Version 13

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

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *JMP® 13 Multivariate Methods*. Cary, NC: SAS Institute Inc.

JMP® 13 Multivariate Methods

ISBN 978-1-62960-478-7 (Hardcopy)
ISBN 978-1-62960-580-7 (EPUB)
ISBN 978-1-62960-581-4 (MOBI)

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

September 2016

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.

Technology License Notices

- Scintilla - Copyright © 1998-2014 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.
- Telerik RadControls: Copyright © 2002-2012, Telerik. Usage of the included Telerik RadControls outside of JMP is not permitted.
- ZLIB Compression Library - Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.
- Packages - 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 - 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.

- bzip2, the associated library "libbzip2", and all documentation, are Copyright © 1996-2010, 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:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

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.

Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

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.

- 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 licence 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 - 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

Contents

Multivariate Methods

Chapter 1

Learn about JMP

Documentation and Additional Resources

This chapter includes the following information:

- book conventions
- JMP documentation
- JMP Help
- additional resources, such as the following:
  - other JMP documentation
  - tutorials
  - indexes
  - Web resources
  - technical support options

Formatting Conventions

The following 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 font.
- Code appears in `Lucida Sans Typewriter` font.
- Code output appears in *`Lucida Sans Typewriter`* italic font and is indented farther than the preceding code.
- **Helvetica bold** formatting 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
  - script output
- Features that are for JMP Pro only are noted with the JMP Pro icon JMP PRO. For an overview of JMP Pro features, visit http://www.jmp.com/software/pro/.

Note: Special information and limitations appear within a Note.

Tip: Helpful information appears within a Tip.

JMP Documentation

JMP offers documentation in various formats, from print books and Portable Document Format (PDF) to electronic books (e-books).

- Open the PDF versions from the **Help > Books** menu.
- All books are also combined into one PDF file, called *JMP Documentation Library*, for convenient searching. Open the *JMP Documentation Library* PDF file from the **Help > Books** menu.
- You can also purchase printed documentation and e-books on the SAS website:

  http://www.sas.com/store/search.ep?keyWords=JMP

JMP Documentation Library

The following table describes the purpose and content of each book 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.
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.
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.

Document Title	Document Purpose	Document Content
Essential Graphing	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Overlay Plot • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Treemap • Scatterplot Matrix • Ternary Plot • Chart 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.
Design of Experiments Guide	Learn how to design experiments and determine appropriate sample sizes.	Covers all topics in the DOE menu and the Specialized DOE Models menu item in the Analyze > Specialized Modeling menu.

Document Title	Document Purpose	Document Content
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: • Modeling Utilities • Neural • Partition • Bootstrap Forest • Boosted Tree • K Nearest Neighbors • Naive Bayes • Model Comparison • Formula Depot Describes these Analyze > Specialized Modeling menu platforms: • Fit Curve • Nonlinear • Gaussian Process • Time Series • Matched Pairs Describes these Analyze > Screening menu platforms: • Response Screening • Process Screening • Predictor Screening • Association Analysis The platforms in the Analyze > Specialized Modeling > Specialized DOE Models menu are described in *Design of Experiments Guide*.

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 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 • Pareto Plot • Diagram

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 and 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 • Multiple Correspondence Analysis • Multidimensional Scaling • Factor Analysis • Choice • MaxDiff • Uplift • Item 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.

Document Title	Document Purpose	Document Content
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.

Note: The **Books** menu also contains two reference cards that can be printed: The *Menu Card* describes JMP menus, and the *Quick Reference* describes JMP keyboard shortcuts.

JMP Help

JMP Help is an abbreviated version of the documentation library that provides targeted information. You can open JMP Help in several ways:

- On Windows, press the F1 key to open the Help system window.
- 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.
- Search and view JMP Help on Windows using the **Help > Help Contents**, **Search Help**, and **Help Index** options. On Mac, select **Help > JMP Help**.
- Search the Help at http://jmp.com/support/help/ (English only).

