
- NIPALS (or SIMPLS) Fit with <N> Factors

If you selected **None** as the CV method, two reports appear:

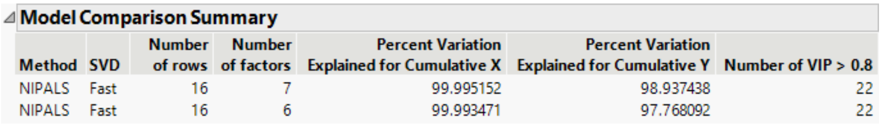
- Model Comparison Summary
- NIPALS (or SIMPLS) Fit with <N> Factors

To fit additional models, specify the desired numbers of factors in the Model Launch panel.

Model Comparison Summary

The Model Comparison Summary shows summary results for each fitted model.

Figure 6.9 Model Comparison Summary



Model Comparison Summary

Method	SVD	Number of rows	Number of factors	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
NIPALS	Fast	16	7	99.995152	98.937438	22
NIPALS	Fast	16	6	99.993471	97.768092	22

The report includes the following summary information:

Method Shows the analysis method that you specified in the Model Launch control panel.

SVD Shows the implementation of the SVD algorithm that you specified.

Number of rows Shows the number of observations used in the training set.

Number of factors Shows the number of extracted factors.

Percent Variation Explained for Cumulative X Shows the percent of variation in X that is explained by the model.

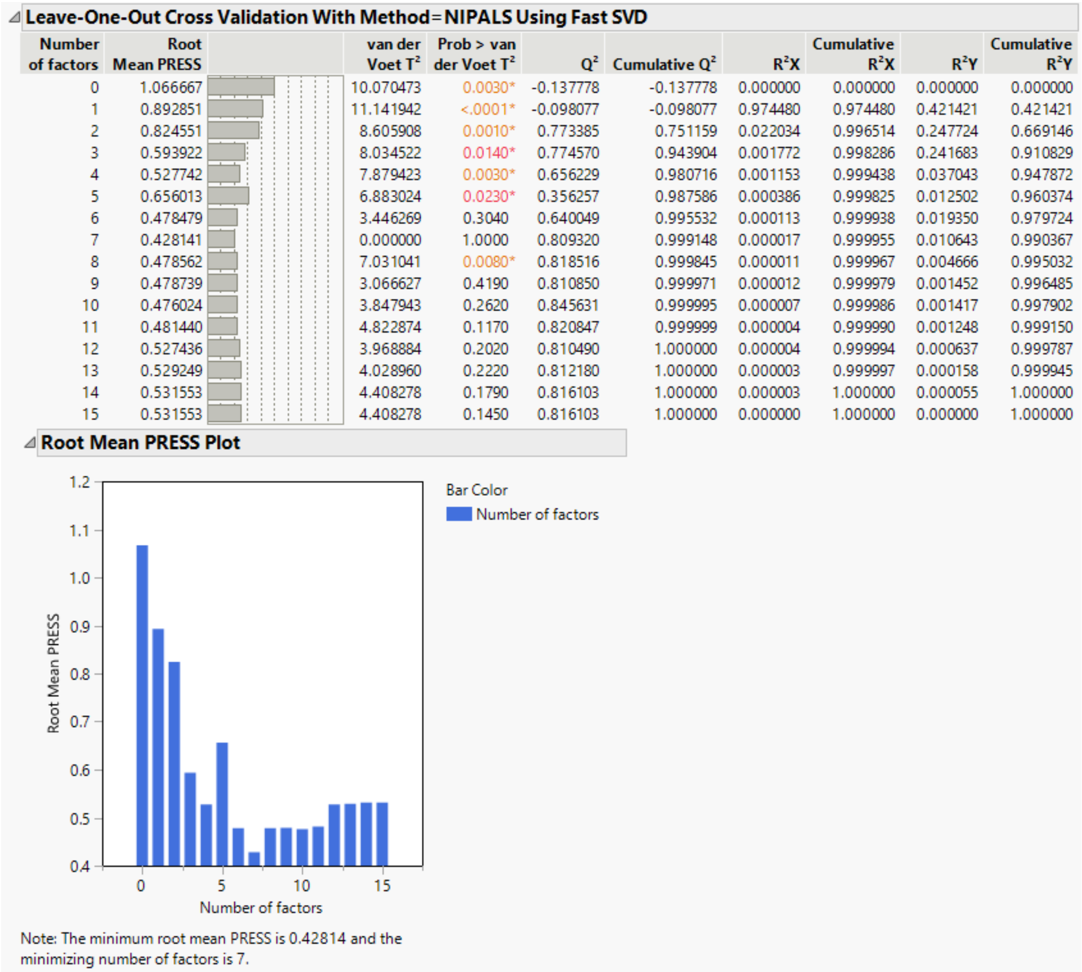
Percent Variation Explained for Cumulative Y Shows the percent of variation in Y that is explained by the model.

Number of VIP>0.8 Shows the number of model effects with VIP (variable importance for projection) values greater than 0.8. The VIP score is a measure of a variable's importance relative to modeling both X and Y (Wold 1994; Eriksson et al. 2006).

Cross Validation Report

This report appears only when a form of cross validation is selected as a Validation Method in the Model Launch control panel. The report title is dynamically named <Cross Validation Method> With <Method = Method Specification>, depending on the cross validation and method options selected in the control panel. It shows summary statistics for models fit, using from 0 to the maximum number of extracted factors, as specified in the Model Launch control panel. The report also provides a plot of Root Mean PRESS values. See "Root Mean PRESS Plot" on page 138. An optimum number of factors is identified using the minimum Root Mean PRESS statistic.

Figure 6.10 Cross Validation Report



Leave-One-Out Cross Validation With Method=NIPALS Using Fast SVD

Number of factors	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2	Q^2	Cumulative Q^2	R^2X	Cumulative R^2X	R^2Y	Cumulative R^2Y
0	1.066667	10.070473	0.0030*	-0.137778	-0.137778	0.000000	0.000000	0.000000	0.000000
1	0.892851	11.141942	<.0001*	-0.098077	-0.098077	0.974480	0.974480	0.421421	0.421421
2	0.824551	8.605908	0.0010*	0.773385	0.751159	0.022034	0.996514	0.247724	0.669146
3	0.593922	8.034522	0.0140*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0030*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0230*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3040	0.640049	0.995532	0.000113	0.999938	0.019350	0.979724
7	0.428141	0.000000	1.0000	0.809320	0.999148	0.000017	0.999955	0.010643	0.990367
8	0.478562	7.031041	0.0080*	0.818516	0.999845	0.000011	0.999967	0.004666	0.995032
9	0.478739	3.066627	0.4190	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2620	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822874	0.1170	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.2020	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2220	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1790	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1450	0.816103	1.000000	0.000000	1.000000	0.000000	1.000000

Root Mean PRESS Plot

Note: The minimum root mean PRESS is 0.42814 and the minimizing number of factors is 7.

JMP PRO When the **Standardize X** option is selected, the standardization is applied once to the entire data table. It is not reapplied to the individual training sets. However, when any combination of the **Centering** or **Scaling** options are selected, this combination of selections is applied to each cross validation training set. Cross validation proceeds by using the training sets, which are individually centered and scaled if these options are selected.

The following statistics are shown in the report. If any form of validation or cross validation is used, the reported results are summaries of the training set statistics.

Number of Factors The number of factors used in fitting the model.

Root Mean PRESS The square root of the average of the PRESS values across all responses. See "Root Mean PRESS" on page 139.

van der Voet T^2 The test statistic for the van der Voet test, which tests whether models with different numbers of extracted factors differ significantly from the optimum model. The null hypothesis for each van der Voet T^2 test states that the model based on the corresponding number of factors does not differ from the optimum model. The alternative hypothesis is that the model does differ from the optimum model. See “van der Voet T^2” on page 149.

Prob > van der Voet T^2 The *p*-value for the van der Voet T^2 test. See “van der Voet T^2” on page 149.

Q^2 Dimensionless measure of predictive ability defined by subtracting the ratio of the PRESS value divided by the total sum of squares for Y from one:

$$1 - PRESS/SSY$$

See “Calculation of Q^2” on page 139.

Cumulative Q^2 Indicator of the predictive ability of models with the given number of factors or fewer. For a given number of factors, *f*, Cumulative Q^2 is defined as follows:

$$1 - \prod_{i=1}^{f} (Press_i/SYY_i)$$

Here $PRESS_i$ and SSY_i correspond to their values for *i* factors.

R^2X Percent of X variation explained by the specified factor. A component with a large R^2X explains a large amount of the variation in the X variables. See “Calculation of $R^{2X \text{ and } R2Y}$ When Validation Is Used” on page 140.

Cumulative R^2X Percent of X variation explained by the model with the given number of factors. This is the sum of the R^2X values for $i = 1$ to the given number of factors.

R^2Y Percent of Y variation explained by the specified factor. A component with a large R^2Y explains a large amount of the variation in the Y variables. See “Calculation of $R^{2X \text{ and } R2Y}$ When Validation Is Used” on page 140.

Cumulative R^2Y Percent of Y variation explained by the model with the given number of factors. This is the Sum of the R^2Y values for $i = 1$ to the given number of factors.

Interpretation of Q^2 and Cumulative R^2Y

The statistics Q^2 and Cumulative R^2Y both measure the predictive ability of the model, but in different ways.

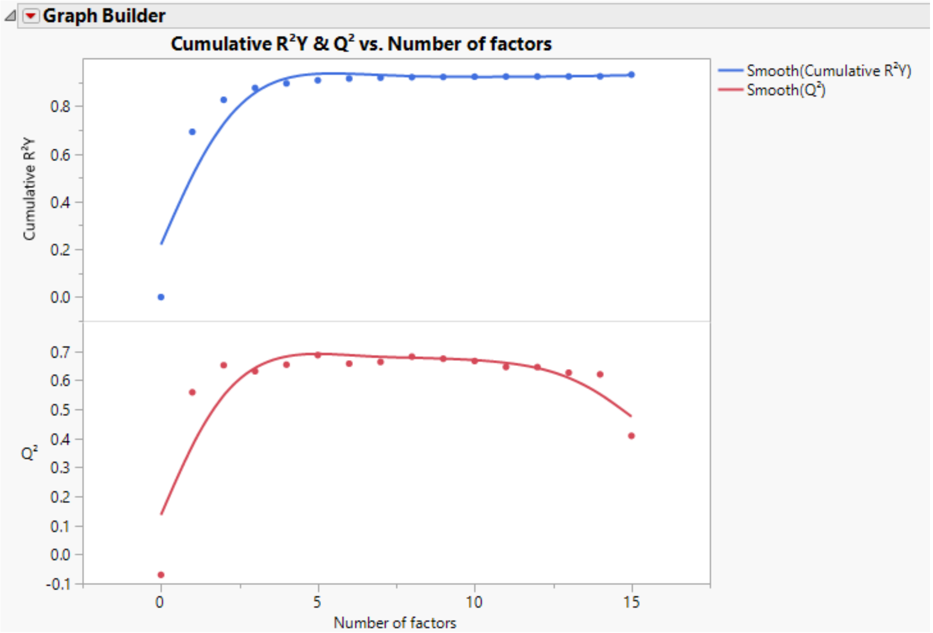
- Cumulative R^2Y increases as the number of factors increases. This is because, as factors are added to the model, more variation is explained.

- Q^2 tends to increase and then decrease, or at least discontinue increasing, as the number of factors increases. This is because, as more factors are added, the model becomes tuned to the training set and does not generalize well to new data, causing the PRESS statistic to decrease.

Analysis of Q^2 and Cumulative R^2Y provides an alternative to using the van der Voet test for determining how many factors to include in your model. Select a number of factors for which Q^2 is large and has not started decreasing. You also want Cumulative R^2Y to be large.

Figure 6.11 shows plots of Cumulative R^2Y and Q^2 against the number of factors for the Penta.jmp data table, using Leave-One-Out as the validation method. Cumulative R^2Y increases and levels off for about four factors. The statistic Q^2 is largest for two factors and then begins to level off. The plot suggests that a model with two factors explains a large portion of the variation in Y without overfitting the data.

Figure 6.11 Cumulative R^2Y and Q^2 for Penta.jmp



Root Mean PRESS Plot

This bar chart shows the number of factors along the horizontal axis and the Root Mean PRESS values on the vertical axis. It is equivalent to the horizontal bar chart that appears to the right of the Root Mean PRESS column in the Cross Validation report (Figure 6.10).

Root Mean PRESS

For a specified number of factors, *a*, the Root Mean PRESS is calculated using the following steps:

1. Fit a model with *a* factors to each training set.
2. Apply the resulting prediction formula to the observations in the validation set.
3. For each Y:
   - For each validation set, compute the squared difference between each observed validation set value and its predicted value (the squared prediction error).
   - For each validation set, average these squared differences and divide the result by a variance estimate for the response. For the KFold and Leave-One-Out validation methods, divide by the variance of the entire response column. For Holdback validation, divide by the variance of the response values in the training set.
   - Sum these means and, in the case of more than one validation set, divide their sum by the number of validation sets minus one. This is the PRESS statistic for the given Y.
4. Root Mean PRESS for *a* factors is the square root of the average of the PRESS values across all responses.
5. The PRESS statistic for multiple Ys is obtained by averaging the PRESS statistic, obtained in step 3, across all responses.

Calculation of Q^2

The statistic Q^2 is defined as 1 - *PRESS* / *SSY*. The *PRESS* statistic is the predicted error sum of squares averaged across all responses for the model developed based on the training data, but evaluated on the validation set. The value of *SSY* is the sum of squares for Y averaged across all responses and based on the observations in the validation set.

The statistic Q^2 in the Cross Validation report is computed in the following ways, depending on the selected Validation Method:

Leave-One-Out Q^2 is the average of the values 1 - *PRESS* / *SSY* computed for the validation sets based on the models constructed by leaving out one observation at a time.

KFold Q^2 is the average of the values 1 - *PRESS* / *SSY* computed for the validation sets based on the *K* models constructed by leaving out each of the *K* folds.

Holdback or Validation Set Q^2 is the value of 1 - *PRESS* / *SSY* computed for the validation set based on the model constructed using the single set of training data.

Calculation of R^2X and R^2Y When Validation Is Used

The statistics R^2X and R^2Y in the Cross Validation report are computed in the following ways, depending on the selected Validation Method:

Note: For all of these computations, R^2Y is calculated analogously.

Leave-One-Out R^2X is the average of the Percent Variation Explained for X Effects for the models constructed by leaving out one observation at a time.

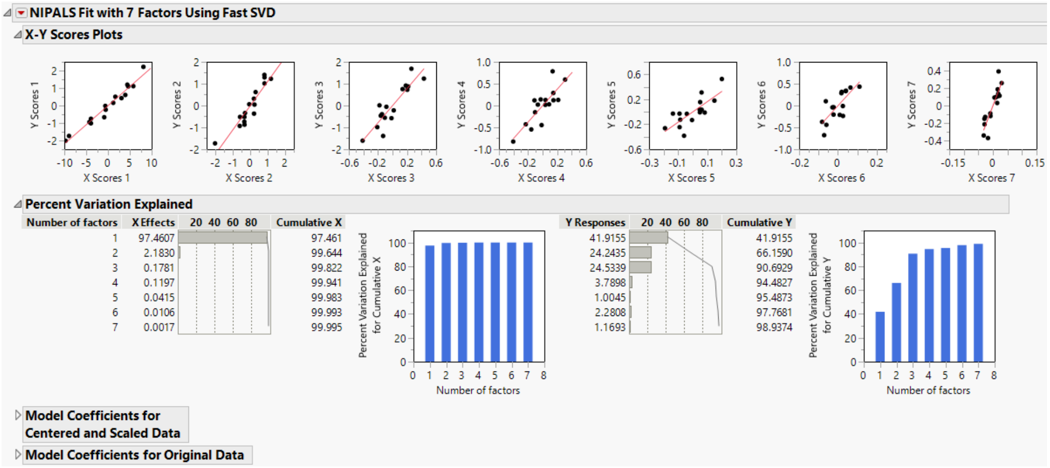
KFold R^2X is the average of the Percent Variation Explained for X Effects for the K models constructed by leaving out each fold.

Holdback or Validation Set R^2X is the Percent Variation Explained for X Effects for the model constructed using the training data.

Model Fit Report

The Model Fit Report shows detailed results for each fitted model. The fit uses either the optimum number of factors based on cross validation, or the specified number of factors if no cross validation methods are specified. The report title indicates whether NIPALS or SIMPLS was used, whether Fast SVD or Classical SVD was used, and gives the number of extracted factors.

Figure 6.12 Model Fit Report



Number of factors	X Effects	Cumulative X	Y Responses	Cumulative Y
1	97.4607	97.461	41.9155	41.9155
2	2.1830	99.644	24.2435	66.1590
3	0.1781	99.822	24.5339	90.6929
4	0.1197	99.941	3.7898	94.4827
5	0.0415	99.983	1.0045	95.4873
6	0.0106	99.993	2.2808	97.7681
7	0.0017	99.995	1.1693	98.9374

The Model Fit report includes the following summary information:

X-Y Scores Plots Scatterplots of the X and Y scores for each extracted factor.

Percent Variation Explained Shows the percent variation and cumulative percent variation explained for both X and Y. Results are given for each extracted factor.

Model Coefficients for Centered and Scaled Data For each Y, shows the coefficients of the Xs for the model based on the centered and scaled data.

Model Fit Options

The Model Fit red triangle menu contains the following options:

Percent Variation Plots Adds two plots entitled Percent Variation Explained for X Effects and Percent Variation Explained for Y Effects. These show stacked bar charts representing the percent variation explained by each extracted factor for the *X*s and *Y*s.

Variable Importance Plot Plots the VIP values for each X variable. VIP scores appear in the Variable Importance Table. See "Variable Importance Plot" on page 143.

VIP vs Coefficients Plots Plots the VIP statistics against the model coefficients. You can show only those points corresponding to your selected *Y*s. Additional labeling options are provided. There are plots for both the centered and scaled data and the original data. See "VIP vs Coefficients Plots" on page 144.

Set VIP Threshold Sets the threshold level for the Variable Importance Plot, Variance Importance Table, and the VIP vs Coefficients Plots.

Coefficient Plots Plots the model coefficients for each response across the X variables. You can show only those points corresponding to your selected *Y*s. There are plots for both the centered and scaled data and the original data.

Loading Plots Plots X and Y loadings for each extracted factor. There are separate plots for the *X*s and *Y*s.

Loading Scatterplot Matrices Shows scatterplot matrices of the X loadings and the Y loadings.

Correlation Loading Plot Shows either a single scatterplot or a scatterplot matrix of the X and Y loadings overlaid on the same plot. When you select the option, you specify how many factors you want to plot.

- If you specify two factors, a single correlation loading scatterplot appears. Select the two factors that define the axes beneath the plot. Click the right arrow button to successively display each combination of factors on the plot.

- If you specify more than two factors, a scatterplot matrix appears with a cell for pair of factors up to the number that you selected.

In both cases, use check boxes to control labeling.

X-Y Score Plots Includes the following options:

Fit Line Shows or hides a fitted line through the points on the X-Y Scores Plots.

Show Confidence Band Shows or hides 95% confidence bands for the fitted lines on the X-Y Scores Plots.

Score Scatterplot Matrices Shows a scatterplot matrix of the X scores and a scatterplot matrix of the Y scores. Each X score scatterplot displays a 95% confidence ellipse, which can be used for outlier detection. For statistical details about the confidence ellipses, see "Confidence Ellipses for X Score Scatterplot Matrix" on page 150.

Distance Plots Shows plots of the following:

- the distance from each observation to the X model
- the distance from each observation to the Y model
- a scatterplot of distances to both the X and Y models

In a good model, both X and Y distances are small, so the points are close to the origin (0,0). Use the plots to look for outliers relative to either X or Y. If a group of points clusters together, then they might have a common feature and could be analyzed separately. When a validation set or a validation and test set are in use, separate reports are provided for these sets and for the training set.

T Square Plot Shows a plot of T^2 statistics for each observation, along with a control limit. An observation's T^2 statistic is calculated based on that observation's scores on the extracted factors. For more information about the computation of T^2 and the control limit, see "$T^{2\ \text{Plot}}$" on page 150.

Diagnostics Plots Shows diagnostic plots for assessing the model fit. Four plot types are available: Actual by Predicted Plot, Residual by Predicted Plot, Residual by Row Plot, and a Residual Normal Quantile Plot. Plots are provided for each response. When a validation set or a validation and test set are in use, separate reports are provided for these sets and for the training set.

Profiler Shows a profiler for each Y variable.

Spectral Profiler Shows a single profiler where all of the response variables appear in the first cell of the plot. This profiler is useful for visualizing the effect of changes in the X variables on the Y variables simultaneously.

Save Columns Includes options for saving various formulas and results. See "Save Columns" on page 144.

Remove Fit Removes the model report from the main platform report.

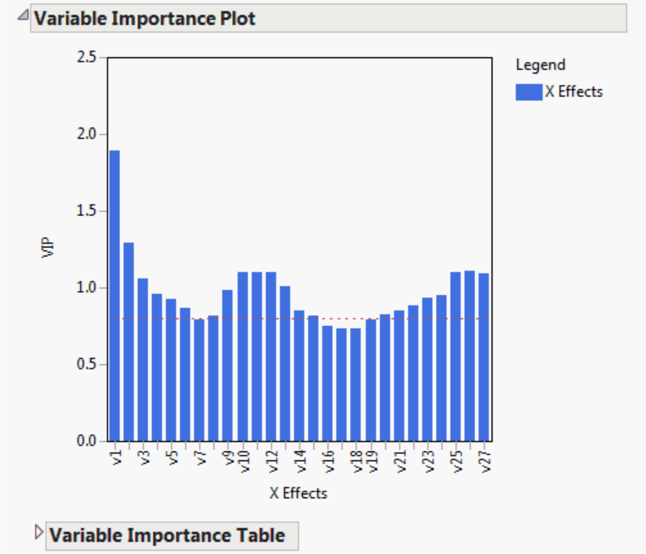
Make Model Using VIP Opens and populates a launch window with the appropriate responses entered as *Y*s and the variables whose VIPs exceed the specified threshold entered as *X*s. Performs the same function as the button in the VIP vs Coefficients for Centered and Scaled Data report. See "VIP vs Coefficients Plots" on page 144.

Model Driven Multivariate Control Chart for Saved X Scores Saves the formulas for each X Score and launches the Model Driven Multivariate Control Chart (MDMCC) launch window. In the MDMCC launch window, the score formulas are assigned as the process columns. You can add or remove processes, add a time ID, or set where historical data ends before clicking OK. See the "Model Driven Multivariate Control Charts" chapter on page 303.

Variable Importance Plot

The Variable Importance Plot graphs the VIP values for each X variable. The Variable Importance Table shows the VIP scores. A VIP score is a measure of a variable's importance in modeling both X and Y. If a variable has a small coefficient and a small VIP, then it is a candidate for deletion from the model (Wold 1994). A value of 0.8 is generally considered to be a small VIP (Eriksson et al. 2006) and a red dashed line is drawn on the plot at 0.8.

Figure 6.13 Variable Importance Plot



VIP vs Coefficients Plots

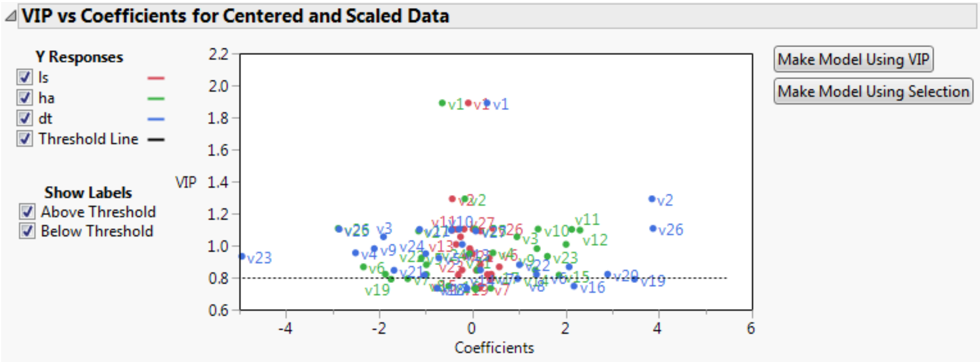
Two options to the right of the plot facilitate variable reduction and model building:

Make Model Using VIP Opens and populates a launch window with the appropriate responses entered as *Y*s and the variables whose VIPs exceed the specified threshold entered as *X*s.

Make Model Using Selection Enables you to select *X*s directly in the plot and then enters the *Y*s and only the selected *X*s into a launch window.

To use another platform based on your current column selection, open the desired platform. Notice in the launch window that the selections are retained. Click the role button and the selected columns are populated.

Figure 6.14 VIP vs Coefficients Plot for Centered and Scaled Data



Save Columns

Save Prediction Formula For each Y variable, saves a column to the data table called Pred Formula <response> that contains the prediction formula.

Save Prediction as X Score Formula For each Y variable, saves a column to the data table called Pred Formula <response> that contains the prediction formula in terms of the X scores.

Save Standard Errors of Prediction Formula For each Y variable, saves a column to the data table called PredSE <response> that contains the standard error of the predicted mean. See "Standard Error of Prediction and Confidence Limits" on page 151.

Save Mean Confidence Limit Formula For each Y variable, saves two columns to the data table called Lower 95% Mean <response> and Upper 95% Mean <response>. These columns contain 95% confidence limits for the response mean. See "Standard Error of Prediction and Confidence Limits" on page 151.

Save Indiv Confidence Limit Formula For each Y variable, saves two columns to the data table called Lower 95% Indiv <response> and Upper 95% Indiv <response>. These columns contain 95% prediction limits for individual values. See "Standard Error of Prediction and Confidence Limits" on page 151.

Save Score Formula Saves two sets of columns to the data table:

- Columns called X Score <N> Formula containing the formulas for each X Score.
- Columns called Y Score <N> Formula containing the formulas for each Y Score

See "Partial Least Squares" on page 148.

Save Y Predicted Values Saves the predicted values for the Y variables to columns in the data table.

Save Y Residuals Saves the residual values for the Y variables to columns in the data table.

Save X Predicted Values Saves the predicted values for the X variables to columns in the data table.

Save X Residuals Saves the residual values for the X variables to columns in the data table.

Save Percent Variation Explained For X Effects Saves the percent variation explained for each X variable across all extracted factors to a new data table.

Save Percent Variation Explained For Y Responses Saves the percent variation explained for each Y variable across all extracted factors to a new data table.

Save Scores Saves the X and Y scores for each extracted factor to the data table.

Save Loadings Saves the X and Y loadings to new data tables.

Save Standardized Scores Saves the X and Y standardized scores used in constructing the Correlation Loading Plot to the data table. For the formulas, see "Standardized Scores and Loadings" on page 152.

Save Standardized Loadings Saves the X and Y standardized loadings used in constructing the Correlation Loading Plot to new data tables. For the formulas, see "Standardized Scores and Loadings" on page 152.

Save T Square Saves a column to the data table that contains the T^2 formulas. The values in this column are also used in the T Square Plot.

Save Distance Saves the Distance to X Model (DModX) and Distance to Y Model (DModY) values to the data table. These are the values used in the Distance Plots.

Save X Weights Saves the weights for each X variable across all extracted factors to a new data table.

Save Validation Saves a new column to the data table describing how each observation was used in validation. For Holdback validation, the column identifies if a row was used for training or validation. For KFold validation, the column identifies the number of the subgroup to which the row was assigned.

Save Imputation If Impute Missing Data is selected, opens a new data table that contains the data table columns specified as X and Y, with missing values replaced by their imputed values. Columns for polynomial terms are not shown. If a Validation column is specified, the validation column is also included.

Publish Prediction Formula Creates prediction formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See *Predictive and Specialized Modeling*.

Publish Score Formula Creates X and Y score formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See *Predictive and Specialized Modeling*.

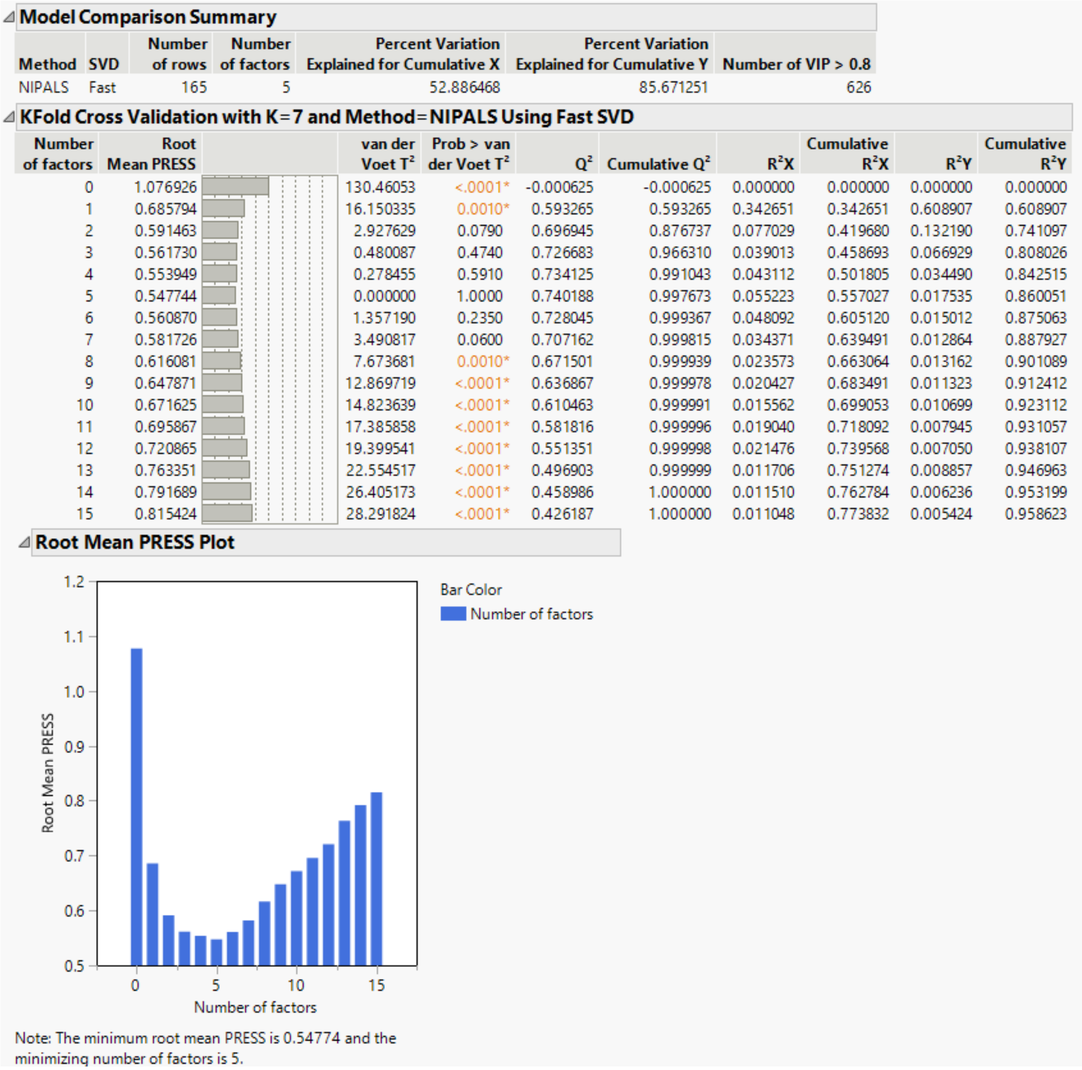
Additional Example of Partial Least Squares

In this example, you will use the Partial Least Squares personality in Fit Model to perform an analysis for a categorical response and a large number of variables. Using Partial Least Squares through Fit Model also enables you to use K Fold Crossvalidation. The example data set consists of serum samples collected from 165 men. 84 of the men have prostate cancer and 81 of the men do not. The goal of the analysis is to be able to determine which men have prostate cancer based on the serum samples.

1. Select **Help > Sample Data Library** and open Prostate Cancer.jmp.
2. Select **Analyze > Fit Model**.
3. Select Status and click **Y**.
4. Select the Proteins column group and click **Add**.
5. From the Personality list, select **Partial Least Squares**.
6. Click **Run**.
7. (Optional) Click the Partial Least Squares red triangle and select **Set Random Seed**.
8. (Optional) Next to Specify random seed, enter 1234.

   Specifying a random seed enables you to reproduce the results shown in this example.
9. (Optional). Click **OK**.
10. In the Model Launch, click **Go**.

Figure 6.15 Partial Least Squares Report



Model Comparison Summary

Method	SVD	Number of rows	Number of factors	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
NIPALS	Fast	165	5	52.886468	85.671251	626

KFold Cross Validation with K=7 and Method=NIPALS Using Fast SVD

Number of factors	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2	Q^2	Cumulative Q^2	R^2X	Cumulative R^2X	R^2Y	Cumulative R^2Y
0	1.076926	130.46053	<.0001*	-0.000625	-0.000625	0.000000	0.000000	0.000000	0.000000
1	0.685794	16.150335	0.0010*	0.593265	0.593265	0.342651	0.342651	0.608907	0.608907
2	0.591463	2.927629	0.0790	0.696945	0.876737	0.077029	0.419680	0.132190	0.741097
3	0.561730	0.480087	0.4740	0.726683	0.966310	0.039013	0.458693	0.066929	0.808026
4	0.553949	0.278455	0.5910	0.734125	0.991043	0.043112	0.501805	0.034490	0.842515
5	0.547744	0.000000	1.0000	0.740188	0.997673	0.055223	0.557027	0.017535	0.860051
6	0.560870	1.357190	0.2350	0.728045	0.999367	0.048092	0.605120	0.015012	0.875063
7	0.581726	3.490817	0.0600	0.707162	0.999815	0.034371	0.639491	0.012864	0.887927
8	0.616081	7.673681	0.0010*	0.671501	0.999939	0.023573	0.663064	0.013162	0.901089
9	0.647871	12.869719	<.0001*	0.636867	0.999978	0.020427	0.683491	0.011323	0.912412
10	0.671625	14.823639	<.0001*	0.610463	0.999991	0.015562	0.699053	0.010699	0.923112
11	0.695867	17.385858	<.0001*	0.581816	0.999996	0.019040	0.718092	0.007945	0.931057
12	0.720865	19.399541	<.0001*	0.551351	0.999998	0.021476	0.739568	0.007050	0.938107
13	0.763351	22.554517	<.0001*	0.496903	0.999999	0.011706	0.751274	0.008857	0.946963
14	0.791689	26.405173	<.0001*	0.458986	1.000000	0.011510	0.762784	0.006236	0.953199
15	0.815424	28.291824	<.0001*	0.426187	1.000000	0.011048	0.773832	0.005424	0.958623

Root Mean PRESS Plot

Note: The minimum root mean PRESS is 0.54774 and the minimizing number of factors is 5.

The Root Mean PRESS Plot shows that Root Mean PRESS is minimized when the number of factors is 5. This is stated in the note beneath the Root Mean PRESS Plot. A report called **NIPALS Fit with 5 Factors Using Fast SVD** is produced.

Statistical Details for the Partial Least Squares Platform

This section provides details about some of the methods used in the Partial Least Squares platform. See Hoskuldsson (1988), Garthwaite (1994), or Cox and Gaudard (2013).

- "Partial Least Squares"
- "van der Voet T^2"
- "T^2 Plot"
- "Confidence Ellipses for X Score Scatterplot Matrix"
- "Standard Error of Prediction and Confidence Limits"
- "Standardized Scores and Loadings"
- "PLS Discriminant Analysis (PLS-DA)"

Partial Least Squares

Partial least squares fits linear models based on linear combinations, called factors, of the explanatory variables (*X*s). These factors are obtained in a way that attempts to maximize the covariance between the *X*s and the response or responses (*Y*s). In this way, PLS exploits the correlations between the *X*s and the *Y*s to reveal underlying latent structures. The factors address the combined goals of explaining response variation and predictor variation. Partial least squares is particularly useful when you have more X variables than observations or when the X variables are highly correlated.

NIPALS

The NIPALS method works by extracting one factor at a time. Let $\mathbf{X} = \mathbf{X}_0$ be the centered and scaled matrix of predictors and $\mathbf{Y} = \mathbf{Y}_0$ the centered and scaled matrix of response values. The PLS method starts with a linear combination $\mathbf{t} = \mathbf{X}_0\mathbf{w}$ of the predictors, where $\mathbf{t}$ is called a *score vector* and $\mathbf{w}$ is its associated *weight vector*. The PLS method predicts both $\mathbf{X}_0$ and $\mathbf{Y}_0$ by regression on $\mathbf{t}$:

$$\hat{\mathbf{X}}_0 = \mathbf{t}\mathbf{p}', \text{ where } \mathbf{p}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{X}_0$$

$$\hat{\mathbf{Y}}_0 = \mathbf{t}\mathbf{c}', \text{ where } \mathbf{c}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{Y}_0$$

The vectors $\mathbf{p}$ and $\mathbf{c}$ are called the *X*- and *Y-loadings*, respectively.

The specific linear combination $\mathbf{t} = \mathbf{X}_0\mathbf{w}$ is the one that has maximum covariance $\mathbf{t}'\mathbf{u}$ with some response linear combination $\mathbf{u} = \mathbf{Y}_0\mathbf{q}$. Another characterization is that the *X*- and *Y*-weights, **w** and **q**, are proportional to the first left and right singular vectors of the covariance matrix $\mathbf{X}_0'\mathbf{Y}_0$. Or, equivalently, the first eigenvectors of $\mathbf{X}_0'\mathbf{Y}_0\mathbf{Y}_0'\mathbf{X}_0$ and $\mathbf{Y}_0'\mathbf{X}_0\mathbf{X}_0'\mathbf{Y}_0$ respectively.

This accounts for how the first PLS factor is extracted. The second factor is extracted in the same way by replacing $\mathbf{X}_0$ and $\mathbf{Y}_0$ with the *X*- and *Y*-residuals from the first factor:

$$\mathbf{X}_1 = \mathbf{X}_0 - \hat{\mathbf{X}}_0$$

$$\mathbf{Y}_1 = \mathbf{Y}_0 - \hat{\mathbf{Y}}_0$$

These residuals are also called the *deflated X* and *Y* blocks. The process of extracting a score vector and deflating the data matrices is repeated for as many extracted factors as desired.

SIMPLS

The SIMPLS algorithm was developed to optimize a statistical criterion: it finds score vectors that maximize the covariance between linear combinations of *X*s and *Y*s, subject to the requirement that the *X*-scores are orthogonal. Unlike NIPALS, where the matrices $\mathbf{X}_0$ and $\mathbf{Y}_0$ are deflated, SIMPLS deflates the cross-product matrix, $\mathbf{X}_0'\mathbf{Y}_0$.

In the case of a single *Y* variable, these two algorithms are equivalent. However, for multivariate *Y*, the models differ. SIMPLS was suggested by De Jong (1993).

van der Voet T^2

The van der Voet T^2 test helps determine whether a model with a specified number of extracted factors differs significantly from a proposed optimum model. The test is a randomization test based on the null hypothesis that the squared residuals for both models have the same distribution. Intuitively, one can think of the null hypothesis as stating that both models have the same predictive ability.

To obtain the van der Voet T^2 statistic given in the Cross Validation report, the calculation below is performed on each validation set. In the case of a single validation set, the result is the reported value. In the case of Leave-One-Out and KFold validation, the results for each validation set are averaged.

Denote by $R_{i,jk}$ the *j*th predicted residual for response *k* for the model with *i* extracted factors. Denote by $R_{opt,jk}$ is the corresponding quantity for the model based on the proposed optimum number of factors, *opt*. The test statistic is based on the following differences:

$$D_{i,jk} = R_{i,jk}^2 - R_{opt,jk}^2$$

Suppose that there are K responses. Consider the following notation:

$$\mathbf{d}_{i,j} = (D_{i,j1}, D_{i,j2}, \ldots, D_{i,jK})'$$

$$\mathbf{d}_{i,.} = \sum_j \mathbf{d}_{i,j}$$

$$\mathbf{S_i} = \sum_j \mathbf{d}_{i,j}\mathbf{d}_{i,j}'$$

The van der Voet statistic for i extracted factors is defined as follows:

$$C_i = \mathbf{d}_{i,.}'\mathbf{S_i}^{-1}\mathbf{d}_{i,.}$$

The significance level is obtained by comparing C_i with the distribution of values that results from randomly exchanging $R^2_{i,jk}$ and $R^2_{opt,jk}$. A Monte Carlo sample of such values is simulated and the significance level is approximated as the proportion of simulated critical values that are greater than or equal to C_i.

T^2 Plot

The T^2 value for the i^{th} observation is computed as follows:

$$T_i^2 = (n-1)\sum_{j=1}^{p}\left(t_{ij}^2 \Big/ \sum_{k=1}^{n} t_{kj}^2\right)$$

where t_{ij} = X score for the i^{th} row and j^{th} extracted factor, p = number of extracted factors, and n = number of observations used to train the model. If validation is not used, n = total number of observations.

The control limit for the T^2 Plot is computed as follows:

$((n\text{-}1)^2/n)$*BetaQuantile(0.95, p/2, $(n\text{-}p\text{-}1)/2$)

where p = number of extracted factors, and n = number of observations used to train the model. If validation is not used, n = total number of observations. See Tracy et al. (1992).

Confidence Ellipses for X Score Scatterplot Matrix

The Score Scatterplot Matrices option adds 95% confidence ellipses to the X Score scatterplots. The X scores are uncorrelated because both the NIPALS and SIMPLE algorithms produce orthogonal score vectors. The ellipses assume that each pair of X scores follows a bivariate normal distribution with zero correlation.

Consider a scatterplot for score *i* on the vertical axis and score *j* on the horizontal axis. The coordinates of the top, bottom, left, and right extremes of the ellipse are defined as follows:

- the top and bottom extremes are +/-sqrt(var(score *i*)*z)
- the left and right extremes are +/-sqrt(var(score *j*)*z)

where $z = ((n\text{-}1)*(n\text{-}1)/n)$*BetaQuantile(0.95, 1, $(n\text{-}3)/2$). For background on the z value, see Tracy et al. (1992).

Standard Error of Prediction and Confidence Limits

Let **X** denote the matrix of predictors and **Y** the matrix of response values, which might be centered and scaled based on your selections in the launch window. Assume that the components of **Y** are independent and normally distributed with a common variance σ^2.

Hoskuldsson (1988) observes that the PLS model for **Y** in terms of scores is formally similar to a multiple linear regression model. He uses this similarity to derive an approximate formula for the variance of a predicted value. See also Umetrics (1995). However, Denham (1997) points out that any value predicted by PLS is a non-linear function of the Ys. He suggests bootstrap and cross validation techniques for obtaining prediction intervals. The PLS platform uses the normality-based approach described in Umetrics (1995).

Denote the matrix whose columns are the scores by **T** and consider a new observation on **X**, $\mathbf{x}_0$. The predictive model for **Y** is obtained by regressing **Y** on **T**. Denote the score vector associated with $\mathbf{x}_0$ by $\mathbf{t}_0$.

Let *a* denote the number of factors. Define s^2 to be the sum of squares of residuals divided by $df = n - a - 1$ if the data are centered and $df = n - a$ if the data are not centered. The value of s^2 is an estimate of σ^2.

Standard Error of Prediction Formula

The standard error of the predicted mean at $\mathbf{x}_0$ is estimated by the following:

$$SE(\bar{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + \mathbf{t}_0(\mathbf{T}'\mathbf{T})^{-1}\mathbf{t}_0'\right)}$$

Mean Confidence Limit Formula

Let $t_{0.975,\,df}$ denote the 0.975 quantile of a *t* distribution with degrees of freedom $df = n - a - 1$ if the data are centered and $df = n - a$ if the data are not centered.

The 95% confidence interval for the mean is computed as follows:

$$\bar{Y}_{x_0} \pm t_{0.975,\,df} SE(\bar{Y}_{x_0})$$

Indiv Confidence Limit Formula

The standard error of a predicted individual response at $\mathbf{x}_0$ is estimated by the following:

$$SE(\hat{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + 1 + \mathbf{t}_0(\mathbf{T}'\mathbf{T})^{-1}\mathbf{t}_0'\right)}$$

Let $t_{0.975,\,df}$ denote the 0.975 quantile of a t distribution with degrees of freedom $df = n - a - 1$ if the data are centered and $df = n - a$ if the data are not centered.

The 95% prediction interval for an individual response is computed as follows:

$$\bar{Y}_{x_0} \pm t_{0.975,\,df} SE(\hat{Y}_{x_0})$$

Standardized Scores and Loadings

Consider the following notation:

- n_{tr} is the number of observations in the training set
- m is the number of effects in X
- k is the number of responses in Y
- $VarX_i$ is the percent variation in X explained by the i^{th} factor
- $VarY_i$ is the percent variation in Y explained by the i^{th} factor
- $\mathbf{XScore}_i$ is the vector of X scores for the i^{th} factor
- $\mathbf{YScore}_i$ is the vector of Y scores for the i^{th} factor
- $\mathbf{XLoad}_i$ is the vector of X loadings for the i^{th} factor
- $\mathbf{YLoad}_i$ is the vector of Y loadings for the i^{th} factor

Standardized Scores

The vector of i^{th} Standardized X Scores is defined as follows:

$$\frac{\mathbf{XScore}_i}{(n_{tr} - 1)\sqrt{mVarX_i / n_{tr}}}$$

The vector of i^{th} Standardized Y Scores is defined as follows:

$$\frac{\mathbf{YScore}_i}{(n_{tr} - 1)\sqrt{kVarY_i / n_{tr}}}$$

Standardized Loadings

The vector of i^{th} Standardized X Loadings is defined as follows:

$$\mathbf{XLoad}_i\sqrt{mVarX_i}$$

The vector of i^{th} Standardized Y Loadings is defined as follows:

$$\mathbf{YLoad}_i\sqrt{kVarY_i}$$

PLS Discriminant Analysis (PLS-DA)

When a categorical variable is entered as Y in the launch window, it is coded using indicator coding. If there are *k* levels, each level is represented by an indicator variable with the value 1 for rows in that level and 0 otherwise. The resulting *k* indicator variables are treated as continuous and the PLS analysis proceeds as it would with continuous *Y*s.

Chapter 7

Multiple Correspondence Analysis

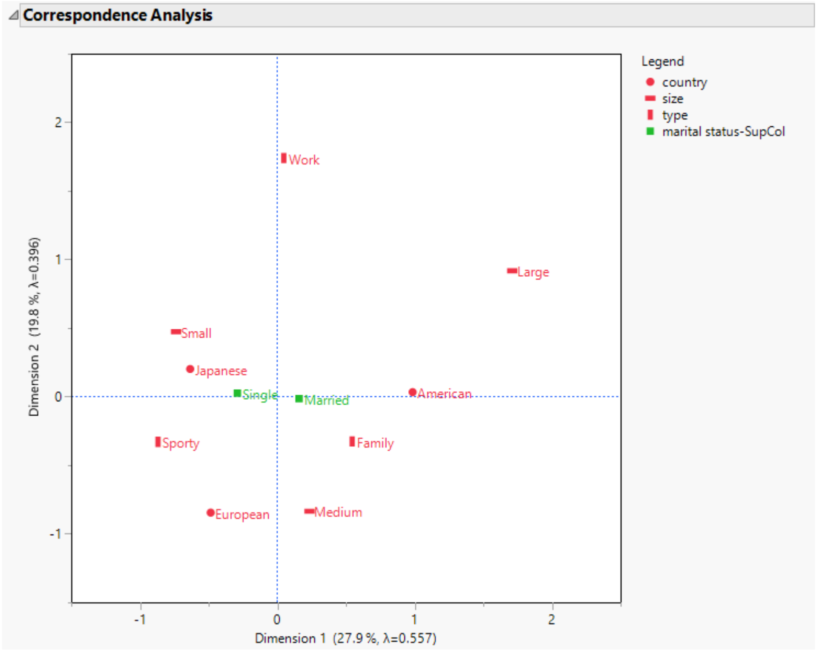
Identify Associations between Levels of Categorical Variables

Multiple Correspondence Analysis (MCA) takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA extends correspondence analysis from two variables to many. It can be thought of as analogous to principal component analysis for quantitative variables. Similar to other multivariate methods, it is a dimension reducing method; it represents the data as points in 2- or 3-dimensional space.

Multiple correspondence analysis is frequently used in the social sciences. It can be used in survey analysis to identify question agreement. It is also used in consumer research to identify potential markets for products.

For more information about multiple correspondence analysis, see LeRoux and Rouanet (2010).

Figure 7.1 Multiple Correspondence Analysis



Contents

Example of Multiple Correspondence Analysis

This example uses data collected from 55 JMP employees about their preferences, or tastes, in these areas:

- Preferred TV program (8 categories): news, comedy, police, nature, sport, film, drama, or soap operas.
- Preferred film (8 categories): action, comedy, costume drama, documentary, horror, musical, romance, or SciFi.
- Preferred type of art (7 categories): performance, landscape, renaissance, still life, portrait, modern, or impressionism.
- Preferred place to eat out (6 categories): fish & chips, pub, Indian restaurant, Italian restaurant, French restaurant, or steak house.

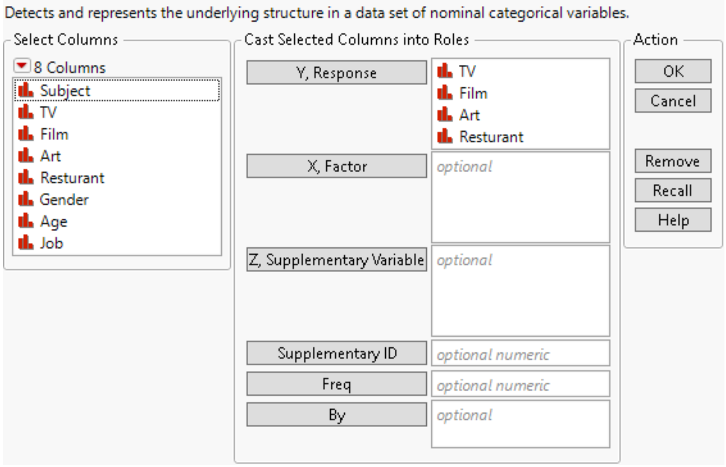
The data also include demographics about the individuals polled. You want to explore the associations between the 55 responses to the four questions. This can lead to an understanding of different groups of employees.

Note: The measures for preferences are based on questions from LeRoux and Rouanet (2010).

1. Select **Help > Sample Data Library** and open Employee Taste.jmp.
2. Select **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.
3. Select TV, Film, Art, and Resturant and click **Y, Response**.

   In MCA, as in principal component analysis, factors are typically considered responses rather than some being responses and others explanatory.
4. Click **OK**.

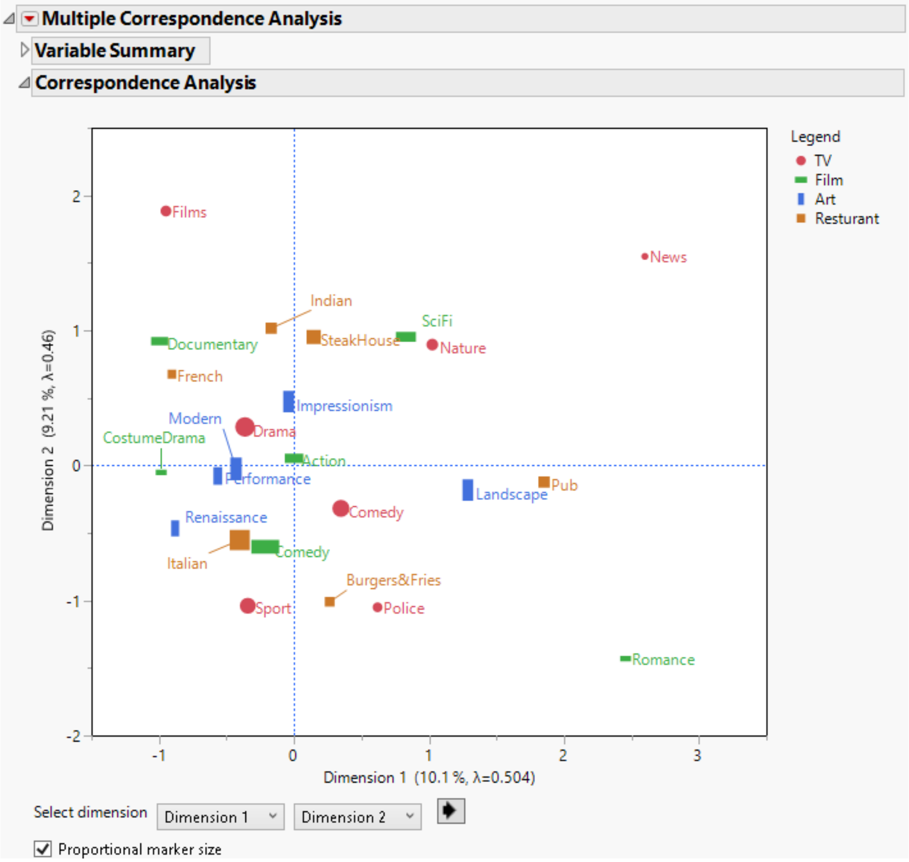
Figure 7.2 Completed Multiple Correspondence Analysis Launch Window



Detects and represents the underlying structure in a data set of nominal categorical variables.

Y, Response: TV, Film, Art, Resturant

Figure 7.3 Part of the Initial Multiple Correspondence Analysis Report



Tip: Click and drag overlapping labels to rearrange them.

5. Below the plot, select **Proportional marker size**.

   The marker size indicates the relative proportion of responses in that category.

   The Correspondence Analysis report shows the projection of categories of the four variables onto the first two principal axes, or dimensions. Use the controls to change the dimensions plotted. The distance between points represents the dissimilarity between response patterns of the employees.

   From this graph, you can use your knowledge of the subject matter to interpret the findings:

   – Grouped near the Burgers and Fries restaurant preference, there is a cluster of the Sport and Police TV program preferences. This cluster could be called the “popular culture” group.

   – Looking at Dimension 2, individual tastes move from the popular culture group to what could be classified as more “sophisticated” taste - those who like films such as documentaries, prefer TV dramas, and specialized restaurants, such as French and Indian cuisines.

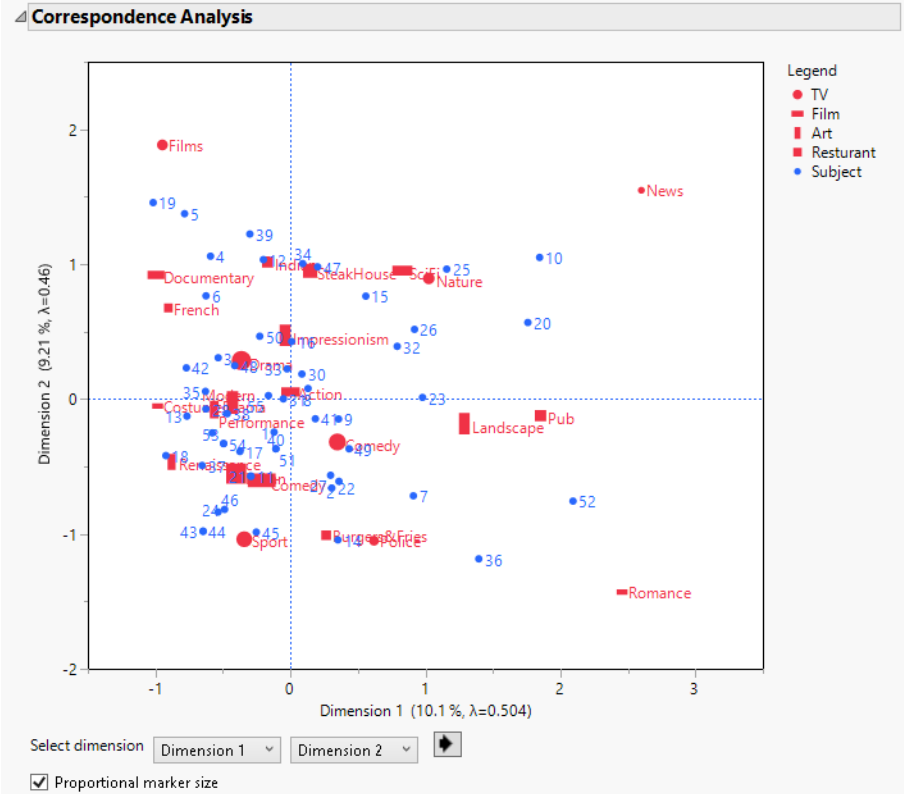
   To obtain the scores for individuals, add the subject as the **X, Factor**:

6. Open the Variable Summary outline.

   The Variable Summary panel allows you to modify your analysis without relaunching the platform. It also provides a concise view of the completed analysis.

7. Select Subject and click **Add X**. The analysis automatically updates.

Figure 7.4 Multiple Correspondence Analysis Report with Subjects

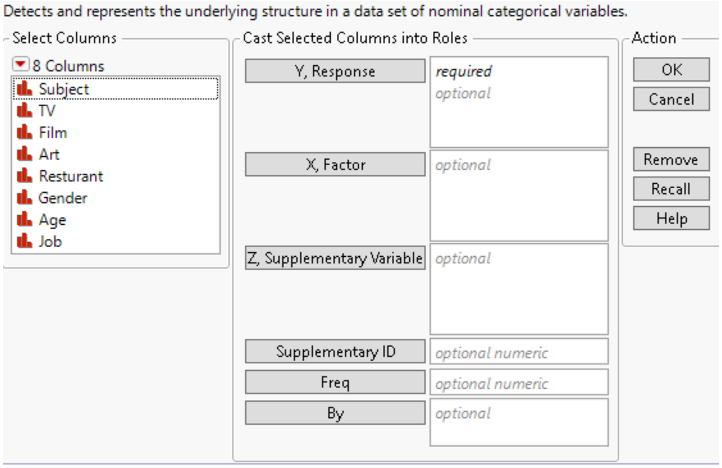


Including the subjects highlights a cluster of employees with similar taste in the lower left quadrant of the plot. These employees fall in the area we identified as popular culture.

Launch the Multiple Correspondence Analysis Platform

Launch the Multiple Correspondence Analysis platform by selecting **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.

Figure 7.5 Multiple Correspondence Analysis Launch Window



Detects and represents the underlying structure in a data set of nominal categorical variables.

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Response Assigns the categorical columns to be analyzed. In MCA, you are generally interested in the associations between variables, but there are not explicit “explanatory” and “response” variables.

X, Factor Assigns one or more categorical columns to be used as factor, or explanatory, variables.

Note: Use a subject ID column as a single X to obtain scores on individuals.

Z, Supplementary Variable Assigns the columns to be used as supplementary variables. These variables are those you are interested in identifying associations with but not include in the calculations. Supplementary variables are used to improve data interpretation.

Supplementary ID Assigns the column that identifies rows to be used as supplementary. A supplementary ID column usually has 1s and 0s. The rows associated with ID 0 are treated as supplementary rows. The Supplementary ID column is ignored if there are levels of the

X or Y variables present in the supplementary rows that are not present in the non-supplementary rows.

Note: Only one of the Supplementary ID and Z, Supplementary Variable roles can be specified at one time.

Freq Assigns a frequency variable to this role. This is useful if your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Note: The Multiple Correspondence Analysis platform handles missing values differently than many other JMP platforms. The analysis uses all nonmissing pairs of cells in a row. It does not remove entire rows from the computation.

The Multiple Correspondence Analysis Report

The initial Multiple Correspondence Analysis report shows the variable summary, correspondence analysis plot, and details of the dimensions of the data in order of importance. From the plot of the cloud of categories or individuals, you can identify associations that exist within the data. The details provide information about whether the two dimensions shown in the plot are sufficient to interpret the relationships within the data.

The Variable Summary in an interactive panel that allows you to modify your analysis without relaunching the platform. The panel shows the columns and roles used in the analysis. If you select the Show Controls check box, a list of the columns in the data table appears to the left. You can change the columns in the analysis either by selecting a column and clicking Add Y, Add X, Add Z, or Add ID. Or you can drag the column to the header in the variable summary table. This enables you to modify the analysis without returning to the launch window.

Figure 7.6 Multiple Correspondence Analysis Report with Show Controls Selected

Y Role	Variable	X Role	Variable	Z Role	Variable	ID Role	Variable
Y	country			Z	marital status		
Y	size						
Y	type						

Multiple Correspondence Analysis Platform Options

The Multiple Correspondence Analysis red triangle menu options give you the ability to customize reports according to your needs. The reports available are determined by the type of analysis that you conduct.

Cross Table Shows or hides a Burt or contingency table as appropriate for variable roles selected. See "Cross Table" on page 169.

Correspondence Analysis Shows or hides the correspondence analysis reports. These reports give the plots, details, coordinates, and summary statistics. See "Additional Examples of the Multiple Correspondence Analysis Platform" on page 170.

Cross Table of Supplementary Rows (Available when one or more supplementary Z and X variables are specified or when an ID is specified.) Shows or hides a Burt or contingency table, dependent on the following rules:

- When one or more X and Z variables are specified, shows or hides a contingency table of the X variable(s) versus the response variable(s).
- When an ID and no X variables are specified, shows or hides a Burt table for the supplementary observations. The supplementary observations are those observations for which the ID variable is equal to 0.
- When an ID and one or more X variables are specified, shows or hides a contingency table for the supplementary observations. The supplementary observations are those observations for which the ID variable is equal to 0.

Cross Table of Supplementary Columns (Available when one or more supplementary Z variables are specified.) Shows or hides a contingency table dependent on the following rules:

- When no X variable is specified, shows or hides a contingency table of the Response variable(s) versus the supplementary variable(s).
- When one or more X variables are specified, shows or hides a contingency table of the X variable(s) versus the supplementary variable(s).

Mosaic Plot (Available only when there is one X variable and one Y variable.) Shows or hides a mosaic plot Y versus X. A mosaic plot is a stacked bar chart where each segment is proportional to its group’s frequency count.

Tests for Independence (Available only when there is one X variable and one Y variable.) Shows or hides the tests for independence between the X and Y variables. There are two versions of this test, the Pearson form and the Likelihood Ratio form, both with chi-square statistics.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Correspondence Analysis Options

The reports available under Correspondence Analysis are determined by the type of multiple correspondence analysis that you conduct. Several of these reports are shown by default.

Show Plot Shows the two-dimensional cloud of categories in the plane described by the first two principal axes. This plot appears by default. The plot uses isometric scaling.

Show Detail Provides the details of the analysis including the singular values, inertias, ChiSquare statistics, percent, and cumulative percent. This report appears by default. See "Show Detail" on page 166.

Show Adjusted Inertia Provides reports of the Benzecri and Greenacre adjusted inertia. See Benzécri (1979) and Greenacre (1984). This option is not available when there are one or more X variables. See "Show Adjusted Inertia" on page 167.

Show Coordinates Provides a report of up to the first three principal coordinates for the categories in the analysis, as appropriate. See "Show Coordinates" on page 167.

Show Summary Statistics Provides a report of the summary statistics, Quality, Mass, and Inertia, for each category in the analysis. See "Show Summary Statistics" on page 168.

Show Partial Contributions to Inertia Provides a report of the contribution of each category to the inertia for each of up to the first three dimensions. See "Show Partial Contributions to Inertia" on page 168.

Show Squared Cosines Provides a report of the squared cosines of each category for each of up to the first three dimensions. The report includes a bar chart that shows, for each level of each Y variable, the squared cosine values for each of up to the first three dimensions. See "Show Squared Cosines" on page 169.

Cochran's Q Test (Available only when all of the Y variables have the same set of only two levels and the X variable has a unique value for each row.) Provides Cochran's *Q* statistic, which tests that the marginal probability of a specific response is unchanged across the Y variables. Cochran's *Q* statistic is a generalization of McNemar's statistic for more than two response variables. See Agresti (2013). See "Cochran's *Q Test*" on page 169.

3D Correspondence Analysis Shows the three-dimensional cloud of categories of the Y, X, and Z variables in the space described by the first three principal axes. This option is not available if there are less than three dimensions.

Save Coordinates Saves the principal coordinates to one or more JMP data tables. Column coordinates, row coordinates, supplementary column coordinates, and supplementary row coordinates are saved to separate JMP data tables. You can choose how many columns to save.

Save Coordinate Formula Saves formula columns to the data table for the principal coordinates in multiple dimensions. The value for each observation is the average of the coordinates for the Y variables scaled by the singular value for each dimension. You can choose how many columns to save.

Show Plot

The plot displays a projection of the projection of categories or individuals onto the plane described by the first two principal axes. The plot uses isometric scaling. You can toggle the dimensions shown in the plot using the Select Dimension controls below the plot. The first control defines the horizontal axis of the plot, and the second control defines the vertical axis of the plot. Click the arrow button to cycle through the dimensions shown in the plot. Use the **Proportional marker size** check box to specify if the size of the points in the plot should be proportional to the count of observations corresponding to each point.

Note: Selecting a point in the correspondence analysis plot also selects the corresponding rows in other tables in the report window. However, rows in the data table are not selected. To select all of the points in the plot associated with a particular variable, select the name of the variable in the plot legend.

Show Detail

Shows the table of singular values.

Singular Value Shows the singular values in the singular value decomposition of the contingency table or Burt table. For the formula, see "Details Report" on page 175.

Inertia Lists the square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

ChiSquare Lists the portion of the overall Chi-square for the Burt or contingency table represented by the dimension.

Percent Portion of inertia with respect to the total inertia.

Cumulative Percent Shows the cumulative portion of inertia. If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

Show Adjusted Inertia

The principal inertias of a Burt table in MCA are the eigenvalues. The problem with these inertias is that they provide a pessimistic indication of fit. Benzécri proposed an inertia adjustment. Greenacre argued that the Benzécri adjustment overestimates the quality of fit and proposed an alternate adjustment. Both adjustments are calculated for your reference. See "Adjusted Inertia" on page 176.

Inertia Lists the square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

Adjusted Inertia Lists the adjusted inertia according to either the Benzécri or Greenacre adjustment.

Percent Portion of adjusted inertia with respect to the total inertia.

Cumulative Percent Shows the cumulative portion of adjusted inertia. If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

Show Coordinates

Shows the Column Coordinates table or the Row and Column Coordinates tables.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Z Lists the columns specified as Z, Supplementary Variables.

Category Lists the levels of the X, Y, or Z variables.

Dimension 1, Dimension 2, Dimension 3 For each level or each response, lists its coordinate along the indicated principal axis. By default, the tables show coordinates for up to the first three dimensions. Coordinate columns for additional dimensions are hidden. To show these optional columns, right-click in a table and select the dimension columns from the **Columns** submenu.

Note: If there are columns specified as X, Factor variables, the Coordinates report displays tables of both X and Y with the same report headings. If a Z, Supplementary Variable is specified, the coordinates are listed below the X and Y coordinates as applicable.

Show Summary Statistics

Shows the Summary Statistics for the Column Points table or the Summary Statistics for the Row and Column Points tables. The Y table gives Quality, Mass, and Inertia for each level of each response, called a *column point*. The X table gives Quality, Mass, and Inertia for each level of the X, Factor variables. See "Summary Statistics" on page 176.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Category Lists the levels of the X or Y variables.

Quality(dim=2) Lists the quality of the representation of the level by the solution.

Mass Lists the row frequency for the level of the response divided by the total frequency. In the Burt table, this is the Total % for each row.

Inertia Lists the proportion of the total inertia accounted for by the level of the response. The inertia values sum to one across the levels and their responses.

Note: If there are columns specified as X, Factor variables, the Summary Statistics report displays tables of both X and Y with the same report headings.

Show Partial Contributions to Inertia

Shows the Partial Contributions to Inertia for the Column Points table or the Partial Contributions to Inertia for the Row and Column Points tables. Also shows the Plot of Partial Contributions to Inertia for the Column Points. This is a bar chart that shows, for each level of each Y variable, its partial contributions to each of the dimensions shown in the table.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Category Lists the levels of the X or Y variables.

Dimension 1, Dimension 2, Dimension 3 Lists the contribution of the response or factor level to the inertia of the indicated dimension. By default, the tables show columns for up to the first three dimensions. Additional columns are hidden. To show these optional columns, right-click a table and select the dimension columns from the **Columns** submenu.

Each level of each response contributes to the inertia of each dimension. The partial contributions within each dimension sum to 1.

Note: If there are columns specified as X, Factor variables, the Partial Contributions to Inertia report displays tables of both X and Y with the same report headings. See “Partial Contributions to Inertia” on page 177.

Show Squared Cosines

Shows the Squared Cosines for the Column Points table or the Squared Cosines for the Row and Column Points. Also shows the Plot of Squared Cosines for the Column Points. This is a bar chart that shows, for each level of each Y variable, the squared cosine values for each of up to the first three dimensions shown.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Category Lists the levels of the X or Y variables.

Dimension 1, Dimension 2, Dimension 3 Lists the *quality* of the representation of the level by the indicated dimension. By default, the tables show results for up to the first three dimensions. Additional columns are hidden. To show these optional columns, right-click a table and select the dimension columns from the **Columns** submenu.

The values indicate the quality of each point for the indicated dimension. The squared cosine can be interpreted as the squared correlation of the point with the dimension. The sum of the squared cosines of the first two dimensions equalsQuality(dim=2) in the Summary Statistics report. See “Summary Statistics” on page 176.

Note: If there are columns specified as X, Factor variables, the Squared Cosines report displays tables of both X and Y with the same report headings.

Cochran’s Q Test

Cochran’s *Q* test is a nonparametric test for homogeneity across matched samples for three or more binary outcomes. You can use it to test for differences in proportions in matched pairs. Cochran’s *Q* test is an extension of the McNemar test used for two outcomes.

Cross Table

The cross table option hides or shows a Burt table or a contingency table. When you select multiple Y, Response columns with no X, Factor columns, the Burt table is created. If you select any X, Factor columns, a traditional contingency table is created instead of a Burt table. The outline node title reflects the structure of the cross table.

The Burt table is the basis of the multiple correspondence analysis. It is a partitioned symmetric table of all pairs of categorical variables. The diagonal partitions are diagonal matrices (a cross-table of a variable with itself). The off-diagonal partitions are ordinary contingency tables.

The red triangle menu for the Burt or contingency table contains the following options of statistics to display in the table.

Count Cell frequency, margin total frequencies, and grand total (total sample size). This appears by default.

Total % Percent of cell counts and margin totals to the grand total. This appears by default.

Cell Chi Square Chi-square values computed for each cell as $(O - E)^2 / E$.

Col % Percent of each cell count to its column total.

Row % Percent of each cell count to its row total.

Expected Expected frequency (E) of each cell under the assumption of independence. Computed as the product of the corresponding row total and column total divided by the grand total.

Deviation Observed cell frequency (O) minus the expected cell frequency (E).

Col Cum Cumulative column total.

Col Cum % Cumulative column percentage.

Row Cum Cumulative row total.

Row Cum % Cumulative row percentage.

Make Into Data Table Creates one data table for each statistic shown in the table.

Additional Examples of the Multiple Correspondence Analysis Platform

- “Example Using a Supplementary Variable”
- “Example Using a Supplementary ID”
- “Example of Cochran’s *Q Test*”

Example Using a Supplementary Variable

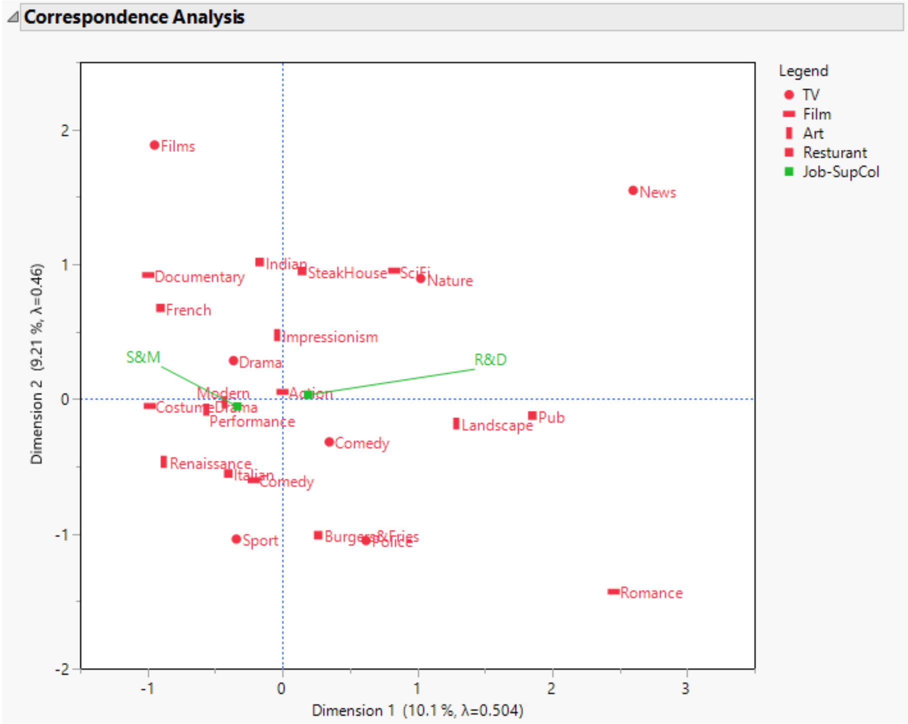
This example uses the sample data on JMP employee preferences described in "Example of Multiple Correspondence Analysis" on page 157.

You want to explore relationships between employee job and their preferences.

1. Select **Help > Sample Data Library** and open Employee Taste.jmp.
2. Select **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.
3. Select TV, Film, Art, and Resturant and click **Y, Response**.
4. Select Job and click **Z, Supplementary Variable**.
5. Click **OK**.

Note: Since Job is a supplementary variable, it is not used in the calculations. Job type is plotted after the calculations are complete.

Figure 7.7 MCA with Supplementary Variable Report



You can see from the plot that the data points for the R&D (research and development) and the S&M (sales and marketing) employees are close together. We interpret this to mean that there is not a strong difference in terms of preferences due to employee job function. In other words, the employee's job role is not a significant factor in their preferences.

Example Using a Supplementary ID

The United States census allows for examining population growth over the last century. The US Regional Population.jmp sample data table contains populations of the 50 US states grouped into regions for each of the census years from 1920 to 2010. Alaska and Hawaii are treated as supplementary regions because they were not states during the entire time, and they are not part of the contiguous United States. You are interested in whether the population growth in these two states differs from the rest of the US.

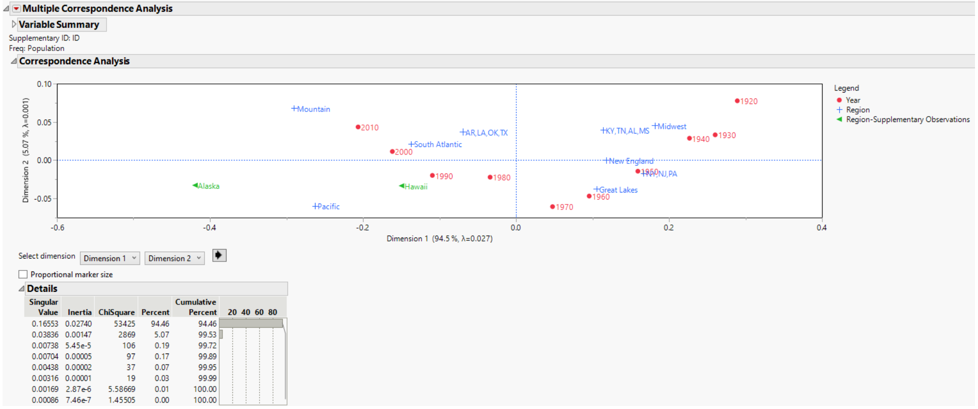
1. Select **Help > Sample Data Library** and open US Regional Population.jmp.
2. Select **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.
3. Select Year and click **Y, Response**.
4. Select Region and click **X, Factor**.
5. Select ID and click **Supplementary ID**.
6. Select Population and click **Freq**.
7. Click **OK**.

The Details report shows that the association between years and regions is almost entirely explained by the first dimension. The plot shows that years are in the correct order on the first dimension. This ordering occurs naturally through the correspondence analysis; there is no information about the order provided to the analysis. The plot highlights the isometric scale used to plot the data.

Notice that the ordering of the regions reflects the population shift from the Midwest to the Northeast to the South and finally to the Mountain and West.

Alaska and Hawaii were not used in the computation of the analysis but are plotted based on the results. Their growth pattern is most similar to the Pacific states. Alaska's growth is even more extreme than the Pacific region.

Figure 7.8 MCA with Supplementary ID Report



Supplementary ID: ID
Freq: Population

☐ Proportional marker size

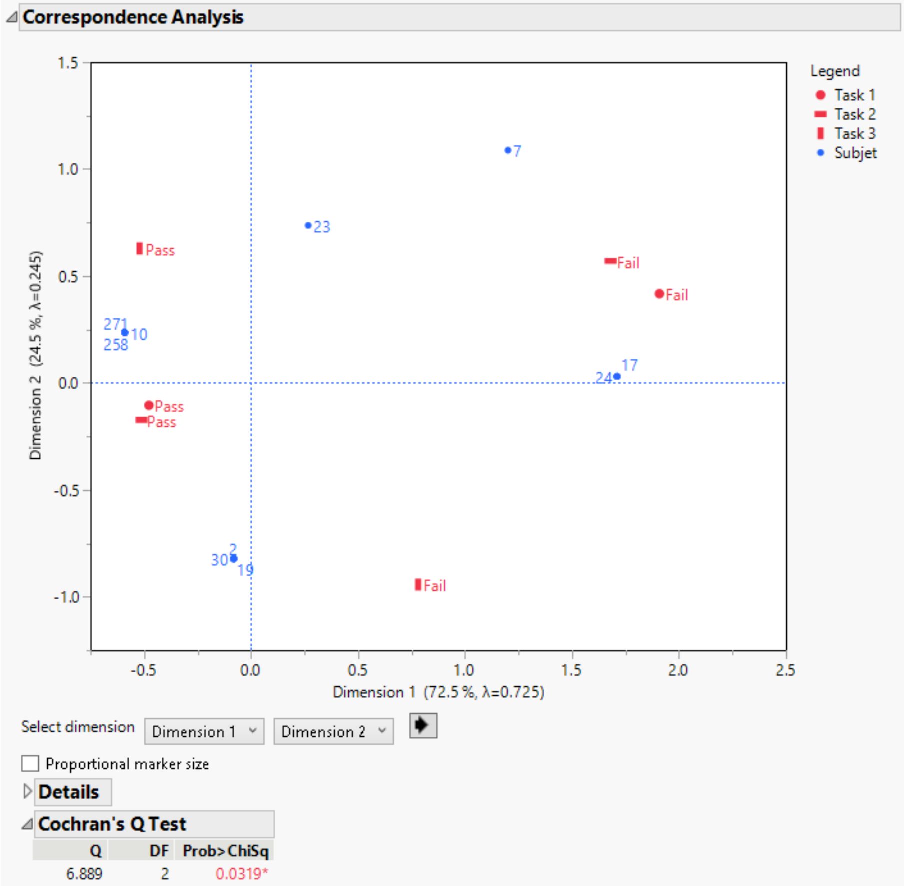
Singular Value	Inertia	ChiSquare	Percent	Cumulative Percent
0.16553	0.02740	53425	94.46	94.46
0.03836	0.00147	2869	5.07	99.53
0.00738	5.45e-5	106	0.19	99.72
0.00704	0.00005	97	0.17	99.89
0.00438	0.00002	37	0.07	99.95
0.00316	0.00001	19	0.03	99.99
0.00169	2.87e-6	5.58669	0.01	100.00
0.00086	7.46e-7	1.45505	0.00	100.00

Example of Cochran's *Q* Test

This example uses a simulated data table to evaluate the ease of completion of three tasks for 30 subjects. Task 1 was simulated to have an 80% completion rate. Tasks 2 and 3 were simulated to have the same result as Task 1 for 95% and 80% of subjects respectively.

1. Select **Help > Sample Data Library** and open Cohcrans Q.jmp.
2. Select **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.
3. Select Task 1, Task 2, and Task 3 and click **Y, Response**.
4. Select Subject and click **X, Factor**.
5. Click **OK**.
6. Click the Multiple Correspondence Analysis red triangle menu and select **Correspondence Analysis > Cochran's Q Test**.

Figure 7.9 Cochran's *Q* Test



Select dimension: Dimension 1, Dimension 2

☐ Proportional marker size

Cochran's Q Test

Q	DF	Prob>ChiSq
6.889	2	0.0319*

The correspondence plot shows the similarity between Task 1 and Task 2 as the Pass and Fail outcomes cluster together. However, Task 3 is farther away from Tasks 1 and 2. Cochran's *Q* test statistic of 6.889 has an associated *p*-value of 0.0319. This *p*-value supports the rejection of the hypothesis that the pass rates are the same across all tasks at a 0.05 significance level.

Statistical Details for the Multiple Correspondence Analysis Platform

- "Details Report"
- "Adjusted Inertia"
- "Summary Statistics"
- "Partial Contributions to Inertia"
- "Cochran's *Q Statistic*"

Details Report

When a simple Correspondence Analysis is performed, the report lists the singular values from the following singular value decomposition:

$$\mathbf{D}_r^{-0.5}(\mathbf{P} - rc')\mathbf{D}_c^{-0.5} = \mathbf{U}Diag(\Lambda)\mathbf{V}'$$

where:

- **P** is the matrix of counts divided by the total frequency
- r and c are the row and column sums of **P**
- the **D** matrices are diagonal matrices of the values of r and c
- Λ is the column vector of singular values

When Multiple Correspondence Analysis is performed, the singular value decomposition extends to the following equation:

$$\mathbf{D}^{-0.5}(\mathbf{C} - dd')\mathbf{D}^{-0.5} = \mathbf{U}Diag(\Lambda)\mathbf{V}'$$

where:

- **C** is the Burt table
- d is a column vector of the column sums of **C** (d is also the row sums, since **C** is symmetric)
- **D** is a diagonal matrix of the values of d

In the Details report, the inertia is the column vector Λ. The singular value is the square root of the inertia vector. The column coordinates are calculated as follows:

$$\mathbf{D}^{-0.5}\mathbf{V}Diag\left(\Lambda^{0.5}\right)$$

Adjusted Inertia

The usual principal inertias of a Burt table constructed from m categorical variables in MCA are the eigenvalues λ_k from Λ^2. These inertias provide a pessimistic indication of fit. Benzécri (1979) proposed the following inertia adjustment; it is also described by Greenacre (1984, p. 145):

$$\left(\frac{m}{m-1}\right)^2 \times \left(\lambda_k - \frac{1}{m}\right)^2 \quad \text{for } \lambda_k > \frac{1}{m}$$

This adjustment computes the percent of adjusted inertia relative to the sum of the adjusted inertias for all inertias greater than $1/m$.

Greenacre (1984, p. 156) argues that the Benzécri adjustment overestimates the quality of fit. Greenacre proposes instead to compute the percentage of adjusted inertia relative to:

$$\frac{m}{m-1}\left(\mathrm{tr}\left(Diag\left(\Lambda^4\right)\right) - \frac{n_c - m}{m^2}\right)$$

for all inertias greater than $1/m$, where $\mathrm{tr}(Diag(\Lambda^4))$ is the sum of squared inertias and n_c is the total number of categories across the m variables.

Summary Statistics

Quality is the ratio of the squared distance of a point from the origin in the space defined by the specified number of dimensions to the distance from the origin in the space with the maximum number of dimensions. For the Chi-Square metric, a point's quality in a given dimension can be obtained from the cosine that its vector makes with the vector that defines the dimension. Quality is also equal to the ratio of the sum of inertias in the specified dimensions to the sum of the inertias in all dimensions. Quality indicates how well the point is represented in the lower-dimensional space.

Mass is the proportion of row or column total frequency to the total frequency.

Inertia is analogous to variance in principal component analysis. The overall inertia is the total Pearson Chi-square for a two-way frequency table divided by the sum of all observations in the table.

Relative inertia is the proportion of the contribution of the point to the overall inertia. In the summary statistics table, the relative inertia is listed in the column labeled Inertia.

Partial Contributions to Inertia

The contribution of a row or column to the inertia of a dimension is calculated as follows:

contribution = (mass)(coordinate)2/(dimension inertia)

Cochran's *Q* Statistic

The Cochran's Q test statistic is calculated as follows:

$$Q = k(k-1)\frac{\sum_{j=1}^{k}\left(\sum_{i=1}^{b} X_{ij} - \frac{N}{k}\right)^{2}}{\sum_{i=1}^{b}\left(\sum_{j=1}^{k} X_{ij}\left(k - \sum_{j=1}^{k} X_{ij}\right)\right)}$$

where

k is the number of treatments

b is the number of blocks

X_{ij} is the response (0 or 1) for the i^{th} subject for the k^{th} treatment

N is the grand total of positive responses across all subjects and treatments

Chapter 8

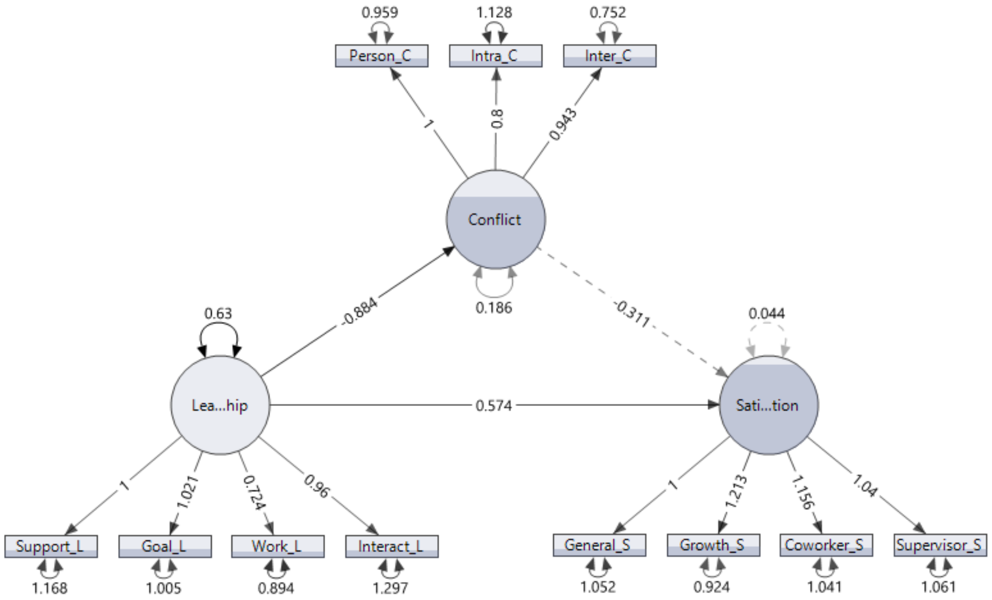
Structural Equation Models

Fit Structural Equation Models

The Structural Equation Models platform is available only in JMP Pro.

The Structural Equation Models platform enables you to fit a variety of models. These models include confirmatory factor analysis, path models with or without latent variables, measurement error models, and latent growth curve models. You can specify the model by selecting *From* and *To* variables and indicating how they are linked. The link is represented by either a unidirectional or a bidirectional arrow in the diagram. These selections populate two alternative views of the model: a dynamic path diagram and a list of model edges. The platform also checks for model identification as you specify the model.

Figure 8.1 Structural Equation Model Path Diagram



Contents

Overview of Structural Equation Models

The Structural Equation Models (SEM) platform enables you to fit a wide variety of models that can be used to test theories of relationships among variables. The variables in the models can be observed (*manifest* variables) or unobserved (*latent* variables). Structural equation modeling is popular in the social and behavioral sciences.

By default, the platform specifies a model with means and variances for all variables. The platform then provides a model-building interface that enables you to see multiple views of the model while it is being built. It also provides some model details during the model construction process that alert you to untenable models prior to running the model.

After you fit one or more models, you can compare the fitted models and two baseline models in the Model Comparison report. The baseline models are an unrestricted model and an independence model. The unrestricted model is a fully saturated model, which fits all means, variances, and covariances of the specified Model Variables without imposing any structure on the data. The independence model fits all means and variances of the specified Model Variables. All covariances among the specified Model Variables are fixed to zero, which leads to a highly restrictive model.

The SEM platform uses the full information maximum likelihood (Finkbeiner 1979) method. This enables you to fully use all available information from the data even when there is a high proportion of observations with random missing values.

For more information about structural equation modeling, see the CALIS Procedure chapter in SAS Institute Inc. (2020a), Bollen (1989), and Kline (2016).

Note: All models in the Structural Equation Models platform are estimated with a mean structure, which means that a Constant term is included. If you do not want to place a structure on the means of the observed variables, then the means should be freely estimated as in the default model specification.

Types of Models

This section describes some of the various types of models that can be fit in the Structural Equation Models platform:

- *Path Analysis* enables you to test alternative explanatory models of the associations between observed variables. This technique is often used when only one variable is available per construct of interest in a study. Perhaps the simplest Path Analysis model is a standard regression model, in which X predicts Y. The SEM platform enables you to fit this simple regression model but you can also specify more interesting models as well. For example, you might have variable Z that is presumed, based on theory or previous research, to be a mediator of the X ⇒ Y relationship. In other words, X predicts Z, which then predicts Y. Thus, the original X ⇒ Y relationship might exist only because Z is excluded in the original model. Path analysis can be carried out in the SEM platform by performing the following steps:

1. Select all the observed variables in the launch window, click Model Variables, and click OK.
2. In the Model Specification report, select the predictors in the From List and the corresponding outcomes in the To List, and click the unidirectional arrow button.

Note: All exogenous variables (those that do not have any unidirectional arrows pointing at them) must be freely correlated in the model, unless a hypothesis of zero correlation is being tested. These covariances are specified with the bidirectional arrow button.

- *Confirmatory Factor Analysis* (CFA) enables you to test alternative measurement models. CFA is often used in survey development and as an initial step prior to fitting structural regression models. The SEM platform enables fitting of confirmatory factor analysis models by performing the following steps:

  1. Select all the observed variables in the launch window, click Model Variables, and click OK.
  2. Using the To List under Model Specification, select the variables that are presumed to load onto a latent variable.
  3. Enter the name for the latent variable in the box below the To List, and click the add latent ⊞ button.
  4. Repeat this process until all the latent variables for the model have been specified.

  Note that the SEM platform always includes a mean structure, so all of the observed variables are listed in the Means/Intercepts list as outcomes of the Constant term. Moreover, all latent variables are automatically identified by setting the loading of their first indicator to 1 (default) or their variance to 1 if the Standardized Latent Variables option was selected in the launch window. Finally, the traditional CFA model allows all latent variables to covary. You can specify these covariances by selecting all the latent variables in the From and To Lists and then clicking the bidirectional arrow button.

- *Structural Regression* (SR) models are also known as *path analysis with latent variables*. These models are often used after having identified an appropriate measurement model through confirmatory factor analysis (CFA). SR models enable you to test specific patterns of relationships between latent variables. In other words, while CFA does not impose any directionality in the effects between latent variables (all latent variables are allowed to freely covary), SR models do. In an example where management Leadership is hypothesized to lead to less team Conflict and more employee Satisfaction in the workplace, the Leadership latent variable can predict the Conflict and Satisfaction latent variables. You can specify these directional effects (regressions) after performing a CFA by performing the following steps:

   1. Select all the observed variables in the launch window, click Model Variables, and click OK.
   2. Using the To List under Model Specification, select the variables that are presumed to load onto a latent variable.
   3. Enter the name for the latent variable next in the box below the To List, and click the add latent ⊞ button.
   4. Repeat this process until all the latent variables for the model have been specified.
   5. In the Model Specification report, select the predictors in the From List and the corresponding outcomes in the To List, and click the unidirectional arrow button.

- *Latent Growth Curve* (LGC) models enable you to fit and test alternative latent trajectories to repeated measures data. These models are very similar to random effects models in the mixed models framework. Often, you want to compare a no-growth model with a linear model. In a no-growth model, individuals can vary in their starting point but have flat trajectories. In a linear model, individuals can vary in both their starting point and linear slope over time. If enough data are available, you can also compare these models with a quadratic model where individuals can vary in their starting point and their linear and quadratic rates of change over time. You can fit LGC models using the SEM platform model specification, but the platform streamlines the fitting of LGC models by performing the following steps:

   1. Select all the observed variables (repeated measures) in the launch window, click Model Variables, and click OK.

      Note: For the Latent Growth model shortcuts to specify the model correctly, the observed variables must be listed in ascending time order and must have equal time intervals.

   2. Using the Model Shortcuts option, select the Longitudinal Analysis > Intercept-Only Growth Curve model, and then click Run.
   3. Using the Model Shortcuts option, select the Longitudinal Analysis > Linear Latent Growth Curve model, and then click Run.
   4. Using the Model Shortcuts option, select the Longitudinal Analysis > Quadratic Latent Growth Curve model, and then click Run.

  The Model Comparison table shows the alternative fit indices and the best model can be selected.

- *Conditional Latent Growth Curve* models can be used after identifying an ideal growth trajectory following the steps above. At this point, predictors of the intercept and change factors can be added to the model. These predictors might prove to be important factors for determining initial scores on a growth process and ensuing changes. To fit a conditional LGC, select all of the observed variables (repeated measures), including the hypothesized predictors of the latent variables, in the launch window. Make sure that the predictors are the last variables in the Model Variables list to facilitate the following steps:

1. Use the Model Shortcuts option to select the appropriate growth trajectory. This option specifies all variables in the LGC model, including the predictors. Thus, you need to exclude the predictors from the growth process and correctly specify them as predictors.
2. Find the predictors in the Loadings list, select all the effects that involve them, and click Remove.
3. Select the predictors in the From List and the Intercept or Slope in the To List.
4. Click the unidirectional arrow to specify the conditional LGC.

Note: If you have more than one predictor, their covariances must be specified by selecting the predictors in the From and To Lists and clicking the bidirectional arrow button.

Example of a Structural Equation Model

In this example, an employee in a human resources department wants to improve job satisfaction. The example uses the Structural Equation Models platform to analyze responses to a survey of 200 individuals regarding aspects of their job satisfaction. The survey contains responses to 11 questions related to their job satisfaction. You seek to build a structural regression model that relates the answers to the survey questions to the latent variables of leadership characteristics, role conflict, and overall job satisfaction.

1. Select **Help > Sample Data Library** and open Job Satisfaction.jmp.
2. Select **Analyze > Multivariate Methods > Structural Equation Models**.
3. Select Support_L through Supervisor_S and click **Model Variables**.
4. Click **OK**.

   The Structural Equation Models report Model Specification outline appears.

Create Latent Variables

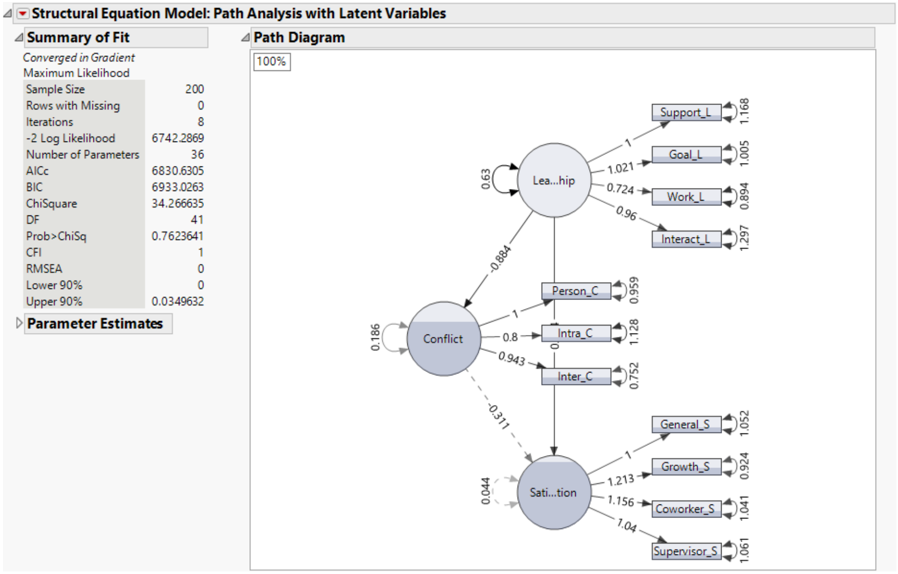
5. Select Support_L through Interact_L in the To List, type Leadership in the box below the To List, and click the add latent ⊞ button.
6. Select Person_C through Inter_C in the To List, type Conflict in the box below the To List, and click the add latent ⊞ button.
7. Select General_S through Supervisor_S in the To List, type Satisfaction in the box below the To List, and click the add latent ⊞ button.

Add Loading and Regression Variables

8. Select Leadership in the From List, select Satisfaction in the To List, and click the unidirectional arrow ➞ button.
9. Select Leadership in the From List, select Conflict in the To List, and click the unidirectional arrow ➞ button.
10. Select Conflict in the From List, select Satisfaction in the To List, and click the unidirectional arrow ➞ button.
11. In the text box below Model Name (in the top left of the Model Specification report), type Path Analysis with Latent Variables.
12. Click **Run**.
13. (Optional.) Click the red triangle next to Structural Equation Model: Path Analysis with Latent Variables and select **Path Diagram Settings > Layout > Top to Bottom**.
14. (Optional.) Click the gray disclosure icon next to Parameter Estimates.

    Closing the Parameter Estimates report enables you to see the entire path diagram.

Figure 8.2 Structural Equation Model Report



Structural Equation Model: Path Analysis with Latent Variables

Summary of Fit

Converged in Gradient

Maximum Likelihood

Sample Size	200
Rows with Missing	0
Iterations	8
-2 Log Likelihood	6742.2869
Number of Parameters	36
AICc	6830.6305
BIC	6933.0263
ChiSquare	34.266635
DF	41
Prob>ChiSq	0.7623641
CFI	1
RMSEA	0
Lower 90%	0
Upper 90%	0.0349632

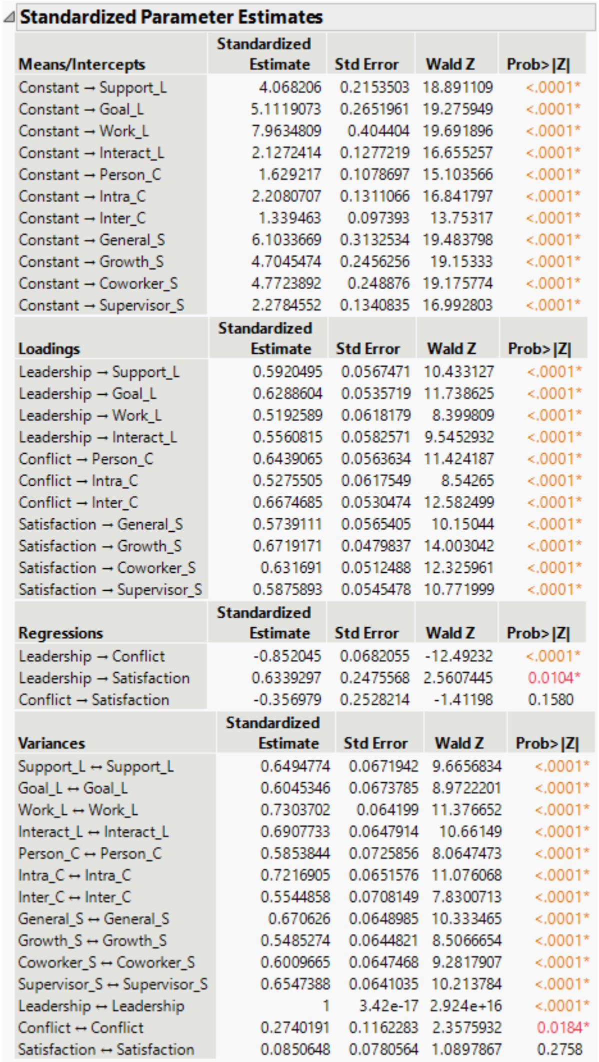
Parameter Estimates

The chi-square statistic for this model, listed in the Summary of Fit report, is 34.27 with 41 degrees of freedom. Note that the corresponding *p*-value is 0.7624, which is not significant. This indicates that there is not evidence to reject the null hypothesis that the model fits well. Therefore, you conclude that this model fits the data reasonably well.

The chi-square value depends on the sample size, and thus, some well-fitting models can still produce a significant chi-square value. The comparative fit index (CFI) and root mean square error of approximation (RMSEA) provide additional guidance for determining model fit. These indices are bounded between 0 and 1. CFI values greater than 0.90 and RMSEA values less than 0.10 are preferred (Browne and Cudeck 1993; Hu and Bentler 1999). Here, the CFI of 1 and RMSEA of 0 indicate an excellent fit.

15. Click the red triangle next to Structural Equation Model: Path Analysis with Latent Variables and select **Standardized Parameter Estimates**.

Figure 8.3 Standardized Parameter Estimates Report



Standardized Parameter Estimates

Means/Intercepts	Standardized Estimate	Std Error	Wald Z	Prob>|Z|
Constant → Support_L	4.068206	0.2153503	18.891109	<.0001*
Constant → Goal_L	5.1119073	0.2651961	19.275949	<.0001*
Constant → Work_L	7.9634809	0.404404	19.691896	<.0001*
Constant → Interact_L	2.1272414	0.1277219	16.655257	<.0001*
Constant → Person_C	1.629217	0.1078697	15.103566	<.0001*
Constant → Intra_C	2.2080707	0.1311066	16.841797	<.0001*
Constant → Inter_C	1.339463	0.097393	13.75317	<.0001*
Constant → General_S	6.1033669	0.3132534	19.483798	<.0001*
Constant → Growth_S	4.7045474	0.2456256	19.15333	<.0001*
Constant → Coworker_S	4.7723892	0.248876	19.175774	<.0001*
Constant → Supervisor_S	2.2784552	0.1340835	16.992803	<.0001*

Loadings	Standardized Estimate	Std Error	Wald Z	Prob>|Z|
Leadership → Support_L	0.5920495	0.0567471	10.433127	<.0001*
Leadership → Goal_L	0.6288604	0.0535719	11.738625	<.0001*
Leadership → Work_L	0.5192589	0.0618179	8.399809	<.0001*
Leadership → Interact_L	0.5560815	0.0582571	9.5452932	<.0001*
Conflict → Person_C	0.6439065	0.0563634	11.424187	<.0001*
Conflict → Intra_C	0.5275505	0.0617549	8.54265	<.0001*
Conflict → Inter_C	0.6674685	0.0530474	12.582499	<.0001*
Satisfaction → General_S	0.5739111	0.0565405	10.15044	<.0001*
Satisfaction → Growth_S	0.6719171	0.0479837	14.003042	<.0001*
Satisfaction → Coworker_S	0.631691	0.0512488	12.325961	<.0001*
Satisfaction → Supervisor_S	0.5875893	0.0545478	10.771999	<.0001*

Regressions	Standardized Estimate	Std Error	Wald Z	Prob>|Z|
Leadership → Conflict	-0.852045	0.0682055	-12.49232	<.0001*
Leadership → Satisfaction	0.6339297	0.2475568	2.5607445	0.0104*
Conflict → Satisfaction	-0.356979	0.2528214	-1.41198	0.1580

Variances	Standardized Estimate	Std Error	Wald Z	Prob>|Z|
Support_L ↔ Support_L	0.6494774	0.0671942	9.6656834	<.0001*
Goal_L ↔ Goal_L	0.6045346	0.0673785	8.9722201	<.0001*
Work_L ↔ Work_L	0.7303702	0.064199	11.376652	<.0001*
Interact_L ↔ Interact_L	0.6907733	0.0647914	10.66149	<.0001*
Person_C ↔ Person_C	0.5853844	0.0725856	8.0647473	<.0001*
Intra_C ↔ Intra_C	0.7216905	0.0651576	11.076068	<.0001*
Inter_C ↔ Inter_C	0.5544858	0.0708149	7.8300713	<.0001*
General_S ↔ General_S	0.670626	0.0648985	10.333465	<.0001*
Growth_S ↔ Growth_S	0.5485274	0.0644821	8.5066654	<.0001*
Coworker_S ↔ Coworker_S	0.6009665	0.0647468	9.2817907	<.0001*
Supervisor_S ↔ Supervisor_S	0.6547388	0.0641035	10.213784	<.0001*
Leadership ↔ Leadership	1	3.42e-17	2.924e+16	<.0001*
Conflict ↔ Conflict	0.2740191	0.1162283	2.3575932	0.0184*
Satisfaction ↔ Satisfaction	0.0850648	0.0780564	1.0897867	0.2758

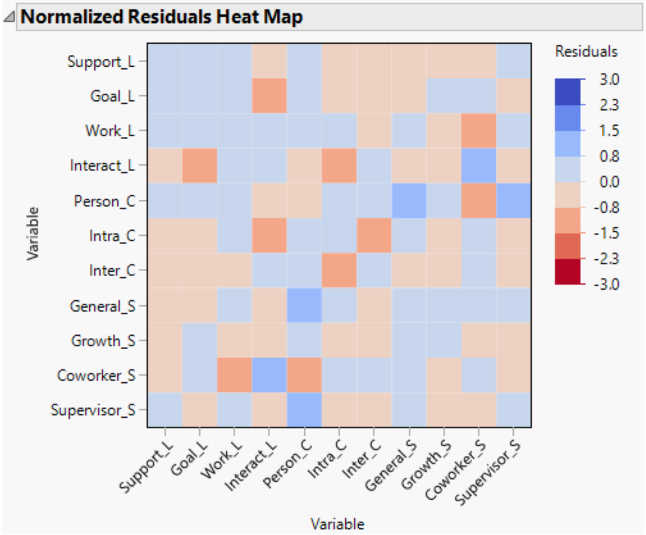
The estimates for the Loadings in the Standardized Parameter Estimates report help explain the measurement model that defines the latent variables. Standardized loadings are the correlation of the observed variable with the unobserved latent variable. Loadings for all latent variables in the report range from 0.52 to 0.67, which suggests well-defined constructs for Leadership, Conflict, and Satisfaction.