Additional Resources for Learning JMP

In addition to JMP documentation and JMP Help, you can also learn about JMP using the following resources:

- Tutorials (see "Tutorials" on page 26)
- Sample data (see "Sample Data Tables" on page 26)
- Indexes (see "Learn about Statistical and JSL Terms" on page 26)
- Tip of the Day (see "Learn JMP Tips and Tricks" on page 26)
- Web resources (see "JMP User Community" on page 27)
- JMPer Cable technical publication (see "JMPer Cable" on page 27)
- Books about JMP (see "JMP Books by Users" on page 27)
- JMP Starter (see "The JMP Starter Window" on page 28)

- Teaching Resources (see "Sample Data Tables" on page 26)

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, then 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\13\Samples\Data

On Macintosh: \Library\Application Support\JMP\13\Samples\Data

In JMP Pro, sample data is installed in the JMPPRO (rather than JMP) directory. In JMP Shrinkwrap, sample data is installed in the JMPSW 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 http://jmp.com/tools.

Learn about Statistical and JSL Terms

The **Help** menu contains the following indexes:

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.

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. See the *Using JMP* book for details.

Tooltips

JMP provides descriptive tooltips when you place your cursor 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 Macintosh.

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/.

JMPer Cable

The JMPer Cable is a yearly technical publication targeted to users of JMP. The JMPer Cable is available on the JMP website:

http://www.jmp.com/about/newsletters/jmpercable/

JMP Books by Users

Additional books about using JMP that are written by JMP users are available on the JMP website:

http://www.jmp.com/en_us/software/books.html

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 the Macintosh) **> 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 Macintosh, select **JMP > Preferences > Initial JMP Starter Window**.

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 http://www.jmp.com/support, including the technical support phone number.

Chapter 2

Introduction to Multivariate Analysis

Overview of Multivariate Techniques

This book 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 to 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 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 7, “Hierarchical Cluster”.
- The KMeans Clustering platform groups observations that share similar values across a number of variables. See Chapter 8, “K Means Cluster”.
- The Normal Mixtures platform enables you to cluster observations when your data come from overlapping normal distributions.See Chapter 9, “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 10, “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 or the most representative variable in the cluster can be used to explain most of the variation in the data. See Chapter 11, “Cluster Variables”.

Chapter 3

Correlations and Multivariate Techniques

Explore the Multidimensional Behavior of Variables

Use the Multivariate platform to explore how many variables relate to each other. The word multivariate simply means involving many variables instead of one (univariate) or two (bivariate). From the Multivariate report, you can:

- summarize the strength of the linear relationships between each pair of response variables using the Correlations table
- identify dependencies, outliers, and clusters using the Scatterplot Matrix
- use other techniques to examine multiple variables, such as partial, inverse, and pairwise correlations, covariance matrices, principal components, and more

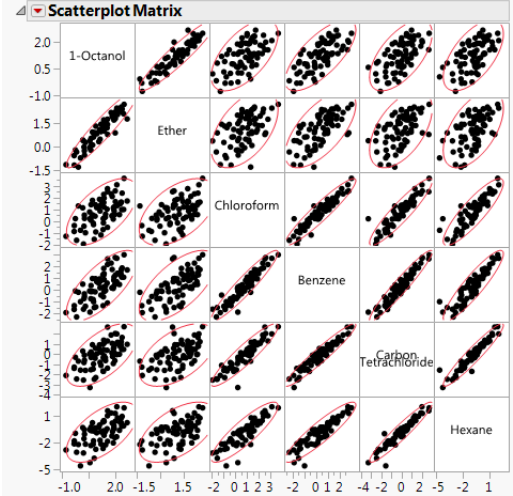
Figure 3.1 Example of a Multivariate Report