In Figure 8.3, the parameter estimates under Regressions suggest a negative effect of Leadership on Conflict and of Conflict on Satisfaction, whereas Leadership has a positive effect on Satisfaction. Thus, higher scores on Leadership are associated with lower Conflict and more Satisfaction, and higher scores in Conflict are associated with lower scores in Satisfaction. The corresponding *p*-values for the parameter estimates are shown under Regressions. The Leadership ⇒ Conflict and Leadership ⇒ Satisfaction regression parameters are both significant at the $\alpha = 0.05$ level. Therefore, you conclude that there are nonzero relationships between the latent variables.

Note: You can also use the Regression parameter estimates in the Standardized Parameter Estimates report as effect sizes. These effect sizes are interpreted as the change in standard deviation units in the outcome for a one standard deviation change in the predictor.

16. Click the red triangle next to Structural Equation Model: Path Analysis with Latent Variables and select **Normalized Residuals Heat Map**.

Figure 8.4 Normalized Residuals Heat Map

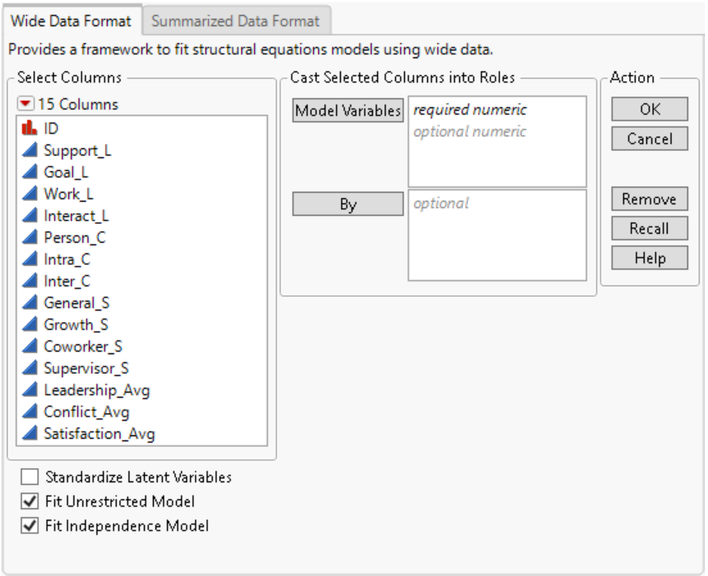


The Normalized Residuals Heat Map has no values that exceed 2.0 units in the positive or negative direction; this is further evidence that the model fits the data well. The residuals can also diagnose ill-fitting models at a more granular level. The normalized residuals in this model do not show evidence of local misfit.

Launch the Structural Equation Models Platform

Launch the Structural Equation Models platform by selecting **Analyze > Multivariate Methods > Structural Equation Models**.

Figure 8.5 Structural Equation Models Launch Window



Provides a framework to fit structural equations models using wide data.

☐ Standardize Latent Variables
☑ Fit Unrestricted Model
☑ Fit Independence Model

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Launch Window includes tabs for two different data formats.

Wide Data Format Select for data tables where each row corresponds to a single observation and the columns contain variables to be used in the model. Rows that contain only missing values are excluded from the analysis.

Summarized Data Format Select for data tables where the data are summarized as a correlation or covariance matrix. Means and standard deviations can also be specified as columns. If they are not specified, the means are assumed to be zero and the standard deviations are assumed to be the square root of the diagonal of the input matrix.

Launch Window Options

Model Variables The columns that you want to include in the model. You must specify at least one column. All columns must have numeric data type and continuous modeling type.

For the Summarized Data Format, the Model Variables columns are the columns that contain the correlation or covariance matrix for the summarized data.

By (Available only for Wide Data Format.) A column that creates a report consisting of separate analyses for each level of the variable. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

Caution: Using a By variable does not perform a multiple-group analysis.

Mean (Available only for Summarized Data Format.) A column of means that correspond to the variables in each row of the correlation or covariance matrix. If they are not specified, the means are assumed to be zero.

Std Dev (Available only for Summarized Data Format.) A column of standard deviations that correspond to the variables in each row of the correlation or covariance matrix. If they are not specified, the standard deviations are assumed to be the square root of the diagonal of the input matrix.

Label (Available only for Summarized Data Format.) A column of labels that correspond to the variables in each row of the correlation or covariance matrix. If they are not specified, the variables in the Model Specification use the names of the columns that contain the input matrix.

Standardize Latent Variables If selected, this option sets the scale of the latent variables by fixing their variance to one and allowing free estimation of all loadings.

Fit Unrestricted Model If selected, the unrestricted model is automatically fit when you fit your first model. You can then compare your fitted model to the unrestricted model in the Model Comparison report. The unrestricted model is a fully saturated model, which fits all means, variances, and covariances of the specified Model Variables without imposing any structure on the data.

Fit Independence Model If selected, the independence model is automatically fit when you fit your first model. You can then compare your fitted model to the independence model in the Model Comparison report. The independence model fits all means and variances of the specified Model Variables. All covariances among the specified Model Variables are fixed to zero, which leads to a highly restrictive model.

Sample Size (Available only for Summarized Data Format.) Specifies the number of observations represented by the summarized data.

The Structural Equation Models Report

- "Model Specification Report"
- "Model Comparison Report"
- "Chi-Square Difference Test Report"
- "Structural Equation Model Fit Report"

Model Specification Report

The Model Specification report contains controls for specifying your model. When you click OK in the launch window, the default model of independence appears in the Diagram and Lists tabs of the Model Specification report. The independence model includes all means and variances of the specified Model Variables.

Specification Panel

The Specification panel contains controls for building your model.

Model Name Enables you to specify a name for the model. When you click Run in the Action panel, a model report that uses the text in Model Name is created.

From List Lists the model variables that were specified in the launch window as well as a Constant term. The Constant term estimates means of any observed or latent variable by regressing each variable on the Constant. Select one or more variables in this list and one or more variables in the To List before clicking one of the arrow buttons to add terms to the model. If you create a latent variable, the latent variable is added to the From List and the To List. You can filter the items in this list using the search filter controls at the top of the list. See "Search Filter Options" on page 195.

To List Lists the model variables that were specified in the launch window as well as a Constant term. The Constant term estimates means of any observed or latent variable by regressing each variable on the Constant. Select one or more variables in the From List and one or more variables in this list before clicking one of the arrow buttons to add terms to the model. If you create a latent variable, the latent variable is added to the From List and the To List. Variables in the To List must be selected to add latent variables to the model. You can filter the items in this list using the search filter controls at the top of the list. See "Search Filter Options" on page 195.

Unidirectional Arrow → Specifies the type of relationship between the selected variables in the From List and the To List. The unidirectional arrow is equivalent to a regression effect.

Bidirectional Arrow ↔ Specifies the type of relationship between the selected variables in the From List and the To List. The bidirectional arrow is equivalent to a covariance effect.

Add Latent ⊞ Adds a latent variable to the From and To lists. The variable is named based on the text in the box to the left of the ⊞ button.

Note: You must select indicators for the latent variable in the To List prior to clicking the ⊞ button.

Remove Latent ⊟ Removes any latent variables that are currently selected in either the From List or the To List. Removing a latent variable from the From and To lists also removes the latent variable from the model.

Model Shortcuts Contains three categories of options to quickly build common models. Note that when you select one of these options, all current model specifications are cleared.

Specification Assist Contains options that shortcut the process of adding covariances for either exogenous or endogenous variables.

Cross-Sectional Classics Contains options for confirmatory factor analysis and mediation analysis.

Longitudinal Analysis Contains options for common latent growth curve models: intercept-only, linear, quadratic, and latent basis growth curve. Additionally, the Fit and Compare Growth Models option enables you to fit no growth, linear growth, and quadratic growth models in one click.

Note: If variables in the To List are selected when you select one of the latent growth curve options, the model is applied only to the selected variables. Otherwise, the model is applied to all of the variables in the To List.

Action Panel

The buttons in the Action panel operate on the lists of specified model variables in the Lists tab. The following operations are available:

Fix To Enables you to fix the parameter value of the currently selected effect to a constant value. When you fix the parameter value for a variable, the fixed value appears in parentheses after the name of the term.

Note: If you keep the default value of 0 when you click OK, the selected effects are removed from the model.

Set Equal Enables you to constrain the parameter values of two or more selected effects to be equal. When you set two or more effects to have equal parameter values, an arbitrary alphanumeric label ("c1") appears in parentheses after the names of the terms. In the case of multiple sets of effects that have parameter values that are set equal to each other, sequential numbering is used ("c1", "c2", and so on).

Note: Equality constraints are allowed only within the same type of parameters.

Free Enables you to remove restrictions on the selected effects in the model. Restrictions include effects that have been set to a fixed value or effects that have been set equal to other effects.

Remove Removes the selected effects from the model.

Undo Undoes the last modification to the model.

Redo Redoes the last undone modification to the model.

Run Fits and creates a report for the currently specified model.

Reset Resets the model specification to the independence model, which is the default.

Diagram Panel

The Diagram panel contains buttons that enable you to adjust the layout of the model diagram. The first button enables you to rotate the manifest variables (represented by rectangles) in the diagram. The second button enables you to cycle through two different arrangements of the diagram. The third button enables you to customize the diagram. See "Customize the SEM Path Diagram" on page 205.

Details Panel

The Details panel contains information about the model that is currently specified in the Model Specification report. Information about the model includes the number of manifest variables, number of latent variables, number of freely estimated parameters, and number of degrees of freedom. You can also adjust the maximum number of iterations.

Diagram Tab

The Diagram tab contains a model diagram that enables you to visualize the currently specified model. Latent variables are represented by circles, and manifest variables are represented by rectangles. Unidirectional arrows represent loadings and regressions. Bidirectional arrows represent variances and covariances. A variance is specified by a bidirectional arrow from one variable to itself. A covariance is specified by a bidirectional arrow between two variables.

The path diagram is interactive, so you can drag items to arrange them. You can also use the arrow keys to move selected items in the path diagram. You can use the buttons in the Diagram panel to rearrange items in the path diagram. You can use the options in the path diagram pop-up menu to add or remove items in the path diagram. The path diagram pop-up menu appears when you right-click in the path diagram itself. The contents of the pop-up menu differs based on whether you click on part of the blank canvas or on an element of the path diagram. You can also zoom in or out in the path diagram by adjusting the zoom setting in the upper left corner of the path diagram. The path diagram also contains a Lock button in the top left corner that enables you to lock the positions of the items in the path diagram. When the path diagram is locked, the Lock button is blue. After making a selection in the path diagram, you can hide the selection by selecting the **Selection > Hide Selection** option in the pop-up menu. When items in the path diagram are hidden, an Unhide All button appears at the top of the path diagram that enables you to show all hidden items.

See “Customize the SEM Path Diagram” on page 205 for more information about path diagram customization options. See “Diagram Panel” on page 192, “Details Panel” on page 192, and “Path Diagram Settings” on page 199.

Tip: To copy the path diagram as an image, right-click the diagram and select **Copy Diagram**. To retain the highest possible quality, paste the clipboard image as a vector graphic.

Lists Tab

The Lists tab contains lists for each type of variable in the model. They are categorized by the arrows that are used to represent them in the model diagram. Unidirectional arrows are used to designate means or intercepts, loadings, and regressions. Bidirectional arrows are used to designate variances and covariances. Means and intercepts for the model appear only in the Lists tab and not in the Diagram tab.

You can filter the items in each list using the search filter controls at the top of each list. See “Search Filter Options” on page 195.

Status Tab

The Status tab contains checks for model identifiability. Structural equation models must be identified to obtain reliable estimates. Identification in the context of SEM means that one can derive unique solutions for every parameter in the model based on the population covariance matrix of the input data. Because you can specify such a variety of models in the SEM framework, there are a variety of rules for identifiability. The Status tab consists of three panels of information: Identification Rules, Model Details, and Data Details.

- The Identification Rules panel contains a list of up to eight rules that are appropriate for a specified model. Some rules are necessary and others are sufficient to guarantee model identification. If necessary rules fail, steps must be taken to correct them before fitting the model. If sufficient rules fail, you do not necessarily need to address any issues. Sometimes sufficient rules can fail without being evidence of anything wrong with the model. Note that all of the rules assume a positive definite covariance matrix. If the covariance matrix is not positive definite, a warning appears below the Model Details panel.

  Tip: For more information about a particular identification rule, click on the row of the table for the rule and a description of the rule appears to the right of the table.

- The Model Details panel contains a list of descriptive values for the currently specified model. These values can be used to detect potential issues with the model.
- The Data Details panel contains a list of descriptive values for the input data. These values can be used to detect potential issues with the data.

If there are singularities in the specified data columns, the report contains a Singularity Details table.

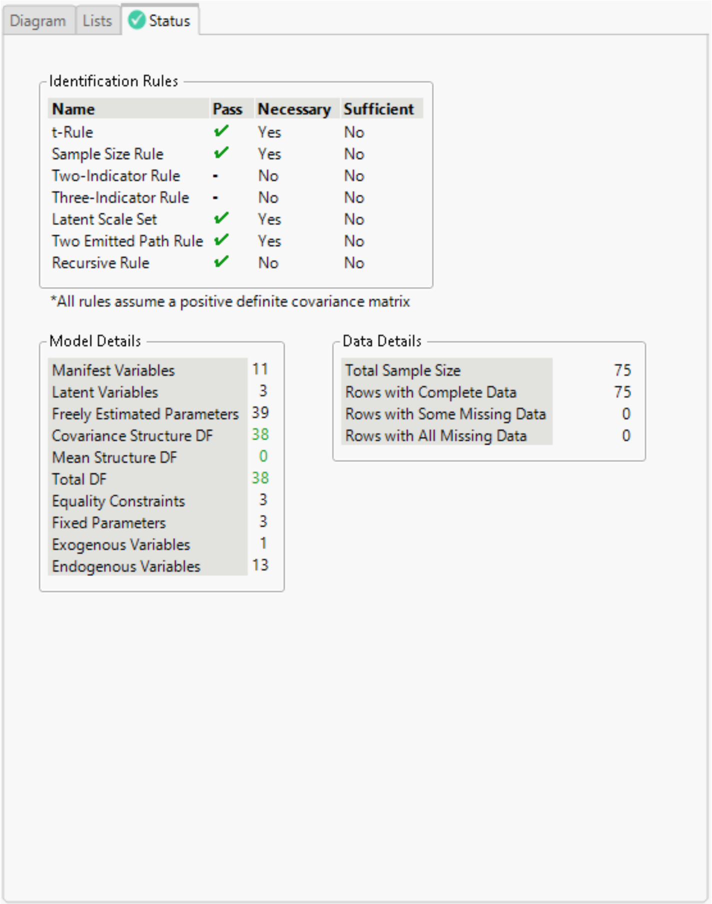
The Status tab itself contains a dynamic icon that shows the current status of the specified model. The icon designates the status:

✓ All applicable identification rules pass.

⊗ At least one necessary identification rule does not pass and you must take steps to correct the issues before fitting the model.

◷ At least one non-necessary identification rule does not pass and further investigation is needed to determine whether the model is correctly specified. Oftentimes, advanced applications of SEM lead to this situation; this does not suggest a problem with the model. Rather, it suggests that the identification rules cannot guarantee that the model is identified.

Figure 8.6 Example Status Tab



Identification Rules

Name	Pass	Necessary	Sufficient
t-Rule	✓	Yes	No
Sample Size Rule	✓	Yes	No
Two-Indicator Rule	-	No	No
Three-Indicator Rule	-	No	No
Latent Scale Set	✓	Yes	No
Two Emitted Path Rule	✓	Yes	No
Recursive Rule	✓	No	No

*All rules assume a positive definite covariance matrix

Model Details

Manifest Variables	11
Latent Variables	3
Freely Estimated Parameters	39
Covariance Structure DF	38
Mean Structure DF	0
Total DF	38
Equality Constraints	3
Fixed Parameters	3
Exogenous Variables	1
Endogenous Variables	13

Data Details

Total Sample Size	75
Rows with Complete Data	75
Rows with Some Missing Data	0
Rows with All Missing Data	0

Search Filter Options

The From List, To List, and the list boxes that appear in the Lists tab in the Model Specification report contain search filters that enable you to filter the items in a specific list.

Click the down arrow button next to the search box to refine your search.

Contains Terms Returns items that contain a part of the search criteria. A search for “ease oom” returns messages such as “Release Zoom”.

Contains Phrase Returns items that contain the exact search criteria. A search for “text box” returns entries that contain “text” followed directly by “box” (for example, “Context Box” and “Text Box”).

Starts With Phrase Returns items that start with the search criteria.

Ends With Phrase Returns items that end with the search criteria.

Whole Phrase Returns items that consist of the entire string. A search for “text box” returns entries that contain only “text box”.

Regular Expression Enables you to use the wildcard (*) and period (.) in the search box. Searching for “get.*name” looks for items that contain “get” followed by one or more words. It returns “Get Color Theme Names”, “Get Name Info”, and “Get Effect Names”, and so on.

Invert Result Returns items that do not match the search criteria.

Match All Terms Returns items that contain both strings. A search for “t test” returns elements that contain either or both of the search strings: “Pat Test”, “Shortest Edit Script” and “Paired t test”.

Ignore Case Ignores the case in the search criteria.

Match Whole Words Returns items that contain each word in the string based on the Match All Terms setting. If you search for “data filter”, and Match All Terms is selected, entries that contain both “data” and “filter” are returned.

Model Comparison Report

The Model Comparison report contains a table of all the models that have been fit. Use the check boxes in the Show column to control which model reports are shown below the Model Comparison report. The rest of the columns to the right of Model Name enable you to compare models based on various criteria.

The AICc Weight value for a model can be interpreted as the probability that a particular model is the true model given that one of the fitted models is the truth. Therefore, the model with the AICc weight closest to one is the best fit. The AICc weights are calculated using only nonmissing AICc values and are defined as follows:

AICcWeight = exp[-0.5(AICc-min(AICc))] / sum(exp[-0.5(AICc-min(AICc))])

where min(AICc) is the smallest AICc value among the fitted models.

For information about the other criteria in the Model Comparison report, see “Structural Equation Model Fit Report” on page 198.

Note: If a model does not converge, an asterisk appears at the beginning of the model name in the Model Comparison report.

To provide context for the performance of the fitted models, the following two models are shown by default in the Model Comparison report:

Unrestricted Fits all means, variances, and covariances of the specified Model Variables without imposing any structure on the data.

Independence Fits all means and variances of the specified Model Variables and fixes all covariances to zero.

Model Comparison Report Options

Two options appear below the Model Comparison table. These options are available after you click on rows to highlight them in the Model Comparison table.

Compare Selected Models (Available only when two or more rows of the table are highlighted.) Computes nested chi-square difference tests between all nested model combinations of the selected rows.

Tip: If any of the models that you select are not nested, a warning appears and the non-nested combination of models does not appear in the Chi-Square Difference Test report.

Clear Selection (Available only when one or more rows of the table are highlighted.) Clears all selections from the rows of the table.

Chi-Square Difference Test Report

The Chi-Square Difference Test report contains a table of nested chi-square tests. The table contains two columns that define the models. The first column contains the more constrained of the two models, and the second column contains the less constrained of the two models. The smaller model is nested within the larger model. The remaining columns show the differences in chi-square values, degrees of freedom, CFI, and RMSEA, as well as the *p*-values for the nested chi-square tests. The Δ notation in the column names indicates differences. A significant Δ chi-square value indicates that the additional constraints in the nested model produce a statistically significant increase in misfit and that the less constrained model should be retained. Because chi-square tests are influenced by sample size, such that they are more likely to be significant as sample sizes increase, the ΔCFI and ΔRMSEA should also be considered; ideally, the ΔCFI should not exceed –0.01 and ΔRMSEA should not exceed 0.015 (Chen 2007).

Caution: Difference tests in this report are meaningful only for nested models.

You can remove any row of the report by clicking the red X button. To remove the report entirely, you must remove all of the rows of the table. If you remove a Structural Equation Model node in the main report, any difference tests involving that model are removed from the table.

Structural Equation Model Fit Report

Each time you click Run, a Structural Equation Model report for the specified model appears. By default, this report contains a Summary of Fit report, a Parameter Estimates report, and a Path Diagram.

Summary of Fit Table of information about the model fit, including the convergence status and the estimation method. The following statistics are reported in this table:

Sample Size The number of observations (rows) used to fit the model.

Rows with Missing The number of observations (rows) that contained at least one missing value. All missing values are handled using full information maximum likelihood (Finkbeiner 1979).

Iterations The number of iterations used to fit the model.

-2 Log Likelihood The log-likelihood of the fitted model multiplied by -2. This value can be used to compare nested models; the difference between two models' -2 Log Likelihood values is chi-square distributed with degrees of freedom equal to the difference of degrees of freedom between the models. See *Fitting Linear Models*.

Number of Parameters The number of freely estimated parameters in the model.

AICc The corrected Akaike information criterion. This value can be used to compare models, where a smaller number indicates a better model fit. See "AICc and BIC" on page 215.

BIC The Bayesian information criterion. This value can be used to compare models, where a smaller number indicates a better model fit. See "AICc and BIC" on page 215.

ChiSquare The chi-square statistic for the model.

DF The degrees of freedom for the chi-square test for model fit.

Prob>ChiSq The p-value of the chi-square statistic for the model.

CFI The Bentler's comparative fit index (CFI) provides additional guidance for determining model fit. The CFI is bounded between 0 and 1. Values greater than 0.90 are preferred (Browne and Cudeck 1993; Hu and Bentler 1999). See "CFI" on page 215.

RMSEA The root mean square error of approximation (RMSEA) provides additional guidance for determining model fit. The RMSEA is bounded between 0 and 1. Values less than 0.10 are preferred (Browne and Cudeck 1993; Hu and Bentler 1999). See "RMSEA" on page 217.

Lower 90% The 90% lower confidence limit for the RMSEA. See "RMSEA" on page 217.

Upper 90% The 90% upper confidence limit for the RMSEA. See "RMSEA" on page 217.

Parameter Estimates Table of estimates for the model parameters. The table is organized in sections for Means/Intercepts, Loadings, Regressions, and Variances. For each estimate, a standard error (Std Error), Wald test statistic (Wald Z), and a corresponding *p*-value (Prob>|Z|) are given.

Tip: The Parameter Estimates table contains hidden columns that identify the From and To components for each model parameter. To show these columns, right-click the table and select **Columns > From** and **Columns > To**.

Path Diagram Shows the path diagram representation of the fitted model. See "Diagram Tab" on page 193.

Structural Equation Models Platform Options

The Structural Equation Models red triangle menu contains the following options:

Path Diagram Settings Contains the following options to modify the path diagram display:

Customize Diagram Enables you to customize many aspects of the path diagram. See "Customize Diagram Appearance Options" on page 207.

Layout Contains two options that change the overall shape of the path diagram. You can choose between a Left to Right layout or a Top to Bottom layout.

Tip: You can also drag items in the path diagram to change the arrangement of specific items.

Copy Diagram Saves an image of the path diagram to the clipboard. To retain the highest possible quality, paste the clipboard image as a vector graphic.

Copy Diagram Properties Copies the current path diagram properties to the clipboard. You can then paste the properties into another SEM path diagram.

Paste Diagram Properties Pastes the path diagram properties from the clipboard into the current SEM path diagram.

Descriptive Statistics Contains the following options to produce descriptive statistics:

Univariate Simple Statistics Shows or hides a univariate simple statistics report. The statistics in this report are estimated for each column independently from other columns that might have missing data.

Full Information Multivariate Statistics Shows or hides a multivariate simple statistics report. The statistics in this report are estimated with full information maximum likelihood to account for missing data.

Launch Explore Outliers Launches the Explore Outliers platform. See *Predictive and Specialized Modeling*.

Launch Explore Missing Values (Not available when there are no missing values.) Launches the Explore Missing Values platform. See *Predictive and Specialized Modeling*.

Add Manifest Variables Enables you to add new manifest variables to the current model. After you select new manifest variables to add, a new Structural Equation Models report is launched. The new report uses the existing model specification with the newly added manifest variables.

Remove Manifest Variables Enables you to launch a new Structural Equation Models report with the specified manifest variables removed. After you select manifest variables to remove, a new report is launched. The new report uses the existing model specification with the specified manifest variables removed.

Copy Model Specification Copies the current structural equation model specifications to the clipboard. You can then paste the model specifications into another SEM platform report.

Paste Model Specification Pastes the model specifications from the clipboard into the current model specifications.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Model Options

After you click Run, a Structural Equation Model report for the specified model appears. This report has a red triangle menu that contains the following options:

Show Path Diagram Shows or hides the path diagram in the model report.

Path Diagram Settings Contains the following options to modify the path diagram for the model:

Customize Diagram Enables you to customize many aspects of the path diagram. See "Customize Diagram Appearance Options" on page 207.

Layout Contains two options that change the overall shape of the path diagram. You can choose between a Left to Right layout or a Top to Bottom layout.

Tip: You can also drag items in the path diagram to change the arrangement of specific items.

Copy Diagram Saves an image of the path diagram to the clipboard. To retain the highest possible quality, paste the clipboard image as a vector graphic.

Copy Diagram Properties Copies the current path diagram properties to the clipboard. You can then paste the properties into another SEM path diagram.

Paste Diagram Properties Pastes the path diagram properties from the clipboard into the current SEM path diagram.

Fit Indices Shows or hides a report that contains a variety of index values that enable you to evaluate the fitted model. In addition to values that appear in the Summary of Fit report, the Fit Indices report also contains the following index values:

TLI The Tucker-Lewis index (TLI) provides additional guidance for determining model fit. This index is also known as the non-normed fit index (NNFI). The TLI is bounded between 0 and 1. Values greater than 0.95 are preferred (West et al. 2012). See "TLI" on page 216.

NFI The normed fit index (NFI) provides additional guidance for determining model fit. The NFI is bounded between 0 and 1. Values greater than 0.95 are preferred (West et al. 2012). See "NFI" on page 216.

Revised GFI The revised goodness-of-fit index provides additional guidance for determining model fit. The revised GFI is bounded between 0 and 1. Values greater than 0.95 are preferred (West et al. 2012). See "Revised GFI and Revised AGFI" on page 216.

Revised AGFI The revised adjusted goodness-of-fit index provides additional guidance for determining model fit. The revised AGFI is bounded between 0 and 1 (West et al. 2012). See "Revised GFI and Revised AGFI" on page 216.

RMR The root mean square residual (RMR) provides additional guidance for determining model fit. The residuals for the RMR are from the differences between the observed and model-implied covariances. The RMR is positive and smaller values are preferred (West et al. 2012). See "RMR and SRMR" on page 218.

SRMR The standardized root mean square residual (SRMR) provides additional guidance for determining model fit. The residuals for the SRMR are from the standardized differences between the observed and model-implied covariances. The SRMR is positive and smaller values are preferred (West et al. 2012). See "RMR and SRMR" on page 218.

Note: For a description of the other values in the Fit Indices report, see "Structural Equation Model Fit Report" on page 198.

Summary of Fit Shows or hides a report that contains details of the model fit.

Parameter Estimates Shows or hides a report that contains the unstandardized parameter estimates for the model.

Standardized Parameter Estimates Shows or hides a report that contains the standardized parameter estimates for the model.

Confidence Intervals Shows or hides confidence intervals in the Parameter Estimates and Standardized Parameter Estimates reports.

Prediction Profiler Enables you to view the effects of a set of predictors on the conditional expected values of a set of outcome variables. When you select this option, a window appears in which you must select one or more predictors and one or more outcomes. The predictions and 95% confidence intervals are based on the model-implied covariance matrix. For more information about the prediction profiler, see *Profilers*.

Note: The initial list of variables in the setup window is limited to variables that are consistent with the model. For example, the Select Predictors list contains only variables that predict something in the model and the Select Outcomes list contains only variables that are predicted by some other variable in the model. Check the **Show All Variables** box to see all model variables in both lists.

Model Implied Covariances Shows or hides a report that contains the covariance matrix that is implied by the model.

Model Implied Means Shows or hides a report that contains the means for each variable that are implied by the model.

Residuals Shows or hides a report that contains a matrix of the residuals for the model. This matrix is the difference between the model implied covariance matrix and the sample covariance matrix.

Normalized Residuals Shows or hides a report that contains a matrix of the normalized residuals for the model.

Normalized Residuals Heat Map Shows or hides a report that contains a heat map of the normalized residuals for the model.

RAM Matrices Shows or hides a report that contains the model matrices used in reticular action model (RAM) notation.

Covariance of Estimates Shows or hides a report that contains the covariance matrix of the parameter estimates for the model.

R^2 for Endogenous Variables (Available only when the model is recursive and contains endogenous variables.) Shows or hides a report that contains the R^2 values for each endogenous variable in the model. This value is calculated as 1 minus the ratio of the residual variance and the model-implied variance for each endogenous variable. The R^2 values represent how much variance is explained by the model in an endogenous variable. An endogenous variable is one that has a path directed at it in the path diagram.

Modification Indices Enables you to show all or a subset of the estimates of model modification indices. These values can be used to determine which parameters might be added to the model to improve model fit. Each table is sorted by the ChiSquare column in descending order.

All Modification Indices Shows or hides a table that contains the estimates of all the model modification indices. This table contains a column that indicates the parameter type for each estimate.

Modification Indices for Means Shows or hides a table that contains the estimates of the model modification indices for the means and intercepts.

Modification Indices for Loadings Shows or hides a table that contains the estimates of the model modification indices for the loading parameters.

Modification Indices for Regressions Shows or hides a table that contains the estimates of the model modification indices for the regression parameters.

Modification Indices for Covariances Shows or hides a table that contains the estimates of the model modification indices for the covariance parameters.

Assess Measurement Model (Available only for confirmatory factor models that do not have covariances among unique factors.) Shows or hides a variety of statistics and graphs for quantifying the reliability and validity of tests and measures, including indicator reliability, coefficients omega and H, and a construct validity matrix.

The construct validity matrix helps you determine whether latent variables are measuring what you think they are measuring:

- The lower triangular entries contain the latent variable correlations. These entries enable you to check how strongly correlated the latent variables are with each other and compare that to the hypothesized strength of correlation.
- The upper triangular entries are the squared latent variable correlations. These entries enable you to focus on the overlap in variance across latent variables. These statistics are particularly valuable when compared against the diagonal entries in the matrix.
- The diagonal entries contain the average amount of variance extracted by each latent variable, which is equivalent to the average of the indicator reliabilities for each latent variable. A good latent variable should have high values in the diagonal because its indicators have sufficient systematic variance to define it properly. Ideally, the diagonal entry for each latent variable should be higher than the entries above and to the right of it.

The visualization of the construct validity matrix enables you to compare the diagonal entries to the upper triangular entries.

Save Columns Enables you to save columns based on the fitted structural equation model to the data table.

Save Factor Scores (Available only when there are latent variables in the model.) Saves a column with the factor score computed using the regression method for each latent variable to the data table. The factor scores are calculated in a hidden column that is also added to the data table. This hidden column uses the `Estimate Factor Score()` JSL function. For more information about this function, see Help > Scripting Index.

Save Bartlett Factor Scores (Available only when there are latent variables in the model.) Saves a column with the factor score computed using the Bartlett method for each latent variable to the data table. The factor scores are calculated in a hidden column that is also added to the data table. This hidden column uses the `Estimate Bartlett Factor Score()` JSL function. For more information about this function, see Help > Scripting Index.

Save Prediction Formulas (Available only when there is at least one endogenous or dependent variable in the model.) Saves a column with a formula for the predicted values of the observed outcomes for each variable to the data table. When there are latent variables in the model, factor scores computed using the Bartlett method are also saved to the data table.

Save Observational Residuals (Available only when there is at least one endogenous or dependent variable in the model.) Saves a column with the residual values of the observed outcomes for each variable to the data table. When there are latent variables in the model, factor scores computed using the Bartlett method are also saved to the data table.

Copy Model Specification Copies the current structural equation model specifications to the clipboard. You can then paste the model specifications into another SEM platform report.

Recall in Model Specification Sets the model in the Model Specification report to the specified model.

Remove Fit Removes the specified model report from the report window.

Customize the SEM Path Diagram

You can customize the path diagram in many ways. Most of the customization options are in one or both of the following places: the pop-up menu in the path diagram and the Customize Diagram Appearance window.

- "Path Diagram Pop-Up Menu Options"
- "Customize Diagram Appearance Options"

Path Diagram Pop-Up Menu Options

The path diagram pop-up menu appears when you right-click in the path diagram itself. The specific set of options that are available in the menu depends on where you right-click in the diagram and if there are any elements of the path diagram selected when you right-click.

Selection Contains options to make selections in the From and To Lists, add unidirectional and bidirectional arrows to the diagram, create new latent variables, hide selected items, and show all hidden items. Many of the options in this submenu are not enabled until you select items in the path diagram.

Tip: When items in the path diagram are hidden, an Unhide All button appears at the top of the path diagram that enables you to show all hidden items.

Show Contains options to show or hide various elements of the path diagram. Select **Show Default** to return to the original set of shown elements.

Note: The Show Default option does not affect the setting of the Show Means/Intercepts option.

For path diagrams in fitted model reports, this submenu also contains options for changing which estimates appear on the arrows in the path diagram. You can choose between showing the unstandardized parameter estimates, the standardized parameter estimates, or no estimates.

Select All of This Type (Available only when you right-click an item in the path diagram.) Selects all other items in the path diagram that are the same type as the item that you right-clicked.

Select Latent Group(s) (Available only when you right-click on a latent variable item in the path diagram.) Selects the entire group of items associated with the selected latent variables. If no latent variables are selected, this option selects only the group of items associated with the latent variable that you right-clicked.

Rename Variable (Available only when you right-click on a latent variable item in the path diagram.) Enables you to change the name of the latent variable in the path diagram.

Note: When you rename a latent variable in the path diagram, the name is updated in the From and To Lists.

Add Regression(s) (Available only when you right-click on a variable in the path diagram.) Enables you to add a unidirectional arrow that represents a regression between the variable that you right-clicked and another variable in the path diagram. When you select this option, click the variable where you want the arrow to point.

Add Covariance(s) (Available only when you right-click on a variable in the path diagram.) Enables you to add a bidirectional arrow that represents a covariance between the variable that you right-clicked and another variable in the path diagram. When you select this option, click the variable where you want the arrow to point.

Customize Diagram Launches the Customize Diagram Appearance window. See "Customize Diagram Appearance Options" on page 207.

Rotate Latent Groups Rotates latent variable indicators clockwise in 90-degree angle increments for the selected latent variables in the path diagram. If no latent variables are selected, this option rotates all latent variable indicators in the path diagram.

Layout Enables you to select one of two arrangement options for the items in the path diagram.

Copy Diagram Saves an image of the path diagram to the clipboard. To retain the highest possible quality, paste the clipboard image as a vector graphic.

Copy Diagram Properties Copies the current path diagram properties to the clipboard. You can then paste the properties into another SEM path diagram.

Paste Diagram Properties Pastes the path diagram properties from the clipboard into the current SEM path diagram.

Undo Undoes the last change to the path diagram.

Redo Redoes the last undone change to the path diagram.

Edit Contains standard options for graphs in JMP. For more information about these options, see *Using JMP*.

Reset Layout Resets the path diagram to the original settings.

Customize Diagram Appearance Options

The Customize Diagram Appearance window appears when you click the customize button in the Diagram panel, select Customize Diagram from the red triangle menu, or select Customize Diagram from the path diagram pop-up menu. The Customize Diagram Appearance window contains four panels as well as a drop-down menu that enables you to select from two preset color themes.

Theme Preset Contains options for preset color themes. You can choose between a Black and White theme or a Blue theme. Select one of these to quickly update the settings to follow either theme.

Variable Appearance Contains options for setting the fill color, border color, text color, and size of the items in the path diagram. You can set the appearance for manifest, latent, and constant variables, as well as the font for each of these items.

Path Settings Contains options for the appearance of the arrows. You can set the thickness and transparency to fixed values or base them on the standardized estimates for each path arrow. You can also represent non-significant p-values with dashed lines, set the α level for non-significance, and change the color and font associated with the arrows. By default, the thickness of the arrows is fixed and the transparency of the arrows is proportional to the standardized parameter estimates.

Other Settings Contains the following options:

Enable Grid Shows or hides a grid in the path diagram that can be used to help align items in the diagram.

Lock Diagram Locks or unlocks the placement of items in the path diagram. When the path diagram is locked, you cannot drag items around and the Lock indicator at the top of the path diagram is highlighted.

Show Means/Intercepts Shows or hides the means and intercepts in the path diagram. The mean structure is not displayed in the path diagram by default.

Show Latent Indicators Shows or hides the latent indicators in the path diagram.

Show Regressions Shows or hides the arrows for regressions in the path diagram.

Show Variances Shows or hides the arrows for variances in the path diagram.

Show Covariances Shows or hides the arrows for covariances in the path diagram.

Show Equality Constraints Shows or hides equality constraints in the path diagram.

Show R Squared Values Shows or hides R^2 values for each node in the path diagram.

Fill Nodes with R Squared Shows or hides a fill coloring for each node in the path diagram that corresponds to the R^2 value for each node.

Preview Contains a preview of the path diagram appearance based on your current settings in the window.

Tip: Use the **Save to Preferences** button at the bottom of the Customize Diagram Appearance window to save the current settings in the window to your platform preferences so that future path diagrams use the same settings.

JMP PRO Additional Example of Structural Equation Models

In this example, you are building the structural equation model for industrialization and political democracy described in Bollen (1989), which uses data from 75 developing countries. The variables in the data table include four measures of democracy in 1960 and 1965, and three measures of industrialization in 1960. These variables are described in the Notes column property in each column of the data table. To view the Notes column property, right-click a column name, select Column Info, and select Notes under Column Properties. The type of structural equation model that you are building is a structural regression model.

There are four main steps to the model specification process: creating latent variables, adding the loading and regression variables, adding the covariance terms, and placing constraints on the loading variables.

1. Select **Help > Sample Data Library** and open Political Democracy.jmp.
2. Select **Analyze > Multivariate Methods > Structural Equation Models**.
3. Select Prod60 through Legis65 and click **Model Variables**.
4. Click **OK**.

   The Structural Equation Models report Model Specification outline appears.

5. Click the **Lists** tab in the View panel box.

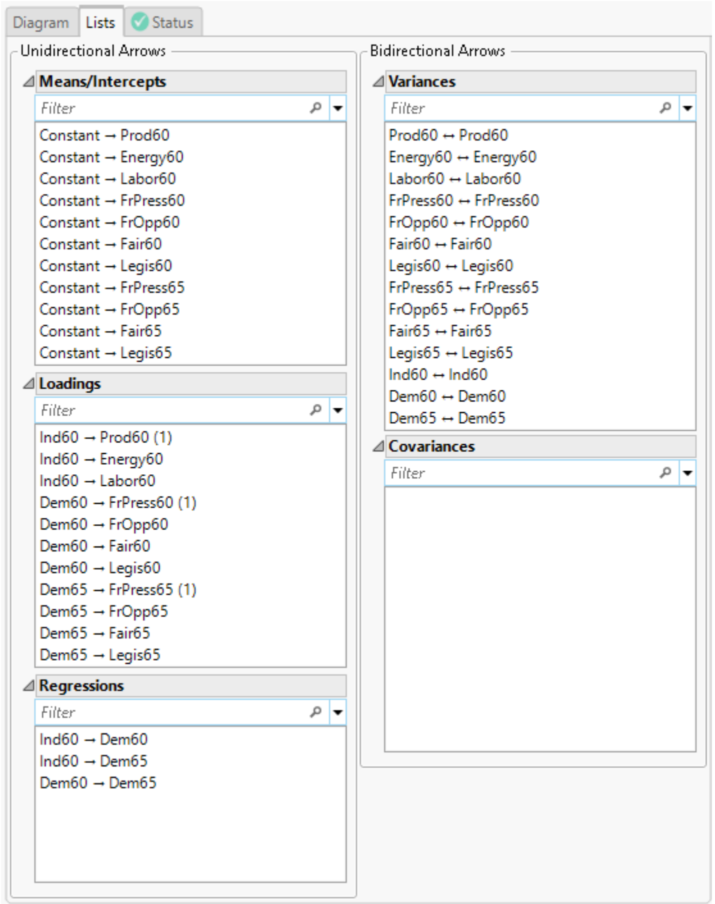
Create Latent Variables

6. Select Prod60 through Labor60 in the To List, type Ind60 in the box below the To List, and click the add latent ⊞ button.
7. Select FrPress60 through Legis60 in the To List, type Dem60 in the box below the To List, and click the add latent ⊞ button.
8. Select FrPress65 through Legis65 in the To List, type Dem65 in the box below the To List, and click the add latent ⊞ button.

Add Loading and Regression Variables

9. Select Ind60 in the From List, select Dem60 in the To List, and click the unidirectional arrow → button.
10. Select Ind60 in the From List, select Dem65 in the To List, and click the unidirectional arrow → button.
11. Select Dem60 in the From List, select Dem65 in the To List, and click the unidirectional arrow → button.

Figure 8.7 Loadings and Regressions



Unidirectional Arrows

Means/Intercepts

Constant → Prod60
Constant → Energy60
Constant → Labor60
Constant → FrPress60
Constant → FrOpp60
Constant → Fair60
Constant → Legis60
Constant → FrPress65
Constant → FrOpp65
Constant → Fair65
Constant → Legis65

Loadings

Ind60 → Prod60 (1)
Ind60 → Energy60
Ind60 → Labor60
Dem60 → FrPress60 (1)
Dem60 → FrOpp60
Dem60 → Fair60
Dem60 → Legis60
Dem65 → FrPress65 (1)
Dem65 → FrOpp65
Dem65 → Fair65
Dem65 → Legis65

Regressions

Ind60 → Dem60
Ind60 → Dem65
Dem60 → Dem65

Bidirectional Arrows

Variances

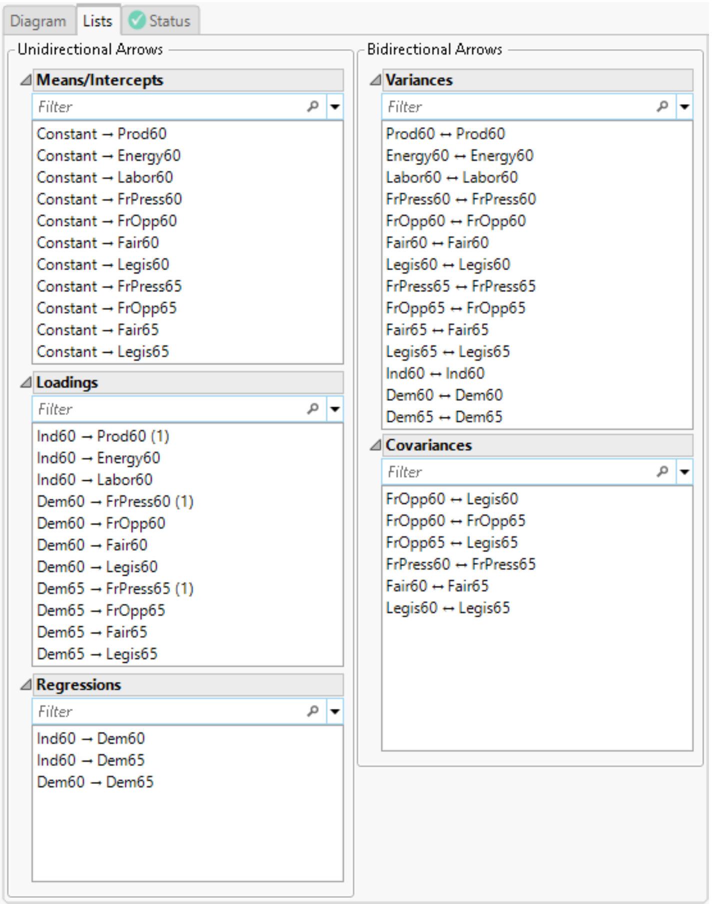
Prod60 ↔ Prod60
Energy60 ↔ Energy60
Labor60 ↔ Labor60
FrPress60 ↔ FrPress60
FrOpp60 ↔ FrOpp60
Fair60 ↔ Fair60
Legis60 ↔ Legis60
FrPress65 ↔ FrPress65
FrOpp65 ↔ FrOpp65
Fair65 ↔ Fair65
Legis65 ↔ Legis65
Ind60 ↔ Ind60
Dem60 ↔ Dem60
Dem65 ↔ Dem65

Covariances

Add Covariances

12. Select FrOpp60 in the From List, select Legis60 and FrOpp65 in the To List, and click the bidirectional arrow ↔ button.
13. Select FrOpp65 in the From List, select Legis65 in the To List, and click the bidirectional arrow ↔ button.
14. Select FrPress60 in the From List, select FrPress65 in the To List, and click the bidirectional arrow ↔ button.
15. Select Fair60 in the From List, select Fair65 in the To List, and click the bidirectional arrow ↔ button.
16. Select Legis60 in the From List, select Legis65 in the To List, and click the bidirectional arrow ↔ button.

Figure 8.8 Covariances



Unidirectional Arrows

Means/Intercepts

Constant → Prod60
Constant → Energy60
Constant → Labor60
Constant → FrPress60
Constant → FrOpp60
Constant → Fair60
Constant → Legis60
Constant → FrPress65
Constant → FrOpp65
Constant → Fair65
Constant → Legis65

Loadings

Ind60 → Prod60 (1)
Ind60 → Energy60
Ind60 → Labor60
Dem60 → FrPress60 (1)
Dem60 → FrOpp60
Dem60 → Fair60
Dem60 → Legis60
Dem65 → FrPress65 (1)
Dem65 → FrOpp65
Dem65 → Fair65
Dem65 → Legis65

Regressions

Ind60 → Dem60
Ind60 → Dem65
Dem60 → Dem65

Bidirectional Arrows

Variances

Prod60 ↔ Prod60
Energy60 ↔ Energy60
Labor60 ↔ Labor60
FrPress60 ↔ FrPress60
FrOpp60 ↔ FrOpp60
Fair60 ↔ Fair60
Legis60 ↔ Legis60
FrPress65 ↔ FrPress65
FrOpp65 ↔ FrOpp65
Fair65 ↔ Fair65
Legis65 ↔ Legis65
Ind60 ↔ Ind60
Dem60 ↔ Dem60
Dem65 ↔ Dem65

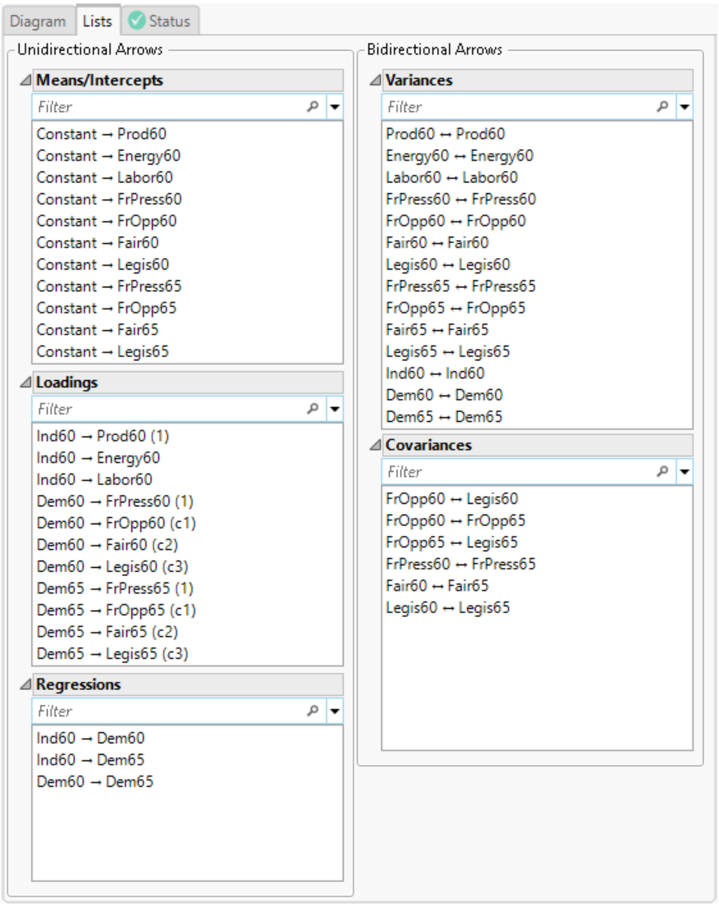
Covariances

FrOpp60 ↔ Legis60
FrOpp60 ↔ FrOpp65
FrOpp65 ↔ Legis65
FrPress60 ↔ FrPress65
Fair60 ↔ Fair65
Legis60 ↔ Legis65

Add Constraints on Loadings

17. Select Dem60->FrOpp60 and Dem65->FrOpp65 in the Loadings list and click **Set Equal**.
18. Select Dem60->Fair60 and Dem65->Fair65 in the Loadings list and click **Set Equal**.
19. Select Dem60->Legis60 and Dem65->Legis65 in the Loadings list and click **Set Equal**.

Figure 8.9 Completed Model Specification



Unidirectional Arrows

Means/Intercepts

Constant → Prod60
Constant → Energy60
Constant → Labor60
Constant → FrPress60
Constant → FrOpp60
Constant → Fair60
Constant → Legis60
Constant → FrPress65
Constant → FrOpp65
Constant → Fair65
Constant → Legis65

Loadings

Ind60 → Prod60 (1)
Ind60 → Energy60
Ind60 → Labor60
Dem60 → FrPress60 (1)
Dem60 → FrOpp60 (c1)
Dem60 → Fair60 (c2)
Dem60 → Legis60 (c3)
Dem65 → FrPress65 (1)
Dem65 → FrOpp65 (c1)
Dem65 → Fair65 (c2)
Dem65 → Legis65 (c3)

Regressions

Ind60 → Dem60
Ind60 → Dem65
Dem60 → Dem65

Bidirectional Arrows

Variances

Prod60 ↔ Prod60
Energy60 ↔ Energy60
Labor60 ↔ Labor60
FrPress60 ↔ FrPress60
FrOpp60 ↔ FrOpp60
Fair60 ↔ Fair60
Legis60 ↔ Legis60
FrPress65 ↔ FrPress65
FrOpp65 ↔ FrOpp65
Fair65 ↔ Fair65
Legis65 ↔ Legis65
Ind60 ↔ Ind60
Dem60 ↔ Dem60
Dem65 ↔ Dem65

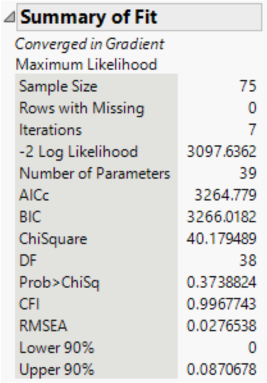
Covariances

FrOpp60 ↔ Legis60
FrOpp60 ↔ FrOpp65
FrOpp65 ↔ Legis65
FrPress60 ↔ FrPress65
Fair60 ↔ Fair65
Legis60 ↔ Legis65

The constraints on the loadings are designated by alphanumeric labels. For example, you can see that Dem60->FrOpp60 and Dem65->FrOpp65 are set equal because they both are labeled “c1”.

20. In the text box below Model Name, type Industrialization and Political Democracy.
21. Click **Run**.

Figure 8.10 Structural Equation Model Summary of Fit Report



Summary of Fit

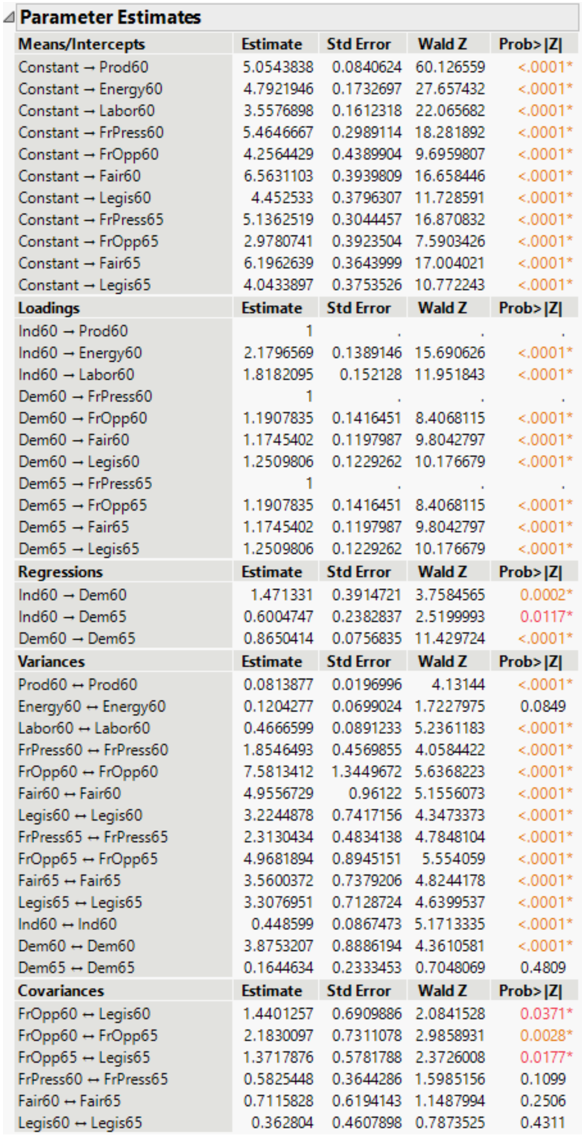
Converged in Gradient
Maximum Likelihood

Sample Size	75
Rows with Missing	0
Iterations	7
-2 Log Likelihood	3097.6362
Number of Parameters	39
AICc	3264.779
BIC	3266.0182
ChiSquare	40.179489
DF	38
Prob>ChiSq	0.3738824
CFI	0.9967743
RMSEA	0.0276538
Lower 90%	0
Upper 90%	0.0870678

The chi-square statistic for this model, which is listed in the Summary of Fit report, is 40.18 with 38 degrees of freedom. Note that the corresponding *p*-value is 0.3739, which is not significant. This indicates that there is not evidence to reject the null hypothesis that the model fits well. Therefore, you conclude that this model fits the data reasonably well.

The chi-square value depends on the sample size, and thus, some well-fitting models can still produce a significant chi-square value. The comparative fit index (CFI) and root mean square error of approximation (RMSEA) provide additional guidance for determining model fit. These indices are bounded between 0 and 1. CFI values greater than 0.90 and RMSEA values less than 0.10 are preferred (Browne and Cudeck 1993; Hu and Bentler 1999). Here, the CFI of 0.9968 and RMSEA of 0.0277 indicate a very good fit.

Figure 8.11 Structural Equation Model Parameter Estimates Report



Parameter Estimates

Means/Intercepts	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
Constant → Prod60	5.0543838	0.0840624	60.126559	<.0001*
Constant → Energy60	4.7921946	0.1732697	27.657432	<.0001*
Constant → Labor60	3.5576898	0.1612318	22.065682	<.0001*
Constant → FrPress60	5.4646667	0.2989114	18.281892	<.0001*
Constant → FrOpp60	4.2564429	0.4389904	9.6959807	<.0001*
Constant → Fair60	6.5631103	0.3939809	16.658446	<.0001*
Constant → Legis60	4.452533	0.3796307	11.728591	<.0001*
Constant → FrPress65	5.1362519	0.3044457	16.870832	<.0001*
Constant → FrOpp65	2.9780741	0.3923504	7.5903426	<.0001*
Constant → Fair65	6.1962639	0.3643999	17.004021	<.0001*
Constant → Legis65	4.0433897	0.3753526	10.772243	<.0001*
Loadings	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
Ind60 → Prod60	1	.	.	.
Ind60 → Energy60	2.1796569	0.1389146	15.690626	<.0001*
Ind60 → Labor60	1.8182095	0.152128	11.951843	<.0001*
Dem60 → FrPress60	1	.	.	.
Dem60 → FrOpp60	1.1907835	0.1416451	8.4068115	<.0001*
Dem60 → Fair60	1.1745402	0.1197987	9.8042797	<.0001*
Dem60 → Legis60	1.2509806	0.1229262	10.176679	<.0001*
Dem65 → FrPress65	1	.	.	.
Dem65 → FrOpp65	1.1907835	0.1416451	8.4068115	<.0001*
Dem65 → Fair65	1.1745402	0.1197987	9.8042797	<.0001*
Dem65 → Legis65	1.2509806	0.1229262	10.176679	<.0001*
Regressions	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
Ind60 → Dem60	1.471331	0.3914721	3.7584565	0.0002*
Ind60 → Dem65	0.6004747	0.2382837	2.5199993	0.0117*
Dem60 → Dem65	0.8650414	0.0756835	11.429724	<.0001*
Variances	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
Prod60 ↔ Prod60	0.0813877	0.0196996	4.13144	<.0001*
Energy60 ↔ Energy60	0.1204277	0.0699024	1.7227975	0.0849
Labor60 ↔ Labor60	0.4666599	0.0891233	5.2361183	<.0001*
FrPress60 ↔ FrPress60	1.8546493	0.4569855	4.0584422	<.0001*
FrOpp60 ↔ FrOpp60	7.5813412	1.3449672	5.6368223	<.0001*
Fair60 ↔ Fair60	4.9556729	0.96122	5.1556073	<.0001*
Legis60 ↔ Legis60	3.2244878	0.7417156	4.3473373	<.0001*
FrPress65 ↔ FrPress65	2.3130434	0.4834138	4.7848104	<.0001*
FrOpp65 ↔ FrOpp65	4.9681894	0.8945151	5.554059	<.0001*
Fair65 ↔ Fair65	3.5600372	0.7379206	4.8244178	<.0001*
Legis65 ↔ Legis65	3.3076951	0.7128724	4.6399537	<.0001*
Ind60 ↔ Ind60	0.448599	0.0867473	5.1713335	<.0001*
Dem60 ↔ Dem60	3.8753207	0.8886194	4.3610581	<.0001*
Dem65 ↔ Dem65	0.1644634	0.2333453	0.7048069	0.4809
Covariances	**Estimate**	**Std Error**	**Wald Z**	**Prob>|Z|**
FrOpp60 ↔ Legis60	1.4401257	0.6909886	2.0841528	0.0371*
FrOpp60 ↔ FrOpp65	2.1830097	0.7311078	2.9858931	0.0028*
FrOpp65 ↔ Legis65	1.3717876	0.5781788	2.3726008	0.0177*
FrPress60 ↔ FrPress65	0.5825448	0.3644286	1.5985156	0.1099
Fair60 ↔ Fair65	0.7115828	0.6194143	1.1487994	0.2506
Legis60 ↔ Legis65	0.362804	0.4607898	0.7873525	0.4311

Next, the parameter estimates under Regressions suggest positive effects of Ind60 on Dem60 and Dem65, as well as a positive effect of Dem60 on Dem65. Thus, higher scores on Ind60 are associated with higher Dem60 and Dem65, and higher scores in Dem60 are associated with higher scores in Dem65. The corresponding *p*-values for the parameter estimates are shown under Regressions. All 3 regression parameters are significant at the $\alpha = 0.05$ level. Therefore, you conclude that there are nonzero relationships between the latent variables.

Statistical Details for the Structural Equation Models Platform

Statistical Details for Summary of Fit Measures

AICc and BIC

The formulas for AICc and BIC are defined as follows:

$$\text{AICc} = \text{-2logL} + 2k + \frac{2k(k+1)}{n-(k+1)}$$

$$\text{BIC} = \text{-2logL} + k\ln(n)$$

where:

- -2logL is twice the negative log-likelihood.
- n is the sample size.
- k is the number of parameters.

For more information about the likelihood-based measures in the Model Comparisons report, see *Fitting Linear Models*.

CFI

The comparative fit index (CFI) is calculated as follows:

$$\text{CFI} = \frac{max(\chi_0^2 - df_0, 0) - max(\chi_{min}^2 - df_{min}, 0)}{max(\chi_0^2 - df_0, 0)}$$

where:

- χ_0^2 is the chi-square statistic of the independence model.
- df_0 is the degrees of freedom of the independence model.
- χ_{min}^2 is the chi-square statistic of the fitted model.
- df_{min} is the degrees of freedom of the fitted model.

For more information about the CFI, see Bentler (1990).

TLI

The Tucker-Lewis index (TLI) is defined as follows:

$$\mathrm{TLI} = \frac{\dfrac{\chi_0^2}{df_0} - \dfrac{\chi_{min}^2}{df_{min}}}{\dfrac{\chi_0^2}{df_0} - 1}$$

where:

- χ_0^2 is the chi-square statistic of the independence model.
- df_0 is the degrees of freedom of the independence model.
- χ_{min}^2 is the chi-square statistic of the fitted model.
- df_{min} is the degrees of freedom of the fitted model.

For more information, see West et al. (2012).

NFI

The Bentler-Bonett normed fit index (NFI) is defined as follows:

$$\mathrm{NFI} = \frac{\chi_0^2 - \chi_{min}^2}{\chi_0^2}$$

where:

- χ_0^2 is the chi-square statistic of the independence model.
- χ_{min}^2 is the chi-square statistic of the fitted model.

For more information, see West et al. (2012).

Revised GFI and Revised AGFI

The revised goodness-of-fit index (Revised GFI) is defined as follows:

$$\text{Revised GFI} = \frac{p}{p + 2\left(\dfrac{\chi_{min}^2 - df_{min}}{n - 1}\right)}$$

where:

- χ^2_{min} is the chi-square statistic of the fitted model.
- df_{min} is the degrees of freedom of the fitted model.
- p is number of observed variables in the fitted model.
- n is the sample size.

The revised adjusted goodness-of-fit index (Revised AGFI) is defined as follows:

$$\text{Revised AGFI} = 1 - \frac{p^*}{df_{min}}(1 - RevisedGFI)$$

where:

- p^* is the number of unique entries in the covariance matrix and the mean vector of the observed variables.
- df_{min} is the degrees of freedom of the fitted model.

For more information, see West et al. (2012).

RMSEA

The root mean square error of approximation (RMSEA) is calculated as follows:

$$\text{RMSEA} = \sqrt{\frac{\max(\chi^2_{min} - df_{min}, 0)}{n \times df_{min}}}$$

where:

- n is the sample size.
- df_{min} is the degrees of freedom of the fitted model.
- χ^2_{min} is the chi-square statistic of the fitted model.

The confidence limits for RMSEA are computed using the cumulative distribution function of the noncentral chi-square distribution $\Phi(x | \lambda, d)$. The 90% confidence limits are computed as follows:

$$\text{Lower limit} = \sqrt{\frac{\lambda_L}{n \times df_{min}}}$$

$$\text{Upper limit} = \sqrt{\frac{\lambda_U}{n \times df_{min}}}$$

where:

- λ_L satisfies $\Phi(\chi^2_{min} \mid \lambda_L, df_{\text{min}}) = 0.95$.
- λ_U satisfies $\Phi(\chi^2_{min} \mid \lambda_U, df_{\text{min}}) = 0.05$.

For more information, see Maydeu-Olivares et al. (2017).

RMR and SRMR

The formulas for RMR and SRMR are defined as follows:

$$\text{RMR} = \sqrt{\frac{1}{b}\left[\sum_{i}^{p}\sum_{j}^{i}(s_{ij} - \hat{\sigma}_{ij})^2 + \sum_{i}^{p}(\bar{x}_i - \hat{\mu}_i)^2\right]}$$

$$\text{SRMR} = \sqrt{\frac{1}{b}\left[\sum_{i}^{p}\sum_{j}^{i}\frac{(s_{ij} - \hat{\sigma}_{ij})^2}{s_{ii}s_{jj}} + \sum_{i}^{p}\frac{(\bar{x}_i - \hat{\mu}_i)^2}{s_{ii}}\right]}$$

where:

- p is the number of manifest variables.
- b is the number of unique entries in the covariance matrix and the mean vector of the observed variables:

  $$b = \frac{p(p+1)}{2} + p$$

- s_{ij} is the $(i, j)^{\text{th}}$ element of the input covariance matrix.
- $\hat{\sigma}_{ij}$ is the $(i, j)^{\text{th}}$ element of the predicted covariance matrix.
- $\bar{x}_i$ is the i^{th} element of the vector of sample means.
- $\hat{\mu}_i$ is the i^{th} element of the vector predicted means.

For more information, see the CALIS Procedure chapter in SAS Institute Inc. (2020a).

Chapter 9

Factor Analysis

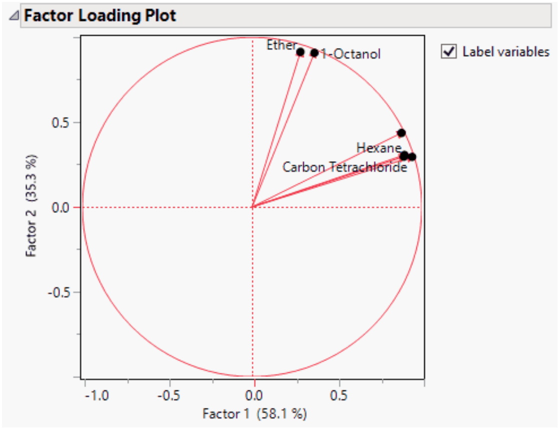
Identify Latent Variables in Your Data

Factor analysis seeks to describe observed variables in terms of a smaller number of (unobservable) latent variables, or factors. Factor analysis is also known as common factor analysis and exploratory factor analysis. The factors are defined as linear combinations of the observed variables (plus error). They are constructed to explain variation that is *common* to the observed variables. A goal of factor analysis is to find a meaningful interpretation of the observed variables in terms of the unobserved factors. Another goal is to reduce the number of variables.

Factor analysis is used in many areas, with roots in psychology, sociology, and education. In these areas, factor analysis is used to understand how observed behavior can be interpreted in terms of underlying patterns and structures. For example, measures of participation in outdoor activities, hobbies, exercise, and travel, might all relate to a factor that can be described as "active versus inactive personality type".

Use factor analysis when you need to explore or interpret underlying patterns and structure in your data. Also consider using it to summarize the information in your variables using a smaller number of latent variables.

Figure 9.1 Rotated Factor Loading



Contents

Overview of the Factor Analysis Platform

Factor analysis models a set of observable variables in terms of a smaller number of unobservable factors. The factors are constructed to account for the correlation or covariance between the observed variables. Factor rotation is used to change the reference axes of the factors to increase their interpretability.

Consider a situation where you have ten observed variables, X_1, X_2, …, X_{10}. Suppose you want to model these ten variables in terms of two latent factors, F_1 and F_2. For convenience, it is assumed that the factors are uncorrelated and that each has mean zero and variance one. The model that you want to derive is of the form:

$$X_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + \varepsilon_i$$

It follows that $Var(X_i) = \beta_{i1}^2 + \beta_{i2}^2 + Var(\varepsilon_i)$. The portion of the variance of X_i that is attributable to the factors, the common variance or *communality*, is $\beta_{i1}^2 + \beta_{i2}^2$. The remaining variance, $Var(\varepsilon_i)$, is the unique variance, and is considered to be a combination of specific and error variances that are unique to X_i.

The platform provides a scree plot for the eigenvalues of the correlation or covariance matrix. You can use this as a guide in determining the number of factors to extract. The platform's default number of factors is the number of eigenvalues that exceed one.

The platform provides two factoring methods for estimating the parameters of this model: Principal Axis and Maximum Likelihood. There are two Prior Communality options for estimating the proportion of variance contributed by common factors for each variable. These options impose assumptions on the diagonal of the correlation (or covariance) matrix. The Principal Components option treats the correlation matrix, which has ones on its diagonal (or the covariance matrix with variances on its diagonal), as the structure to be analyzed. The Common Factor Analysis option sets the diagonal entries to square multiple correlations. These are values that reflect the proportion of the variation shared with other variables.

Factor rotation is used to support interpretability of the extracted factors. The Factor Analysis platform provides a variety of rotation methods that encompass both orthogonal and oblique rotations.

In contrast to factor analysis, which looks at common variance, principal component analysis accounts for the total variance of the observed variables. See the "Principal Components" chapter on page 53.

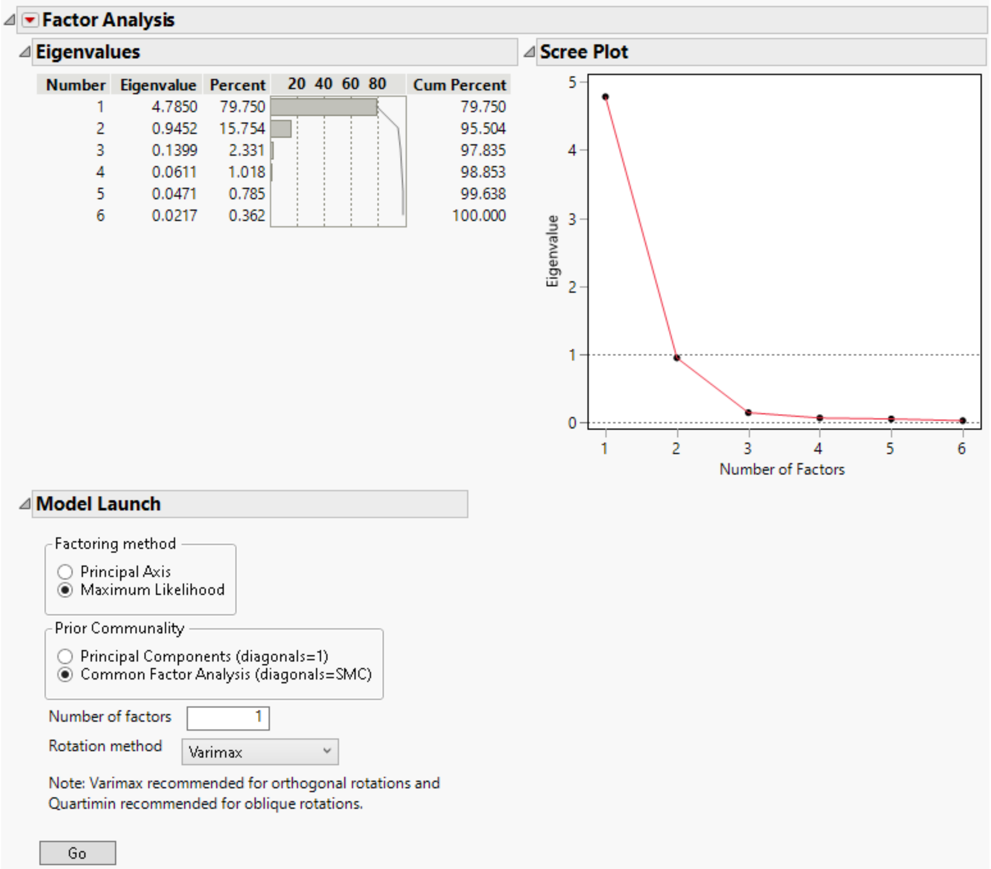
For more information about factor analysis, see Jöreskog (1977) or Cudeck and MacCallum (2007).

Example of the Factor Analysis Platform

This example uses the Solubility.jmp sample data table to extract two factors explained by six solvents.

1. Select **Help > Sample Data Library** and open Solubility.jmp.
2. Select **Analyze > Multivariate Methods > Factor Analysis**.
3. Select 1-Octanol through Hexane and click **Y, Columns**.
4. Click **OK**.

Figure 9.2 Initial Factor Analysis Report



Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.7850	79.750		79.750
2	0.9452	15.754		95.504
3	0.1399	2.331		97.835
4	0.0611	1.018		98.853
5	0.0471	0.785		99.638
6	0.0217	0.362		100.000

Factoring method
○ Principal Axis
● Maximum Likelihood

Prior Communality
○ Principal Components (diagonals=1)
● Common Factor Analysis (diagonals=SMC)

Number of factors: 1

Rotation method: Varimax

Note: Varimax recommended for orthogonal rotations and Quartimin recommended for oblique rotations.

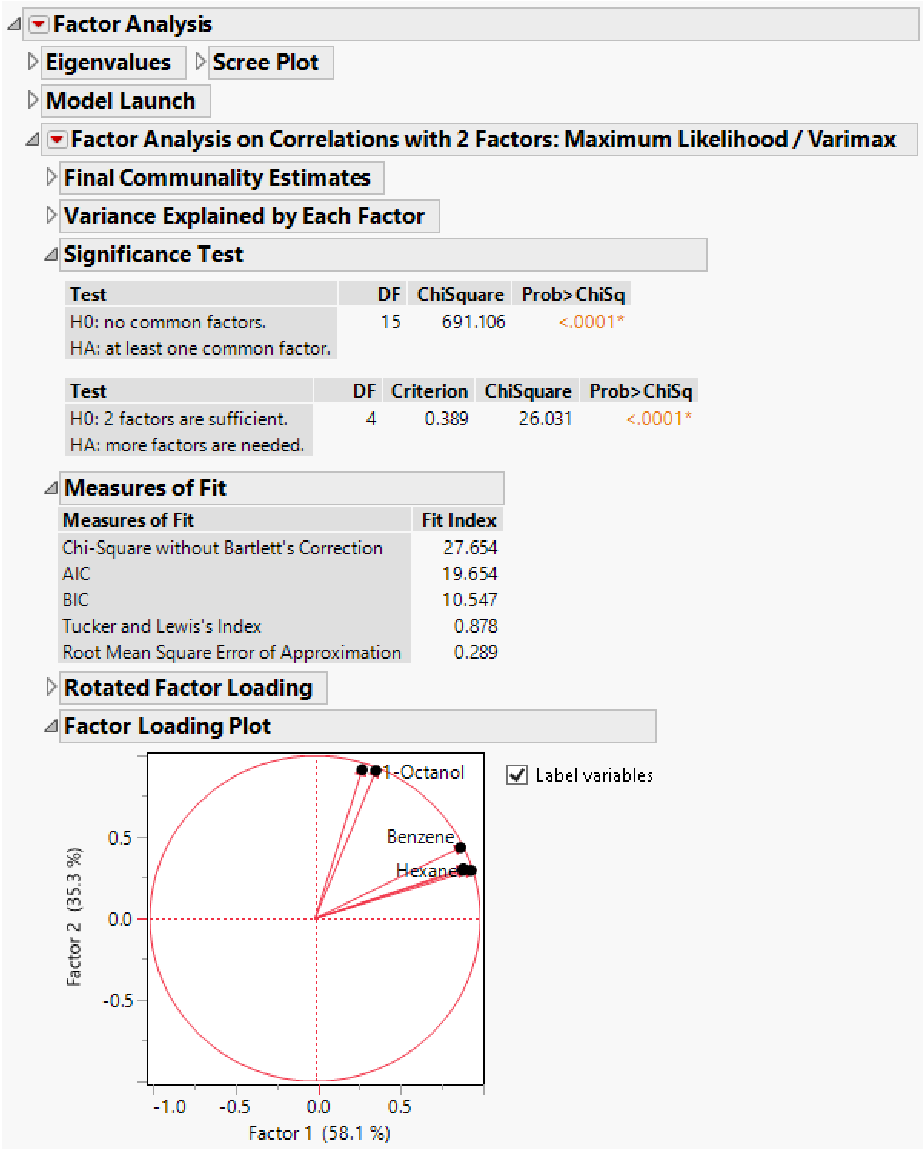
The elbow on the scree plot indicates that a model with two factors is appropriate.

5. In the Model Launch outline, make the following selections:
   - Factoring Method= **Maximum Likelihood**

  - Prior Communality= **Common Factor Analysis**
  - Number of factors = 2
  - Rotation Method= **Varimax**

6. Click **Go**.

Figure 9.3 Factor Analysis Report



Factor Analysis

Eigenvalues

Scree Plot

Model Launch

Factor Analysis on Correlations with 2 Factors: Maximum Likelihood / Varimax

Final Communality Estimates

Variance Explained by Each Factor

Significance Test

Test	DF	ChiSquare	Prob>ChiSq
H0: no common factors. HA: at least one common factor.	15	691.106	<.0001*

Test	DF	Criterion	ChiSquare	Prob>ChiSq
H0: 2 factors are sufficient. HA: more factors are needed.	4	0.389	26.031	<.0001*

Measures of Fit

Measures of Fit	Fit Index
Chi-Square without Bartlett's Correction	27.654
AIC	19.654
BIC	10.547
Tucker and Lewis's Index	0.878
Root Mean Square Error of Approximation	0.289

Rotated Factor Loading

Factor Loading Plot

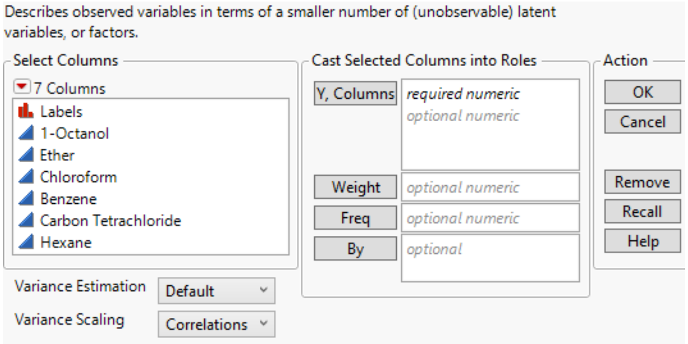
The report lists the communality estimates, variance estimates, significance tests, measures of fit, rotated factor loadings, and a factor loading plot. The Rotated Factor Loadings and Factor Loading Plot suggest that Factor 1 relates to the Carbon Tetrachloride-Chloroform-Benzene-Hexane cluster of variables, and that Factor 2 relates to the Ether–1-Octanol cluster of variables. See "Factor Analysis Model Fit Options" on page 230 for details of the information shown in the report.

Tip: Click the points in the Factor Loading Plot to select and move the labels. Click in the lower right corner to increase the plot size to more easily view the labels.

Launch the Factor Analysis Platform

Launch the Factor Analysis platform by selecting **Analyze > Multivariate Methods > Factor Analysis**.

Figure 9.4 Factor Analysis Launch Window



Describes observed variables in terms of a smaller number of (unobservable) latent variables, or factors.

Variance Estimation: Default

Variance Scaling: Correlations

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns The columns to be analyzed. These must have a Numeric data type.

Weight A column containing a weight for each observation in the data table. A row is included in the analyses only when its value is greater than zero.

Freq Assigns a frequency to each row in the analysis. This is useful when your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Variance Estimation Lists the methods for estimating the variance-covariance matrix for the analysis. For more information about the methods, see "The Multivariate Report" on page 35.

Variance Scaling Lists the scaling methods for performing the factor analysis.

Correlations Default method that enables analysis on correlations.

Covariances Enables the analysis on a weighted correlation matrix where the weights are the variables' variances.

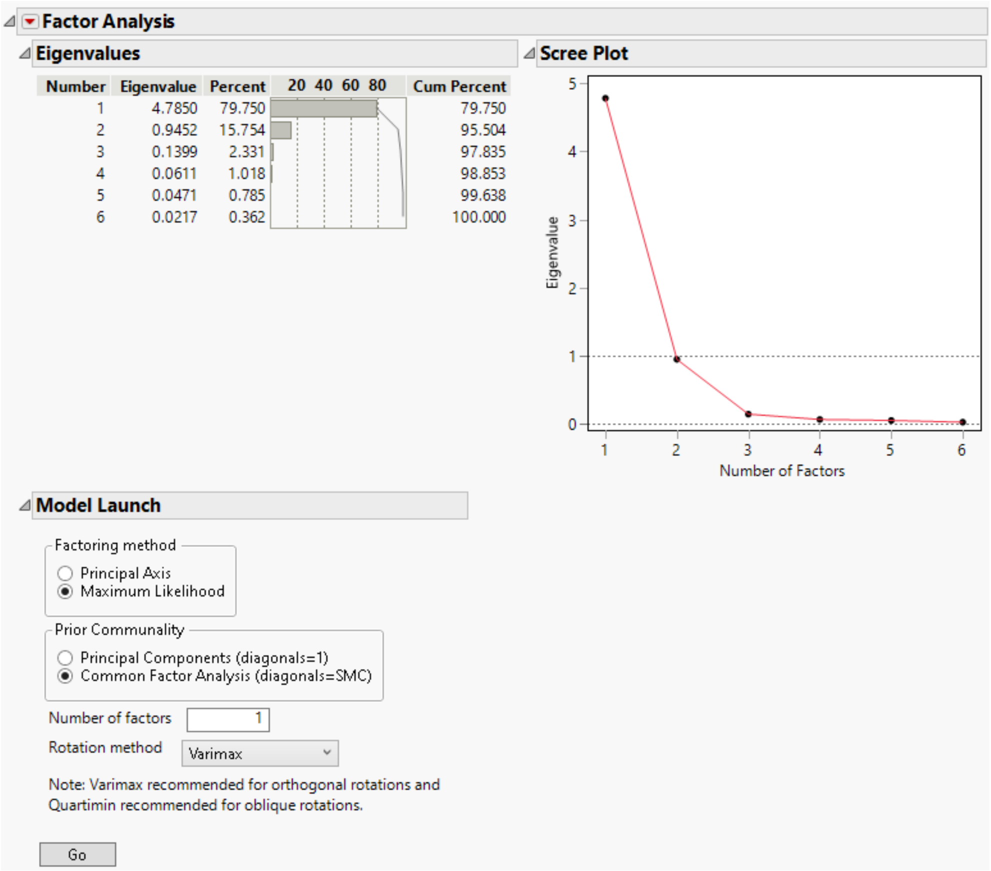
Unscaled Enables the analysis of variables that are already centered or scaled.

The Factor Analysis Report

The initial Factor Analysis report shows Eigenvalues and the Scree Plot. The eigenvalues are obtained from a principal components analysis. The scree plot graphs these eigenvalues. The initial number of factors in the Model Launch equals the number of eigenvalues that exceed 1.0. You can change the number of factors to extract.

Use the scree plot as a guide for the number of factors. The number of eigenvalues before the scree plot levels out provides an upper bound on the number of factors. For this example, two components (factors) are suggested by the scree plot.

Figure 9.5 Factor Analysis Report



Factor Analysis

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	4.7850	79.750	79.750
2	0.9452	15.754	95.504
3	0.1399	2.331	97.835
4	0.0611	1.018	98.853
5	0.0471	0.785	99.638
6	0.0217	0.362	100.000

Model Launch

Factoring method
○ Principal Axis
● Maximum Likelihood

Prior Community
○ Principal Components (diagonals=1)
● Common Factor Analysis (diagonals=SMC)

Number of factors: 1

Rotation method: Varimax

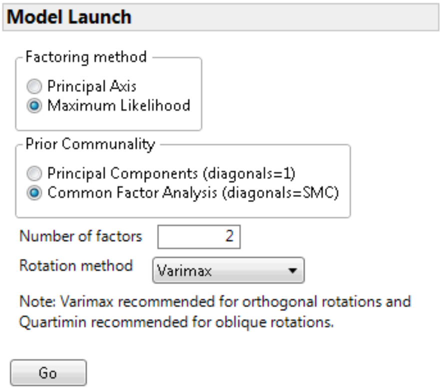
Note: Varimax recommended for orthogonal rotations and Quartimin recommended for oblique rotations.

The Eigenvalues table indicates that the first eigenvalue accounts for 79.75% of the variation and the second eigenvalue accounts for 15.75%. Therefore, the first two eigenvalues account for 95.50% of the total variation. The third eigenvalue explains only 2.33% of the variation, and the contributions from the remaining eigenvalues are negligible. Although the **Number of factors** box is initially set to 1, this analysis suggests that it is appropriate to extract 2 factors.

Model Launch

Configure the Factor Analysis from the Model Launch control panel. Click **Go** to obtain the results of the factor analysis.

Figure 9.6 Model Launch



Model Launch

Factoring method
○ Principal Axis
● Maximum Likelihood

Prior Communality
○ Principal Components (diagonals=1)
● Common Factor Analysis (diagonals=SMC)

Number of factors: 2

Rotation method: Varimax

Note: Varimax recommended for orthogonal rotations and Quartimin recommended for oblique rotations.

Factoring method Defines the method for extracting factors:

Principal Axis Performs eigenvalue decomposition on a reduced correlation or covariance matrix, where the diagonal of the matrix is replaced by an estimate of the communality of the variables. This method is a computationally efficient method, but it does not allow for hypothesis testing.

Maximum Likelihood Enables you to test hypotheses about the number of common factors and to obtain model fit statistics.

Note: The Maximum Likelihood method requires a positive definite correlation matrix. If your correlation matrix is not positive definite, select the Principal Axis method.

Prior Communality Defines the method for estimating the proportion of variance contributed by common factors for each variable.

Principal Components (diagonals = 1) Sets all communalities equal to 1, indicating that 100% of each variable’s variance is explained by all of the factors.

Tip: Using this option with Factoring Method set to Principal Axis results in principal component analysis.

Common Factor Analysis (diagonals = SMC) Sets the communalities equal to squared multiple correlation (SMC) coefficients. For a given variable, the SMC is the RSquare for a regression of that variable on all other variables.

Number of factors Specifies the number of factors to extract from the analysis. The default is the number of eigenvalues that are greater than or equal to 1.0. You can set the number of factors to at least one and no more than the number of variables.

Rotation method Defines the method for rotation. The default is Varimax. See "Rotation Methods" on page 228 for a description of the available rotation methods.

Rotation Methods

Rotations are used to change the reference axes of the factors to make the factors more interpretable. Rotations are applied to the factors extracted from the data. Rotation methods are based on various complexity or simplicity functions. For more information about rotations see the FACTOR Procedure chapter in SAS Institute Inc. (2020c), Browne (2001), or Frank and Todeschini (1994).

After the initial extraction, the factors are uncorrelated with each other. If the factors are rotated by an orthogonal transformation, the rotated factors are also uncorrelated. If the factors are rotated by an oblique transformation, the rotated factors become correlated. Oblique rotations often produce more interpretable factors than do orthogonal rotations. However, a consequence of correlated factors is that there is no single unambiguous measure of the importance of a factor in explaining a variable.

Orthogonal Rotation Methods

Varimax Maximizes the sum of the variances of the squared loadings of a factor on all variables. This common method results in each variable having either a small or large loading on each factor. (Orthomax with $\gamma = 1$.)

Biquartimax An equally weighted solution of the Varimax and Quartimax rotations. (Orthomax with $\gamma = 0.5$.)

Equamax A weighted solution between the Varimax rotation and the Quartimax rotation. (Orthomax with $\gamma = N/2$, where N = number of factors.)

Factorparsimax A solution that aims to minimize the complexity of factors. This method might result in cross-loadings as variable complexity is not considered in the algorithm. (Orthomax with $\gamma = N$, where N = number of factors.)

Orthomax A general weighted rotation method where the weight is denoted by γ. Many specific orthogonal rotation methods are Orthomax rotations with a specific γ.

Parsimax Balances the variable and the factor complexity. (Orthomax with $\gamma = (I(N-1))/(I+N-2)$, where I = the number of items and N = number of factors.)

Quartimax Minimizes the number of factors needed to explain each variable. (Orthomax with $\gamma = 1$.)

Oblique Rotation Methods

Biquartimin A rotation to minimize the ratio of the covariances (Oblimin with $\tau = 0.5$.).

Covarimin Oblique Varimax rotation. (Oblimin with $\tau = 1$.).

Obbiquartimax Oblique Biquartimax rotation.

Obequamax Oblique Equamax rotation.

Obfactorparsimax Oblique factor Parsimax rotation.

Oblimin A general weighted oblique rotation method where the weight is denoted by τ. Many specific oblique rotation methods are Oblimin rotations with a specific τ.

Obparsimax Oblique Parsimax rotation

Obquartimax Oblique Quartimax rotation, equivalent to the Quartimin method.

Obvarimax Oblique Varimax rotation.

Quartimin Oblique Quartimin rotation, equivalent to oblique Quartimax (Oblimin with $\tau = 0$.)

Promax A two step rotation in which Varimax is performed first and then the Procrustes rotation is used to attain simple structure. This is a computationally efficient method that is an alternative to Oblimin.

Factor Analysis Platform Options

The Factor Analysis red triangle menu contains the following options:

Eigenvalues Shows or hides a table of the eigenvalues of the original correlation, covariance, or unscaled matrix. The table includes the percent of the total variance represented by each eigenvalue, a bar chart illustrating the percent contribution, and the cumulative percent contributed by each successive eigenvalue. The number of eigenvalues that are greater than or equal to 1.0 can be taken as a guideline of the number of factors for analysis.

Scree Plot Shows or hides a plot of the eigenvalues versus the number of components (or factors). The plot can be used as an additional guideline to determine the number of factors that contribute to the maximum amount of variance. The point at which the plotted line levels out can be used as the number of sufficient factors for analysis.

Kaiser-Meyer-Olkin Test Shows or hides the results of the Kaiser-Meyer-Olkin (KMO) test. The test is an indicator of the proportion of variance that might be common variance, potentially due to underlying factors. The test statistic is calculated for each individual variable and for the set of all variables and is referred to as the Measure of Sampling Adequacy (MSA) in JMP. The following values indicate the suitability of the variables for factor analysis:

- 0.00 to 0.49 unacceptable
- 0.50 to 0.59 miserable
- 0.60 to 0.69 mediocre
- 0.70 to 0.79 middling
- 0.80 to 0.89 meritorious
- 0.90 to 1.00 marvelous

Note: The KMO test is not available if the correlation matrix is singular.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Factor Analysis Model Fit Options

The Factor Analysis Model Fit red triangle menu contains the following options:

Prior Communality (Available only for Common Factor Analysis.) Shows or hides an initial estimate of the communality for each variable. For a given variable, this estimate is the squared multiple correlation coefficient (SMC), or RSquare, for a regression of that variable on all other variables.

Eigenvalues (Available only for Common Factor Analysis.) Shows or hides the eigenvalues of the reduced correlation matrix and the percent of the common variance for which they account. The reduced correlation matrix is the correlation matrix with its diagonal entries replaced by the communality estimates. The eigenvalues indicate the common variance explained by the factors. The Cum Percent can exceed 100% because the reduced correlation matrix is not necessarily positive definite and can have negative eigenvalues.

Note the table indicates the number of factors retained for analysis.

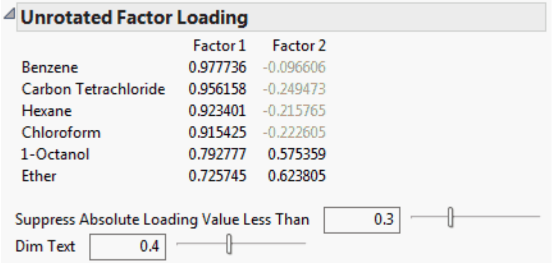
Unsorted and Unrotated Factor Loading Shows or hides the factor loading matrix before sorting and rotation.

Unrotated Factor Loading Shows or hides the factor loading matrix before rotation. Factor loadings measure the influence of a common factor on a variable. Because the unrotated factors are orthogonal, the factor loading matrix is the matrix of correlations between the variables and the factors. The closer the absolute value of a loading is to 1, the stronger the effect of the factor on the variable.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient for factor values less in absolute value than the specified **Suppress Absolute Loading Values Less Than** value.

Figure 9.7 Unrotated Factor Loading with Dim Text Controls



Unrotated Factor Loading

	Factor 1	Factor 2
Benzene	0.977736	-0.096606
Carbon Tetrachloride	0.956158	-0.249473
Hexane	0.923401	-0.215765
Chloroform	0.915425	-0.222605
1-Octanol	0.792777	0.575359
Ether	0.725745	0.623805

Suppress Absolute Loading Value Less Than 0.3

Dim Text 0.4

Note: The Unrotated Factor Loading matrix is ordered so that variables associated with the same factor appear next to each other.

Rotation Matrix Shows or hides the values used for rotating the factor loading plot and the factor loading matrix.

Interfactor Correlations (Available only for oblique rotations.) Shows or hides the matrix of correlations between factors.

Target Matrix (Available only for the Promax rotation.) Shows or hides the matrix to which the varimax factor pattern is rotated.

Factor Structure (Available only for oblique rotations.) Shows or hides the matrix of correlations between variables and common factors.

Final Communality Estimates Shows or hides estimates of the communalities after the factor model has been fit. When the factors are orthogonal, the final communality estimate for a variable equals the sum of the squared loadings for that variable.

Standard Score Coefficients Shows or hides a table of the multipliers used to estimate factor scores when saving rotated factors to the source data table.

Variance Explained by Each Factor (Available only for orthogonal rotations.) Shows or hides the variance, percent, and cumulative percent, of common variance explained by each rotated factor.

Variance Explained by Each Factor Ignoring Other Factors (Available only for oblique rotations.) Shows or hides the variance and percent of common variance explained by each rotated factor regardless of other factors.

Significance Test (Available only for the Maximum Likelihood factoring method.) Provides the results of two Chi-square tests.

The first test is for H_0: No common factors. This null hypothesis indicates that none of the common factors explain the intercorrelations among the variables. This test is Bartlett's Test for Sphericity, whose null hypothesis is that the correlation matrix of the factors is an identity matrix (Bartlett, 1954).

The second test is for H_0: N factors are sufficient, where N is the specified number of factors. Rejection of this null hypothesis indicates that more factors might be required to explain the intercorrelations among the variables (Bartlett, 1954). The Criterion is the log-likelihood objective function value.

Measures of Fit (Available only for the Maximum Likelihood factoring method.) Shows or hides measures of fit: Chi-Squared without Bartlett's Correction, AIC, BIC, Tucker-Lewis's Index, and the root mean square error of approximation.

Measures of Factor Scores Shows or hides measures of factor score determinacy: Multiple R, Multiple R Square, and Minimum Correlation. These measures are used to evaluate if the factor scores might be useful for secondary analyses.

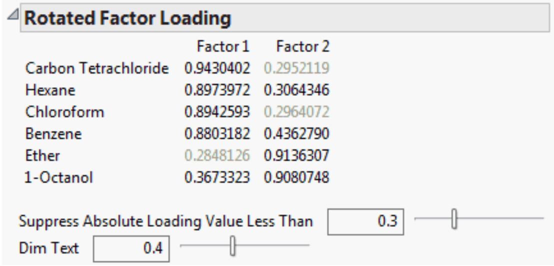
Unsorted and Rotated Factor Loading Shows or hides the unsorted factor loading matrix after rotation.

Rotated Factor Loading Shows or hides the factor loading matrix after rotation. If the rotation is orthogonal, these values are the correlations between the variables and the rotated factors.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient. The lower the value, the more transparency the font for factor values less in absolute value than the specified **Suppress Absolute Loading Values Less Than** value.

Figure 9.8 Rotated Factor Loading with Dim Text Controls



Rotated Factor Loading

	Factor 1	Factor 2
Carbon Tetrachloride	0.9430402	0.2952119
Hexane	0.8973972	0.3064346
Chloroform	0.8942593	0.2964072
Benzene	0.8803182	0.4362790
Ether	0.2848126	0.9136307
1-Octanol	0.3673323	0.9080748

Suppress Absolute Loading Value Less Than 0.3

Dim Text 0.4

Note: The Rotated Factor Loading matrix is ordered so that variables associated with the same factor appear next to each other.

Factor Loading Plot Shows or hides a plot of the rotated factor loadings. When more than 2 factors are modeled, the loading plot is a matrix of plots.

Score Plot Shows or hides a scatterplot of the estimated factor scores. When more than 2 factors are modeled, the score plot is a matrix of plots.

Score Plot with Imputation (Available only if there are missing values.) Shows or hides a scatterplot of the estimated factor scores with imputed values for missing values.

Display Options Enables you to show or hide arrows on the loading plots.

Save Rotated Factors Saves the rotated factor scores and formulas to the data table.

Note: The formula cannot evaluate rows with missing values.

Remove Fit Removes the fit model results from the Factor Analysis report. This option enables you to change the Model Launch configuration for a new report.

Chapter 10

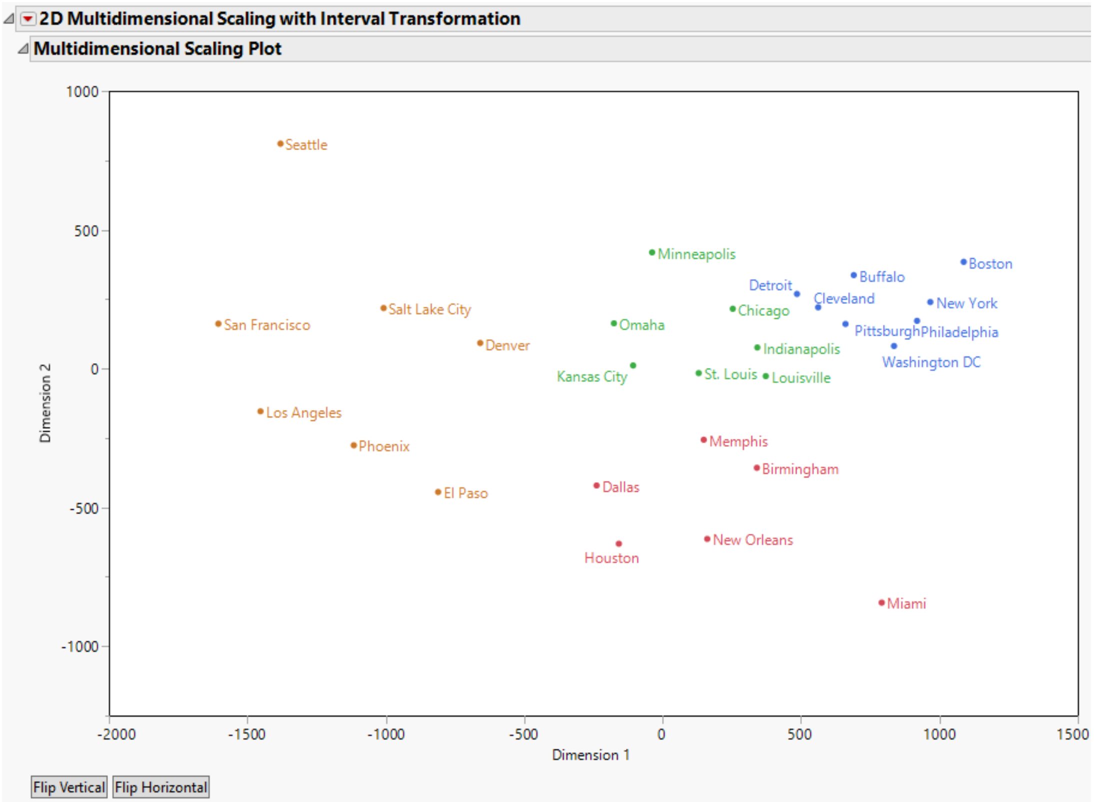
Multidimensional Scaling

Visualize Proximities among a Set of Objects

Multidimensional Scaling (MDS) is a technique that is used to create a visual representation of the pattern of proximities (similarities, dissimilarities, or distances) among a set of objects. For example, given a matrix of distances between cities, MDS can be used to generate a map of the cities in two dimensions.

Multidimensional Scaling is frequently used in consumer research where researchers have measures of perceptions about brands, tastes, or other product attributes. MDS is applicable to many other areas where one is interested in visualizing the proximity of objects based on a set of attributes or proximities.

Figure 10.1 Multidimensional Scaling Example



Contents

Overview of the Multidimensional Scaling Platform

The Multidimensional Scaling platform generates a plot of proximities among a set of objects. This plot can be used to visually explore structure in a data set. MDS is a multivariate technique that is used to visualize the patterns of proximities (distances, similarities) among a set of objects in a small number of dimensions. MDS is applied to a distance matrix. The coordinates for the MDS plot are obtained by minimizing a stress function (the difference between the actual and predicted proximities).

The term distance can refer to a measure of physical distance, such as between cities. More often distance is a subjective assessment rather than a precise measurement. Proximities can measure perceived similarities between brands of a product, correlations of crime rates, or economic similarities for a sample of countries. Distance can also be called proximity or similarity (dissimilarity). If the data are given as an attribute list, then a distance matrix is first constructed from the attribute list.

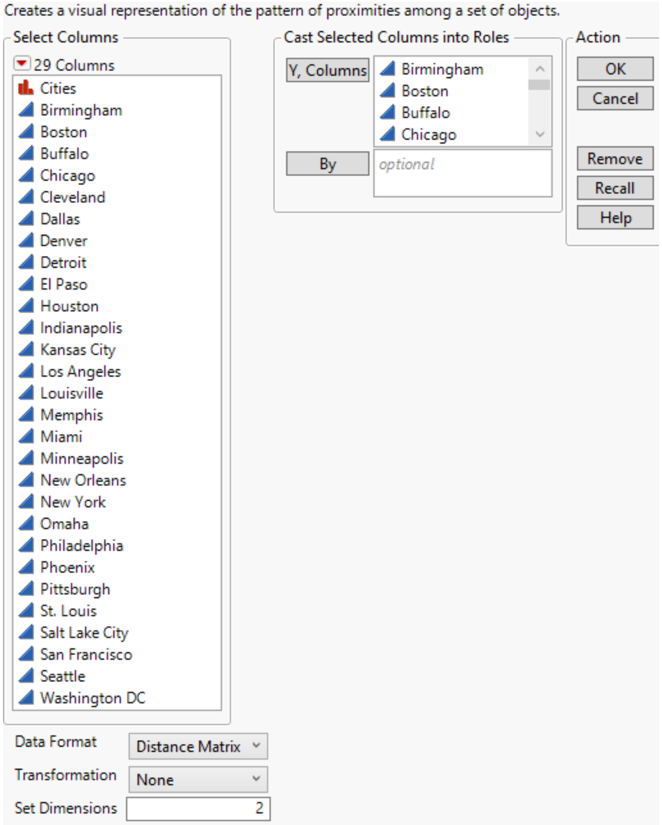
For more information about multidimensional scaling, see Borg and Groenen (2005) or Jackson (2003).

Example of Multidimensional Scaling

This example uses the Flight Distances.jmp sample data table, which is a distance matrix of flight distances between 28 US cities. You can use MDS to construct a map of the cities in two dimensions that is based on the pairwise distances in the data table.

1. Select **Help > Sample Data Library** and open Flight Distances.jmp.
2. Select **Analyze > Multivariate Methods > Multidimensional Scaling.**
3. Select Birmingham through Washington DC and click **Y, Columns.**

Figure 10.2 Completed Multidimensional Scaling Launch Window



Creates a visual representation of the pattern of proximities among a set of objects.

Y, Columns: Birmingham, Boston, Buffalo, Chicago

Data Format: Distance Matrix

Transformation: None

Set Dimensions: 2

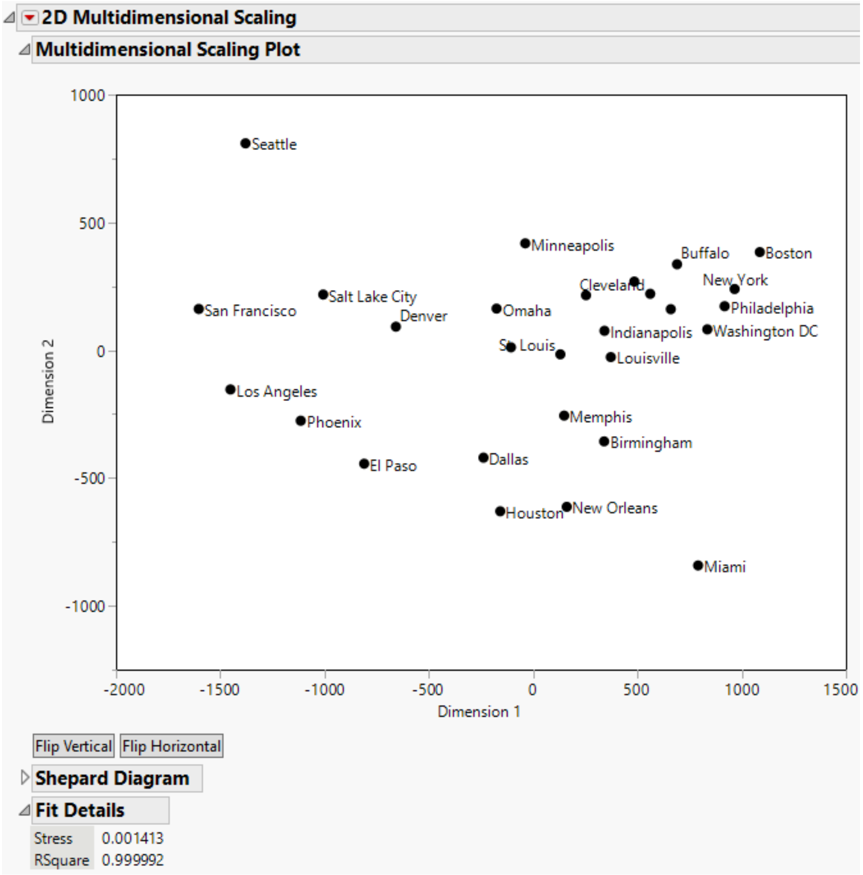
4. Click **OK.**

   In the Multidimensional Scaling Plot, hover over data points to view the row number or row label. The next 7 steps are to label and rotate your MDS Plot.