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

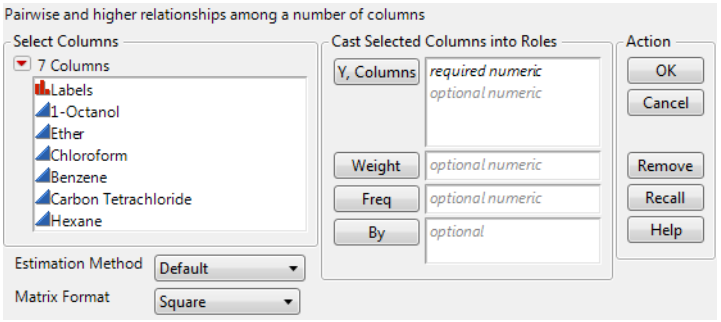
Scatterplot Matrix



Launch the Multivariate Platform

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

Figure 3.2 The Multivariate Launch Window



Y, Columns Defines one or more response columns.

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 Performs a separate multivariate analysis for each level of the By variable.

Estimation Method Select from one of several estimation methods for the correlations. With the **Default** option, Row-wise is used for data tables with no missing values. Pairwise is used for data tables that have more than 10 columns or more than 5000 rows, and that have missing values. Otherwise, the default estimation method is REML. For details, see "Estimation Methods" on page 32.

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.

Estimation Methods

Several estimation methods for the correlations options are available to provide flexibility and to accommodate personal preferences. **REML** and **Pairwise** are the methods used most frequently. You can also estimate missing values by using the estimated covariance matrix, and then using 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** is used for data tables with no missing values.
- **Pairwise** is used in these circumstances:
  - the data table has more than 10 columns or more than 5000 rows and has missing values
  - the data table has more columns than rows and has missing values
- **REML** is used otherwise.

REML

REML (restricted maximum likelihood) estimates are less biased than the ML (maximum likelihood) estimation method. The REML method maximizes marginal likelihoods based upon error contrasts. The REML method is often used for estimating variances and covariances.The REML method in the Multivariate 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. REML uses all of your data, even if missing cells are present, and is most useful for smaller datasets. Because of the bias-correction factor, this method is slow if your dataset is large and there are many missing data values. If there are no missing cells in the data, then the REML estimate is equivalent to the sample covariance matrix.

ML

The maximum likelihood estimation method (ML) is useful for large data tables with missing cells. The ML estimates are similar to the REML estimates, but the ML estimates are generated faster. Observations with missing values are not excluded. For small data tables, REML is preferred over ML because REML's variance and covariance estimates are less biased.

Robust

Note: If you select Robust, and your data table contains more columns than rows, JMP switches the Estimation Method to Row-wise.

Robust estimation is useful for data tables that might have outliers. For statistical details, see "Robust" on page 47.

Row-wise

Rowwise estimation does not use observations containing missing cells. This method is useful in the following situations:

- checking compatibility with JMP versions earlier than JMP 8. Rowwise estimation was the only estimation method available before JMP 8.
- excluding any observations that have missing data.