5. Select the Flight Distances data table.
6. Right-click the column Cities and select **Label/Unlabel**.
7. Select **Rows > Row Selection> Select all Rows**.
8. Select **Rows > Label/Unlabel**.
9. Select the Multidimensional Scaling Plot.
10. Click the **Flip Vertical** button.
11. Click the **Flip Horizontal** button.

    The Flip Vertical and Flip Horizontal buttons enable you to change the orientation of the MDS Plot. The MDS results are invariant to orientation. When the results have a known orientation, such as physical locations, then you might want to rotate or flip your plot.

Figure 10.3 Multidimensional Scaling Plot

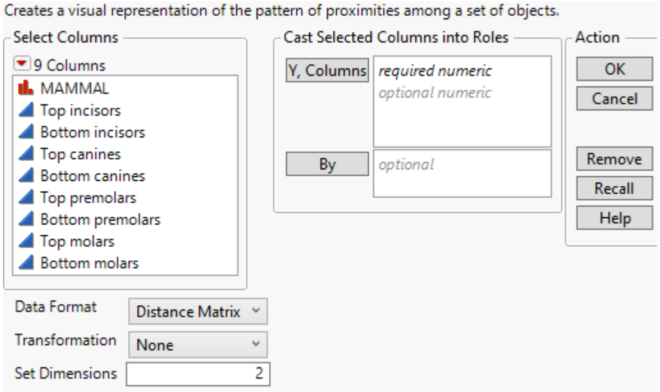


Stress	0.001413
RSquare	0.999992

Launch the Multidimensional Scaling Platform

Launch the Multidimensional Scaling Platform by selecting **Analyze > Multivariate Methods > Multidimensional Scaling**.

Figure 10.4 Multidimensional Scaling Launch Window



Creates a visual representation of the pattern of proximities among a set of objects.

Data Format: Distance Matrix
Transformation: None
Set Dimensions: 2

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns The columns to be analyzed. These must have a Numeric data type.

By A column or columns whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed using the other variables that you have specified. The results are presented in separate reports. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Note: When using a distance matrix, the By variable requires a full matrix for each level of the By variable.

Data Format MDS supports two data formats:

Distance Matrix A full symmetric, lower, or upper triangular matrix where the number of rows equals the number of columns.The diagonal entries can either be zeros or missing.

Attribute List A set of columns that contain measures of a quality or characteristic of an object. The objects are typically named in a column. The object column is not used in the analysis but rather is used as a label for the data points on the MDS plot.

Transformation Supported transformations are Ratio, Interval, and Ordinal.

None No transformation used.

Ratio Data has an ordering from smallest to largest, the differences between values have meaning, and the scale has a true zero. Used to scale the MDS plot.

Interval Data has an ordering from smallest to largest and the differences between values have meaning. Used to scale and shift the MDS plot.

Ordinal Data has an ordering from smallest to largest. Used for ordinal data.

Set Dimensions The number of dimensions for the visual representation of the proximities among your objects. Typically, two or three dimensions are used. With greater than three dimensions, the visualization becomes complex.

Note: The dimension selected can be between 1 to n - 1 where n = the number of objects, otherwise the dimension is set to 2.

The Multidimensional Scaling Report

The initial Multidimensional Scaling report shows these reports: Multidimensional Scaling Plot, the Shepard Diagram, and the Fit Details. If you specify three or more dimensions for the fit in the launch window, then the Multidimensional Scaling Plot provides controls for selecting the dimensions that you view.

Objects that are close together on the MDS plot share similar attributes. Adding labels and colors to the plot can help in the identification of similar groups. The Shepard diagram and summary of fit statistics provide measures of how well the MDS plot represents the proximities of the objects.

Multidimensional Scaling Plot

The MDS plot displays the multidimensional scaling in two dimensions. Below the plot are two buttons to flip the axis either in the vertical or horizontal direction. The MDS solution can be reflected, rotated, or translated without changing the inter-point proximities. The rotating or reflection of the axes is most common when working with geographical objects that have a known map orientation.

If more than two dimensions were used in the analysis, then you can toggle the dimensions shown in the plot using the **Select Dimension** controls below the plot. The first control defines the horizontal axis of the plot, and the second control defines the vertical axis of the plot.

Shepard Diagram

The Shepard plot is a plot of the actual or transformed proximities versus the predicted proximities. The plot indicates how well the Multidimensional Scaling Plot reflects the actual proximities. The Shepard is analogous to an Actual by Predicted plot. Ideally the points fall on the Y = X line, which is shown in red.

Fit Details

The Fit Details provides statistics that summarize how well the MDS proximities match the actual proximities as well as details about transformations when used.

Stress The value of the stress function (Stress1) that was minimized in the fitting procedure. Stress can be between 0 and 1 with lower values indicating a better fit.

RSquare The R^2 value for linear fit of the actual or transformed proximities versus the predicted proximities.

Slope If a ratio or interval transformation was used, the slope for the transformation is provided. It is the slope of the linear regression of the actual against transformed proximities.

Intercept If an interval transformation was used, the intercept for the transformation is provided. It is the intercept of the linear regression of the actual against transformed proximities.

Multidimensional Scaling Platform Options

The Multidimensional Scaling red triangle menu options give you the ability to customize reports according to your needs. The options available are determined by the type of data and the number of dimensions that you use for your analysis.

MDS Plot Shows or hides the MDS Plot.

Diagnostics Provides diagnostics for the MDS.

Shepard Diagram Shows a plot of actual proximity (or transformed proximity if a transformation is used) versus the predicted proximity. This report appears by default. See Shepard Diagram.

Waern Links Displays the Waern links on the MDS plot. Controls for the portion (smallest or largest) are available when this option is selected. See Waern Links.

Show Coordinates Provides a report of the solution coordinates. These are the coordinates of the points on the Multidimensional Scaling Plot. The report shows the coordinates of up to three dimensions. Right-click in the report and select **Columns** to add additional dimensions to the report. The maximum number of dimensions is the number of dimensions set in the launch window.

Show Proximity Provides a report of the proximities. The original and derived proximities (distances) are provided between each pair of objects. The pairs are identified in the From and To object columns. If a transformation was used, the transformed proximities are also included in the table.

Save Proximity (Available only if Attribute List is the data format.) Saves the distance matrix to the data table.

3D Plot (Available only if three or more dimensions are specified for Set Dimensions in the launch window.) Shows a 3-D plot of the first three dimensions.

Save Coordinates Saves the solution coordinates to the data table in separate columns.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Waern Links

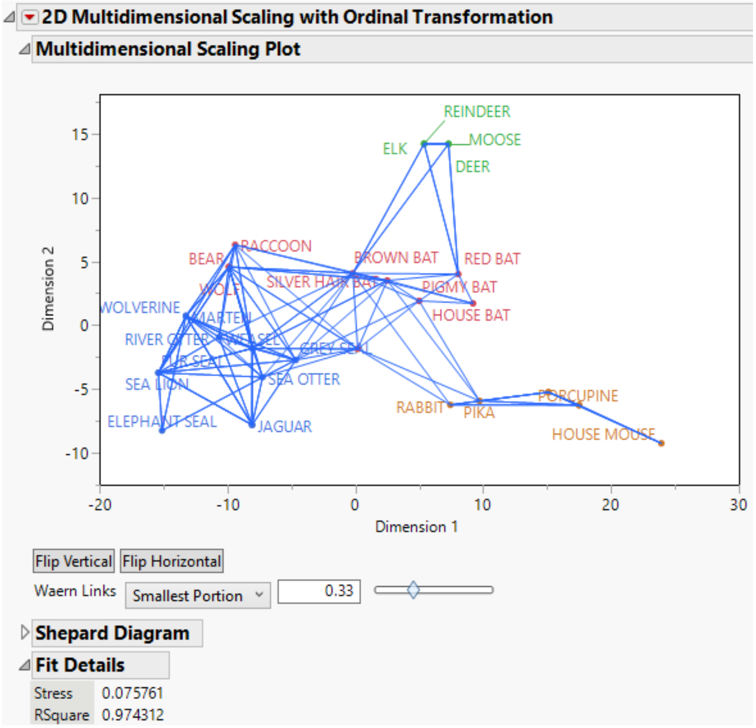
Waern links provide a visual check of the MDS results by comparing actual proximities to predicted proximities. The links join points on the Multidimensional Scaling Plot based on their actual proximities. The objects with the smallest (largest) proximities are connected. A typical scenario to consider is the smallest 33% of the proximities between objects. If the MDS Plot is a good representation of the proximities, then the links for the smallest actual proximities should connect the closest objects in the plot. If a link for a small proximity stretches across the plot connecting distant objects, then the MDS fit would be questioned.

Waern Link Controls

There is a list from which you can choose to show the Smallest Portion or the Largest Portion of links on the plot. The portion of links shown is controlled by entering a value in the box or by using the slider. Figure 10.5 shows Waern links for the Teeth.jmp data table for the 33% smallest portion.

For more information about Waern links, see Waern (1972).

Figure 10.5 MDS Plot with Waern Links



Waern Links Smallest Portion 0.33

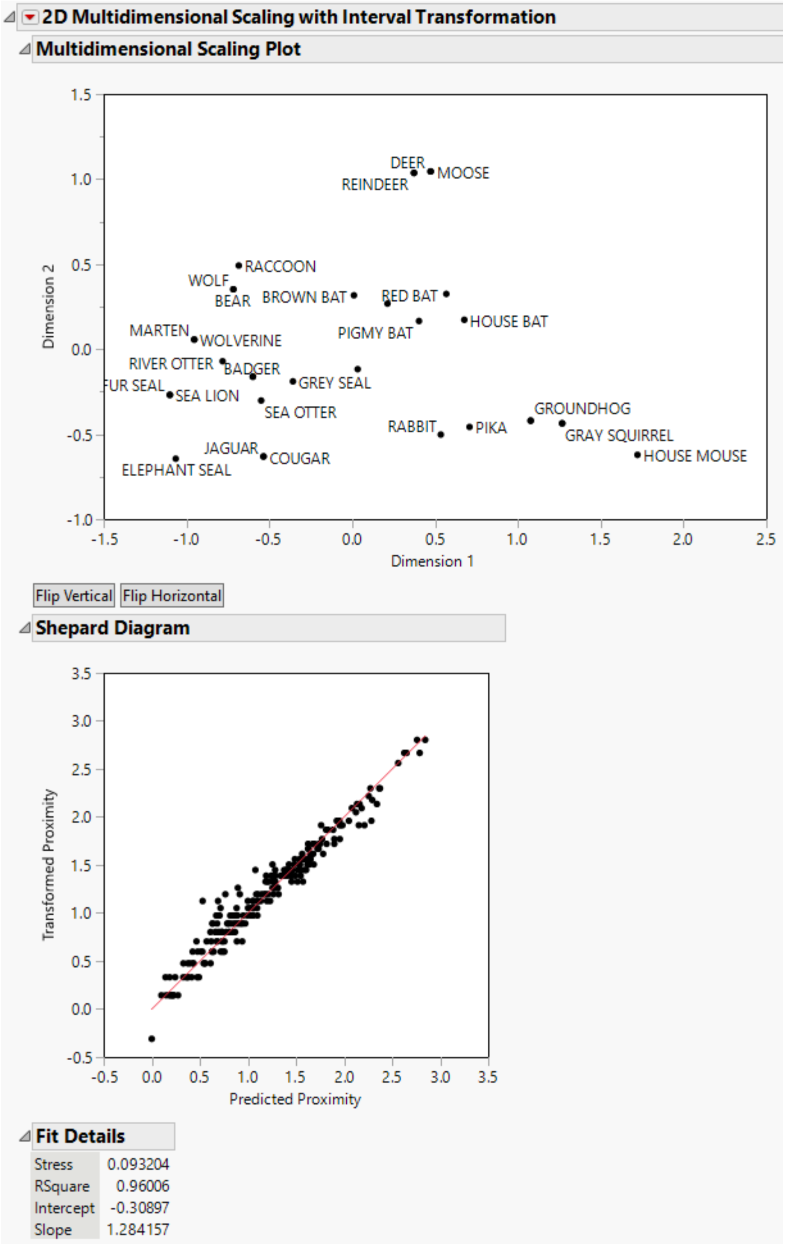
Stress	0.075761
RSquare	0.974312

Additional Example of the Multidimensional Scaling Platform

This example uses the Teeth.jmp sample data table, which is an attribute list of the counts of eight teeth types in 32 mammals. You can use MDS to explore the similarities of mammals based on their teeth. An interval transformation is used to illustrate the output from that transformation. The data do have an ordering that has a meaning (four teeth are twice as many as two teeth). One might explore other transformations such as the ordinal transformation.

1. Select **Help > Sample Data Library** and open Teeth.jmp.
2. Right-click the column MAMMAL and select **Label/Unlabel**.
3. Select **Rows > Row Selection> Select all Rows.**
4. Select **Rows > Label/Unlabel**
5. Select **Analyze > Multivariate Methods > Multidimensional Scaling**.
6. Select Top incisors through Bottom molars and click **Y, Columns**.
7. Select **Data Format > Attribute List**.
8. Select **Transformation > Interval**.
9. Click **OK**.

Figure 10.6 Multidimensional Scaling Report



Stress	0.093204
RSquare	0.96006
Intercept	-0.30897
Slope	1.284157

The Shepard Diagram and the Fit Details indicate that the MDS Plot is a good representation of similarities of animals due to similarities in their teeth. The Stress statistics of 0.093 is low and the R^2 fit of the transformed versus predicted proximities is high at 0.96. In addition, the Fit Details provides the intercept and slope for the transformation of the actual proximities.

Statistical Details for the Multidimensional Scaling Platform

JMP uses a quasi Newton optimization method to minimize the Stress function to determine the MDS coordinates. This minimization leads to a set of coordinates in a predetermined number of dimensions that minimize the derived proximity measures for each pairwise set of the dimensions. When the data is ordinal, monotonic regression is used. Otherwise, standard least squares regression is used.

- "Stress"
- "Transformations"
- "Attributes List Format"

Stress

The following notation is used to define Stress:

- i, j - indexes for the number of objects
- d_{ij} - the derived distance between objects i and j

  δ_{ij} - the observed relative distance between objects i and j
- $f(\delta_{rs})$ - transformation function for the distance

The Stress function is defined as follows:

$$\text{Stress} = \left[\frac{\sum_{i<j} [f(\delta_{ij}) - d_{ij}]^2}{\sum_{i<j} d_{ij}^2}\right]^{\frac{1}{2}}$$

This measure of stress is also known as Kruskal's Stress, Type I, or simply Stress1.

Transformations

The section uses the notation described in “Stress” on page 247. Transformations are used to scale the actual proximities. Transformations would be considered to improve the MDS representation of the actual proximities by taking into account specific structures in the data. The parameters in the transformation functions become additional parameters in the minimization algorithm.

Ratio Transformation

For ratio data:

$$f(\delta_{rs}) = b\delta_{rs}$$

Interval Transformation

For interval data:

$$f(\delta_{rs}) = a + b\delta_{rs}$$

Ordinal Transformation

For ordinal data the data is not transformed, rather the algorithm uses monotone regression rather than least squares regression.

Attributes List Format

When the data is an attributes list, it is converted to a distance matrix and then MDS is applied. The distance matrix is calculated using Euclidean distance. For each pair of items define the distance between the items by:

$$\delta_{ij} = \sqrt{\sum_{k} \frac{(x_{ki} - x_{kj})^2}{k}}$$

where *k* is the number of attributes.

Note: For an advanced example of the MDS platform, see the San Francisco Crime Distances.jmp sample data table and the source script for that table. The script creates the distance matrix using pairwise correlations. The resulting distance matrix is then used to explore the relationships between crime categories.

Chapter **11**

Item Analysis

Analyze Test Results by Item and Subject

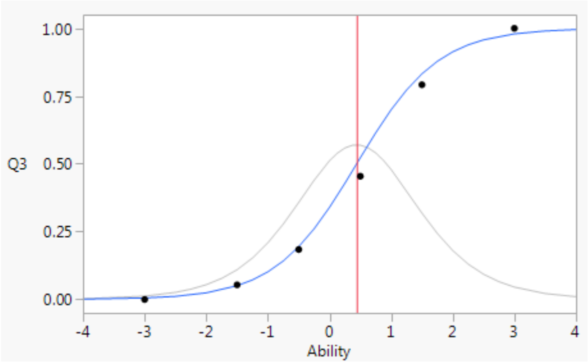
The Item Analysis platform enables you to fit *item response theory* models. The Item Response Theory (IRT) method is used for the analysis and scoring of measurement instruments such as tests and questionnaires. Item response theory uses a system of models to relate a trait or ability to an individual's probability of endorsing or correctly responding to an item. Frequently, the trait or ability of interest is not directly measurable and is therefore called *latent*. IRT can be used to study standardized tests, cognitive development, and consumer preferences. IRT is an alternative method to classical test theory (CTT) where the focus is on the total observed score rather than the item scores.

The Item Analysis platform implements the IRT method with the following outcomes:

- Measurement instruments are scored at the item level, providing insight into the contributions of each item on the latent response.
- Scores for both the responders and the items are produced on the same scale.
- Respondent and item scores are shown on a single plot.
- Item characteristic curves are shown. These curves can be used to explore the relationship between items and respondent's underlying trait or ability.

For more information about item response theory, see de Ayala (2009).

Figure 11.1 Item Analysis Characteristic Plot



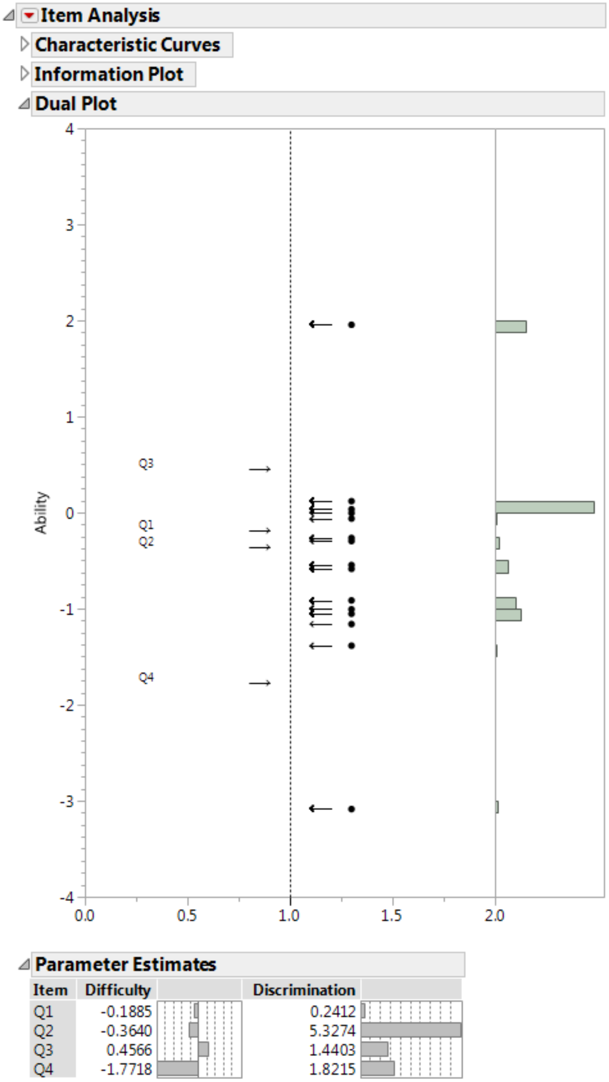
Contents

Example of Item Analysis

This example uses the MathScienceTest.jmp sample data table, which is a subset of the data from the Third International Mathematics and Science Study (TIMSS) conducted in 1996. The data table contains scores (1 = correct or 0 = incorrect) for 1263 students on 14 questions. You examine the first four questions to understand the relationship between questions and respondent's mathematical ability. The questions on the test are the items that are used to measure the latent mathematical ability. Fit a 2PL model to this data.

1. Select **Help > Sample Data Library** and open MathScienceTest.jmp.
2. Select **Analyze > Multivariate Methods > Item Analysis**.
3. Select Q1 through Q4, click **Y, Test Items** and click **OK**.

Figure 11.2 Item Response Report



Item	Difficulty	Discrimination
Q1	-0.1885	0.2412
Q2	-0.3640	5.3274
Q3	0.4566	1.4403
Q4	-1.7718	1.8215

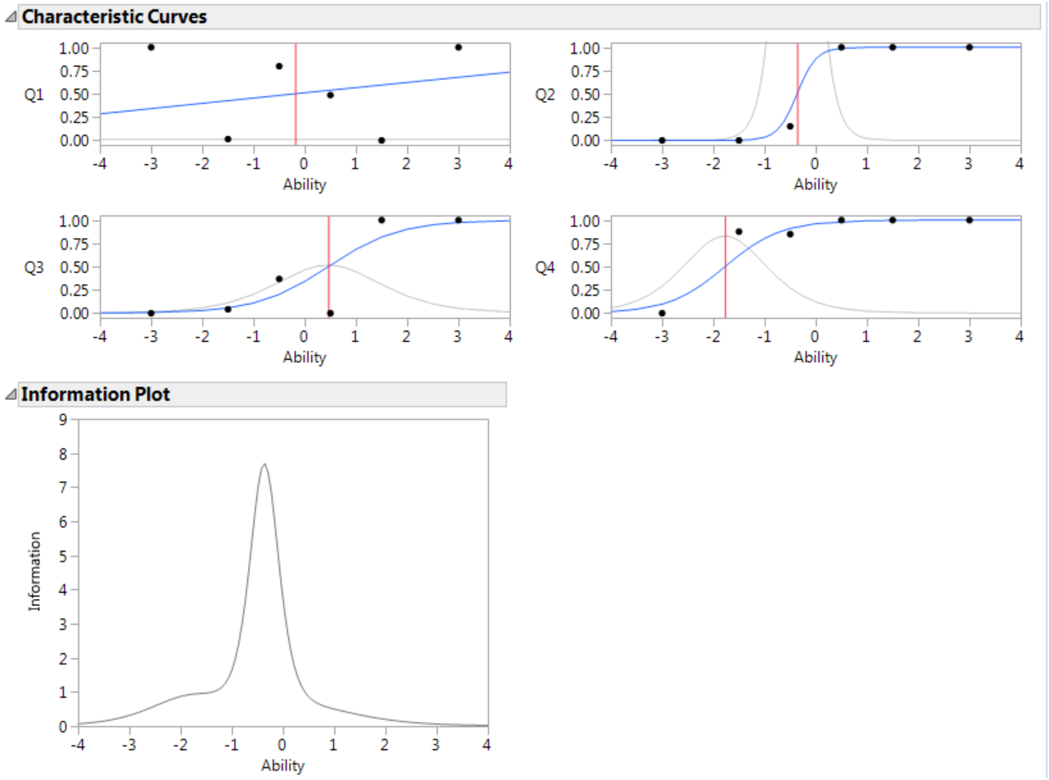
From the dual plot you note that Q4 is the easiest of the four questions to answer as it has the lowest difficulty score at -1.78. Q3 is the hardest with a difficulty score of 0.46. Most of the respondents fall in the middle to lower end of the ability scale as shown by the data points in the center part of the graph. In the histogram, you can see that approximately 40% of the respondents fall slightly above zero on the ability scale.

Note: Ability scores are not computed for individuals with all incorrect or all correct answers. See "Fitting the IRT Model" on page 263.

---

4. Click the gray Characteristic Curves report disclosure icon to open.
5. Click the Item Analysis red triangle and select **Number of Plots Across**.
6. Enter 2 and click **OK**.
7. Click the gray Information Plot report disclosure icon to open.

Figure 11.3 Item Response Example



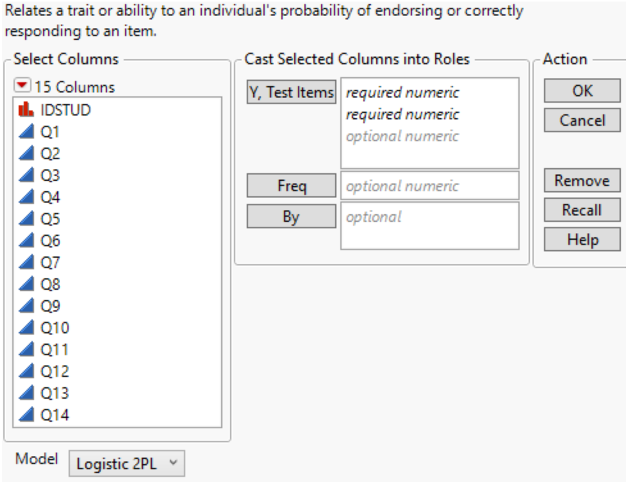
Q1 has a flat characteristic curve and a flat information curve. This suggests that Q1 does not provide much information to discriminate respondents' mathematical ability. The characteristic curve for Q2 is steep, which indicates that Q2 is useful for discriminating respondent ability. The vertical line in each plot is at the inflection point for the characteristic curve. This vertical line is the ability level at which the respondent has a 50% probability of answering the specified question correctly.

The information plot indicates that together the four questions analyzed provide the most information about ability levels between about -1 and 0. Including more questions of higher difficulty in the model could broaden the information curve.

Launch the Item Analysis Platform

Launch the Item Analysis platform by selecting **Analyze > Multivariate Methods > Item Analysis**.

Figure 11.4 Item Analysis Launch Window



Relates a trait or ability to an individual's probability of endorsing or correctly responding to an item.

Model: Logistic 2PL

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Test Items Assigns two or more columns to be analyzed. The columns must be numeric, continuous, and contain only 0s and 1s.

Tip: Use **Cols > Recode** if you need to recode your data to 0s and 1s. See *Using JMP*.

Freq Assigns a frequency variable to this role. This is useful if your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variable.

Model Specifies the desired model from the following options:

Logistic 2PL The 2-parameter logistic model.

Logistic 3PL The 3-parameter logistic model.

Logistic 1PL The 1-parameter logistic model with a Rasch parameterization.

Logistic 3PL Model Details

If you select Logistic 3PL for Model, you are prompted to enter a penalty for the guessing parameters after you click OK. For a 3PL model, the default value of the penalty is zero. However, you can enter a non-zero penalty for the *c* parameters (the guessing for each item). This penalty is similar to the type of penalty parameter that you would use in ridge regression. The penalty is on the variance of the estimated guessing parameters. The use of the penalty has the following benefits:

- Stabilizes the estimation of model parameters.
- Speeds up computations.
- Reduces the variability of the guessing parameter across items at the expense of some bias.

Large values of the penalty force the guessing parameters to zero while smaller values help reduce the variability of the guessing parameter across items. A value of zero can be used for no penalty.

Data Format

The Item Analysis platform requires a data table that contains a row for each individual and a column for each item. The item columns must be numeric and contain only 0s and 1s to indicate incorrect or correct responses, respectively. The MathScienceTest.jmp sample data table illustrates the required data format for an item response analysis of 1,263 individuals responses to 14 test questions.

The Item Analysis Report

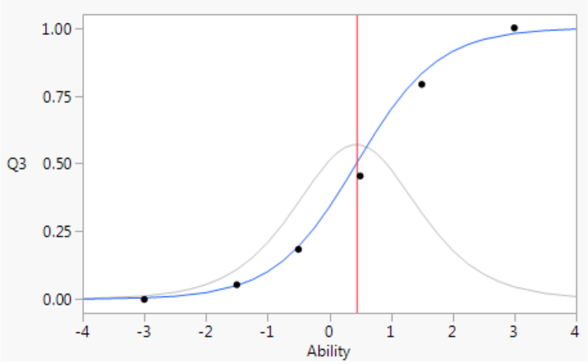
- "Characteristic Curves"
- "Information Plot"
- "Dual Plot"
- "Parameter Estimates"

Characteristic Curves

The Characteristic Curves contains an item characteristic curve (ICC) for each item that you specified in the launch window. The Characteristic Curves are initially closed.

The item characteristic curve plots the probability of answering an item correctly versus ability. Ability is measured on a standardized scale, so a respondent with ability equal to 0 is a respondent of average ability. Data points for the observed probability of correct answers for fixed ability levels are plotted. Comparing the fitted characteristic curve to the data points provides a visual measure of goodness of fit of the model for each individual item. In addition, the characteristic plots have a background information curve and a vertical line at the characteristic curve inflection point. The background information curve is a plot of the slope of the item characteristic curve, which is maximized at the inflection point.

Figure 11.5 Item Characteristic Curve

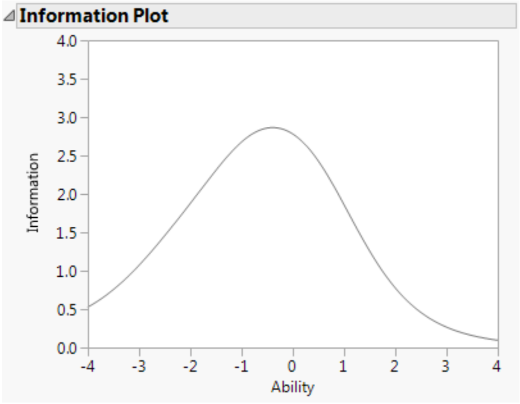


Tip: You can adjust the number of characteristic curves that appear in each row of the report using the Number of Plots Across option in the Item Analysis red triangle menu.

Information Plot

The Information Plot report contains a plot of the overall information curve, which is constructed by summing the individual item information curves. The information plot provides insight into the appropriate ability levels that the test is able to measure. Figure 11.6 describes a test with items that are appropriate for assessing individuals with average to low levels of the ability more so than individuals with high levels of ability. This plot is initially closed.

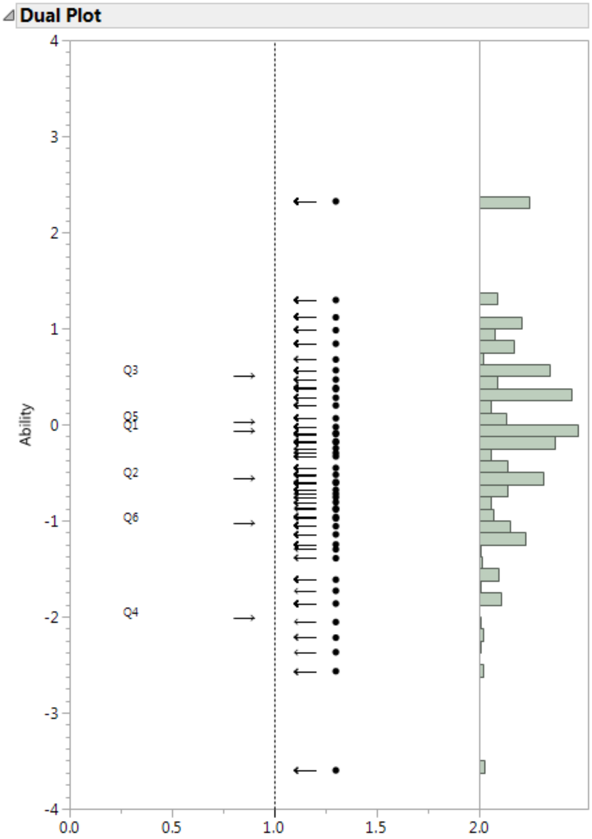
Figure 11.6 Information Plot



Dual Plot

The Dual Plot report contains a plot that shows item difficulty and subject ability in one plot. Difficulty and ability use a common standardized scale shown on the *y*-axis. The items are plotted by their difficulty on the left side of the plot. The subjects are plotted to the right with data points and a histogram. The dual plot enables you to relate the difficulty of each item to the ability of each respondent.

Figure 11.7 Dual Plot



Parameter Estimates

The Parameter Estimates report contains a table of estimated parameters for each item. The parameters provided depend on the model used in your analysis (1PL, 2PL, or 3PL).

Item The test item.

Difficulty The *b* parameter or the measure of the difficulty of the item. A histogram of the difficulty parameters is shown beside the difficulty estimates.

Discrimination (Available only for 2PL and 3PL models.) The *a* parameter or the measure of the item discrimination. A histogram of the discrimination parameters is shown beside the discrimination estimates.

Lower Asymptote (Available only for 3PL models.) The *c* parameter or a measure of guessing.

Item Analysis Platform Options

Number of Plots Across Enables you to specify how many ICC plots to display in each row of plots in the Characteristic Curve report. The default is one ICC plot per row.

Save Ability Formula Saves the ability formula to a new column in the data table. This options uses the `IRT Ability()` JSL function. For more information about this function, see Help > Scripting Index.

See *Using JMP* for more information about the following options:

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Statistical Details for the Item Analysis Platform

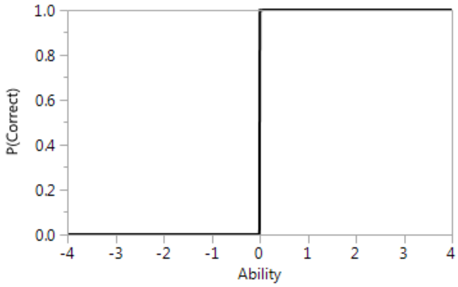
Item response theory (IRT) uses a series of equations to relate items to an unobserved (latent) trait or ability. Items, or questions, are indicators of an underlying latent construct that cannot be directly observed. At the time of data collection, both the subject abilities and the item characteristics are unknown.

- “Item Response Curves”
- “Item Response Curve Models”
- “IRT Model Assumptions”
- “Fitting the IRT Model”
- “Ability Formula”

Item Response Curves

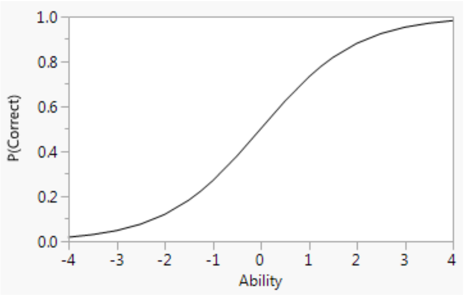
Item response curves (item characteristic curves) are used to describe the relationship between the ability, defined on an *ability* scale, and each item. An item response curve plots the probability of correctly answering an item against different levels of ability. An item with perfect discrimination has a 0% probability of correct answers for respondents with ability below a threshold and a 100% probability of a correct response for subjects with ability above the threshold.

Figure 11.8 Characteristic Curve for an Item with Perfect Discrimination



A typical relationship between the probability of correctly answering an item and ability is an S-shaped function with lower and upper asymptotes. As a respondent's ability increases, their probability of correctly answering the item increases to 100%. The shape of the curve for a specific item is related to the difficulty and discriminatory properties of the item.

Figure 11.9 Typical Item Response Curve



Item Response Curve Models

One-, two-, and three-parameter logistic models can be used to model the item response curves. The three-parameter logistic (3PL) model is defined as follows:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-(a)(\theta - b)}}$$

- $P(\theta)$ is the probability of answering the item correctly for an ability level θ. For more information about fitting the item response theory model, see "Fitting the IRT Model" on page 263.
- The *a* parameter defines the steepness of the curve at its inflection point. It provides an estimate of the discriminatory power of the item.
- The *b* parameter defines the location of the inflection point on the Ability axis. It provides an estimate of the difficulty of an item.
- The *c* parameter is the lower asymptote. It provides an estimate of the probability that an item is answered correctly by guessing.
- For the 2PL model, the *c* parameter is set to 0.

$$P(\theta) = \frac{1}{1 + e^{-(a)(\theta - b)}}$$

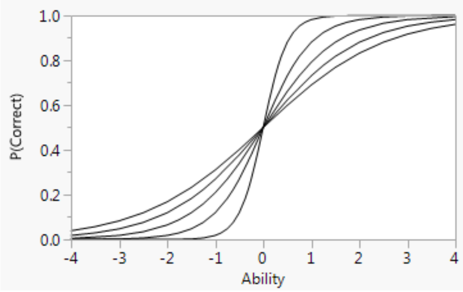
- For the 1PL model, the *c* parameter is set to 0 and the *a* parameter is set to 1. This parameterization is also known as the Rasch model (Rasch 1980).

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}}$$

The *a* Parameter: Item Discrimination

In the 2PL and 3PL models, the *a* parameter, or the steepness of the curve at its inflection point, provides a measure of the discriminatory power of an item. The discriminatory power, or discrimination, of an item refers to how well an item can distinguish between respondents with low ability levels versus those with high ability levels. A steep item response curve indicates that the item has strong discrimination. Respondents with low ability levels have a low probability of a correct response to the item while respondents with high ability have a high probability of a correct response. Items whose curves are relatively flat have low discrimination. Items with low discrimination are candidates to be dropped from the measurement instrument.

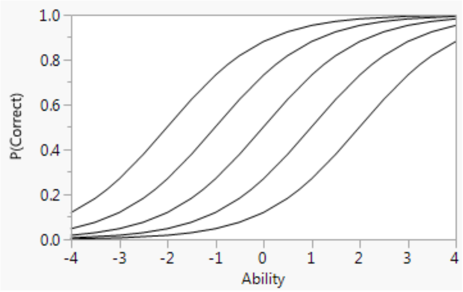
Figure 11.10 Logistic Model for Several Values of *a*



The *b* Parameter: Item Difficulty

The *b* parameter, or the location of the inflection point with respect to ability, provides a measure of item difficulty. Item response curves with inflection points farther to the right on the ability scale are indicative of items that are more difficult to answer than items with inflection points to the left. In the 1PL and 2PL models, the *b* parameter provides an estimate of the ability level required for a 50% probability of correctly answering the item.

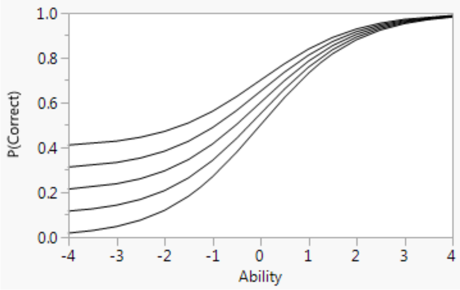
Figure 11.11 Logistic Curve for Several Values of *b*



The *c* Parameter: Guessing

In the 3PL model, the *c* parameter, or the lower asymptote of the item response curve, provides a measure of the guessing parameter. A nonzero lower asymptote represents the nonzero probability of a person with a very low ability level answering an item correctly.

Figure 11.12 Logistic Model for Several Values of *c*



IRT Model Assumptions

The 2PL model is the default model in the Item Analysis platform. The 1PL model is appropriate when you can assume that all items have equal discriminating power. When this assumption is not appropriate, the 2PL or 3PL model should be used. The 2PL model has greater numerical stability than the 3PL model, especially for small data sets. In addition, in the 2PL model, b can be interpreted as the ability level required for a 50% chance of a responder answering an item correctly.

The IRT model assumes that the underlying trait is unidimensional. That is, there is a single underlying latent construct. If there are several traits that have complex interactions with each other being measured, then a unidimensional model is not appropriate. The IRT model is appropriate for continuous latent variables. For a categorical latent variable, you should consider a latent class model. See the “Latent Class Analysis” chapter on page 325. IRT models are assumed to be item-invariant. Item-invariance means that $P(\theta)$ is interpreted as the probability of a correct response for a set of individuals with ability level θ. If a large group of individuals with equal ability levels answered the item, $P(\theta)$ predicts the proportion who would answer the item correctly. This implies that IRT models would have the same parameters regardless of the group of subjects tested. In addition, the IRT model assumes local independence, which means that once the latent construct has been accounted for, the items are independent of one another.

Fitting the IRT Model

The IRT model is fit using Marginal Maximum Likelihood estimation (MMLE). MMLE is an alternative method to Joint Maximum Likelihood estimation (JLE). MMLE treats the subjects as random effects. The items and abilities are related as conditional probabilities. The formulas are defined as follows:

$$p(\mathbf{x}|\theta, \vartheta) = \prod_{j=1}^{L} p_j(\theta)^{x_j}(1 - p_j(\theta))^{1 - x_j}$$

where $p(\mathbf{x} | \theta, \vartheta)$ is the probability of a response vector **x** given the subject ability θ and the vector of item parameters ϑ. The number of item parameters depends on the model used (1PL, 2PL, or 3PL).

MMLE integrates out the subject effects using Gaussian quadrature to obtain item parameter estimates. The probability of response vector **x** is calculated as follows:

$$p(\mathbf{x}) = \int_{-\infty}^{\infty} p(\mathbf{x}|\theta, \vartheta) g(\theta|\upsilon) d\theta$$

where $g(\theta | \nu)$ is the distribution of the subjects and ν is a vector of the population location and scale parameters. The normal distribution with mean 0 and standard deviation 1 is used for $g(\theta | \nu)$ in JMP.

Note: A missing value for a test question is treated as an incorrect response. Ability scores are not calculated for individuals with all incorrect or all correct answers. The patterns of the responses for these subjects are included in the model estimation.

The MMLE procedure for fitting the IRT model can be compared to fitting a random effects model in two stages. The ability parameters are treated as random effects with variance of 1. In the first step, these random effects are integrated out using Gaussian quadrature. The item parameters are treated as fixed effects that are estimated using ML from the marginal likelihood with the ability parameters integrated out. The ability parameters are in essence best linear unbiased predictions that are estimated using the full unintegrated (joint) likelihood, treating the item parameters as known and held fixed at the values obtained in the first stage.

There are 2^L patterns of responses for L items. The ability level for each pattern can be calculated by finding the ability level with the highest probability for the response pattern by applying the following until θ converges:

$$\theta_i^{t+1} = \theta_i^t - \frac{X_i - \sum_{j=1}^{L} p_{ij}(t)}{-\sum_{j=1}^{L} p_{ij}(t)(1 - p_{ij}(t))}$$

where:

θ maximizes the likelihood of obtaining the response pattern

t is the number of iterations

L is the number of items

X_i is the observed score

p_{ij} is the probability of a correct response on the j^{th} item by the i^{th} person based on the item parameters.

Ability Formula

The Save Ability Formula option from the Item Analysis red triangle menu saves the ability formula to a new column in the data table. This formula can be used to score additional subjects added to the data table or it can be copied to a new table to score a new set of subjects.

The function saved to the data table is called the `IRT Ability` function. The item parameter estimates are stored in a matrix in this function. For more information about this function, see Help > Scripting Index.

Chapter 12

Hierarchical Cluster

Group Observations Using a Tree of Clusters

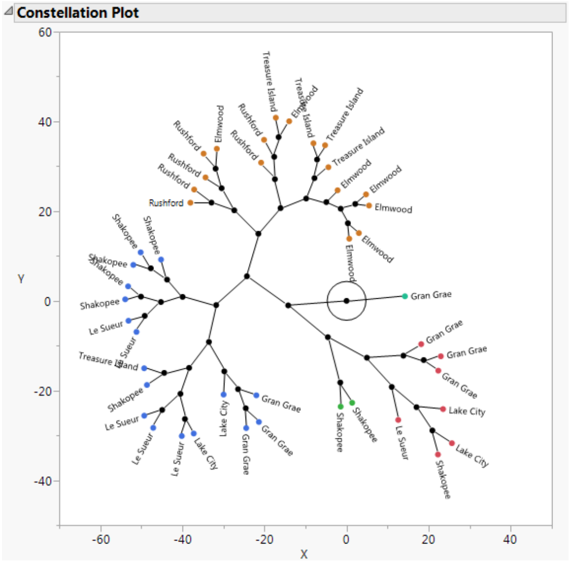
Clustering is a multivariate technique that groups together observations that share similar values across a number of variables. Use it to understand the clumping structure of your data.

Hierarchical clustering combines clusters successively. The method begins by treating each observation as its own cluster. Then, at each step, the two clusters that are closest in terms of distance are combined into a single cluster. The result is depicted as a tree, called a *dendrogram*.

Use hierarchical clustering for small data tables with no more than several tens of thousands of rows. The algorithm is time-intensive and can run slowly for larger data tables. For larger data tables, use K Means Cluster or Normal Mixtures.

Note: Hierarchical cluster supports character columns; K Means Cluster or Normal Mixtures require numeric columns.

Figure 12.1 Example of a Constellation Plot



Contents

Overview of the Hierarchical Clustering Platform

Hierarchical Clustering is one of four platforms that JMP provides for clustering observations. For a comparison of all four methods, see "Overview of Platforms for Clustering Observations" on page 269.

The hierarchical clustering method starts with each observation forming its own cluster. At each step, the clustering process calculates the distance between all pairs of clusters and combines the two clusters that are closest together. This process continues until all the points are contained in one cluster. Hierarchical clustering is also called *agglomerative clustering* because of the combining approach that it uses.

The agglomerative process is portrayed as a tree, called a dendrogram. To help you decide on a number of clusters, JMP provides a distance graph. You can select a number of clusters by determining when the distances between clusters no longer appear to be of practical importance.

Hierarchical clustering also supports character columns. There are two ways to define the distances.

- If a column is ordinal, then the value used for clustering is the index of the ordered category, treated as if it were continuous data. These values are standardized as if they were continuous data.
- If a column is nominal, then the distance between two observations where the categories match is zero. If the categories differ, the distance is one.

Hierarchical clustering enables you to choose among five rules for defining distances between clusters: Average, Centroid, Ward, Single, and Complete. Each rule can generate a different sequence of clusters.

Tip: The hierarchical clustering process starts with $n(n + 1)/2$ distances for n observations, except when the Fast Ward method is used. For this reason, this method can take a long time to run when n is large. For large numbers of numeric observations, consider K Means Cluster or Normal Mixtures.

Overview of Platforms for Clustering Observations

Clustering is a multivariate technique that groups together observations that share similar values across a number of variables. Typically, observations are not scattered evenly through p-dimensional space, where p is the number of variables. Instead, the observations form clumps, or clusters. Identifying these clusters provides you with a deeper understanding of your data.

Note: JMP also provides a platform that enables you to cluster variables. See the "Cluster Variables" chapter on page 341.

JMP provides four platforms that you can use to cluster observations:

- Hierarchical Cluster is useful for smaller tables with up to several tens of thousands of rows and allows character data. Hierarchical clustering combines rows in a hierarchical sequence that is portrayed as a tree. You can choose the number of clusters that is most appropriate for your data after the tree is built.
- K Means Cluster is appropriate for larger tables with up to millions of rows and allows only numerical data. You need to specify the number of clusters, *k*, in advance. The algorithm guesses at cluster seed points. It then conducts an iterative process of alternately assigning points to clusters and recalculating cluster centers.
- Normal Mixtures is appropriate when your data come from a mixture of multivariate normal distributions that might overlap and allows only numerical data. For situations where you have multivariate outliers, you can use an outlier cluster with an assumed uniform distribution.

  You need to specify the number of clusters in advance. Maximum likelihood is used to estimate the mixture proportions and the means, standard deviations, and correlations jointly. Each point is assigned a probability of being in each group. The EM algorithm is used to obtain estimates.
- Latent Class Analysis is appropriate when most of your variables are categorical. You need to specify the number of clusters in advance. The algorithm fits a model that assumes a multinomial mixture distribution. A maximum likelihood estimate of cluster membership is calculated for each observation. An observation is classified into the cluster for which its probability of membership is the largest.

Table 12.1 Summary of Clustering Methods

Method	**Data Type or Modeling Type**	**Data Table Size**	**Specify Number of Clusters**
Hierarchical Cluster	Any	With Fast Ward, up to 200,000 rows With other methods, up to 5,000 rows	No
K Means Cluster	Numeric	Up to millions of rows	Yes
Normal Mixtures	Numeric	Any size	Yes
Latent Class Analysis	Nominal or Ordinal	Any size	Yes

Some of the clustering platforms have options to handle outliers in the data. However, if your data has outliers, it is best to explore them first prior to analyzing. This can be done using the Explore Outliers Utility. For more information, see the Modeling Utilities chapter in *Predictive and Specialized Modeling*.

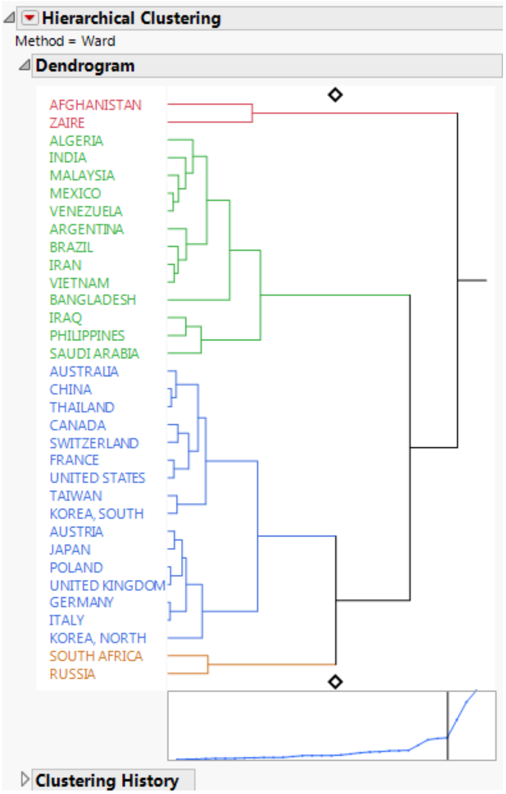
Example of Clustering

In this example, we group together countries by their 2009 crude birth and death rates per 1,000 people.

1. Select **Help > Sample Data Library** and open Birth Death Subset.jmp
2. Select **Analyze > Clustering > Hierarchical Cluster**.
3. Select birth and death and click **Y, Columns**.
4. Select country and click **Label**.

   This selection ensures that the country column, rather than the row number, is used to label the dendrogram that appears when you click OK.
5. Click **OK**.
6. Click the Hierarchical Clustering red triangle and select **Color Clusters**.

Figure 12.2 Hierarchical Clustering Report



Method = Ward

The dendrogram shows how the clustering is conducted. The clustering process can be viewed by reading the dendrogram from left to right. Each step consists of combining the two *closest* clusters into a single cluster.

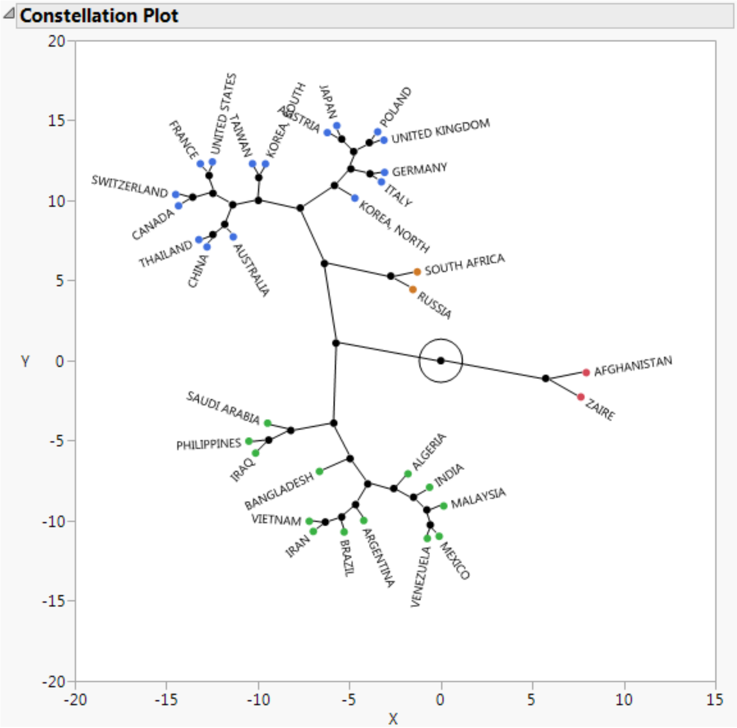
In the dendrogram, the relative distances between clusters are given by the horizontal distances between vertical lines that join the clusters. For example, Afghanistan and Zaire differ more than Malaysia differs from the cluster consisting of Mexico and Venezuela.

The plot that appears beneath the dendrogram has a point for each step where two clusters are joined into a single cluster. The horizontal coordinates represent the numbers of clusters and they decrease from left to right. The vertical coordinate of the point is the distance between the two clusters that are joined to form the specified number of clusters. You can click either diamond in the dendrogram and drag the line to choose the number of clusters that best represent the data. You can also use the Number of Clusters option in the Hierarchical Clustering red triangle menu to choose the number of clusters.

The distance graph has a noticeable change in slope at four clusters. The change in slope indicates that the differences in clusters that are joined up to the point where four clusters remain, are comparatively small. This suggests that four is a good choice for the number of clusters. Note that this is the number of clusters that was shown by default.