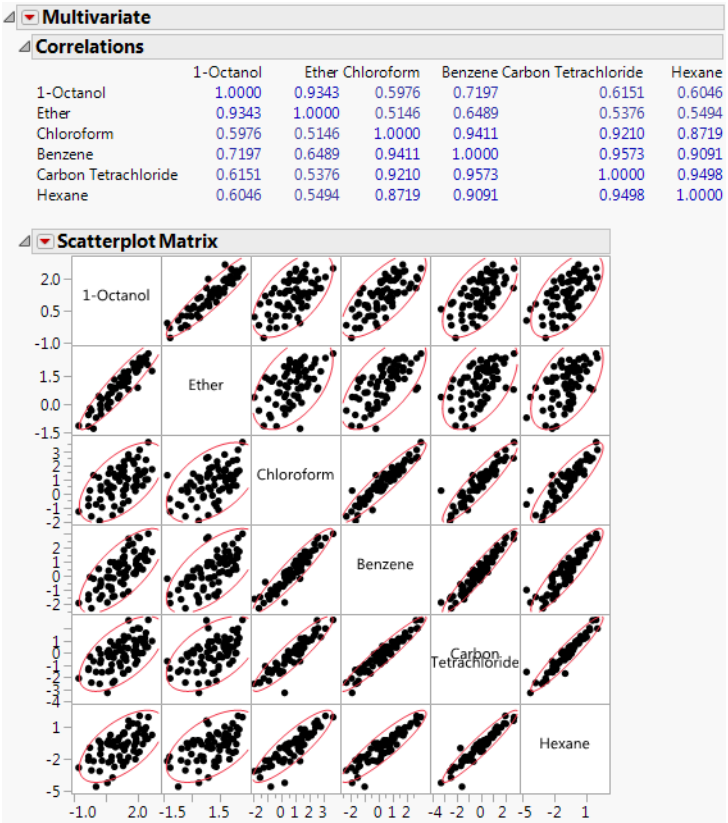
Pairwise

Pairwise estimation performs correlations for all rows for each pair of columns with nonmissing values.

The Multivariate Report

The default multivariate report shows the standard correlation matrix and the scatterplot matrix. The platform menu lists additional correlation options and other techniques for looking at multiple variables. See "Multivariate Platform Options" on page 35.

Figure 3.3 Example of a Multivariate Report



To Produce the Report in Figure 3.3

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

About Missing Values

In most of the analysis options, a missing value in an observation does not cause the entire observation to be deleted. However, the **Pairwise Correlations** option excludes rows that are missing for either of the variables under consideration. The **Simple Statistics > Univariate** option calculates its statistics column-by-column, without regard to missing values in other columns.

Multivariate Platform Options

Correlations Multivariate	Shows or hides the Correlations table, which 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 47. This correlation matrix is calculated by the method that you select in the launch window.
Correlation Probability	Shows the Correlation Probability report, which 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 Correlation	Shows the two-tailed confidence intervals of the correlations. This option is off by default. The default confidence coefficient is 95%. Use the **Set α Level** option to change the confidence coefficient.

Inverse Correlations	Shows or hides the inverse correlation matrix (Inverse Corr table). This option is off by default. 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, the diagonal is $1/(1 - R^2)$ for the fit of that variable by 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 49.
Partial Correlations	Shows or hides the partial correlation table (Partial Corr), which shows the measure of the relationship between a pair of variables after adjusting for the effects of all the other variables. This option is off by default. This table is the negative of the inverse correlation matrix, scaled to unit diagonal.
Covariance Matrix	Shows or hides the covariance matrix which measures the degree to which a pair of variables change together. This option is off by default.
Pairwise Correlations	Shows or hides the Pairwise Correlations table, which lists the Pearson product-moment correlations for each pair of Y variables. This option is off by default. 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.

Hotelling's T^2 Test

Allows you to conduct a one-sample test for the mean of the multivariate distribution of the variables that you entered as Y. Specify the mean vector under the null hypothesis in the window that appears by entering 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:

Variable Lists the variables entered as Y.

Mean Gives the sample mean for each variable.

Hypothesized Mean Shows the null hypothesis means that you specified.

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

F Ratio Gives the value of the test statistic. If you have n rows and k variables, the F ratio is given 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 n and n - k degrees of freedom.

Simple Statistics	This menu contains two options that each show or hide simple statistics (mean, standard deviation, and so on) 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 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. If the **REML**, **ML**, or **Robust** method is selected, the mean vector and covariance matrix are estimated by that selected method. If the **Row-wise** method is selected, all rows with at least one missing value are excluded from the calculation of means and variances. If the **Pairwise** method is selected, the mean and variance are calculated for each column. These options are off by default.
Nonparametric Correlations	This menu contains three nonparametric measures: Spearman's Rho, Kendall's Tau, and Hoeffding's D. These options are off by default. For details, see "Nonparametric Correlations" on page 40.
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. For details, see "Scatterplot Matrix" on page 41.