7. Click the Hierarchical Clustering red triangle and select **Constellation Plot**.

Figure 12.3 Constellation Plot

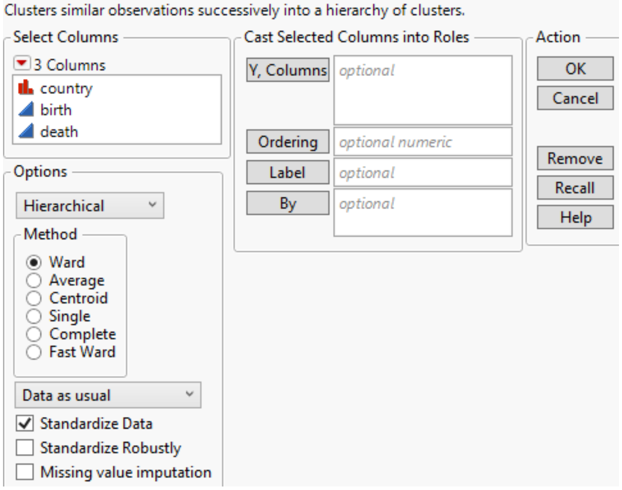


This constellation plot arranges the countries as endpoints and each cluster join as a new point. The lines represent membership in a cluster. The length of a line between cluster joins approximates the distance between the clusters that were joined. The constellation plot indicates that the cluster that contains Afghanistan and Zaire is about as distant from the cluster of remaining countries as are the two clusters that consist of the remaining countries in the upper half of the plot and those in the lower half of the plot.

Launch the Hierarchical Cluster Platform

Launch the Hierarchical Cluster platform by selecting **Analyze > Clustering > Hierarchical Cluster**.

Figure 12.4 Hierarchical Cluster Launch Dialog



Clusters similar observations successively into a hierarchy of clusters.

Hierarchical

Method
● Ward
○ Average
○ Centroid
○ Single
○ Complete
○ Fast Ward

Data as usual

☑ Standardize Data
☐ Standardize Robustly
☐ Missing value imputation

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns The variables used for clustering observations.

Ordering Sorts clusters by their mean values based on the specified column.

Tip: Use the first principal component obtained by conducting a principal components analysis as an Ordering column. The clusters are sorted by these values.

Attribute ID (Available only if **Data is stacked** is selected as the data structure.) Specifies the variables that are stacked.

Object ID (Available only if **Data are summarized** or **Data is stacked** is selected as the data structure.) A column or columns that provide a unique identifier for each unit for which measurements are stacked.

Label A column of values used to label the dendrogram in the report.

Note: If the selected data structure is **Data is distance matrix**, the Label column must have a character data type.

By A column whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed. The results are presented in separate reports. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

The launch window has the following menus and options:

- "Clustering Method"
- "Method for Distance Calculation"
- "Data Structure"
- "Transformations to Y, Columns Variables"

Clustering Method

Hierarchical is the default clustering method, but the dialog enables you to switch to KMeans or Normal Mixtures. If you select KMeans or Normal Mixtures, when you click OK, a Control Panel appears where you can select any of the following as Method:

K-Means Clustering See the "K Means Cluster" chapter on page 295.

Normal Mixtures See the "Normal Mixtures" chapter on page 313.

Robust Normal Mixtures See "Normal Mixtures Report" on page 321 in the "Normal Mixtures" chapter.

Self Organizing Map See "Self Organizing Map" on page 308 in the "K Means Cluster" chapter.

Method for Distance Calculation

Select a method used to calculate distances. For distance formulas, see "Distance Method Formulas" on page 293.

Ward In Ward's minimum variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters summed over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give the proportions of variance (squared semipartial correlations).

Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the assumptions of multivariate normal mixtures, spherical covariance matrices, and equal sampling probabilities.

Ward's method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with approximately the same number of observations. It is also very sensitive to outliers. See Milligan (1980).

Average The distance between two clusters is the average distance between pairs of observations. Average linkage tends to join clusters with small variances and is slightly biased toward producing clusters with the same variance. See Sokal and Michener (1958).

Centroid The distance between two clusters is defined as the squared Euclidean distance between their means. The centroid method is more robust to outliers than most other hierarchical methods but in other respects might not perform as well as Ward's method or average linkage. See Milligan (1980).

Single The distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster. Single linkage has many desirable theoretical properties but has performed poorly in Monte Carlo studies. See Jardine and Sibson (1971), Fisher and Van Ness (1971), Hartigan (1981), and Milligan (1980). Single linkage was originated by Florek et al. (1951a, 1951b) and later reinvented by McQuitty (1957) and Sneath (1957).

By imposing no constraints on the shape of clusters, single linkage sacrifices performance in the recovery of compact clusters in return for the ability to detect elongated and irregular clusters. Single linkage tends to chop off the tails of distributions before separating the main clusters. See Hartigan (1981).

Complete The distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. Complete linkage is strongly biased toward producing clusters with approximately equal diameters and can be severely distorted by moderate outliers. See Milligan (1980).

Fast Ward Applies an algorithm that computes Ward's method more quickly for large numbers of rows. The computation time is shorter because this algorithm does not require the calculation of a distance matrix. It is used automatically whenever there are more than 2,000 rows.

Data Structure

These options describe the form of the data that is used in calculating multivariate distances:

Data as usual Data that are rectangular with one row for each observation and one column for each variable.

Data as summarized Data that are summarized by the levels of one or more identifying columns. When you select this option, an Object ID text box appears in the launch window. Specify the identifying columns as the Object ID. The **Data as summarized** option calculates level means and treats these means as your input data.

Data is distance matrix Data that consist of distances between observations. For *n* observations, the distance table should have *n* rows and *n* + *1* columns. One column (usually the first) must contain a unique identifier for each of the *n* observations. The remaining columns contain distances between that observation and the *n* observations. Note the following:

- The diagonal elements of the table should be zero or missing, because the distance between a point and itself is zero. Values that are not zero or missing are treated as zero, and a note appears in the report.
- The distance columns can be a symmetric square matrix, or they can be upper or lower triangular with missing entries in the lower or upper portion. If the distances are given as a square matrix, a warning appears in the report if the table is not symmetric.
- You can begin with a different data structure and then save a distance matrix. See “Save Distance Matrix” on page 285.

When you select the **Data is distance matrix** option, enter the distance columns as Y, Columns and the identifier column as Label. The Label column must have the Character data type. For an example, see “Example of a Distance Matrix” on page 286.

Data is stacked Data that have a single response of interest and multiple rows for each object.

When you select the **Data is stacked** option, Attribute ID and Object ID text boxes appear in the launch window.

- Enter a *single* column as Y, Columns.
- Enter columns that describe groupings of the Y, Columns variable as Attribute ID. If only two columns are entered and if you select Add Spatial Measures, then you can add spatial components to be used in the cluster analysis. See “Add Spatial Measures” on page 279.

- Enter the identifying columns for objects as Object ID.

  The analysis that is conducted is equivalent to splitting the Y, Column variable by the Attribute ID columns and then performing hierarchical clustering without standardizing the response columns.

Tip: Use this option together with the Add Spatial Measures option to perform two-dimensional spatial clustering. For example, wafer data are often recorded using one row for each die. Interest centers around clustering wafers. See "Example of Wafer Defect Classification Using Spatial Measures" on page 288.

Caution: Because there is a single measurement column, the Standardize Data option is not appropriate for stacked data.

Not Enough Nonmissing Data Alert

The JMP alert **Not enough nonmissing data** can be difficult to understand when you are using the **Data as summarized** or **Data is stacked** data structures. The alert occurs in the following situations:

- For **Data as usua**l, when all rows or all but one row are missing at least one value for a Y, Columns variable.
- For **Data as summarized**, when your data are summarized across the Object ID columns, all rows or all but one row are missing at least one value of the summarized Y, Column variables. To see the data structure that the Cluster platform is analyzing, select Tables > Summary, enter the Object ID columns as Group and the Y, Columns variables as Statistics > Mean.
- For **Data is stacked**, when your data are split across the Attribute ID columns, all rows or all but one row are missing at least one value of the split Y, Column values. To see the data structure that the Cluster platform is analyzing, select **Tables > Split**, enter the Attribute ID columns as Split By, the Y, Columns variable as Split Columns, and the Object ID columns as Group.

Transformations to Y, Columns Variables

The following options specify the form of the Y, Columns variables to be used in the cluster analysis:

Standardize Data Addresses the issue of different measurement scales for continuous and ordinal columns. Except when the **Data is stacked** option is selected, the values in each column are standardized by subtracting the column mean and dividing by the column standard deviation. Deselect the Standardize Data check box if you do not want the cluster distances computed on standardized values.

Standardize Robustly Reduces the influence of outliers on estimates of the mean and standard deviation for continuous and ordinal columns. This option uses Huber M-estimates of the mean and standard deviation (Huber 1964; Huber 1973; Huber and Ronchetti 2009). For columns with outliers, this option gives the standardized values greater representation in determining multivariate distances.

Note: If both Standardize Data and Standardize Robustly are selected, each column is standardized by subtracting its robust column mean and dividing by its robust standard deviation. This option is useful when columns represent different measurement scales *or* when observations tend to be outliers in only specific dimensions.

Note: If Standardize Data is unchecked and Standardize Robustly is selected, the robust mean and robust standard deviation for the values in all columns combined are used to standardize each column. This option can be useful when columns all represent the same measurement scale *and* when observations tend to be outliers in all dimensions.

Missing value imputation Imputes missing values. If the number of variables is either 50 or less, or less than half the number of rows, multivariate normal imputation is used. Otherwise, multivariate SVD imputation is used.

Multivariate normal imputation calculates pairwise covariances to construct a covariance matrix for the response columns. Then each missing value is imputed by a method that is equivalent to regression prediction using all the predictors with no missing values for the given observation. If the constructed covariance matrix is not positive definite, missing values are imputed using their column means.

Multivariate SVD imputation avoids constructing a covariance matrix by using the singular value decomposition. See *Predictive and Specialized Modeling*.

Caution: Missing value imputation assumes that there are no clusters, that the data come from a single multivariate normal distribution, and that the values are missing completely at random. Because these assumptions are usually not reasonable in practice, use this feature with caution. However, the feature can produce more informative results than discarding most of your data.

Add Spatial Measures (Available only if **Data is stacked** is selected as the data structure.) Select the **Add Spatial Measures** option when your data are stacked and contain two attribute columns that correspond to spatial coordinates (horizontal and vertical coordinates, for example). This option opens a window in which you can select and weight spatial components to aid in clustering defect patterns. This is a specialty method and is applicable in only very specific settings. See “Spatial Measures” on page 291 and “Example of Wafer Defect Classification Using Spatial Measures” on page 288.

Hierarchical Cluster Report

The Hierarchical Cluster report displays the method used, a dendrogram, and the Clustering History table. If you assigned a column as a Label in the launch window, the column's values identify each observation in the dendrogram.

- "Dendrogram Report"
- "Illustration of Dendrogram and Distance Graph"
- "Clustering History Report"

Dendrogram Report

The dendrogram is a tree diagram that represents the agglomeration of observations into clusters. The dendrogram also gives information about the degree of dissimilarity of clusters.

The clustering process can be viewed by reading the dendrogram from left to right. Each step consists of combining the two *closest* clusters into a single cluster.

- The joining of clusters is indicated by horizontal lines that are connected by vertical lines.
- The horizontal position of the vertical line represents the distance between the two clusters that are most recently joined to form the specified number of clusters.

Note: When the number of observations is less than 256, the distances are proportional to the distances shown in the Distance Graph. Otherwise, Geometric Spacing is used. See "Dendrogram Scale" on page 283.

You can perform the following tasks:

- Click and drag the diamond-shaped handle at either the top or bottom of the dendrogram to identify a given number of clusters.
- Click any cluster stem to select all the members of the cluster in the dendrogram and in the data table.

Distance Graph

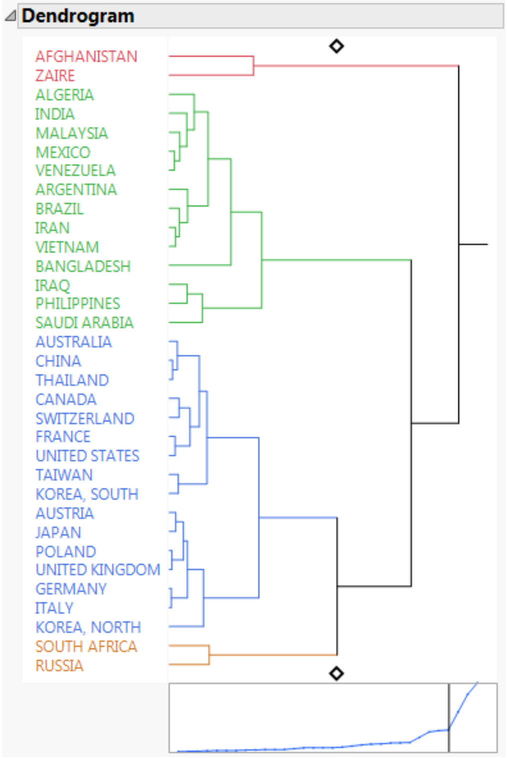
The Distance Graph is the plot that appears beneath the dendrogram. This graph has a point for each step where two clusters are joined into a single cluster. The horizontal coordinates represent the numbers of clusters, which decrease from left to right. The vertical coordinate of the point is the distance between the clusters that were joined at the given step.

You can click and drag either diamond-shaped handle in the dendrogram to control the chosen number of clusters. When you click and drag the diamond, a vertical line appears in the plot that moves to correspond to the number of clusters. Often there is a point where the slope of the distance graph levels off. Such a point suggests a natural break and helps you determine the number of clusters.

Illustration of Dendrogram and Distance Graph

Consider the dendrogram report for Birth Death Subset.jmp in “Example of Clustering” on page 271.

Figure 12.5 Dendrogram Report for Birth Death Subset.jmp



The diamonds are set at four clusters. The two clusters that are most recently joined to form the four cluster model are the cluster consisting of Algeria to Bangladesh and the cluster consisting of Iraq to Saudi Arabia. The distance between these two clusters is the point on the distance plot indicated by the vertical line when the diamond is set to 4. The distance is given in the Clustering History report next to Number of Clusters equal to 4. There, it is shown that the distance is 1.618708760 and that clusters beginning with Algeria and Iraq are combined to yield four clusters.

The two clusters that are combined to yield five clusters are the cluster that consists of Australia to Korea, South, and the cluster that consists of Austria to Korea, North. The vertical join line in the dendrogram for these clusters is at about the same horizontal distance from the left as the vertical join line for the clusters that were joined to form the four-cluster model, Algeria to Bangladesh and Iraq to Saudi Arabia. It follows that there is not much difference in terms of distance between the clusters joined to form the four-cluster model and those joined to form the five-cluster model.

The fact that the distance plot levels off starting with the four-cluster model indicates that the cluster groupings up to that point do not account for much distance between clusters. However, the four-cluster model shows good separation between clusters.

Clustering History Report

The Clustering History table contains the clustering history.

Number of Clusters Lists the numbers of clusters that result *after* the joining indicated by the Leader and Joiner is performed. The number of clusters begins with the first join, when there are *n - 1* clusters, where *n* is the number of objects. The report lists the number of clusters in decreasing order until all objects are contained in one cluster. In this way, the Clustering History follows the order of the dendrogram from left to right.

Distance The distance between clusters, calculated according to the distance method that you select on the launch window. See "Method for Distance Calculation" on page 275.

Leader A representative of the first cluster in the dendrogram being joined. The cluster order and the representative shown in the Leader column is a consequence of how the data are sorted and has no intrinsic meaning.

Joiner A representative of the second cluster in the dendrogram being joined. The cluster order and the representative shown in the Joiner column is a consequence of how the data are sorted and has no intrinsic meaning.

Hierarchical Cluster Options

The Hierarchical Clustering red triangle menu contains the following options:

Color Clusters Colors the labels for dendrogram and their associated join bars according to cluster membership. Also assigns the corresponding colors to the rows of the data table The colors update if you change the number of clusters. If you deselect this option, the colors are no longer updated based on the number of clusters.

Mark Clusters Assigns markers to the rows of the data table corresponding to the cluster to which the row belongs. The markers update if you change the number of clusters. If you deselect this option, the markers are no longer updated based on the number of clusters.

Number of Clusters Prompts you to enter a number of clusters and positions the dendrogram slider to that number.

Cluster Criterion Gives the Cubic Clustering Criterion (CCC) for the entire range of number of clusters. The CCC is used to estimate the number of clusters. It can be used with any distance-based clustering algorithm. Larger values of the CCC indicate better fit in terms of number of clusters. See SAS Institute Inc. (1983). (Not available when **Data is distance matrix** is selected.)

Show Dendrogram Shows or hides the Dendrogram report.

Dendrogram Scale Contains the following options for scaling the dendrogram:

Distance Scale Shows the horizontal distances between any two join points as the distances between the two clusters joined at that point, based on the distance method specified on the launch window. The distance scale is the same scale as used in the Distance Graph and is the default scale for the dendrogram.

Even Spacing Shows the horizontal distances between any two join points as equal.

Geometric Spacing Increases the horizontal distances between join points as the number of clusters increases. This option is useful when there are many objects and you want the smaller clusters to be more visible than the larger clusters.

Distance Graph Shows or hides the distance plot beneath the dendrogram.

Show NCluster Handle Shows or hides the handles on the dendrogram used to manually change the number of clusters.

Zoom to Selected Rows Selects and enlarges a particular cluster after you select the cluster in the dendrogram. Alternatively, you can double-click the cluster to zoom in on it. Use Release Zoom to return to the original view.

Release Zoom Returns the dendrogram to the original view after zooming.

Pivot on Selected Cluster Reverses the order of the two sub-clusters of the currently selected cluster.

Color Map Gives the option to add a color map, or heat map, showing each Y, Column variable colored by value. Several color theme choices are available in a submenu.

Two Way Clustering Clusters by the variables specified in Y, Columns as well as rows. A color map is added with a dendrogram for the Y, Column variables at its base. Typically, for two-way clustering, your variables are measured on the same scale and you do not select Standardize Data. (Not available when **Data is stacked** is selected.)

Positioning Provides options for changing the positions of labels and other parts of the dendrogram.

Legend Shows or hides a legend for the colors used in color maps. This option is available only if a color map is enabled.

More Color Map Columns Adds a color map for specified columns. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.)

Constellation Plot Shows or hides an alternative way to present the information in the hierarchical clustering dendrogram. Each observation (row) is represented by an endpoint and each cluster join is represented by a new point. The lines that are drawn represent cluster membership. The lengths of the lines represent the distance between clusters. Longer lines represent greater distances between clusters.

You can hover over the lines in the constellation plot to see their length. However, the length values are meaningful only with respect to each other. The axis scaling, orientation of points, and angles of the lines are arbitrary. They are determined such that the ends of the nodes are spaced out and the plot does not appear cluttered, which is important with larger data sets.

To turn off the labels on the endpoints, right-click inside the Constellation Plot and deselect **Show Labels**.

Save Constellation Coordinates Saves the coordinates of the constellation plot to the data table. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.)

Save Clusters Creates a data table column that contains the cluster number. If Add Spatial Measures is selected on the launch window, the cluster numbers are also saved to the Hough Data Table.

Save Formula for Closest Cluster Creates a data table column that contains a formula for the closest cluster. This option calculates the squared Euclidean distance to each cluster's centroid and selects the cluster that is closest. Note that this formula does not always reproduce the cluster assignment given by Hierarchical Clustering since the clusters are determined differently. However, the cluster assignment is very similar. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.)

Save Display Order Creates a data table column that contains the order in which the row appears in the dendrogram.

Save Cluster Hierarchy Creates a data table that contains the information needed to write a script for a custom dendrogram. For each cluster join, there are three rows: the first for the joiner, the second for the leader, and the third for the result, giving the cluster centers, size, and other information.

Save Cluster Tree Creates a new data table that contains information needed to compare cluster trees between JMP and SAS. For each cluster join, there is one row for each new cluster, with the cluster's size and other information.

Save Distance Matrix Creates a new data table that contains the distances between the observations.

Save Cluster Means Creates a new data table that contains the number of rows and the means of each column in each cluster.

Cluster Summary (Not available when **Data is distance matrix** is selected.) Displays the following information:

Cluster Means A table that gives, for each cluster, the number of observations (or Object IDs, if the data are stacked) and means for each variable.

Cluster Standard Deviations A table that gives, for each cluster, the number of observations (or Object IDs, if the data are stacked) and standard deviations for each variable.

Cluster Means Plot Either a parallel plot or a two-dimensional heat map of the cluster means.

The plot is a parallel plot unless **Data is stacked** is selected and there are two Attribute ID variables. For the parallel plot, the axis for each variable is scaled.

- If Standardize Data were selected, the axis ranges from two standard deviations above and below the mean, where the standard deviation and mean are computed for the raw data. If a cluster mean falls beyond this range, the axis is extended to include it.
- If Standardize Data were not selected, there is a common vertical axis whose scaling is displayed. (The scaling is equivalent to the Scale Uniformly option in Graph Builder).

When **Data is stacked** is selected and there are two Attribute ID variables, two-dimensional plots of the mean of the Y variable at each location are shown for each cluster. These plots are colored using a Blue to Gray to Red color gradient.

Column Summary For each variable, gives the RSquare value that represents the proportion of variation explained by the clusters. This number is the RSquare value for a regression of the variable on the clusters. The option also gives a bar graph of RSquare values.

Scatterplot Matrix Creates a scatterplot matrix using all the variables. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.)

Parallel Coord Plots Creates a parallel coordinate plot for each cluster. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.) The axes are scaled as described for the Cluster Means Plot. See "Cluster Means Plot" on page 285.

Cluster Treatment Comparisons (Available only if you hold Shift and click the Hierarchical Clustering red triangle.) Select a response column and a two-level treatment column. Creates a Hierarchically Clustered Differences report.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Additional Examples of the Hierarchical Clustering Platform

- "Example of a Distance Matrix"
- "Example of Wafer Defect Classification Using Spatial Measures"

Example of a Distance Matrix

The proper data table structure for a distance matrix consists of the following:

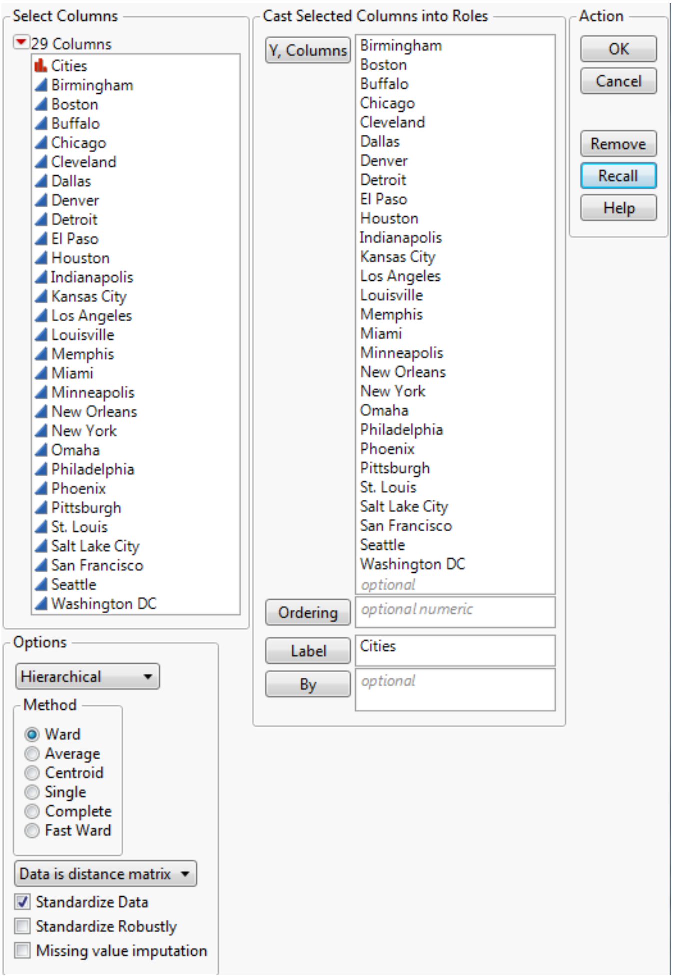
- An identifier column (usually the first column) that has a Character data type.
- A set of *n* columns, where *n* is also the number of rows. These *n* columns define a symmetric matrix with zero or missing values on the diagonal.

Notice that the distance matrix in Flight Distances.jmp follows the preceding format.

1. Select **Help > Sample Data Library** and open Flight Distances.jmp.
2. Select **Analyze > Clustering > Hierarchical Cluster**.

3. In the list at the bottom left corner of the launch window, change **Data as usual** to **Data is distance matrix**.
4. Select Cities and click **Label**.
5. Select all remaining columns and click **Y, Columns**.

Figure 12.6 Completed Distance Matrix Launch Window



Y, Columns: Birmingham, Boston, Buffalo, Chicago, Cleveland, Dallas, Denver, Detroit, El Paso, Houston, Indianapolis, Kansas City, Los Angeles, Louisville, Memphis, Miami, Minneapolis, New Orleans, New York, Omaha, Philadelphia, Phoenix, Pittsburgh, St. Louis, Salt Lake City, San Francisco, Seattle, Washington DC, *optional*

Ordering: *optional numeric*

Label: Cities

By: *optional*

Hierarchical

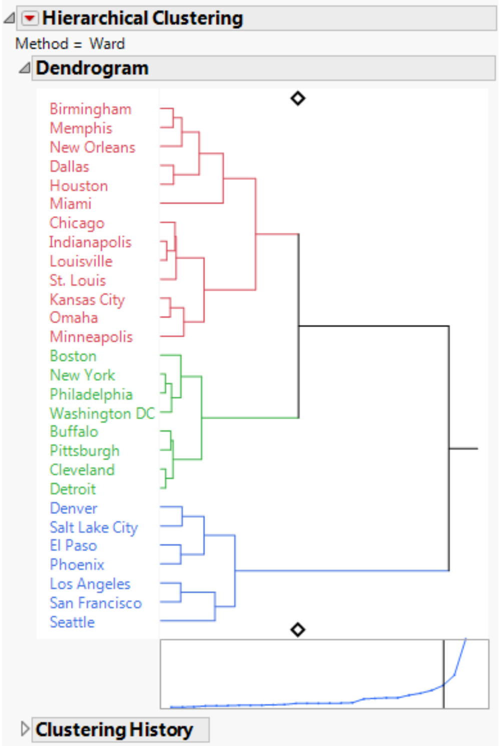
Method
● Ward
○ Average
○ Centroid
○ Single
○ Complete
○ Fast Ward

Data is distance matrix

☑ Standardize Data
☐ Standardize Robustly
☐ Missing value imputation

6. Click **OK**.
7. Click the Hierarchical Clustering red triangle and select **Color Clusters**.

Figure 12.7 Dendrogram Report for Flight Distances



In the Dendrogram report for the flight distances, the placement of the diamonds indicates that the model has grouped the cities into three clusters. The clusters are color-coded on the dendrogram. For more information about how to interpret the report, see "Dendrogram Report" on page 280.

Example of Wafer Defect Classification Using Spatial Measures

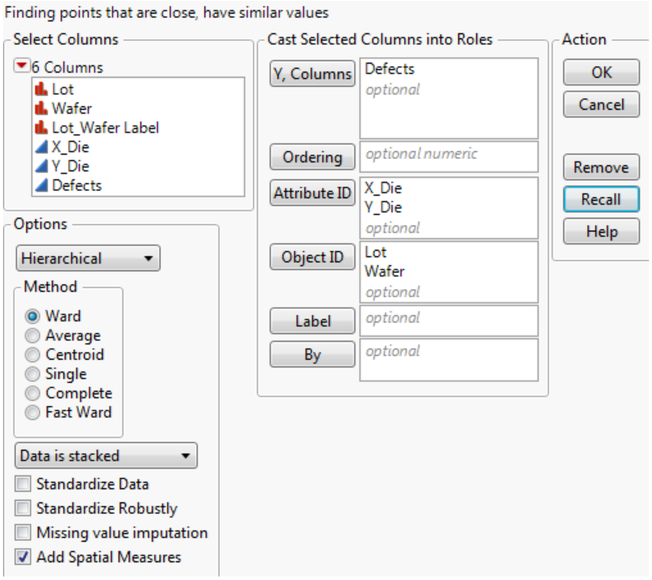
A specialty clustering option called Spatial Measures is available in the Hierarchical Cluster platform. In this example, you use this option to cluster wafers. For information about the option, see "Spatial Measures" on page 291.

1. Select **Help > Sample Data Library** and open Wafer Stacked.jmp.
2. Select **Analyze > Clustering > Hierarchical Cluster**.
3. In the list in the lower left corner, change **Data as usual** to **Data is stacked**.

   Additional options for stacked data appear in the launch window.
4. Select Defects and click **Y, Columns**.
5. Select X_Die and Y_Die and click **Attribute ID**.

6. Select Lot and Wafer and click **Object ID**.
7. Select **Add Spatial Measures** from the list of options in the lower left corner.

Figure 12.8 Completed Clustering Launch Window



Finding points that are close, have similar values

Select Columns

6 Columns
- Lot
- Wafer
- Lot_Wafer Label
- X_Die
- Y_Die
- Defects

Cast Selected Columns into Roles

Y, Columns	Defects optional
Ordering	optional numeric
Attribute ID	X_Die Y_Die optional
Object ID	Lot Wafer optional
Label	optional
By	optional

Action

Options

Hierarchical

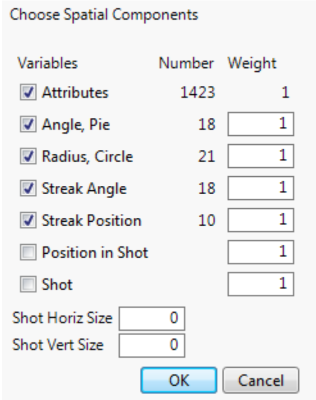
Method
- ● Ward
- ○ Average
- ○ Centroid
- ○ Single
- ○ Complete
- ○ Fast Ward

Data is stacked

- ☐ Standardize Data
- ☐ Standardize Robustly
- ☐ Missing value imputation
- ☑ Add Spatial Measures

8. Click **OK**.

Figure 12.9 Spatial Components Window



Choose Spatial Components

Variables	Number	Weight
☑ Attributes	1423	1
☑ Angle, Pie	18	1
☑ Radius, Circle	21	1
☑ Streak Angle	18	1
☑ Streak Position	10	1
☐ Position in Shot		1
☐ Shot		1

Shot Horiz Size 0
Shot Vert Size 0

Because Defects is measured at 1423 locations, there are 1423 Attributes variables.

9. Click **OK** to accept the selections in the Spatial window.

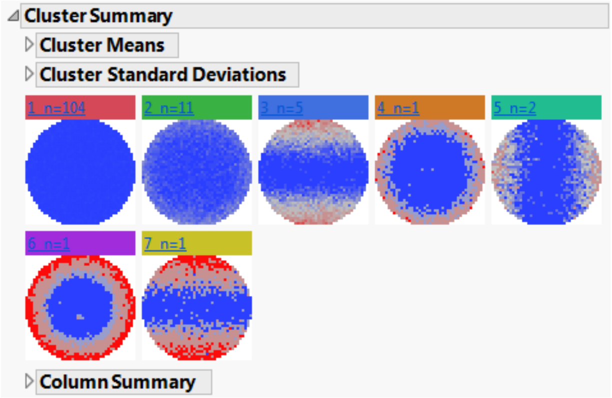
Two windows open: the Hierarchical Clustering report and the Wafer Stacked Defects Spatial data table.

10. In the Dendrogram plot, click and drag the diamond-shaped handle at the top to explore various numbers of clusters.

    As you drag the handle, the vertical line in the distance graph below the dendrogram moves to the corresponding number of clusters. The vertical coordinate gives the distance between the clusters that were joined at the given step. The graph seems to level off when the number of clusters is 7.

11. Click the Hierarchical Clustering red triangle and select **Number of Clusters**.
12. Enter 7 and click **OK**.
13. Click the Hierarchical Clustering red triangle and select **Cluster Summary**.

Figure 12.10 Cluster Summary Report



The wafer maps indicate the spatial nature of defects for each cluster. Cluster 1 contains 104 wafers with relatively few defects that are spread throughout the wafers. Cluster 3 has 5 wafers with defects concentrated at the extremes of the top and bottom hemispheres. You can view the maps for individual wafers and their Hough space maps in the data table produced by the cluster analysis. See "Spatial Measures" on page 291.

Statistical Details for the Hierarchical Clustering Platform

- "Spatial Measures"
- "Distance Method Formulas"

Spatial Measures

To use the Add Spatial Measures option, your data must be stacked and contain two attribute columns that correspond to spatial coordinates. Some of the spatial measures are constructed using the Hough transform. See White et al. (2008) and Ballard (1981). See "Example of Wafer Defect Classification Using Spatial Measures" on page 288.

Choose Spatial Components Window

The Choose Spatial Components window appears if you do the following in the launch window:

- Select the Data is stacked data structure
- Specify two columns as Attribute ID that correspond to spatial coordinates
- Specify an Object ID
- Select Add Spatial Measures

In the Choose Spatial Components window, you select and weight spatial components for your cluster analysis. These components are used to construct the variables used in the cluster analysis. A new table with a row for each object opens. This table contains the calculated spatial components for each object.

Variables The types of variables that are constructed and used in the cluster analysis. The variables are constructed using spatial components and the response, Y.

Attributes The value of the Y variable calculated at each location for each object, as defined by the two Attribute ID variables.

Angle, Pie Variables that reflect wedge shapes or hemispherical shapes.

Radius, Circle The variables that reflect circular shapes.

Streak Angle The variables that reflect streaks that have the same angle.

Streak Position The variables that reflect streaks with the same spatial position.

Position in Shot The variables that are based on the position of the die in the shot. Position in Shot variables are represented as ShotPos[vShotSize, hShotSize], where vShotSize and hShotSize are the defined vertical and horizontal shot sizes.

Shot The variables that identify which rectangle an object is in, where you specify the number of horizontal and vertical positions of objects in the rectangle. The term *shot* is used in semiconductor wafer data to identify which dies are imaged together across a wafer.

Enter values for Shot Horizontal Size and Shot Vertical Size. Specifying a horizontal shot size of 4 and a vertical shot size of 5 indicates that there are up to 20 dies in a shot. The total number of identifiers created is calculated as follows:

floor[(hSize+hShotSize-1)/hShotSize]* floor[(vSize+vShotSize-1)/vShotSize]

where hSize and vSize are the maximum numbers of horizontal and vertical positions, respectively, hShotSize = Shot Horizontal Size, and vShotSize = Shot Vertical Size.

Note: Shot variables are represented as Shot[vert, horiz], where vert and horiz represent the vertical and horizontal die locations, respectively.

Number The total number of variables of the given type that are constructed.

Weight A measure of importance for the given type of variable used in determining the clusters.

Spatial Measures Reports

When you click OK in the Choose Spatial Components window, two windows appear.

Hierarchical Clustering Report

When you conduct an analysis with stacked data and two Attribute IDs, the Cluster Summary report shows spatial maps of the Y variable. Each plot is a two-dimensional plot that displays the cluster mean for each location defined by the Attribute ID variables. The plot uses a Blue to Gray to Red color gradient with a Quantile scale. Using the quantile scale mitigates the effect of outliers.

Spatial Data Table

The data table for Spatial measures has a row for each unique Object ID. Columns are displayed using a Blue to Gray to Red default color gradient to show the Y variable. The table contains the following columns:

Object An expression column that shows a heat map of the Y variable at each spatial location defined by the two Attribute ID variables.

Hough An expression column that shows a heat map of the Hough space for each object. See White et al. (2008).

Spatial Measures A column for each spatial measure that shows the computed values for each object. Cells are colored by value.

Distance Method Formulas

This section provides the formulas used in calculating distances based on the Method that you select on the launch window. For a description of the methods, see "Method for Distance Calculation" on page 275.

The formulas use the following notation, where lowercase symbols generally pertain to observations and uppercase symbols to clusters:

n is the number of observations

v is the number of variables

x_i is the ith observation

C_K is the Kth cluster, subset of $\{1, 2, \ldots, n\}$

N_K is the number of observations in C_K

$\bar{x}$ is the sample mean vector

$\bar{x}_K$ is the mean vector for cluster C_K

$\|x\|$ is the square root of the sum of the squares of the elements of **x** (the Euclidean length of the vector **x**)

$d(x_i, x_j)$ is $\|x_i - x_j\|^2$

Average Linkage The distance for the average linkage cluster method is:

$$D_{KL} = \sum_{i \in C_K} \sum_{j \in C_L} \frac{d(x_i, x_j)}{N_K N_L}$$

Centroid Method The distance for the centroid method of clustering is:

$$D_{KL} = \|\overline{x_K} - \overline{x_L}\|^2$$

Ward's The distance for Ward's method is:

$$D_{KL} = \frac{\|\overline{x_K} - \overline{x_L}\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

Single Linkage The distance for the single linkage cluster method is:

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} \mathrm{d}(x_i, x_j)$$

Complete Linkage Distance for the Complete linkage cluster method is:

$$D_{KL} = \max_{i \in C_K} \max_{j \in C_L} \mathrm{d}(x_i, x_j)$$

Chapter 13

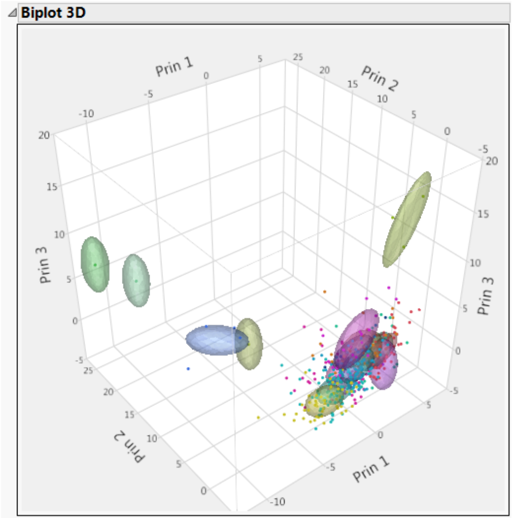
K Means Cluster

Group Observations Using Distances

Use the K Means Cluster platform to group observations that share similar values across a number of variables. Use the *k*-means method with larger data tables, ranging from approximately 200 to 100,000 observations.

The K Means Cluster platform constructs a specified number of clusters using an iterative algorithm that partitions the observations. The method, called *k-means*, partitions observations into clusters so as to minimize distances to cluster centroids. You must specify the number of clusters, *k*, in advance. However, you can compare the results of different values of *k* to select an optimal number of clusters for your data.

Figure 13.1 3D Biplot



Contents

Overview of the K Means Cluster Platform

K Means Cluster is one of four platforms that JMP provides for clustering observations. For a comparison of all four methods, see "Overview of Platforms for Clustering Observations" on page 297.

The K Means Cluster platform forms a specified number of clusters using an iterative fitting process. The *k*-means algorithm first selects a set of *k* points called *cluster seeds* as an initial guess for the means of the clusters. Each observation is assigned to the nearest cluster seed to form a set of temporary clusters. The seeds are then replaced by the cluster means, the points are reassigned, and the process continues until no further changes occur in the clusters.

The *k*-means algorithm is a special case of the *EM algorithm*, where *E* stands for Expectation, and *M* stands for maximization. In the case of the *k*-means algorithm, the calculation of temporary cluster means represents the Expectation step, and the assignment of points to the closest clusters represents the Maximization step.

K-Means clustering supports only numeric columns. *K*-Means clustering ignores modeling types (nominal and ordinal) and treats all numeric columns as continuous.

You must specify the number of clusters, *k*, or a range of values for *k*, in advance. However, you can compare the results of different values of *k* to select an optimal number of clusters for your data.

For background on *K*-Means clustering, see the FASTCLUS Procedure chapter in SAS Institute Inc. (2020d) and Hastie et al. (2009).

Overview of Platforms for Clustering Observations

Clustering is a multivariate technique that groups together observations that share similar values across a number of variables. Typically, observations are not scattered evenly through *p*-dimensional space, where *p* is the number of variables. Instead, the observations form clumps, or clusters. Identifying these clusters provides you with a deeper understanding of your data.

Note: JMP also provides a platform that enables you to cluster variables. See the "Cluster Variables" chapter on page 341.

JMP provides four platforms that you can use to cluster observations:

- Hierarchical Cluster is useful for smaller tables with up to several tens of thousands of rows and allows character data. Hierarchical clustering combines rows in a hierarchical sequence that is portrayed as a tree. You can choose the number of clusters that is most appropriate for your data after the tree is built.

- K Means Cluster is appropriate for larger tables with up to millions of rows and allows only numerical data. You need to specify the number of clusters, *k*, in advance. The algorithm guesses at cluster seed points. It then conducts an iterative process of alternately assigning points to clusters and recalculating cluster centers.
- Normal Mixtures is appropriate when your data come from a mixture of multivariate normal distributions that might overlap and allows only numerical data. For situations where you have multivariate outliers, you can use an outlier cluster with an assumed uniform distribution.

  You need to specify the number of clusters in advance. Maximum likelihood is used to estimate the mixture proportions and the means, standard deviations, and correlations jointly. Each point is assigned a probability of being in each group. The EM algorithm is used to obtain estimates.
- Latent Class Analysis is appropriate when most of your variables are categorical. You need to specify the number of clusters in advance. The algorithm fits a model that assumes a multinomial mixture distribution. A maximum likelihood estimate of cluster membership is calculated for each observation. An observation is classified into the cluster for which its probability of membership is the largest.

Table 13.1 Summary of Clustering Methods

Method	**Data Type or Modeling Type**	**Data Table Size**	**Specify Number of Clusters**
Hierarchical Cluster	Any	With Fast Ward, up to 200,000 rows With other methods, up to 5,000 rows	No
K Means Cluster	Numeric	Up to millions of rows	Yes
Normal Mixtures	Numeric	Any size	Yes
Latent Class Analysis	Nominal or Ordinal	Any size	Yes

Some of the clustering platforms have options to handle outliers in the data. However, if your data has outliers, it is best to explore them first prior to analyzing. This can be done using the Explore Outliers Utility. For more information, see the Modeling Utilities chapter in *Predictive and Specialized Modeling*.

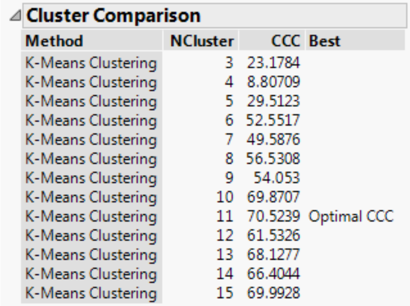
Example of K Means Cluster

In this example, you use the Cytometry.jmp sample data table to cluster observations using K Means Cluster. Cytometry is used to detect markers of the surface of cells and the readings from these markers help diagnose certain diseases. In this example, the observations are grouped based on readings of four markers in a cytometry analysis.

1. Select **Help > Sample Data Library** and open Cytometry.jmp
2. Select **Analyze > Clustering > K Means Cluster**.
3. Select CD3, CD8, CD4, and MCB and click **Y, Columns**.
4. Click **OK**.
5. Enter 3 next to **Number of Clusters.**
6. Enter 15 next to **Range of Clusters** (Optional).

   Because the Range of Clusters is set to 15, the platform provides fits for 3 to 15 clusters. You can then determine your preferred number of clusters.
7. Click **Go**.

Figure 13.2 Cluster Comparison Report



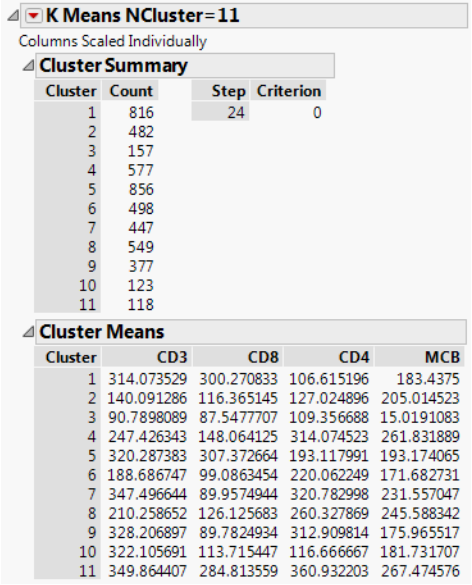
Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	3	23.1784	
K-Means Clustering	4	8.80709	
K-Means Clustering	5	29.5123	
K-Means Clustering	6	52.5517	
K-Means Clustering	7	49.5876	
K-Means Clustering	8	56.5308	
K-Means Clustering	9	54.053	
K-Means Clustering	10	69.8707	
K-Means Clustering	11	70.5239	Optimal CCC
K-Means Clustering	12	61.5326	
K-Means Clustering	13	68.1277	
K-Means Clustering	14	66.4044	
K-Means Clustering	15	69.9928	

The Cluster Comparison report appears at the top of the report window. The best fit is determined by the highest CCC value. In this case, the best fit occurs when you fit 11 clusters.

8. Scroll to the K Means NCluster=11 report.

Figure 13.3 K Means NCluster=11 Report



K Means NCluster=11

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	816
2	482
3	157
4	577
5	856
6	498
7	447
8	549
9	377
10	123
11	118

Step	Criterion
24	0

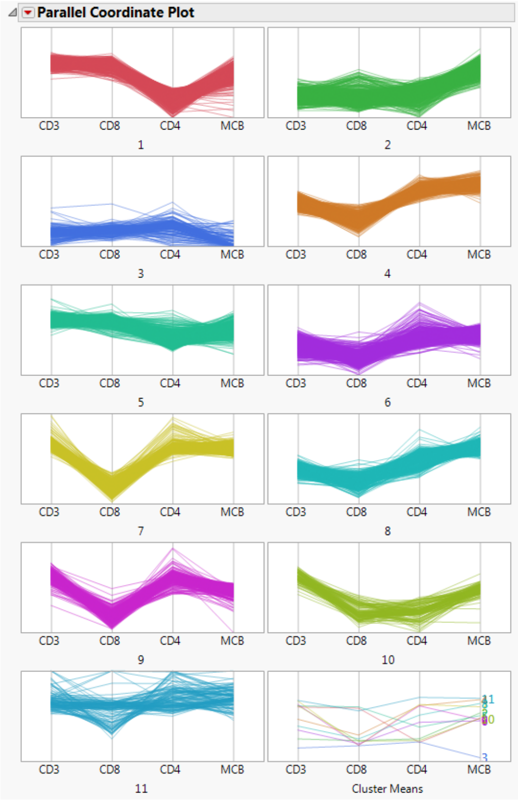
Cluster Means

Cluster	CD3	CD8	CD4	MCB
1	314.073529	300.270833	106.615196	183.4375
2	140.091286	116.365145	127.024896	205.014523
3	90.7898089	87.5477707	109.356688	15.0191083
4	247.426343	148.064125	314.074523	261.831889
5	320.287383	307.372664	193.117991	193.174065
6	188.686747	99.0863454	220.062249	171.682731
7	347.496644	89.9574944	320.782998	231.557047
8	210.258652	126.125683	260.327869	245.588342
9	328.206897	89.7824934	312.909814	175.965517
10	322.105691	113.715447	116.666667	181.731707
11	349.864407	284.813559	360.932203	267.474576

The Cluster Summary report shows the number of observations in each of the eleven clusters. The Cluster Means report shows the means of the four marker readings for each cluster.

9. Click the K Means NCluster=11 red triangle and select **Parallel Coord Plots**.

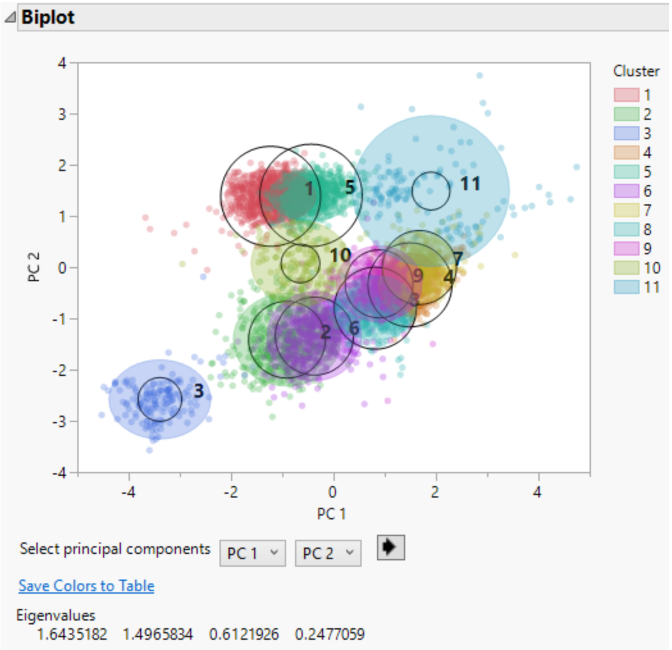
Figure 13.4 Parallel Coordinate Plots for Cytometry Data



The Parallel Coordinate Plots display the structure of the observations in each cluster. Use these plots to see how the clusters differ. Clusters 4, 6, 7, 8, and 9 tend to have comparatively low CD8 values and high CD4 values. Cluster 1, on the other hand, has higher CD8 values and lower CD4 values.

10. Click the K Means NCluster=11 red triangle and select **Biplot**.

Figure 13.5 Biplot for Cytometry Data



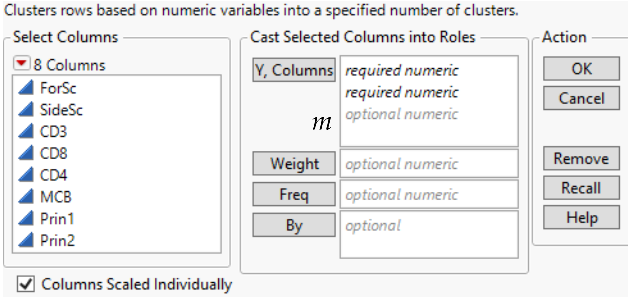
Eigenvalues
1.6435182 1.4965834 0.6121926 0.2477059

A legend that identifies the colors of the clusters is shown to the right of the plot. The clusters that appear to be most separated from the others based on their first two principal components are clusters 3, 10, and 11. This is supported by their parallel coordinate plots in Figure 13.4, which differ from the plots for the other clusters. Use the list below the plot to see the biplot for other combinations of principal components.

Launch the K Means Cluster Platform

Launch the K Means Cluster platform by selecting **Analyze > Clustering > K Means Cluster**.

Figure 13.6 K Means Cluster Launch Window



Clusters rows based on numeric variables into a specified number of clusters.

☑ Columns Scaled Individually

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns The variables used for clustering observations.

Note: *K*-Means clustering supports only numeric columns.

Weight A column whose numeric values assign a weight to each row in the analysis.

Freq A column whose numeric values assign a frequency to each row in the analysis.

By A column whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed. The results are presented in separate reports. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

Launch Window Options

Columns Scaled Individually Scales each column independently of the other columns. Use when variables do not share a common measurement scale, and you do not want one variable to dominate the clustering process. For example, one variable might have values that are between 0 and 1000, and another variable might have values between 0 and 10. In this situation, you can use the option so that the clustering process is not dominated by the first variable.

When you click OK, a Control Panel appears. See "Iterative Clustering Control Panel" on page 304.

Iterative Clustering Report

When you click OK in the launch window, the Iterative Clustering report window appears, showing a Control Panel for fitting models. See "Iterative Clustering Control Panel" on page 304. As you fit models, additional reports are added to the window. See "K Means Report" on page 305.

Iterative Clustering Options

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

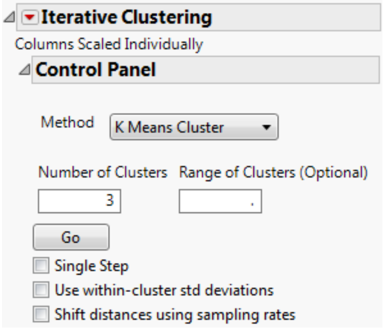
Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Iterative Clustering Control Panel

The Control Panel for the Cytometry.jmp data table is shown in Figure 13.7. You can iteratively fit different numbers of clusters or you can specify a range using the Range of Clusters option.

Figure 13.7 Iterative Clustering Control Panel



Columns Scaled Individually

Method: K Means Cluster

Number of Clusters: 3

Range of Clusters (Optional): .

☐ Single Step

☐ Use within-cluster std deviations

☐ Shift distances using sampling rates

The Control Panel has the following options:

Method The following clustering methods are available:

KMeans Clustering Described in this chapter.

Self Organizing Map Described in “Self Organizing Map Control Panel” on page 308.

Number of Clusters Designates the number of clusters to form.

Range of Clusters (Optional) Provides an upper bound for the number of clusters to form. If a number is entered here, the platform creates separate analyses for every integer between Number of Clusters and the value entered as Range of Clusters (Optional).

Go Unless Single Step is selected, fits the clusters automatically.

Single Step Enables you to step through the clustering process one iteration at a time. When you select Single Step and click Go, a K Means Cluster report appears with no cluster assignments but containing a Go and a Step button.

- Click the Step button to step through the iterations one at a time.
- Click the Go button to fit the clusters automatically.

Use within-cluster std deviations Scales distances using the estimated standard deviation of each variable for observations within each cluster. If you do not select this option, distances are scaled by an overall estimate of the standard deviation of each variable.

Shift distances using sampling rates Adjusts distances based on the sizes of clusters. If you have unequally sized clusters, an observation should have a higher probability of being assigned to larger clusters because there is a higher prior probability that the observation comes from a larger cluster.

K Means Report

When you click Go in the Control Panel, the following reports appear:

- A Cluster Comparison report. See "Cluster Comparison Report" on page 305.
- One or more K Means reports. The reports are dynamically named K Means NCluster=<k> depending on, *k*, the number of clusters fit. A K Means NCluster=<k> report appears for each fit that you conduct.

The Cluster Comparison report and the KMeans NCluster=11 report for the Cytometry.jmp data table, with the variables CD3 through MCB as Y, Columns, are shown in Figure 13.2 and "K Means NCluster=11 Report" on page 300.

Cluster Comparison Report

The Cluster Comparison report gives fit statistics to compare the various models. The fit statistic is the Cubic Clustering Criterion (CCC). Larger values of CCC indicate better fit. The best fit is indicated with the designation Optimal CCC in a column called Best. See SAS Institute Inc. (1983). Constant columns are not included in the CCC calculation.

K Means Report

Each K Means report gives the following summary statistics for each cluster:

- The Cluster Summary report gives the number of clusters and the observations in each cluster, as well as the number of iterations required.

- The Cluster Means report gives means for the observations in each cluster for each variable.
- The Cluster Standard Deviations report gives standard deviations for the observations in each cluster for each variable.

K Means Report Options

Each K Means report contains the following options:

Biplot Shows a plot of the points and clusters in the first two principal components of the data, along with a legend identifying the cluster colors. Circles are drawn around the cluster centers and the size of the circles is proportional to the count inside the cluster. The shaded area is the density contour around the mean. By default, this area indicates where 90% of the observations in that cluster would fall (Mardia et al. 1980). Use the list below the plot to change the plot axes to other principal components. Alternatively, use the arrow button to cycle through all possible axes combinations. An option to save the cluster colors to the data table is also located below the plot. See "Save Colors to Table" on page 307. The eigenvalues are shown in decreasing order.

Note: If Columns Scaled Individually is checked in the launch window, the biplot uses a correlation matrix. If Columns Scaled Individually is not checked, the biplot uses a covariance matrix.

Biplot Options Contains the following options for controlling the appearance of the Biplot:

Show Biplot Rays Shows the biplot rays. The labeled rays show the directions of the covariates in the subspace defined by the principal components. They represent the degree of association of each variable with each principal component.

Biplot Ray Position Enables you to specify the position and radius scaling of the biplot rays. By default, the rays emanate from the point (0,0). In the plot, you can drag the rays or use this option to specify coordinates. You can also adjust the scaling of the rays to make them more visible with the radius scaling option.

Biplot Contour Density Enables you to specify the confidence level for the density contours. The default confidence level is 90%.

Mark Clusters Assigns markers that identify the clusters to the rows of the data table.

Biplot 3D Shows a three-dimensional biplot of the data. Available only when there are three or more variables.

Parallel Coord Plots Creates a parallel coordinate plot for each cluster. The plot report has options for showing and hiding the data and means. See *Essential Graphing*.

Scatterplot Matrix Creates a scatterplot matrix using all of the Y variables.

Save Colors to Table Assigns colors that identify the clusters to the rows of the data table. If there is a Biplot in the report window, the colors saved to the data table match the colors of the clusters in the Biplot. If the colors are changed in the Biplot and the Save Colors To Table option is selected again, the colors in the table update to match those in the Biplot.

Note: When any of the Save options are selected, each saved column contains a Notes column property that specifies the number of clusters for that particular column's data. This enables you to save columns from more than one cluster fit and use the column property to identify which clustering fit the saved column is from.

Save Clusters Saves the following two columns to the data table:

- The Cluster column contains the number of the cluster to which the given row is assigned.
- (Not available for Self Organizing Maps.) The `Distance` column contains the squared Euclidean distance between the given observation and its cluster mean. For each variable, the difference between the observation's value and the cluster mean on that variable is divided by the overall standard deviation for the variable. These scaled differences are squared and summed across the variables.

Save Cluster Distance (Not available for Self Organizing Maps.) Saves a Distance column to the data table. This column is the same as the Distance column obtained from the **Save Clusters** option.

Save Cluster Formula Saves a formula column called Cluster Formula to the data table. This is the formula that identifies cluster membership for each.

Save Distance Formula (Not available for Self Organizing Maps.) Saves a formula column called Distance Formula to the data table. This is the formula that calculates the distance to the assigned cluster.

Save K Cluster Distances (Not available for Self Organizing Maps.) Saves *k* columns containing the squared Euclidean distances to each cluster center.

Save K Distance Formulas (Not available for Self Organizing Maps.) Saves *k* columns containing the formulas for the squared Euclidean distances to each cluster center.

Publish Cluster Formulas Publishes to the Formula Depot the same scoring code used in the Save Cluster Formula option.

Simulate Clusters Creates a new data table containing simulated cluster observations on the Y variables, using the cluster means and standard deviations.

Remove Removes the clustering report.

Self Organizing Map

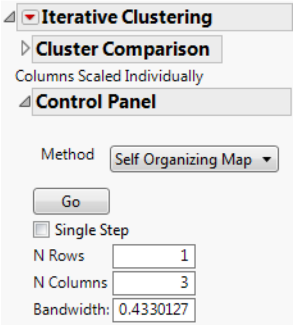
The *Self-Organizing Map* (SOM) technique was developed by Teuvo Kohonen (1989, 1990) and extended by other neural network enthusiasts and statisticians. The original SOM was cast as a learning process, like the original neural net algorithms, but the version implemented here is a variation on *k*-means clustering. In the SOM literature, this variation is called a *batch algorithm* using a *locally weighted linear smoother*.

The goal of a SOM is not only to form clusters in a particular layout on a cluster grid, such that points in clusters that are near each other in the SOM grid are also near each other in multivariate space. In classical *k*-means clustering, the structure of the clusters is arbitrary, but in SOMs the clusters have a grid structure. The grid structure helps interpret the clusters in two dimensions: clusters that are close are more similar than distant clusters. See "Description of SOM Algorithm" on page 309.

Self Organizing Map Control Panel

Select the Self Organizing Map option from the Method list in the Iterative Clustering Control Panel.

Figure 13.8 Self Organizing Map Control Panel



Columns Scaled Individually

Method: Self Organizing Map

☐ Single Step

N Rows: 1

N Columns: 3

Bandwidth: 0.4330127

Some of the options on the panel are described in "Iterative Clustering Control Panel" on page 304. The remaining options are described below.

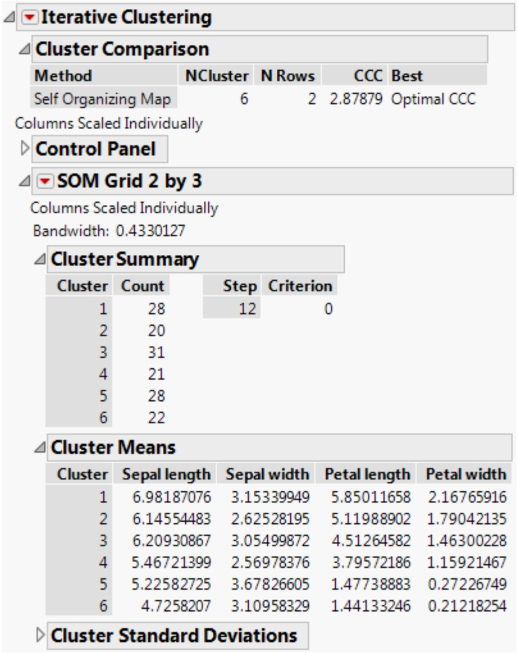
N Rows The number of rows in the cluster grid.

N Columns The number of columns in the cluster grid.

Bandwidth Specifies the effect of neighboring clusters for predicting centroids. A smaller bandwidth results in putting more weight on closer clusters.

Self Organizing Map Report

Figure 13.9 Self Organizing Map Report



Iterative Clustering

Cluster Comparison

Method	NCluster	N Rows	CCC	Best
Self Organizing Map	6	2	2.87879	Optimal CCC

Columns Scaled Individually

Control Panel

SOM Grid 2 by 3

Columns Scaled Individually

Bandwidth: 0.4330127

Cluster Summary

Cluster	Count
1	28
2	20
3	31
4	21
5	28
6	22

Step	Criterion
12	0

Cluster Means

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	6.98187076	3.15339949	5.85011658	2.16765916
2	6.14554483	2.62528195	5.11988902	1.79042135
3	6.20930867	3.05499872	4.51264582	1.46300228
4	5.46721399	2.56978376	3.79572186	1.15921467
5	5.22582725	3.67826605	1.47738883	0.27226749
6	4.7258207	3.10958329	1.44133246	0.21218254

Cluster Standard Deviations

The Cluster Comparison Report shows the total number of clusters and also the number of rows requested. The SOM report is named according to the Grid size requested. The Bandwidth is given at the top of the SOM Grid report. The report itself is analogous to the K Means NCluster report. See "K Means Report" on page 305.

For more information about the red triangle options for Self Organizing Maps, see "K Means Report Options" on page 306.

Description of SOM Algorithm

This section contains the steps of the SOM implementation in JMP.

- Initial cluster seeds are selected in a way that provides a good coverage of the multidimensional space. JMP uses principal components to determine the two directions that capture the most variation in the data.
- JMP then lays out a grid in this principal component space with its edges 2.5 standard deviations from the middle in each direction. The clusters seeds are determined by translating this grid back into the original space of the variables.

- The cluster assignment proceeds as with *k*-means. Each point is assigned to the cluster closest to it.
- The means are estimated for each cluster as in *k*-means. JMP then uses these means to set up a weighted regression with each variable as the response in the regression, and the SOM grid coordinates as the regressors. The weighting function uses a kernel function that gives large weight to the cluster whose center is being estimated. Smaller weights are given to clusters farther away from the cluster in the SOM grid. The new cluster means are the predicted values from this regression.
- These iterations proceed until the process has converged.

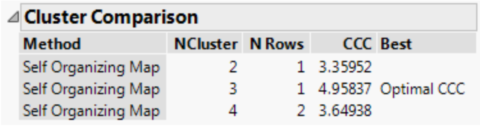
Additional Example of K Means Cluster Platform

Example of a Self-Organizing Map

This example uses the Iris.jmp sample data table, which includes measurements of sepal length, sepal width, petal length, and petal width for three species of irises.

1. Select **Help > Sample Data Library** and open Iris.jmp.
2. Select **Analyze > Clustering > K Means Cluster**.
3. Select Sepal length, Sepal width, Petal length, and Petal width and click **Y, Columns**.
4. Click **OK**.
5. Select **Self Organizing Map** from the Method menu on the Control Panel.
6. Set **N Rows** equal to 1 and **N Columns** equal to 2.
7. Click **Go**.
8. Open the Control Panel Report.
9. Set **N Rows** equal to 1 and **N Columns** equal to 3.
10. Click **Go**.
11. Open the Control Panel Report.
12. Set **N Rows** equal to 2 and **N Columns** equal to 2.
13. Click **Go**.

Figure 13.10 SOM Cluster Comparison



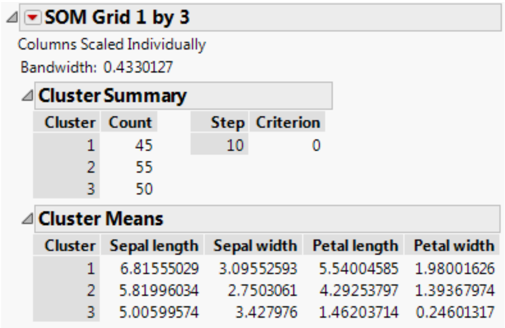
Cluster Comparison

Method	NCluster	N Rows	CCC	Best
Self Organizing Map	2	1	3.35952	
Self Organizing Map	3	1	4.95837	Optimal CCC
Self Organizing Map	4	2	3.64938	

The Cluster Comparison report appears at the top of the report window. The best fit is determined by the highest CCC value. Notice the number of clusters that gives the largest CCC is 3, which is the number of species.

14. Scroll to the SOM Grid 1 by 3 report. We can see the classification was not perfect; each cluster should represent each species, with 50 rows for each.

Figure 13.11 Self-Organizing Map Report for Iris.jmp



SOM Grid 1 by 3

Columns Scaled Individually

Bandwidth: 0.4330127

Cluster Summary

Cluster	Count
1	45
2	55
3	50

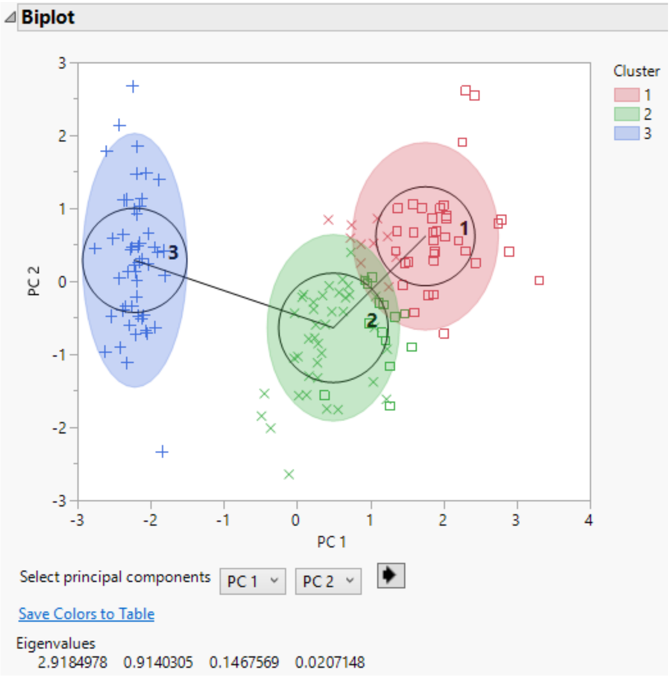
Step	Criterion
10	0

Cluster Means

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	6.81555029	3.09552593	5.54004585	1.98001626
2	5.81996034	2.7503061	4.29253797	1.39367974
3	5.00599574	3.427976	1.46203714	0.24601317

15. In the data table, select the Species column and select **Rows > Color or Mark by Column**.
16. Select the **Classic** option under Markers.
17. Click **OK**.
18. Click the red triangle menu next to SOM Grid 1 by 3 and select **Biplot**.

Figure 13.12 SOM Biplot



Eigenvalues
2.9184978 0.9140305 0.1467569 0.0207148

We can see that all rows from Cluster 3 are correctly identified as the setosa species. The other two species, virginica and versicolor, overlap slightly and can be mistaken for each other.

Chapter 14

Normal Mixtures

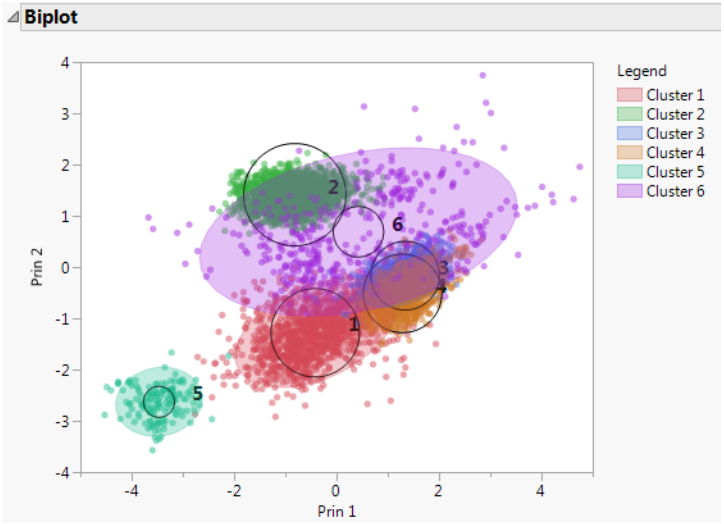
Group Observations Using Probabilities

Use Normal Mixtures for clustering when your data come from overlapping normal distributions. You need to specify the number of clusters in advance.

Normal mixtures is an iterative technique based on the assumption that the joint probability distribution of the observations is approximated using a mixture of multivariate normal distributions. These mixtures represent different clusters. The individual clusters have multivariate normal distributions.

When clusters are well separated, hierarchical and *k*-means clustering work well. But when clusters overlap, normal mixtures provides a better alternative, because it is based on cluster membership probabilities, rather than arbitrary cluster assignments based on borders.

Figure 14.1 Normal Mixtures Biplot



Contents

Overview of the Normal Mixtures Clustering Platform

Normal Mixtures is one of four platforms that JMP provides for clustering observations. For a comparison of all four methods, see "Overview of Platforms for Clustering Observations" on page 315.

Normal mixtures is an iterative clustering technique for numerical variables. However, it also predicts the proportion of responses expected within each cluster. Normal mixtures assumes that the joint probability distribution of the measurement columns can be approximated using a mixture of multivariate normal distributions, which represent different clusters. Mean vectors and covariance matrices are estimated for each cluster. See McLachlan and Krishnan (1997) and Section 9.6 in Hand et al. (2001).

Note: The Normal Mixtures algorithm involves iterating through random guesses for the cluster centers. Because of this, results from different runs of the analysis might differ slightly.

If you suspect that you have multivariate outliers, you have two options. You can use an outlier cluster or the Explore Outliers Utility. The outlier cluster option assumes a uniform distribution and is less sensitive to outliers than the standard Normal Mixtures method. The Explore Outliers Utility enables you to explore and handle outliers prior to analysis. See "Outlier Cluster" on page 321 and *Predictive and Specialized Modeling*.

Overview of Platforms for Clustering Observations

Clustering is a multivariate technique that groups together observations that share similar values across a number of variables. Typically, observations are not scattered evenly through p-dimensional space, where p is the number of variables. Instead, the observations form clumps, or clusters. Identifying these clusters provides you with a deeper understanding of your data.

Note: JMP also provides a platform that enables you to cluster variables. See the "Cluster Variables" chapter on page 341.

JMP provides four platforms that you can use to cluster observations:

- Hierarchical Cluster is useful for smaller tables with up to several tens of thousands of rows and allows character data. Hierarchical clustering combines rows in a hierarchical sequence that is portrayed as a tree. You can choose the number of clusters that is most appropriate for your data after the tree is built.
- K Means Cluster is appropriate for larger tables with up to millions of rows and allows only numerical data. You need to specify the number of clusters, *k*, in advance. The

algorithm guesses at cluster seed points. It then conducts an iterative process of alternately assigning points to clusters and recalculating cluster centers.

- Normal Mixtures is appropriate when your data come from a mixture of multivariate normal distributions that might overlap and allows only numerical data. For situations where you have multivariate outliers, you can use an outlier cluster with an assumed uniform distribution.

  You need to specify the number of clusters in advance. Maximum likelihood is used to estimate the mixture proportions and the means, standard deviations, and correlations jointly. Each point is assigned a probability of being in each group. The EM algorithm is used to obtain estimates.

- Latent Class Analysis is appropriate when most of your variables are categorical. You need to specify the number of clusters in advance. The algorithm fits a model that assumes a multinomial mixture distribution. A maximum likelihood estimate of cluster membership is calculated for each observation. An observation is classified into the cluster for which its probability of membership is the largest.

Table 14.1 Summary of Clustering Methods

Method	Data Type or Modeling Type	Data Table Size	Specify Number of Clusters
Hierarchical Cluster	Any	With Fast Ward, up to 200,000 rows With other methods, up to 5,000 rows	No
K Means Cluster	Numeric	Up to millions of rows	Yes
Normal Mixtures	Numeric	Any size	Yes
Latent Class Analysis	Nominal or Ordinal	Any size	Yes

Some of the clustering platforms have options to handle outliers in the data. However, if your data has outliers, it is best to explore them first prior to analyzing. This can be done using the Explore Outliers Utility. For more information, see the Modeling Utilities chapter in *Predictive and Specialized Modeling*.

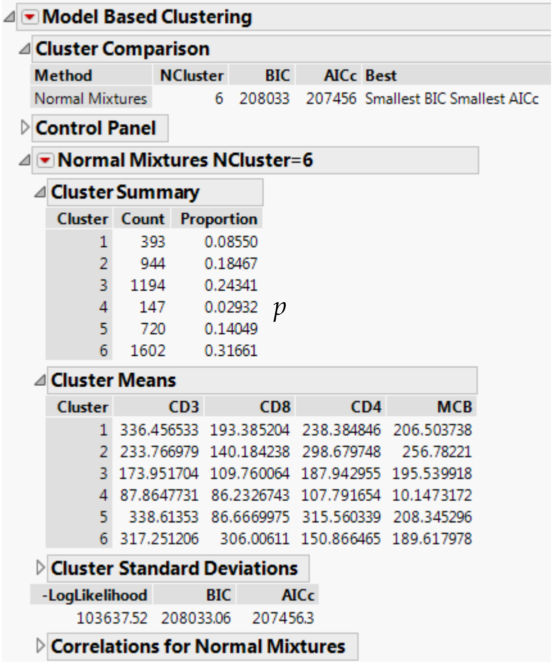
Example of Normal Mixtures Clustering

Cytometry is used to measure various characteristics of cells. Measurements of cell markers help diagnose certain diseases. In this example, you cluster observations based on readings of four markers in a cytometry analysis.

1. Select **Help > Sample Data Library** and open Cytometry.jmp
2. Select **Analyze > Clustering > Normal Mixtures**
3. Select CD3, CD8, CD4, and MCB and click **Y, Columns**.
4. Click **OK**.
5. Enter 6 next to **Number of Clusters**.
6. Click **Go**.

Note: Your results might differ because the algorithm has a random starting value.

Figure 14.2 Normal Mixtures NCluster=6 Report



Method	NCluster	BIC	AICc	Best
Normal Mixtures	6	208033	207456	Smallest BIC Smallest AICc

Cluster	Count	Proportion
1	393	0.08550
2	944	0.18467
3	1194	0.24341
4	147	0.02932
5	720	0.14049
6	1602	0.31661

Cluster	CD3	CD8	CD4	MCB
1	336.456533	193.385204	238.384846	206.503738
2	233.766979	140.184238	298.679748	256.78221
3	173.951704	109.760064	187.942955	195.539918
4	87.8647731	86.2326743	107.791654	10.1473172
5	338.61353	86.6669975	315.560339	208.345296
6	317.251206	306.00611	150.866465	189.617978

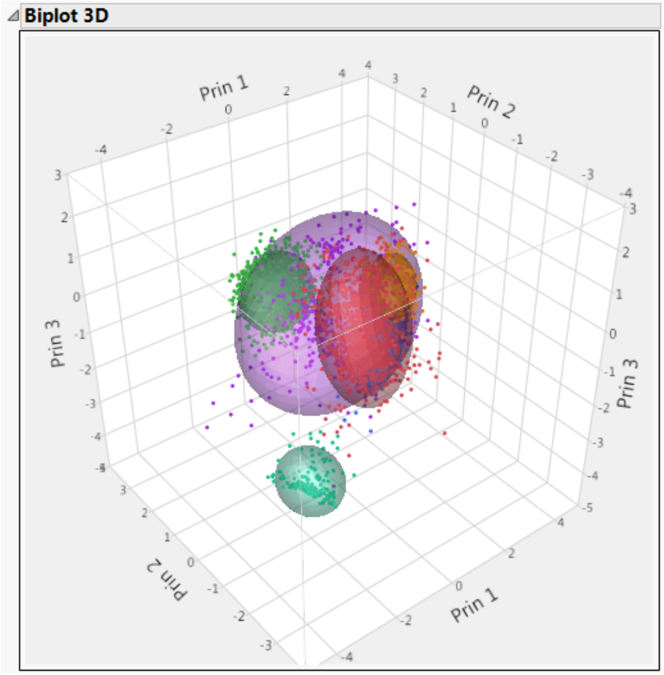
-LogLikelihood	BIC	AICc
103637.52	208033.06	207456.3

The Cluster Summary report shows the number of observations in each of the six clusters. The Cluster Means report shows the means of the four marker readings for each cluster.

7. Click the red triangle next to Normal Mixtures NCluster=6 and select **Biplot 3D**.

Note: Your biplot 3D might appear differently because the algorithm has a random starting value.

Figure 14.3 3D Biplot of Cytometry Data

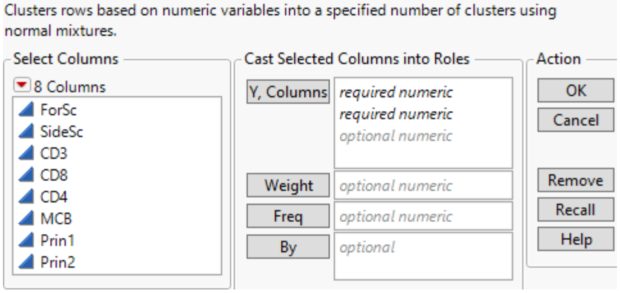


The plot shows contours for the normal densities that are fit to the clusters. Note that one cluster appears to be distinctly separated from the other clusters based on the first three principal components.

Launch the Normal Mixtures Clustering Platform

Launch the Normal Mixtures Clustering platform by selecting Analyze > Clustering > Normal Mixtures.

Figure 14.4 Normal Mixtures Launch Window



Clusters rows based on numeric variables into a specified number of clusters using normal mixtures.

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns The variables used for clustering observations.

> **Note:** Normal Mixtures clustering supports only numeric columns.

Weight A column whose numeric values assign a weight to each row in the analysis.

Freq A column whose numeric values assign a frequency to each row in the analysis.

By A column whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed. The results are presented in separate reports. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

When you click OK, a Control Panel appears. See "Model Based Clustering Control Panel" on page 320.

Model Based Clustering Report

When you click OK in the launch window, the Model Based Clustering report window opens, showing a Control Panel for fitting models. See "Model Based Clustering Control Panel" on page 320. As you fit models, additional reports are added to the window. See "Normal Mixtures Report" on page 321.

Model Based Clustering Options

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

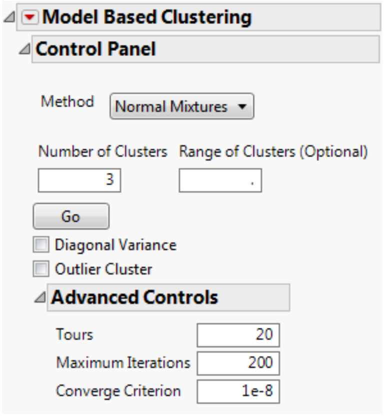
Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Model Based Clustering Control Panel

The Control Panel for the Cytometry.jmp data table, with the variables CD3 through MCB as **Y, Columns**, is shown in Figure 14.5. You can fit various numbers of clusters using the Control Panel iteratively or you can specify a range using the Range of Clusters option.

Figure 14.5 Control Panel for Normal Mixtures Method



Model Based Clustering
Control Panel
Method: Normal Mixtures
Number of Clusters: 3
Range of Clusters (Optional): .

☐ Diagonal Variance
☐ Outlier Cluster
Advanced Controls
Tours: 20
Maximum Iterations: 200
Converge Criterion: 1e-8

The Model Based Clustering Control Panel has these options:

Number of Clusters Designates the number of clusters to form.

Range of Clusters (Optional) Provides an upper bound for the number of clusters to form. If a number is entered here, the platform creates separate analyses for every integer between **Number of clusters** and the value entered as **Range of Clusters (Optional)**.

Go Fits the clusters.

Diagonal Variance Constrains the off-diagonal elements of the covariance matrix to zero. The platform fits multivariate normal distributions that have no correlations between the variables.

> **Note:** The Diagonal Variance option is sometimes necessary to avoid obtaining a singular covariance matrix when there are fewer observations than variables. It can also be used to avoid estimating very large covariance matrices for large numbers of variables.

Outlier Cluster Fits a cluster to catch outliers that do not fall into any of the normal clusters. If this cluster is created, it is designated Cluster 0, and the count of observations appears in the Cluster Summary report. The distribution of observations that fall in the outlier cluster is assumed to be uniform over the hypercube that encompasses the observations.

Advanced Controls The following advanced controls are available:

Tours The number of independent restarts of the estimation process. Each restart has a different starting value. Independent starts help guard against finding local solutions.

Maximum Iterations The maximum number of iterations of the convergence stage of the EM algorithm.

Converge Criterion The difference in the likelihood at which the EM iterations terminate.

Normal Mixtures Report

When you click Go in the Control Panel, the following reports appear:

- A Cluster Comparison report. See “Cluster Comparison Report” on page 322
- One or more Normal Mixtures reports. The reports are dynamically named Normal Mixtures NCluster=<k>, depending on *k*, the number of clusters fit. A Normal Mixtures NCluster=<k> report appears for each fit that you conduct.

The Cluster Comparison report and the Normal Mixtures NCluster=6 report for the Cytometry.jmp data table, with the variables CD3 through MCB as **Y, Columns**, are shown in Figure 14.2 and "Normal Mixtures NCluster=6 Report" on page 317.

Cluster Comparison Report

The Cluster Comparison report gives fit statistics to compare the various models. The fit statistics are BIC and AICc. Smaller values of each indicate better fit. The best fit is indicated in a column called Best.

Normal Mixtures Report

The Normal Mixtures report gives summary statistics for each cluster:

- The Cluster Summary report gives the number of observations and proportion for each cluster.
- The Cluster Means report gives means for the observations in each cluster for each variable.
- The Cluster Standard Deviations report gives standard deviations for the observations in each cluster for each variable.
- The -LogLikelihood table gives the negative log-likelihood, BIC, and AICc. See *Fitting Linear Models*.
- The Correlations for Normal Mixtures report gives the estimated correlation matrix for each cluster

Normal Mixtures Report Options

Biplot Shows a plot of the points and clusters in the first two principal components of the data, along with a legend identifying the cluster colors. Circles are drawn around the cluster centers and the size of the circles is proportional to the count inside the cluster. The shaded area is the density contour around the mean. By default, this area indicates where 90% of the observations in that cluster would fall (Mardia et al. 1980). Use the list below the plot to change the plot axes to other principal components. Alternatively, use the arrow button to cycle through all possible axes combinations. An option to save the cluster colors to the data table is also located below the plot. See "Save Colors to Table" on page 323. The eigenvalues are shown in decreasing order.

Note: The biplot always uses the correlation matrix to calculate the principal components.

Biplot Options Contains options for controlling the appearance of the Biplot.

Show Biplot Rays Shows the biplot rays. The labeled rays show the directions of the covariates in the subspace defined by the principal components. They represent the degree of association of each variable with each principal component.

Biplot Ray Position Enables you to specify the position and radius scaling of the biplot rays. By default, the rays emanate from the point (0,0). In the plot, you can drag the rays or use this option to specify coordinates. You can also adjust the scaling of the rays to make them more visible with the radius scaling option.

Biplot Contour Density Enables you to specify the confidence level for the density contours. The default confidence level is 90%.

Mark Clusters Assigns markers that identify the clusters to the rows of the data table.

Biplot 3D Shows a three-dimensional biplot of the data. Available only when there are three or more variables.

Parallel Coord Plots Creates a parallel coordinate plot for each cluster. The plot report has options for showing and hiding the data and means. See *Essential Graphing*.

Scatterplot Matrix Creates a scatterplot matrix using all of the Y variables.

Save Colors to Table Assigns colors that identify the clusters to the rows of the data table. If there is a Biplot in the report window, the colors saved to the data table match the colors of the clusters in the Biplot. If the colors are changed in the Biplot and the Save Colors To Table option is selected again, the colors in the table update to match those in the Biplot.

Note: When any of the Save options are selected, each saved column contains a Notes column property that specifies the number of clusters for that particular column's data. This enables you to save columns from more than one cluster fit and use the column property to identify which clustering fit the saved column is from.

Save Clusters Adds a column called Cluster that contains the number of the cluster to which the given row is assigned to the data table. For normal mixtures, this is the cluster that is most likely.

Save Cluster Formula Adds a formula column called Cluster Formula to the data table. This formula identifies which cluster the row belongs to.

Publish Cluster Formulas Publishes to the Formula Depot the same scoring code used in the Save Cluster Formula option. If Publish Cluster Formulas is selected and Run Script is chosen from the model within the Formula Depot, the columns saved to the data table should match those that are saved when Save Cluster Formula is selected.

Save Mixture Probabilities Adds a column called Prob Cluster <k> for each cluster that contains the probability an observation belongs to that cluster.

Save Mixture Formulas Adds columns to the data table that contain the formulas used to calculate the mixture probabilities. Use these formula columns to score probabilities for excluded data, or data that you add to the table.

Dist Formula <k> The estimated multivariate normal density function for Cluster <k> evaluated at the observation.

Dist Total The sum of the distance formula columns. The formula in this column is equivalent to the formula in the Mixture Density column created by the Save Density Formula option.

Prob Formula <k> The probability that the observation belongs to Cluster <k>. These columns contain the formulas that give the values in the Prob Cluster <k> columns created by the Save Mixture Probabilities option. The column formula for calculating the mixture probabilities is:

$$\text{Prob Formula <k>} = \frac{\text{Dist Formula <k>}}{\text{Dist Total}}$$

Save Density Formula Adds a column called Mixture Density that contains the estimated density function for the normal mixture to the data table.

Simulate Clusters Uses the mixture density to simulate predictor values. Saves these and the clusters into which they are classified in a new data table.

Remove Removes the clustering report.

Statistical Details for the Normal Mixtures Clustering Platform

Normal Mixtures uses the EM algorithm to do fitting because it is more stable than the Newton-Raphson algorithm. In addition, JMP uses a Bayesian regularized version of the EM algorithm, which allows smooth handling of cases where the covariance matrix is singular. Since the estimates are heavily dependent on initial guesses, the platform iterates through a number of tours. Each tour has randomly selected points for the initial center.

Doing multiple tours makes the estimation process somewhat expensive, so considerable patience is required for large problems. Controls enable you to specify the tour and iteration limits.

Chapter 15

Latent Class Analysis

Group Observations of Categorical Variables

Latent class analysis enables you to find clusters of observations for categorical response variables. A latent variable is an unobservable grouping variable. Each level of the latent variable is called a latent class. The Latent Class Analysis platform fits a latent class model and determines the most likely cluster or latent class for each observation. In most situations, a subject matter expert uses the results of a latent class analysis to create definitions for each latent class based on the characteristics of the class.

Figure 15.1 Example of Latent Class Analysis

Parameter Estimates

		sex		marital status		country			size			type		
Cluster	Overall	Female	Male	Married	Single	American	European	Japanese	Large	Medium	Small	Family	Sporty	Work
Cluster 1	0.28553	0.3763	0.6237	0.4160	0.5840	0.1555	0.2306	0.6139	0.0009	0.2839	0.7152	0.0190	0.9635	0.0175
Cluster 2	0.25943	0.4066	0.5934	0.6740	0.3260	0.0271	0.0562	0.9168	0.0029	0.2942	0.7029	0.6419	0.0368	0.3213
Cluster 3	0.20712	0.6793	0.3207	0.7526	0.2474	0.5618	0.1376	0.3005	0.0025	0.9914	0.0062	0.8013	0.1950	0.0036
Cluster 4	0.19697	0.4707	0.5293	0.9922	0.0078	0.8532	0.0826	0.0642	0.4965	0.2047	0.2989	0.7651	0.0020	0.2329
Cluster 5	0.05095	0.1837	0.8163	0.0324	0.9676	0.8477	0.1468	0.0054	0.7766	0.1180	0.1054	0.4366	0.0957	0.4677

Cluster	Overall	sex	marital status	country	size	type
Cluster 1	0.28553					
Cluster 2	0.25943					
Cluster 3	0.20712					
Cluster 4	0.19697					
Cluster 5	0.05095					

Contents

Overview of the Latent Class Analysis Platform

The Latent Class Analysis platform fits a latent class model to categorical response variables and determines the most likely cluster or latent class for each observation. A *latent variable* is an unobservable grouping variable. Each level of the latent variable is called a *latent class*. For example, latent classes could be clusters of survey respondents that are grouped by their preference for risk.

The model takes the form of a multinomial mixture model. There are two sets of parameters in the model: the γ parameters and the ρ parameters. The γ parameters represent the overall probabilities of cluster membership. The ρ parameters represent the probabilities of observing a given response conditional on cluster membership. A latent class is characterized by a pattern of these conditional probabilities.

In order for the analysis results to be meaningful, a subject matter expert must interpret the clusters that the platform generates. This subject matter expert examines characteristics of the latent classes and constructs a definition for each class based on those characteristics.

Note: Rows with missing values in any of the response columns are excluded from the analysis.

For more information about latent class models, see Collins and Lanza (2010) and Goodman (1974).

Example of Latent Class Analysis

This example uses the Latent Class Analysis platform to analyze responses to a 2005 survey of US high school students. The survey asked students a variety of multiple choice questions regarding health-risk behaviors.

In this example, you fit a latent class model to identify clusters of students based on their responses to 12 questions. The columns that you analyze were obtained from multiple choice survey questions by binning the responses into two classes (Yes/No).

1. Select **Help > Sample Data Library** and open Health Risk Survey.jmp.
2. In the Health Risk Survey data table, click the green triangle next to the Launch LCA Platform script.

   The script selects the 12 columns of interest, opens the Latent Class Analysis launch window, and enters the 12 columns of interest as Y.

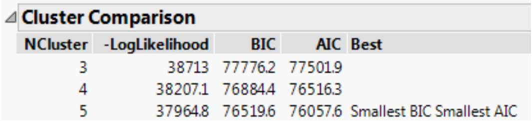
Note: To launch the LCA Platform on your own, select Analysis > Clustering > Latent Class Analysis.

3. Type 5 in the box next to **Up to**.

   This option fits latent class models for 3 and up to 5 clusters.

4. Click **OK**.

Figure 15.2 Cluster Comparison Report



Cluster Comparison

NCluster	-LogLikelihood	BIC	AIC	Best
3	38713	77776.2	77501.9	
4	38207.1	76884.4	76516.3	
5	37964.8	76519.6	76057.6	Smallest BIC Smallest AIC

The Latent Class Analysis outline contains a Cluster Comparison report and three separate Latent Class Model reports. The Latent Class Model reports show the models for three, four, and five clusters. In the Cluster Comparison report, the model with five clusters has the smallest BIC and AIC, which indicates that it is the best fitting model out of the three. This is the model that you analyze.

5. In the Latent Class Model for 5 Clusters report, examine the bar charts under Parameter Estimates. Note the following:
   - Cluster 1 has mostly No answers to all of the risk behaviors.
   - Cluster 2 has high numbers of Yes answers for the four risk behaviors before the age of 13.
   - Cluster 3 has high numbers of Yes answers for driving when drinking and five or more drinks in the past 30 days.
   - Cluster 4 has high numbers of Yes answers for most of the risk behaviors except for the ones before the age of 13.
   - Cluster 5 has the highest number of Yes answers for most of the risk behaviors.

   Use this information to give the clusters meaningful names.

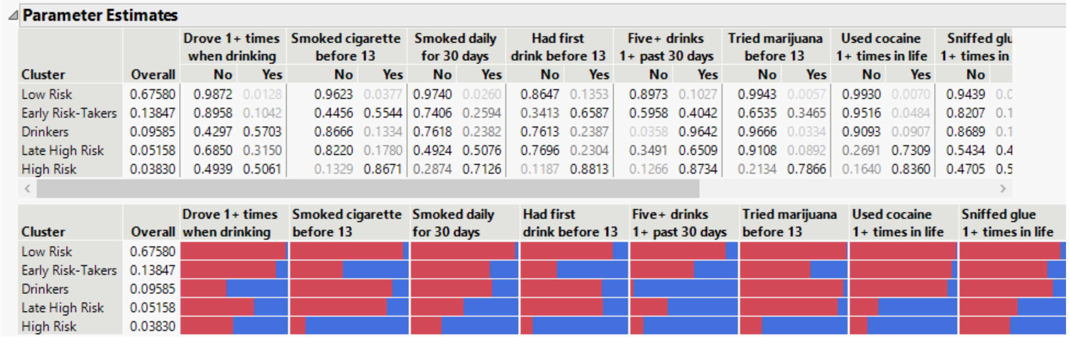
6. Click the red triangle next to Latent Class Model for 5 Clusters and select **Rename Clusters**:
   - Enter Low Risk for Cluster 1.
   - Enter Early Risk-Takers for Cluster 2.
   - Enter Drinkers for Cluster 3.
   - Enter Late High Risk for Cluster 4.
   - Enter High Risk for Cluster 5.

7. Click **OK**.

8. Click **OK** in the JMP Alert that appears.

Note: The new cluster names are not saved to scripts.

Figure 15.3 Partial Parameter Estimates Report



Parameter Estimates

Cluster	Overall	Drove 1+ times when drinking No	Drove 1+ times when drinking Yes	Smoked cigarette before 13 No	Smoked cigarette before 13 Yes	Smoked daily for 30 days No	Smoked daily for 30 days Yes	Had first drink before 13 No	Had first drink before 13 Yes	Five+ drinks 1+ past 30 days No	Five+ drinks 1+ past 30 days Yes	Tried marijuana before 13 No	Tried marijuana before 13 Yes	Used cocaine 1+ times in life No	Used cocaine 1+ times in life Yes	Sniffed glu 1+ times in No	
Low Risk	0.67580	0.9872	0.0128	0.9623	0.0377	0.9740	0.0260	0.8647	0.1353	0.8973	0.1027	0.9943	0.0057	0.9930	0.0070	0.9439	0.0
Early Risk-Takers	0.13847	0.8958	0.1042	0.4456	0.5544	0.7406	0.2594	0.3413	0.6587	0.5958	0.4042	0.6535	0.3465	0.9516	0.0484	0.8207	0.1
Drinkers	0.09585	0.4297	0.5703	0.8666	0.1334	0.7618	0.2382	0.7613	0.2387	0.0358	0.9642	0.9666	0.0334	0.9093	0.0907	0.8689	0.1
Late High Risk	0.05158	0.6850	0.3150	0.8220	0.1780	0.4924	0.5076	0.7696	0.2304	0.3491	0.6509	0.9108	0.0892	0.2691	0.7309	0.5434	0.4
High Risk	0.03830	0.4939	0.5061	0.1329	0.8671	0.2874	0.7126	0.1187	0.8813	0.1266	0.8734	0.2134	0.7866	0.1640	0.8360	0.4705	0.5

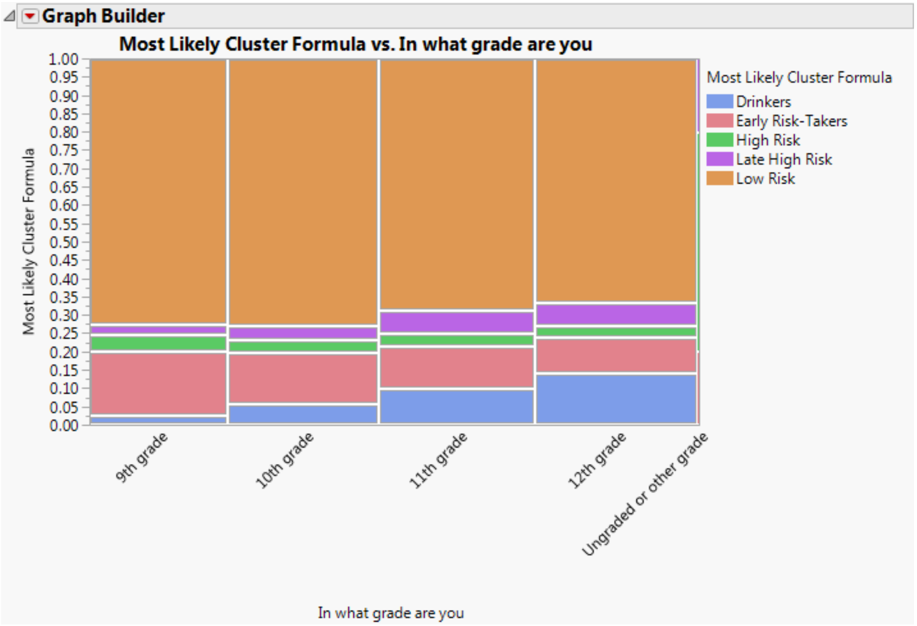
Cluster	Overall
Low Risk	0.67580
Early Risk-Takers	0.13847
Drinkers	0.09585
Late High Risk	0.05158
High Risk	0.03830

Figure 15.3 shows parameter estimates for the first eight variables in the analysis. The new cluster names appear in the report window.

Next, compare cluster membership to the demographic question "In what grade are you".

9. Click the red triangle next to Latent Class Model for 5 Clusters and select **Save Mixture and Cluster Formulas**.
10. Select **Graph > Graph Builder**.
11. Enter In what grade are you as **X**.
12. Enter Most Likely Cluster Formula as **Y**.
13. Select the Mosaic element.
14. Click **Done**.

Figure 15.4 Mosaic Plot of Cluster Membership versus Grade Level

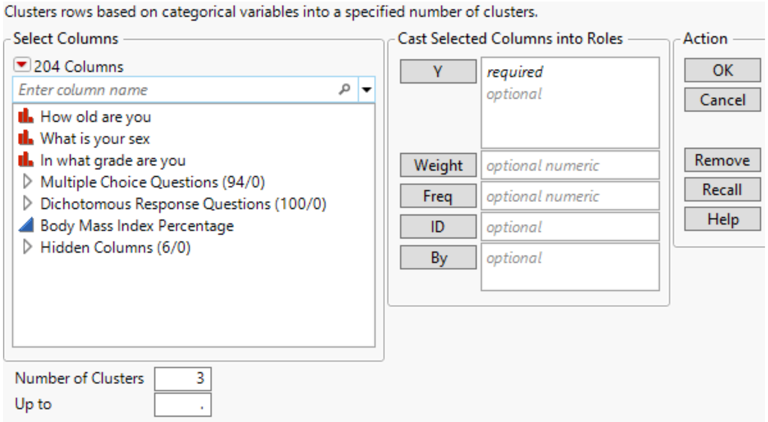


Observe that most of the respondents fall into the Low Risk cluster. The class labeled Drinkers includes more respondents as the grade level increases.

Launch the Latent Class Analysis Platform

Launch the Latent Class Analysis platform by selecting **Analyze > Clustering > Latent Class Analysis**.

Figure 15.5 Latent Class Analysis Launch Window



Clusters rows based on categorical variables into a specified number of clusters.

Number of Clusters 3

Up to .

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Latent Class Analysis platform launch window contains the following options:

Y The column or columns that you want to analyze. You can analyze columns with nominal, ordinal, or multiple response modeling types. To analyze nominal or ordinal responses, two or more columns are required. Only one column is required if it contains multiple responses and has a multiple response modeling type.

Weight A column whose numeric values assign a weight to each row in the analysis.

Freq A column whose numeric values assign a frequency to each row in the analysis.

ID A column used to identify separate respondents. This identification is used in some output tables.

By A column that creates a report consisting of separate analyses for each level of the variable. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

Number of Clusters The number of clusters to be computed in the analysis.

Up to Specifies a maximum number of clusters. If this number exceeds the value specified for Number of Clusters, a model report is produced with a number of clusters equal to each integer value in the range between Number of Clusters and Up to. These reports appear as part of the Latent Class Analysis report outline.

Caution: There is a maximum number of clusters that an LCA model can adequately fit. If you request more clusters than the maximum, a warning message appears in the report window. The LCA Platform fits up to the maximum number of clusters supported by the columns. For more information about determining the maximum number of clusters, see "Maximum Number of Clusters" on page 339.

The Latent Class Analysis Report

The initial Latent Class Analysis report contains a Cluster Comparison and Latent Class Model reports for each specified number of clusters.

Cluster Comparison Report

The Cluster Comparison report gives fit statistics to compare the various models. The fit statistics are the negative log-likelihood (-LogLikelihood), BIC, and AIC. Smaller values of each indicate better fit. The best fit is indicated in a column called Best. See *Fitting Linear Models*.

Latent Class Model Report

Each Latent Class Model Report is dynamically named Latent Class Model for <k> Clusters, depending on *k*, the number of clusters fit. The reports contains the following results and outlines:

- "Model Summary" on page 333
- "Parameter Estimates" on page 333
- "Transposed Parameter Estimates" on page 334
- "Effect Sizes" on page 334
- "MDS Plot" on page 334
- "Mixture Probabilities" on page 334

Model Summary

By default, a summary of the model for the specified number of clusters appears at the top of each Latent Class Model report. The model summary contains the -LogLikelihood, Number of Parameters, BIC, and AIC. These summary values can be used to determine how well the model fits the data. Lower values of -LogLikelihood, BIC, and AIC indicate better fits. See *Fitting Linear Models*. The Number of Parameters value gives the number of unique parameters in the latent class model. See "Statistical Details for the Latent Class Analysis Platform" on page 338.

Parameter Estimates

The Parameter Estimates report contains tabular and graphical summaries of the parameter estimates and is displayed by default. Each summary contains rows corresponding to the model clusters.

The Overall column shows the probability of an observation belonging to each cluster. (These are the γ parameters. See "Statistical Details for the Latent Class Analysis Platform" on page 338.)

The remaining columns in the displays are grouped with vertical dividers according to the Y columns specified in the Latent Class Analysis launch window:

- Each group of categorical response columns has a column for each level within the respective response. In each group, the value in a given row and column is the conditional probability of the response indicated by the column, given that the observation belongs to the cluster identified by the row. (These are the ρ parameters.)
- Each group of multiple response columns has a column for each category within the multiple response. In each group, the value in a given row and column is the conditional probability of a response at the lower level of the indicated category, given that the observation belongs to the cluster identified by the row. (These are the ρ parameters.)

The graphical display shows the conditional probability values as *share charts*. For each cluster and each Y, the conditional probabilities given cluster membership are plotted as a horizontal stacked bar chart. For a binary or nominal response column, the percentages in these charts sum to one for each response. For a multiple response column, the percentages are of the lower level of each of the categories and do not sum to one. The stacking of bars follows the order of appearance of the variables in the table of values. You can also hover over the bars to view the levels or categories of the variable.

Tip: You can select one or more rows in either table in the Parameter Estimates report to select the observations assigned to the corresponding clusters.

Transposed Parameter Estimates

The Transposed Parameter Estimates report contains a table that is the transpose of the Parameter Estimates report table. Here the clusters are shown as columns. The conditional probabilities for each cluster are shown for each response category of each Y column in the analysis.

Note: The estimates from the Overall column are not included in the transposed table.

Effect Sizes

The Effect Sizes table compares the Y columns across clusters and is displayed by default. The statistics in each row of this table are obtained from a contingency table analysis of expected counts for cluster membership by levels or categories of a Y column. The expected counts are obtained by multiplying the number of observations in each cluster by the conditional probabilities for each level or category of the Y column.

For each response, the Pearson chi-square statistic, χ^2, is calculated for the contingency table of expected counts for levels by clusters. Let n represent the number of observations. The value in the Effect Size column is defined as follows:

$$\text{Effect Size} = \sqrt{\frac{\chi^2}{n}}$$

Each value in the LR Logworth column shows -log10(p_{LR}) where p_{LR} is the likelihood ratio test p-value for the contingency table of expected counts. A Logworth value above 2 corresponds to significance at the 0.01 significance level.

Tip: You can select one or more rows in the Effect Sizes table to select the corresponding columns in the data table.

MDS Plot

The MDS Plot contains one point for each cluster and is displayed by default. It is a two-dimensional representation of cluster proximity. Clusters that are closer together are more similar. The plot is created from a dissimilarity matrix of the ρ parameters. For more information about MDS plots, see the “Multidimensional Scaling” chapter on page 235.

Mixture Probabilities

The Mixture Probabilities table displays probabilities of cluster membership for each row. The Most Likely Cluster column indicates the cluster with the highest probability of membership for each row.

Note: Rows that contain a missing value for one or more of the Y columns are excluded from the analysis and do not appear in the Mixture Probabilities table.

Latent Class Analysis Platform Options

- "Latent Class Analysis Options"
- "Latent Class Model Options"

Latent Class Analysis Options

The Latent Class Analysis red triangle menu contains the following options:

New Number of Clusters Enables you to run another analysis using a different number of clusters. The new analysis report is appended to the current report.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Latent Class Model Options

The Latent Class Model for <k> Clusters red triangle menu contains the following options:

Model Reports Enables you to show or hide the available model reports. For more information about the model reports, see "The Latent Class Analysis Report" on page 332.

Color by Cluster Colors each row in the data table according to its most likely cluster. For an example, see "Additional Example of the Latent Class Analysis Platform" on page 336.

Save Mixture and Cluster Formulas Saves a formula column to the data table for each cluster as well as a formula column for the most likely cluster.

Save Cluster Formula Only Saves a column to the data table with a formula that determines the most likely cluster.

Publish Probability Formulas Creates probability formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See *Predictive and Specialized Modeling*.

Save Mixture Probabilities Saves the values in the Mixture Probabilities table to the corresponding rows in the data table.