Color Maps	The **Color Map** menu contains three types of color maps. **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 may be different. These options are off by default.
Parallel Coord Plot	Shows or hides a parallel coordinate plot of the variables. This option is off by default.
Ellipsoid 3D Plot	Shows or hides a three-dimensional scatterplot with a 95% confidence ellipsoid. You are prompted to specify the three variables when you select this option. This option is off by default.
Principal Components	This menu contains options to show or hide a principal components report. You can select correlations, covariances, or unscaled. Selecting one of these options when another of the reports is shown changes the report to the new option. Select **None** to remove the report. This option is off by default. *Principal components* is a technique to take 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. For details, see the chapter "Principal Components" on page 53.
Outlier Analysis	This menu contains options that show or hide plots that measure distance in the multivariate sense using one of these methods: the Mahalanobis distance, jackknife distances, and the T^2 statistic. For details, see "Outlier Analysis" on page 43.

Item Reliability	This menu contains options that each shows or hides 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. These options are off by default. For details, see "Item Reliability" on page 45.
Impute Missing Data	Produces a new data table that duplicates your data table and replaces all missing values with estimated values. This option is available only if your data table contains missing values. For details, see "Impute Missing Data" on page 45.
Save Imputed Formula	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 the JMP Reports chapter in the *Using JMP* book 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. Each Nonparametric correlation report gives the significance probability for the measure of association and displays the association value on a bar chart.

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

Kendall's Tau is 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 (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.

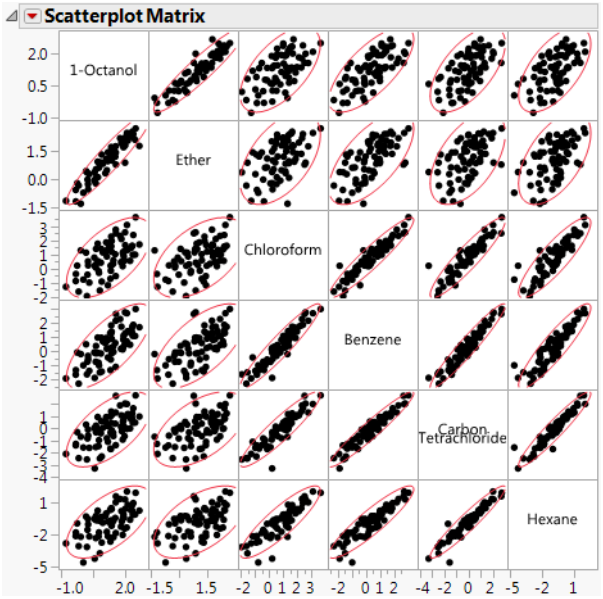
Note: When a Weight variable is specified, missing and zero-valued weights are excluded from the nonparametric correlation calculations. All other weight values are treated as 1.

For statistical details about these three methods, see the "Nonparametric Measures of Association" on page 47.

Scatterplot Matrix

A scatterplot matrix helps you visualize the correlations between each pair of response variables. The scatterplot matrix is shown by default, and can be hidden or shown by selecting **Scatterplot Matrix** from the red triangle menu for Multivariate.

Figure 3.4 Scatterplot Matrix



By default, a 95% bivariate normal density ellipse is shown 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.

Working with the Scatterplot Matrix

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).

Options for Scatterplot Matrix

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

Table 3.1 Options for the Scatterplot Matrix

Show Points	Shows or hides the points in the scatterplots.

Table 3.1 Options for the Scatterplot Matrix *(Continued)*

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 histogram in the upper left corner of each scatterplot.
Show Histogram	Shows either 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 one.
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 one. 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

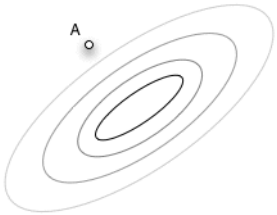
The **Outlier Analysis** menu contains options that show or hide plots that measure distance in the multivariate sense using one of these methods:

- Mahalanobis distance
- jackknife distances
- T^2 statistic

These methods all measure distance in the multivariate sense, with respect to the correlation structure. Testing is done at the alpha level that appears at the bottom of the plot.