Save Cluster Only Saves a new column to the data table that contains the most likely cluster for each row. This column does not contain a formula.

Rename Clusters Enables you to give meaningful names to the clusters in the report.

Note: The new cluster names are not saved to a script unless you have specified a random seed for the report. Setting a random seed is available only when you launch the report via a script.

Remove Fit Removes the specified clustering report from the report window.

Additional Example of the Latent Class Analysis Platform

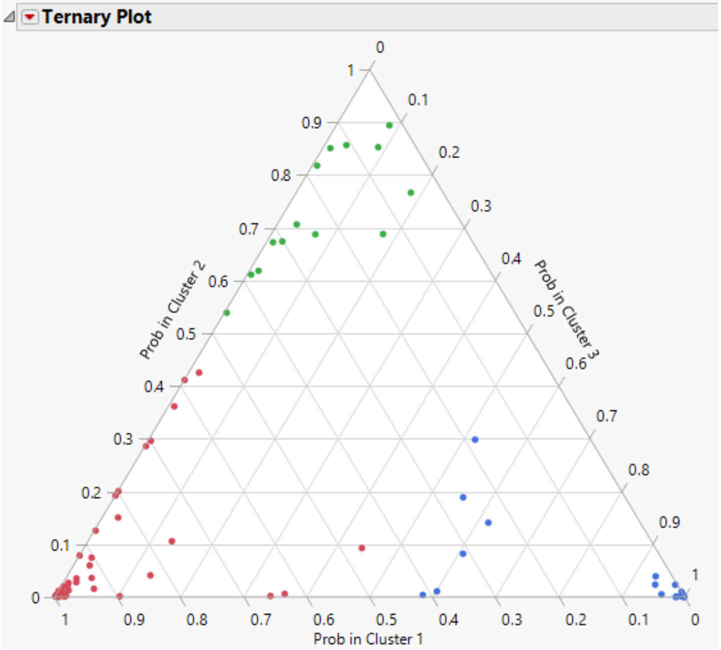
Plot Probabilities of Cluster Membership

This example uses the Car Poll.jmp sample data table, which contains survey data for car owners and car makes. You are interested in classifying the car owners into three clusters and producing a plot to visualize the probabilities of cluster membership. A ternary plot provides a good visualization when you have three clusters.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Clustering > Latent Class Analysis**.
3. Select all of the columns except age and click **Y**.
4. Click **OK**.
5. Click the red triangle next to Latent Class Model for 3 Clusters and select **Color by Cluster**.
6. Click the red triangle next to Latent Class Model for 3 Clusters and select **Save Mixture Probabilities**.

7. In the Car Poll data table window, select the LCA Cluster Probabilities column group from the column list.
8. Select **Graph > Ternary Plot**.
9. Click **X, Plotting**.
10. Click **OK**.

Figure 15.6 Ternary Plot of Cluster Membership Probabilities



In the ternary plot of cluster probabilities for each observation, most of the cluster membership probabilities fall near the vertices. This indicates that they have high values for one cluster and lower values for the other two. However, there are some points in the middle of the plot, indicating that these observations do not have high probabilities of cluster membership for any of the clusters. These observations might warrant closer inspection or they might indicate that more clusters are needed to better represent the data.

Note: Your results might be different because a random seed was not specified.

Statistical Details for the Latent Class Analysis Platform

- "Latent Class Model Fit"
- "Maximum Number of Clusters"

Latent Class Model Fit

This section describes the latent class model that is fit in the Latent Class Analysis platform. For more information about latent class models, see Collins and Lanza (2010) and Agresti (2013).

Note: The LCA algorithm that is used in the Text Explorer platform takes advantage of the sparsity of the document term matrix. For this reason, the LCA results in the Text Explorer platform do not exactly match the results in the Latent Class Analysis platform.

Let $j = 1, \ldots, J$ represent the observed columns of responses. These are the Y columns in the Latent Class Analysis platform launch window. Denote the number of levels for column j by R_j.

A multidimensional contingency table of the J variables contains $W = R_1 * \ldots * R_J$ cells. Each of these cells is defined by its response pattern for the J variables. Therefore, each response pattern is a J-length vector of the form $\mathbf{y} = y_1, \ldots, y_j$. Define $\mathbf{Y}$ to be the W by J array of all the response patterns considered as row vectors. Each element, $\mathbf{y}_w$, in $\mathbf{Y}$ has a probability $\Pr(\mathbf{y}_w)$. These probabilities sum to 1:

$$\sum_{w=1}^{W} Pr(y_w) = 1$$

Consider the following notation:

- C is the number of clusters in the latent class model.
- γ_c is the probability of membership in cluster c. (The γ_c are the latent class prevalences.) These parameters sum to 1.
- $r_{j,k}$ is the k^{th} level of the j^{th} response.
- $\rho_{j,k|c}$ is the probability of observing response $r_{j,k}$ in column j conditional on membership in class c. (The $\rho_{j,k|c}$ are the item-response probabilities.) For a given cluster and response variable j, the sum of the $\rho_{j,k|c}$ is 1.
- $I(y_j = r_{j,k})$ is an indicator function that equals 1 when the y_j response is the k^{th} level of the j^{th} response, and 0 otherwise.

The probability of observing a specific vector of responses $\mathbf{y}_w = y_1, \ldots, y_j$ is the sum of the conditional probabilities of observing that vector of responses for each of the C latent classes:

$$\Pr(y) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{k=1}^{R_j} \rho_{j,k|c}^{I(y_j = r_{j,k})}$$

This equation is the denominator in the Prob Formula Cluster formulas that you can save to the data table by selecting the Save Mixture and Cluster Formulas option from the Latent Class Analysis red triangle menu. The formula in the Prob Formula Cluster column gives $\Pr(Cluster = c \mid \mathbf{y}_w)$, which equals $\Pr(\mathbf{y}_w, Cluster = c) / \Pr(\mathbf{y}_w)$.

The γ and ρ parameters for latent class models are estimated using the iterative Expectation-Maximization (EM) algorithm. The number of unique parameters in a latent class model is defined as follows:

$$(C-1) + C\sum_{j=1}^{J}(R_j - 1)$$

Maximum Number of Clusters

The maximum number of clusters that can be fit in an LCA model depends on the model degrees of freedom. The degrees of freedom in an LCA model are based on the size of the contingency table created by the columns. The size of the contingency table is the number of cells in the table that contain at least one observation and is denoted as K. If all cells contain at least one observation, K is the product of the number of levels of the response columns. The formula for degrees of freedom is defined as follows:

$\text{DF} = K - \{\text{nCluster} - 1 + \text{nCluster}(\text{nTotalLevels} - \text{nCols})\} - 1$

where

nCluster = the number of clusters

nTotalLevels = the sum of the levels of the response columns

nCols = the number of response columns

In order for the LCA model to be adequately fit, the degrees of freedom must be positive. Therefore, to ensure DF > 0, the maximum number of clusters is defined as follows:

$\max(\text{nCluster}) < \text{floor}[K/(1 + \text{nTotal Levels} - \text{nCols})]$

Chapter 16

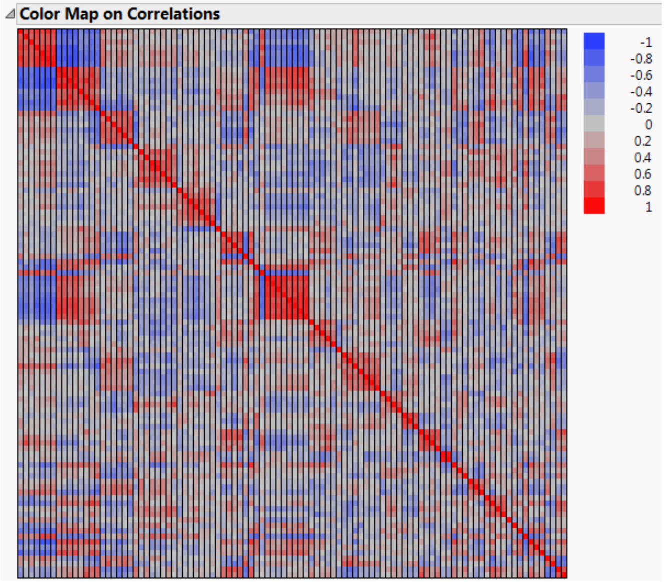
Cluster Variables

Group Similar Variables into Representative Groups

Variable clustering provides a method for grouping similar variables into representative groups. Each cluster can be represented by a single component or variable. The component is a linear combination of all variables in the cluster. Alternatively, the cluster can be represented by the variable identified to be the most representative member in the cluster.

You can use Cluster Variables as a dimension-reduction method. Instead of using a large set of variables in modeling, either the cluster components or the most representative variable in the cluster can be used to explain most of the variation in the data. In addition, dimension reduction using Cluster Variables is often more interpretable than dimension reduction using principal components.

Figure 16.1 Example of Correlation Map for Variables



Contents

Overview of the Cluster Variables Platform

Principal components analysis constructs components that are linear combinations of all the variables in the analysis. In contrast, the Cluster Variables option constructs components that are linear combinations of variables in a cluster of similar variables. The entire set of variables is partitioned into clusters. For each cluster, a *cluster component* is constructed using the first principal component of the variables in that cluster. This cluster component is the linear combination that explains as much of the variation as possible among the variables in that cluster.

You can use the Cluster Variables option as a dimension-reduction method. A substantial part of the variation in a large set of variables can often be represented by cluster components or by the most representative variable in the cluster. These new variables can then be used in predictive or other modeling techniques. The new cluster-based variables are usually more interpretable than principal components based on all the variables.

Principal components constructed from a common set of variables are orthogonal. However, cluster components are not orthogonal because they are constructed from distinct sets of variables.

When you have a large set of variables, the Cluster Variables platform uses an algorithm based on the singular value decomposition to shorten computation time. For additional background, see “Wide Linear Methods and the Singular Value Decomposition” on page 357 in the “Statistical Details” appendix.

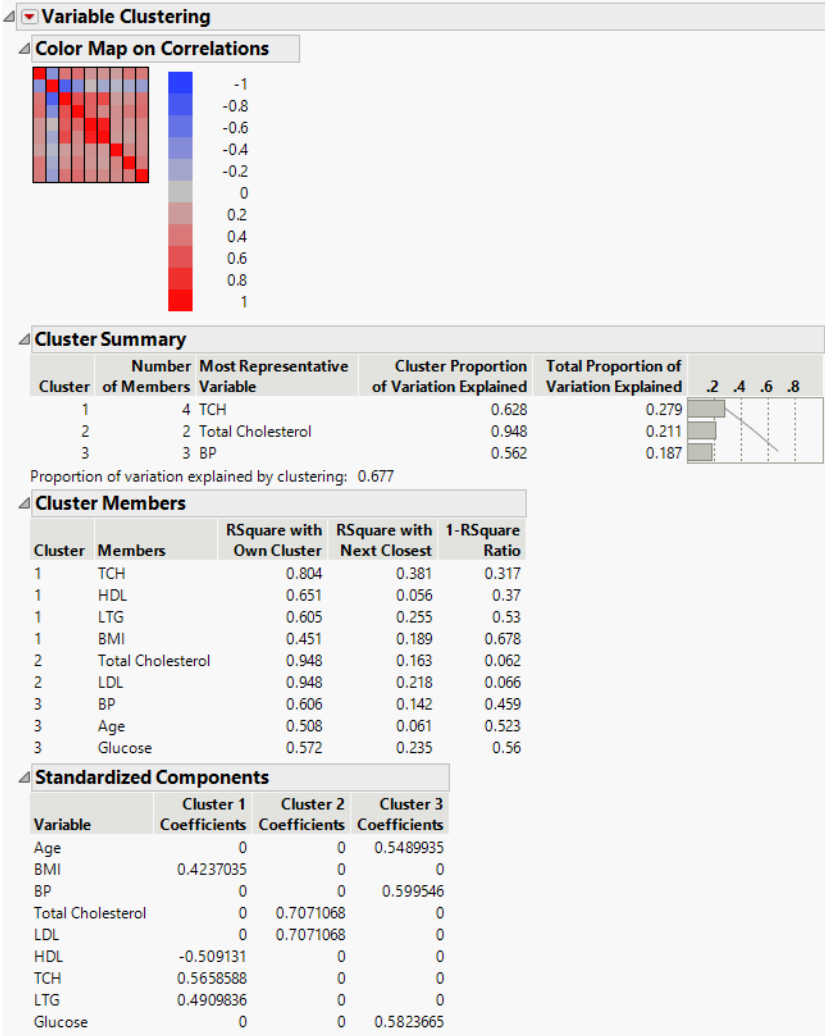
Example of the Cluster Variables Platform

The Diabetes.jmp sample data table contains ten baseline variables used in modeling disease progression. In this example, you cluster the continuous baseline variables.

1. Select **Help > Sample Data Library** and open Diabetes.jmp.
2. Select **Analyze > Clustering > Cluster Variables.**
3. Select the columns Age through Glucose except for Gender (Age, BMI, BP, Total Cholesterol, LDL, HDL, TCH, LTG, and Glucose) and click **Y, Columns**.

   The Gender column cannot be included because Cluster Variables requires numeric continuous variables.
4. Click **OK**.

Figure 16.2 Cluster Variables Report for Diabetes Data



Cluster Summary

Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained
1	4	TCH	0.628	0.279
2	2	Total Cholesterol	0.948	0.211
3	3	BP	0.562	0.187

Proportion of variation explained by clustering: 0.677

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	TCH	0.804	0.381	0.317
1	HDL	0.651	0.056	0.37
1	LTG	0.605	0.255	0.53
1	BMI	0.451	0.189	0.678
2	Total Cholesterol	0.948	0.163	0.062
2	LDL	0.948	0.218	0.066
3	BP	0.606	0.142	0.459
3	Age	0.508	0.061	0.523
3	Glucose	0.572	0.235	0.56

Standardized Components

Variable	Cluster 1 Coefficients	Cluster 2 Coefficients	Cluster 3 Coefficients
Age	0	0	0.5489935
BMI	0.4237035	0	0
BP	0	0	0.599546
Total Cholesterol	0	0.7071068	0
LDL	0	0.7071068	0
HDL	-0.509131	0	0
TCH	0.5658588	0	0
LTG	0.4909836	0	0
Glucose	0	0	0.5823665

The Cluster Summary report shows that the variables were grouped into three clusters:

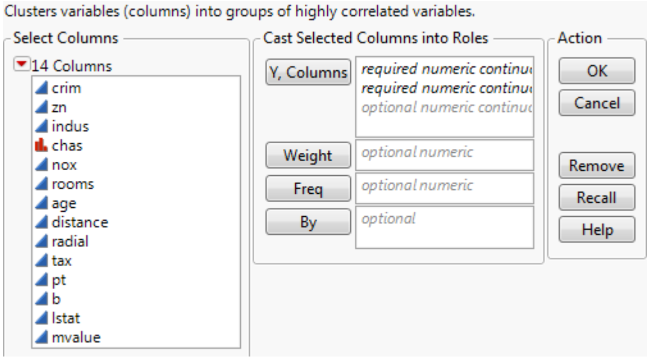
- Cluster 1 consists of TCH, HDL, LTG, and BMI, as shown in the Cluster Members report. The Cluster Summary report shows that TCH is the most representative variable for Cluster 1 and that for the variables in Cluster 1, 62.8% of the variation is explained by the first principal component.
- Cluster 2 consists of Total Cholesterol and LDL. The Cluster summary report shows that Total Cholesterol is the most representative variable for Cluster 2 and that for the variables in Cluster 2, 94.8% of the variation is explained by the first principal component.

- Cluster 3 consists of BP, Age, and Glucose. The Cluster Summary report shows that the most representative variable is BP and that for the variables in Cluster 3, 56.2% of the variation is explained by the first principal component.

Launch the Cluster Variables Platform

Launch the Cluster Variables platform by selecting **Analyze > Clustering > Cluster Variables**.

Figure 16.3 Cluster Variables Launch Dialog



Clusters variables (columns) into groups of highly correlated variables.

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns The variables to be clustered. Variables must be numeric and continuous.

Weight A column whose numeric values assign a weight to each row in the analysis.

Freq A column whose numeric values assign a frequency to each row in the analysis.

By A column whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed. The results are presented in separate reports. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

The Cluster Variables Report

By default, the Cluster Variables report displays the following:

- "Color Map on Correlations" on page 346
- "Cluster Summary" on page 346

- "Cluster Members" on page 347
- "Standardized Components" on page 347

Tip: In any of the Cluster Variables reports, select rows in order to select the corresponding columns in the data table. Press Ctrl and click the row to deselect the column in the data table.

Color Map on Correlations

The Color Map on Correlations report displays a color map of the correlations between variables. The variables are arranged in the order in which they are listed in the Cluster Members report. This arrangement ensures that members of the same cluster are adjacent in the correlation plot. See "Example of Color Map on Correlations" on page 349.

Tip: Hover over a square on the color map to see the variables involved in that square and their correlation.

Variables in the same cluster tend to have higher absolute correlations (deeper red or blue colors) than variables in different clusters. Therefore, the squares formed by the cells of the correlation map that correspond to the variables for a given component often stand out along the diagonal.

Correlations are computed using the row-wise method. This method excludes any observation with missing data on any of the variables from the correlation calculation. For more information about the row-wise estimation method, see "Estimation Methods" on page 47 in the "Correlations and Multivariate Techniques" chapter.

Cluster Summary

The Cluster Summary report gives the following information:

Cluster The cluster identifier.

Number of Members The number of variables in the cluster.

Most Representative Variable The cluster variable that has the largest squared correlation with its cluster component.

Cluster Proportion of Variance Explained The cluster's proportion of variance explained by the first principal component among the variables in the cluster. If there is only one variable in the cluster, then this is 1. This statistic is based only on variables within the cluster rather than on all variables.

Total Proportion of Variation Explained The overall proportion of variance explained by the cluster component. This is equivalent to using only the variables within each cluster to calculate the first principal component.

A note beneath the table gives the total proportion of variation explained by all the cluster components.

Cluster Members

The Cluster Members report gives the following:

Cluster The cluster identifier.

Members The variables included in the cluster.

RSquare with Own Cluster The squared correlation of the variable with its cluster component.

RSquare with Next Closest The squared correlation of the variable with the cluster component for its next closest cluster. The next closest cluster is the cluster for which the squared correlation of the variable with the cluster component is the second highest.

1 - RSquare Ratio A measure of the relative closeness between the cluster to which a variable belongs and its next closest cluster. It is defined as follows:

(1 - RSquare with Own Cluster)/(1 - RSquare with Next Closest)

Standardized Components

The Standardized Components report gives the coefficients that define the cluster components. These coefficients are the eigenvectors of the first principal component within each cluster.

Cluster Variables Platform Options

The Variable Clustering red triangle menu contains the following options:

Color Map on Correlations Shows or hides the Color Map on Correlations plot. See “Color Map on Correlations” on page 346.

Cluster Summary Shows or hides the Cluster Summary report. See “Cluster Summary” on page 346.

Cluster Members Shows or hides the Cluster Members report. See "Cluster Members" on page 347.

Cluster Components Shows or hides the Standardized Components report. See "Standardized Components" on page 347.

Save Cluster Components Saves columns to the data table as a group called Cluster Components. Each column is named Cluster <i> Components and contains a formula that expresses the cluster component in terms of the uncentered and unscaled variables.

Launch Fit Model Opens a Model Specification window with the Most Representative Variables for each cluster entered in the Construct Model Effects list. Use this option to construct models based on the Most Representative variables.

Tip: To fit a model using the components, first select the **Save Cluster Components** option. Then replace the Most Representative variables for each cluster in the Construct Model Effects list with the desired Cluster Components columns.

See *Using JMP* for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Additional Examples of the Cluster Variables Platform

- "Example of Color Map on Correlations"
- "Example of Cluster Variables Platform for Dimension Reduction"

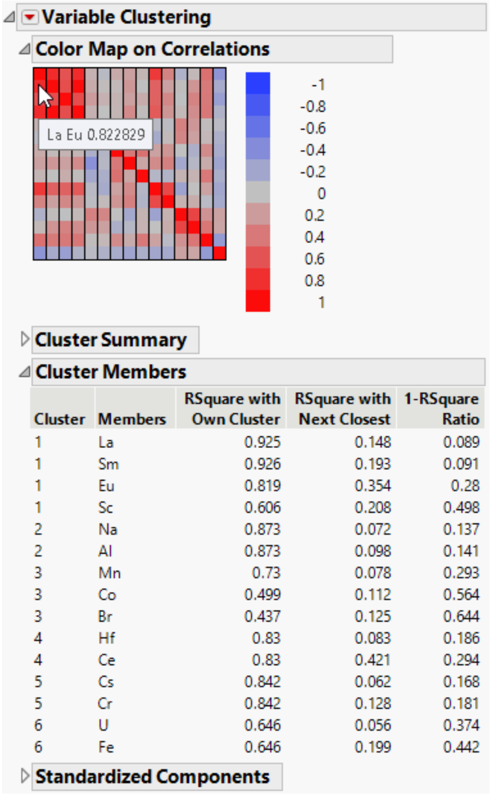
Example of Color Map on Correlations

In this example, you construct and examine a Color Map on Correlations.

1. Select **Help > Sample Data Library** and open Cherts.jmp.
2. Select **Analyze > Clustering > Cluster Variables.**
3. Select all continuous variables and click **Y, Columns**.
4. Click **OK**.
5. Close the Cluster Summary and the Standardized Components reports.
6. Hover over the cell in the second row and first column of the Color Map.

   A tooltip appears, showing that the variables corresponding to this cell are La and Eu, and that their correlation is 0.822829.

Figure 16.4 Color Map on Correlations for Cherts.jmp



Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	La	0.925	0.148	0.089
1	Sm	0.926	0.193	0.091
1	Eu	0.819	0.354	0.28
1	Sc	0.606	0.208	0.498
2	Na	0.873	0.072	0.137
2	Al	0.873	0.098	0.141
3	Mn	0.73	0.078	0.293
3	Co	0.499	0.112	0.564
3	Br	0.437	0.125	0.644
4	Hf	0.83	0.083	0.186
4	Ce	0.83	0.421	0.294
5	Cs	0.842	0.062	0.168
5	Cr	0.842	0.128	0.181
6	U	0.646	0.056	0.374
6	Fe	0.646	0.199	0.442

The Cluster Members report shows that there are four variables in Cluster 1. In the Color Map on Correlations, the four-by-four square of cells in the upper left corner that corresponds to these five variables shows a distinct pattern of positive correlations. The color map also shows patterns of positive correlations for the variables in Clusters 2, 4, and 5. The two-by-two square of cells in the lower right corner of the color map that corresponds to the two Cluster 6 variables shows that they are negatively correlated. See "Color Map on Correlations" on page 346.

Example of Cluster Variables Platform for Dimension Reduction

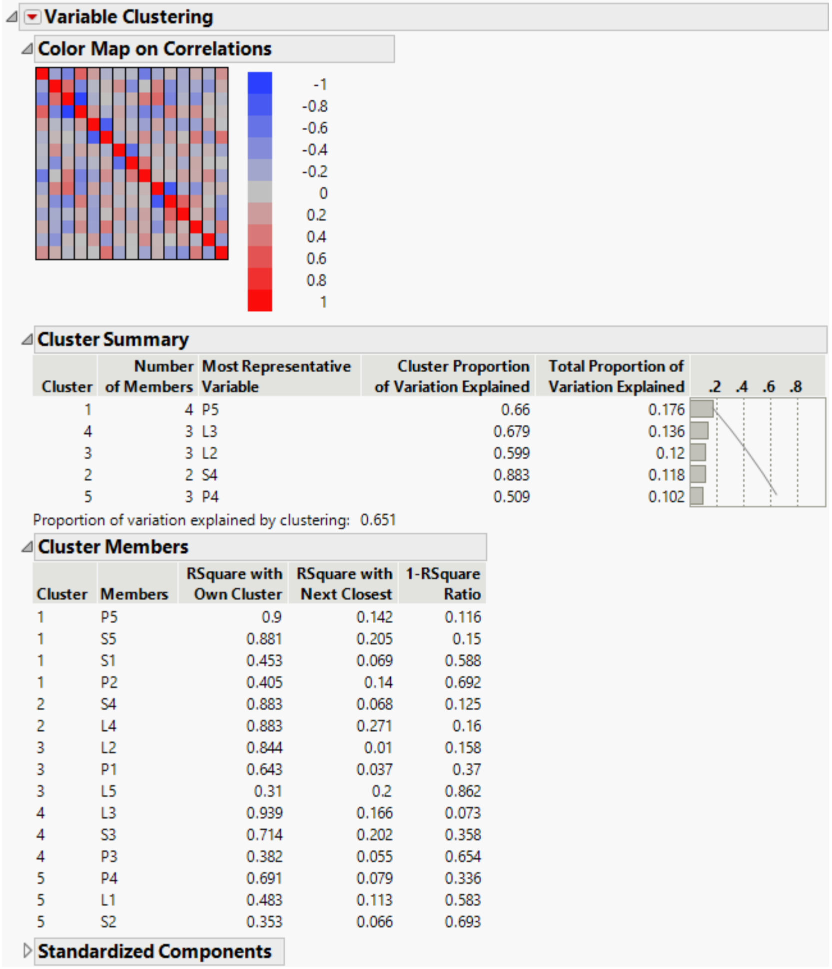
In this example, you use the Cluster Variables platform as a dimension-reduction tool for modeling. The Penta.jmp sample data table contains 15 variables used to predict the response variable, log RAI. Use Cluster Variables to reduce this number.

Cluster Variables

1. Select **Help > Sample Data Library** and open Penta.jmp.
2. Select **Analyze > Clustering > Cluster Variables**.
3. Select all of the continuous variables, except logRAI and click **Y, Columns**.
4. Click **OK**.
5. Click the Variable Clustering red triangle and select **Save Cluster Components**.

   Five grouped formula columns are added to the data table.

Figure 16.5 Cluster Variables Report for Penta.jmp



Variable Clustering

Color Map on Correlations

Cluster Summary

Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained
1	4	P5	0.66	0.176
4	3	L3	0.679	0.136
3	3	L2	0.599	0.12
2	2	S4	0.883	0.118
5	3	P4	0.509	0.102

Proportion of variation explained by clustering: 0.651

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	P5	0.9	0.142	0.116
1	S5	0.881	0.205	0.15
1	S1	0.453	0.069	0.588
1	P2	0.405	0.14	0.692
2	S4	0.883	0.068	0.125
2	L4	0.883	0.271	0.16
3	L2	0.844	0.01	0.158
3	P1	0.643	0.037	0.37
3	L5	0.31	0.2	0.862
4	L3	0.939	0.166	0.073
4	S3	0.714	0.202	0.358
4	P3	0.382	0.055	0.654
5	P4	0.691	0.079	0.336
5	L1	0.483	0.113	0.583
5	S2	0.353	0.066	0.693

Standardized Components

The Cluster Summary and Cluster Members reports show that the variables are clustered into five groups, so there are five Cluster Component variables.

Fit Models

Next, fit and compare two models to predict logRAI:

- A model using all continuous variables as predictors.
- A model using the Cluster Components as predictors.

1. Click the Variable Clustering red triangle and select **Launch Fit Model**.
2. Select logRAI and click **Y**.

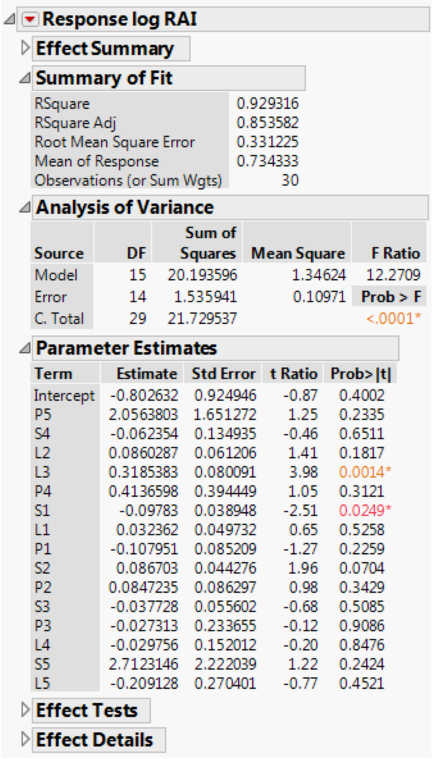
Notice that the Most Representative Variables the five clusters have been entered in the Construct Model Effects list. However, you want to enter all predictors.

3. Select all of the continuous variables from S1 to P5 and click **Add**.

   Be careful not to include Obs Name.

4. Select the box next to **Keep dialog open**.
5. Click **Run**.

Figure 16.6 Fit Least Squares Report for Model with All Continuous Predictors



Summary of Fit

RSquare	0.929316
RSquare Adj	0.853582
Root Mean Square Error	0.331225
Mean of Response	0.734333
Observations (or Sum Wgts)	30

Analysis of Variance

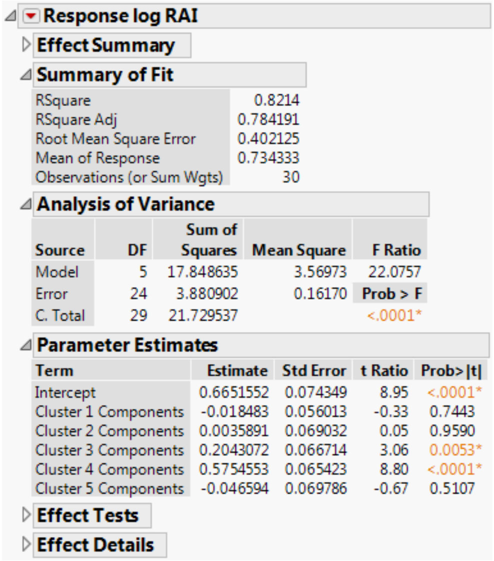
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	15	20.193596	1.34624	12.2709
Error	14	1.535941	0.10971	Prob > F
C. Total	29	21.729537		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.802632	0.924946	-0.87	0.4002
P5	2.0563803	1.651272	1.25	0.2335
S4	-0.062354	0.134935	-0.46	0.6511
L2	0.0860287	0.061206	1.41	0.1817
L3	0.3185383	0.080091	3.98	0.0014*
P4	0.4136598	0.394449	1.05	0.3121
S1	-0.09783	0.038948	-2.51	0.0249*
L1	0.032362	0.049732	0.65	0.5258
P1	-0.107951	0.085209	-1.27	0.2259
S2	0.086703	0.044276	1.96	0.0704
P2	0.0847235	0.086297	0.98	0.3429
S3	-0.037728	0.055602	-0.68	0.5085
P3	-0.027313	0.233655	-0.12	0.9086
L4	-0.029756	0.152012	-0.20	0.8476
S5	2.7123146	2.222039	1.22	0.2424
L5	-0.209128	0.270401	-0.77	0.4521

6. In the Fit Model window, select all variables in the Construct Model Effects window and click **Remove**.
7. Select the Cluster Components group and click **Add**.
8. Click **Run**.

Figure 16.7 Fit Least Squares Report for Model with Cluster Components as Predictors



Response log RAI

Effect Summary

Summary of Fit

RSquare	0.8214
RSquare Adj	0.784191
Root Mean Square Error	0.402125
Mean of Response	0.734333
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	17.848635	3.56973	22.0757
Error	24	3.880902	0.16170	Prob > F
C. Total	29	21.729537		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.6651552	0.074349	8.95	<.0001*
Cluster 1 Components	-0.018483	0.056013	-0.33	0.7443
Cluster 2 Components	0.0035891	0.069032	0.05	0.9590
Cluster 3 Components	0.2043072	0.066714	3.06	0.0053*
Cluster 4 Components	0.5754553	0.065423	8.80	<.0001*
Cluster 5 Components	-0.046594	0.069786	-0.67	0.5107

Effect Tests

Effect Details

The model that includes the five Cluster Components as the only predictors explains a substantial amount of the variation in the response, with an adjusted RSquare of 0.784. The model that uses all fifteen predictors has only a slightly higher adjusted RSquare of 0.853 (Figure 16.6).

Statistical Details for the Cluster Variables Platform

Variable Clustering Algorithm

The clustering algorithm iteratively splits clusters of variables and reassigns variables to clusters until no more splits are possible. The initial cluster consists of all variables. The algorithm was developed by SAS and is implemented in PROC VARCLUS (SAS Institute Inc. 2020g).

Note: The algorithm uses only observations for which there are no missing values for any variable in the Y, Columns list.

These are the iterative steps in the algorithm:

1. For all clusters, do the following:
   a. Compute the principal components for the variables in each cluster.

   b. If the second eigenvalues for all of the clusters are less than one, then terminate the algorithm.

2. Partition the cluster whose second eigenvalue is the largest (and greater than 1) into two new clusters using the following steps:
   a. Rotate the principal components for the variables in the current cluster using an orthoblique rotation.
   b. Define one cluster to consist of the variables in the current cluster whose squared correlations to the first rotated principal component are higher than their squared correlations to the second principal component.
   c. Define the other cluster to consist of the remaining variables in the original cluster. These are the variables that are more highly correlated with the second principal component.
   d. Compute the principal components of the two new clusters.
3. Test to see whether any variable in the data set should be assigned to a different cluster. For each variable, do the following:
   a. Compute the variable's squared correlation with the first principal component for each cluster.
   b. Place the variable in the cluster for which its squared correlation is the largest.

Note: An orthoblique rotation is also known as a raw quartimax rotation. See Harris and Kaiser (1964).

Appendix A

Statistical Details

Multivariate Methods

This appendix discusses Wide Linear methods and the use of the singular value decomposition.

Contents

Wide Linear Methods and the Singular Value Decomposition

Wide Linear methods in the Cluster, Principal Components, and Discriminant platforms enable you to analyze data sets with thousands (or even millions) of variables. Most multivariate techniques require the calculation or inversion of a covariance matrix. When your multivariate analysis involves a large number of variables, the covariance matrix can be prohibitively large so that calculating it or inverting it is problematic and computationally expensive.

Suppose that your data consist of n rows and p columns. The rank of the covariance matrix is at most the smaller of n and p. In wide data sets, p is often much larger than n. In these cases, the inverse of the covariance matrix has at most n nonzero eigenvalues. Wide Linear methods use this fact, together with the singular value decomposition, to provide efficient calculations. See "Calculating the SVD" on page 359.

The Singular Value Decomposition

The *singular value decomposition* (SVD) enables you to express any linear transformation as a rotation, followed by a scaling, followed by another rotation. The SVD states that any n by p matrix **X** can be written as follows:

$$\mathbf{X} = \mathbf{U}Diag(\Lambda)\mathbf{V}'$$

Let r be the rank of **X**. Denote the r by r identity matrix by $\mathbf{I}_r$.

The matrices **U**, $Diag(\Lambda)$, and **V** have the following properties:

U is an n by r semi-orthogonal matrix with $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$

V is a p by r semi-orthogonal matrix with $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$

$Diag(\Lambda)$ is an r by r diagonal matrix with positive diagonal elements given by the column vector $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_r)'$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$.

The λ_i are the nonzero *singular values* of **X.**

The following statements relate the SVD to the spectral decomposition of a square matrix:

- The squares of the λ_i are the nonzero eigenvalues of $\mathbf{X}'\mathbf{X}$.
- The r columns of **V** are eigenvectors of $\mathbf{X}'\mathbf{X}$.

Note: There are various conventions in the literature regarding the dimensions of the matrices **U**, **V**, and the matrix containing the singular values. However, the differences have no practical impact on the decomposition up to the rank of **X**.

For more information about singular value decomposition, see Press et al. (1998, Section 2.6).

The SVD and the Covariance Matrix

This section describes how the eigenvectors and eigenvalues of a covariance matrix can be obtained using the SVD. When the matrix of interest has at least one large dimension, calculating the SVD is much more efficient than calculating its covariance matrix and its eigenvalue decomposition.

Let n be the number of observations and p the number of variables involved in the multivariate analysis of interest. Denote the n by p matrix of data values by $\mathbf{X}$.

The SVD is usually applied to standardized data. To standardize a value, subtract its mean and divide by its standard deviation. Denote the n by p matrix of standardized data values by $\mathbf{X}_s$. Then the covariance matrix of the standardized data is the correlation matrix for $\mathbf{X}$ and is defined as follows:

$$Cov = \mathbf{X}_s'\mathbf{X}_s/(n-1)$$

The SVD can be applied to $\mathbf{X}_s$ to obtain the eigenvectors and eigenvalues of $\mathbf{X}_s'\mathbf{X}_s$. This allows efficient calculation of eigenvectors and eigenvalues when the matrix $\mathbf{X}$ is either extremely wide (many columns) or tall (many rows). This technique is the basis for Wide PCA. See "Principal Components Report" on page 60 in the "Principal Components" chapter.

The SVD and the Inverse Covariance Matrix

Some multivariate techniques require the calculation of inverse covariance matrices. This section describes how the SVD can be used to calculate the inverse of a covariance matrix.

Denote the standardized data matrix by $\mathbf{X}_s$ and define $\mathbf{S} = \mathbf{X}_s'\mathbf{X}_s$. The singular value decomposition allows you to write $\mathbf{S}$ as follows:

$$\mathbf{S} = (\mathbf{U}Diag(\Lambda)\mathbf{V}')'(\mathbf{U}Diag(\Lambda)\mathbf{V}') = \mathbf{V}Diag(\Lambda)^2\mathbf{V}'$$

If $\mathbf{S}$ is of full rank, then $\mathbf{V}$ is a p by p orthonormal matrix, and you can write $\mathbf{S}^{-1}$ as follows:

$$\mathbf{S}^{-1} = (\mathbf{V}Diag(\Lambda)^2\mathbf{V}')^{-1} = \mathbf{V}Diag(\Lambda)^{-2}\mathbf{V}'$$

If $\mathbf{S}$ is not of full rank, then $Diag(\Lambda)^{-1}$ can be replaced with a generalized inverse, $Diag(\Lambda)^{+}$, where the diagonal elements of $Diag(\Lambda)$ are replaced by their reciprocals. This defines a generalize inverse of $\mathbf{S}$ as follows:

$$\mathbf{S}^{-} = \mathbf{V}(Diag(\Lambda)^{+})^2\mathbf{V}'$$

This generalized inverse can be calculated using only the SVD.

For more information about the application of the SVD for wide linear discriminant analysis, see "Wide Linear Discriminant Method" on page 115 in the "Discriminant Analysis" chapter.

Calculating the SVD

In the Multivariate Methods platforms, JMP calculates the SVD of a matrix following the method suggested in Golub and Kahan (1965). Golub and Kahan's method involves a two-step procedure. The first step consists of reducing the matrix **M** to a bidiagonal matrix **J**. The second step consists of computing the singular values of **J**, which are the same as the singular values of the original matrix **M**. The columns of the matrix **M** are usually standardized in order to equalize the effect of the variables on the calculation. The Golub and Kahan method is computationally efficient.

Appendix B

References

Agresti, A. (2013). *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Baglama, J., and Reichel, L. (2005). "Augmented implicitly restarted Lanczos bidiagonalization methods." *SIAM Journal on Scientific Computing* 27:19–42.

Ballard, D. H. (1981). "Generalizing the Hough Transform to Detect Arbitrary Shapes." *Pattern Recognition* 13:111–122.

Bartlett, M. S. (1937). "Properties of sufficiency and statistical tests." *Proceedings of the Royal Society of London, Series A* 160:268–282.

Bartlett, M. S. (1954). "A Note on the Multiplying Factors for Various Chi Square Approximations." *Journal of the Royal Statistical Society, Series B* 16:296–298.

Benzécri, J. P. (1979). "Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.]." *Cahiers de l'Analyse des Données* 4:377–378.

Bentler, P. M. (1990). "Comparative Fit Indexes in Structural Models." *Psychological Bulletin* 107:238.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications.* 2nd ed. New York: Springer.

Boulesteix, A.-L., and Strimmer, K. (2007). "Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data." *Briefings in Bioinformatics* 8:32–44.

Browne, M. (2001). "An Overview of Analytic Rotation in Exploratory Factor Analysis." *Multivariate Behavioral Research* 36:111 -150.

Browne, M. W., and Cudeck, R. (1993). "Alternative Ways of Assessing Model Fit." In *Testing Structural Equation Models*, edited by K. A. Bollen, and J. S. Long, 136–162. Newbury Park, CA: Sage Publications.

Chen, F. F. (2007). "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 14:464–504.

Collins, L., and Lanza, S. (2010). *Latent Class and Latent Transition Analysis*. Hoboken NJ: John Wiley & Sons.

Cox, I., and Gaudard, M. (2013). *Discovering Partial Least Squares with JMP*. Cary, NC: SAS Institute Inc.

Cronbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16:297–334.

Cudeck, R., and MacCallum, R. C., eds. (2007). Factor Analysis at 100, Historical Developments and Future Directions. Mahwah, NJ: Lawrence Erlbaum Associates.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

De Jong, S. (1993). “SIMPLS: An Alternative Approach to Partial Least Squares Regression.” *Chemometrics and Intelligent Laboratory Systems* 18:251–263.

Denham, M. C. (1997). “Prediction Intervals in Partial Least Squares.” *Journal of Chemometrics* 11:39–52.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., and Wold, S. (2006). *Multi- and Megavariate Data Analysis Basic Principles and Applications (Part I).* Chapter 4. Umetrics.

Finkbeiner, C. (1979). “Estimation for the Multiple Factor Model when Data are Missing.” *Psychometrika* 44:409–420.

Fisher, L., and Van Ness, J. W. (1971). “Admissible Clustering Procedures.” *Biometrika* 58:91–104.

Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951a). “Sur la liaison et la division des points d’un ensemble fini.” *Colloquium Mathematicae* 2:282–285.

Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951b). “Taksonomia Wroclawska.” *Przeglad Antropologiczny* 17:193–211.

Frank, I. E., and Todeschini, T. (1994). *The Data Analysis Handbook*. New York: Elsevier.

Friedman, J. H. (1989). “Regularized Discriminant Analysis.” *Journal of the American Statistical Association* 84:165–175.

Garthwaite, P. (1994). “An Interpretation of Partial Least Squares.” *Journal of the American Statistical Association* 89:122–127.

Golub, G. H., and Kahan, W. (1965). “Calculating the singular values and pseudo-inverse of a matrix.” *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 2:205–224.

Goodman, L. A. (1974). “Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models.” *Biometrika* 61:215–231.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.

Harris, C. W., and Kaiser, H. F. (1964). “Oblique Factor Analytic Solutions by Orthogonal Transformation.” *Psychometrika* 32:363–379.

Hartigan, J. A. (1981). “Consistency of Single Linkage for High–Density Clusters.” *Journal of the American Statistical Association* 76:388–394.

Hastie, T., Tibshirani, R., and Friedman, J. H.(2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Verlag.

Hoskuldsson, A. (1988). “PLS Regression Methods.” *Journal of Chemometrics* 2:211–228.

Hoeffding, W. (1948). “A Non-Parametric Test of Independence.” *Annals of Mathematical Statistics* 19:546–557.

Hu, L.-T., and Bentler, P. M. (1999). "Cutoff Criteria for Fit Indices in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling* 6:1–55.

Huber, P. J. (1964). "Robust Estimation of a Location Parameter." *Annals of Mathematical Statistics* 35:73–101.

Huber, P. J. (1973). "Robust Regression: Asymptotics, Conjecture, and Monte Carlo." *Annals of Statistics* 1:799–821.

Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics.* 2nd ed. Hoboken, NJ: John Wiley & Sons.

Jackson, J. E. (2003). *A User's Guide to Principal Components*. Hoboken, NJ: John Wiley & Sons.

Jardine, N., and Sibson, R. (1971). *Mathematical Taxonomy*. New York: John Wiley & Sons.

Jöreskog, K. G. (1977). "Factor Analysis by Least-Squares and Maximum Likelihood Methods." In *Statistical Methods for Digital Computers*, edited by K. Enslein, A. Ralston, and H. Wilf, 125 - 165. New York: John Wiley & Sons.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Methodology*. 4th ed. New York: The Guilford Press.

Kohonen, T. (1989). *Self-Organization and Associative Memory*. 3rd ed. Vol. 8 of Springer Series in Information. Berlin: Springer-Verlag.

Kohonen, T. (1990). "The Self-Organizing Map." *Proceedings of the IEEE* 78:1464–1480.

Lindberg, W., Persson, J.-A., and Wold, S. (1983). "Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate." *Analytical Chemistry* 55:643–648.

LeRoux, B., and Rouanet, H. (2010). Multiple Correspondence Analysis. Vol.07–163 of Sage University Paper Series on Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage Publications.

Mardia, K., Kent, J., and Bibby, J. (1980). *Multivariate Analysis.* New York: Academic Press.

Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications.* Philadelphia: SIAM.

Maydeu-Olivares, A., Shi, D., and Rosseel, Y. (2017). "Assessing Fit in Structural Equation Models: A Monte-Carlo Evaluation of RMSEA Versus SRMR Confidence Intervals and Tests of Close Fit." *Structural Equation Model: A Multidisciplinary Journal* 25:389–402.

McLachlan, G. J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons.

McQuitty, L. L. (1957). "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies." *Educational and Psychological Measurement* 17:207–229.

Milligan, G. W. (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika* 45:325–342.

Nelson, P. R. C., Taylor, P. A., and MacGregor, J. F. (1996). "Missing Data Methods in PCA and PLS: Score calculations with incomplete observations." *Chemometrics and Intelligent Laboratory Systems* 35:45–65.

Nunnally, J. C. (1978). *Psychometric theory*. 2nd ed. New York: McGraw-Hill.

Penny, K. I. (1996). “Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance.” *Journal of the Royal Statistical Society, Series C* 45:73–81.

Press, W. H, Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1998). *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. Cambridge, England: Cambridge University Press.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

SAS Institute Inc. (1983). *SAS Technical Report A-108: Cubic Clustering Criterion*. Cary, NC: SAS Institute Inc. Accessed December 16, 2015. https://support.sas.com/kb/22/addl/fusion_22540_1_a108_5903.pdf.

SAS Institute Inc. (2020a). “The CALIS Procedure.” *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/calis.pdf.

SAS Institute Inc. (2020b). “The CANDISC Procedure.” *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/candisc.pdf.

SAS Institute Inc. (2020c). “The FACTOR Procedure.” *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/factor.pdf.

SAS Institute Inc. (2020d). “The FASTCLUS Procedure.” *SAS/STAT 15.2 User’s Guide.* Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/fastclus.pdf.

SAS Institute Inc. (2020e). “The MIXED Procedure.” *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/mixed.pdf.

SAS Institute Inc. (2020f). “The PLS Procedure.” *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/pls.pdf.

SAS Institute Inc. (2020g). “The VARCLUS Procedure.” *SAS/STAT 15.2 User’s Guide*. Cary, NC: SAS Institute Inc. https://go.documentation.sas.com/api/docsets/statug/15.2/content/varclus.pdf.

Schafer, J., and Strimmer, K. (2005). “A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.” *Statistical Applications in Genetics and Molecular Biology* 4 Article 32.

Sneath, P. H. A. (1957). “The Application of Computers to Taxonomy.” *Journal of General Microbiology* 17:201–226.

Sokal, R. R., and Michener, C. D. (1958). “A Statistical Method for Evaluating Systematic Relationships.” *University of Kansas Science Bulletin* 38:1409–1438.

Steiger, J. H. (1989). *EzPATH: A Supplementary Module for SYSTAT and SYGRAPH*. Evanston, IL: Systat.

Steiger, J. H. (1990). “Structural Model Evaluation and Modification: An Interval Estimation Approach.” *Multivariate Behavioral Research* 25:173–180.

Tobias, R. D. (1995). "An Introduction to Partial Least Squares Regression." In *Proceedings of the Twentieth Annual SAS Users Group International Conference*, 1250–1257. Cary, NC: SAS Institute Inc. http://www.sascommunity.org/sugi/SUGI95/Sugi-95-210%20Tobias.pdf.

Tracy, N. D., Young, J. C., and Mason, R. R. (1992). "Multivariate Control Charts for Individual Observations." *Journal of Quality Technology* 24:88–95.

Umetrics. (1995). *Multivariate Analysis (3-day course)*. Winchester, MA.

Waern, Y. (1972). "Structure in Similarity Matrices: A Graphic Approach." *Scandinavian Journal of Psychology* 13:5–16.

West, S. G., Taylor, A. B., and Wu, W. (2012). "Model Fit and Model Selection in Structural Equation Modeling." In *Handbook of Structural Equation Modeling*, edited by R. H. Hoyle, 209–231. New York: The Guilford Press.

White, K. P., Jr., Kundu, B., and Mastrangelo, C. M. (2008). "Classification of Defect Clusters on Semiconductor Wafers Via the Hough Transform." *IEEE Transactions on Semiconductor Manufacturing* 21:272–278.

Wold, S. (1994). "PLS for Multivariate Linear Modeling." In *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry*, edited by H. van de Waterbeemd, pp. 195–218. Weinheim, Germany: Verlag-Chemie.

Wold, S., Sjostrom, M., and Eriksson, L. (2001). "PLS-Regression: A Basic Tool of Chemometrics." *Chemometrics and Intelligent Laboratory Systems* 58:109–130.

Appendix C

Technology License Notices

- Scintilla is Copyright © 1998–2017 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

  NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.
- Progressë® Telerik® UI for WPF: Copyright © 2008-2019 Progress Software Corporation. All rights reserved. Usage of the included Progress® Telerik® UI for WPF outside of JMP is not permitted.
- ZLIB Compression Library is Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.
- Packages is Copyright © 2009–2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  Neither the name of the WhiteBox nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES

(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- iODBC software ia Copyright © 1995–2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  – Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL OPENLINK OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- This program, “bzip2”, the associated library “libbzip2”, and all documentation, are Copyright © 1996–2019 Julian R Seward. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.

  3. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

4. The name of the author may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Julian Seward, jseward@acm.org

bzip2/libbzip2 version 1.0.8 of 13 July 2019

- R software is Copyright © 1999–2012, R Foundation for Statistical Computing.
- MATLAB software is Copyright © 1984-2012, The MathWorks, Inc. Protected by U.S. and international patents. See www.mathworks.com/patents. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.
- libopc is Copyright © 2011, Florian Reuter. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  - Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.
  - Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF

USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- libxml2 - Except where otherwise noted in the source code (e.g. the files hash.c, list.c and the trio files, which are covered by a similar license but with different Copyright notices) all the files are:

  Copyright © 1998–2003 Daniel Veillard. All Rights Reserved.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

  THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

- Regarding the decompression algorithm used for UNIX files:

  Copyright © 1985, 1986, 1992, 1993

  The Regents of the University of California. All rights reserved.

  THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

- Snowball is Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton.

  All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS \”AS IS\” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- Pako is Copyright © 2014–2017 by Vitaly Puzrin and Andrei Tuputcyn.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

- HDF5 (Hierarchical Data Format 5) Software Library and Utilities are Copyright 2006–2015 by The HDF Group. NCSA HDF5 (Hierarchical Data Format 5) Software Library and Utilities Copyright 1998-2006 by the Board of Trustees of the University of Illinois. All rights reserved. DISCLAIMER: THIS SOFTWARE IS PROVIDED BY THE HDF GROUP AND THE CONTRIBUTORS “AS IS” WITH NO WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED. In no event shall The HDF Group or the Contributors be liable for any damages suffered by the users arising out of the use of this software, even if advised of the possibility of such damage.
- agl-aglfn technology is Copyright © 2002, 2010, 2015 by Adobe Systems Incorporated. All Rights Reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  - Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
  - Neither the name of Adobe Systems Incorporated nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- dmlc/xgboost is Copyright © 2019 SAS Institute.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libzip is Copyright © 1999–2019 Dieter Baron and Thomas Klausner.

  This file is part of libzip, a library to manipulate ZIP archives. The authors can be contacted at <libzip@nih.at>.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  3. The names of the authors may not be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE AUTHORS ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- OpenNLP 1.5.3, the pre-trained model (version 1.5 of en-parser-chunking.bin), and dmlc/xgboost Version .90 are licensed under the Apache License 2.0 are Copyright © January 2004 by Apache.org.

  You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

  - You must give any other recipients of the Work or Derivative Works a copy of this License; and
  - You must cause any modified files to carry prominent notices stating that You changed the files; and
  - You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
  - If the Work includes a "NOTICE" text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.
  - You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.

- LLVM is Copyright © 2003–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- clang is Copyright © 2007–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS", WITHOUT WARRANTIES OR CONDITIONS OF

ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- lld is Copyright © 2011–2019 by the University of Illinois at Urbana-Champaign.

  Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS", WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libcurl is Copyright © 1996–2020, Daniel Stenberg, daniel@haxx.se, and many contributors, see the THANKS file. All rights reserved.

  Permission to use, copy, modify, and distribute this software for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

  THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.