In Figure 3.5, Point A is an outlier because it is outside the correlation structure rather than because it is an outlier in any of the coordinate directions.

Figure 3.5 Example of an Outlier



Mahalanobis Distance

The Mahalanobis Outlier Distance plot shows 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 for more information.

Jackknife Distances

The Jackknife Distances plot shows 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 50 for more information.

T^2 Statistic

The T^2 plot shows 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 for more information.

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. Nunnally (1979) suggests a Cronbach's α of 0.7 as a rule-of-thumb acceptable level of agreement.

See "Cronbach's α" on page 51 for details about computations.

Impute Missing Data

To impute missing data, select **Impute Missing Data** from the red triangle menu for Multivariate. A new data table is created that duplicates your data table and replaces all missing values with estimated values.

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

This option is available only if your data table contains missing values.

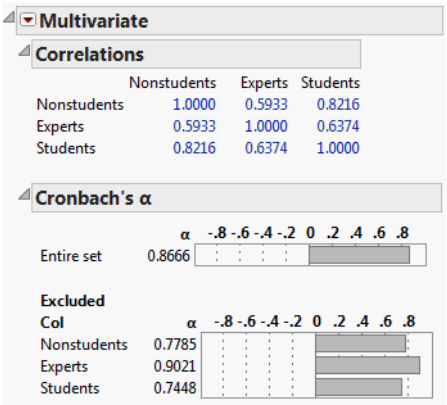
Example of Item Reliability

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, nonstudents, 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 nonstudents, 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. From the red triangle menu for Multivariate, select **Item Reliability > Cronbach's α**.
6. (Optional) From the red triangle menu for Multivariate, select **Scatterplot Matrix** to hide that plot.

Figure 3.6 Cronbach's α Report



The Cronbach's α results in Figure 3.6 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

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 weight:

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

where 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, and

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

where 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

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

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

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

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, with 1 indicating 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 $\mathbf{R}$ and the inverse correlation matrix is denoted $\mathbf{R}^{-1}$, the diagonal element is denoted r^{ii} and is computed as

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

where $\mathbf{R}_i^2$ 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

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

where:

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

$\overline{Y}$ 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

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

where:

n = number of observations

p = number of variables (columns)

$\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}$ = (1–αth) 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

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

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

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

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–αth) 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

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

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

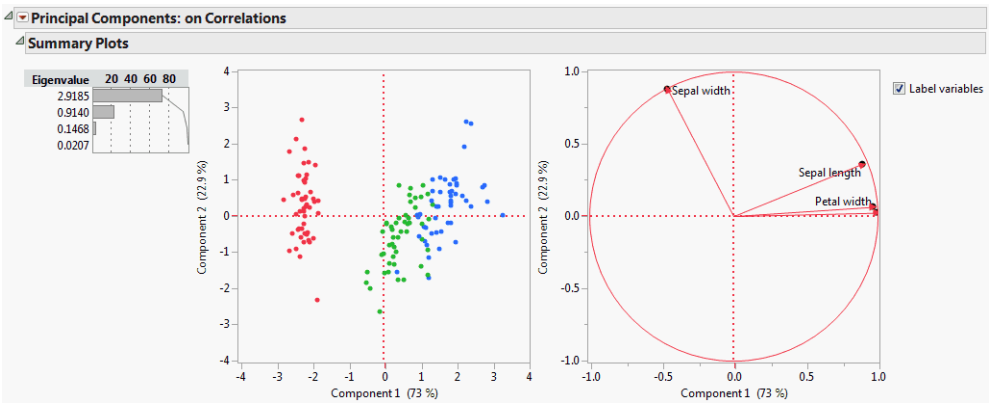
The purpose of principal component analysis is to derive a small number of independent linear combinations (*principal components*) of a set of measured variables that capture as much of the variability in the original variables as possible. Principal component analysis is a dimension-reduction technique, 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 in the *Consumer Research* book.

Figure 4.1 Example of Principal Components



Overview of Principal Component Analysis

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 representation is important in visualizing multivariate data by reducing it to graphable dimensions. 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, p principal components are formed as follows:

- 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 in the *Consumer Research* book for details.

Example of Principal Component Analysis

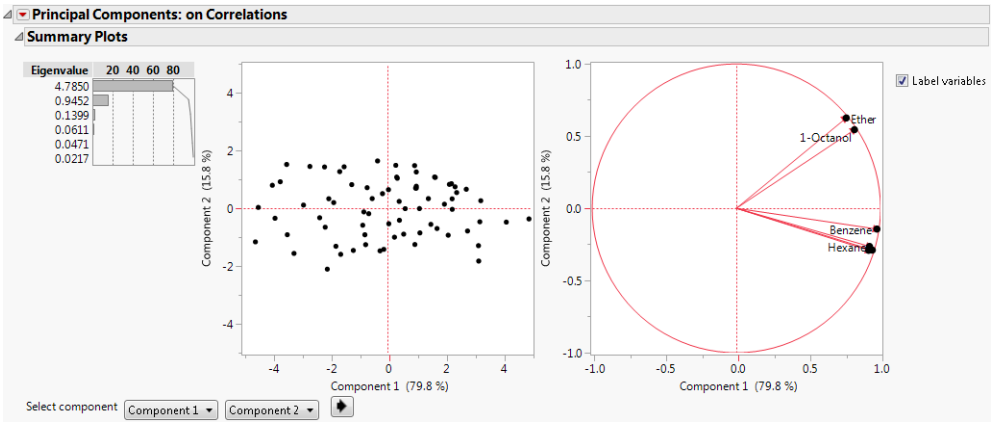
To view an example Principal Component Analysis report for a data table for two factors:

1. Select **Help > Sample Data Library** and open Solubility.jmp.
2. Select **Analyze > Multivariate Methods > Principal Components**.

   The Principal Components launch window appears.
3. Select all of the continuous columns and click **Y, Columns**.
4. Keep the default Estimation Method.
5. Click **OK**.

   The Principal Components on Correlations report appears.

Figure 4.2 Principal Components on Correlations Report



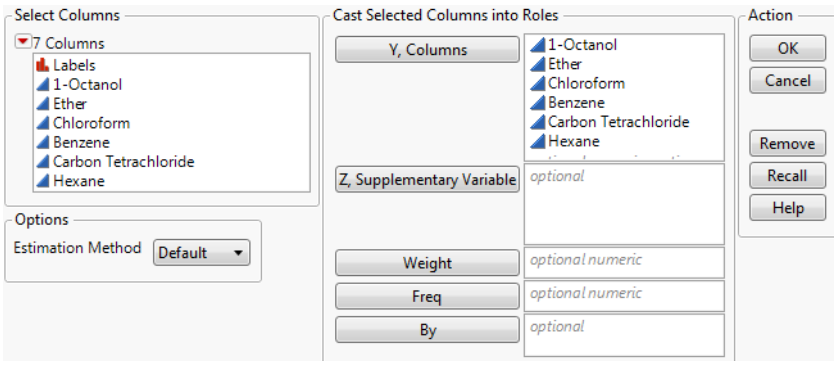
The report gives the eigenvalues and a bar chart of the percent of the variation accounted for by each principal component. There is a Score Plot and a Loadings Plot as well. See "Principal Components Report" on page 59 for details.

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 54 uses all of the continuous variables from the Solubility.jmp sample data table.

Figure 4.3 Principal Components Launch Window



Y, Columns Lists the variables to analyze for components.

Z, Supplementary Variable Lists 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. The supplementary variables can be projected on to the loading plot and used to enhance interpretation.

Weight and Freq Enables you to weight the analysis to account for pre-summarized data.

Note: The Weight and Freq roles are 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 Lists different methods for calculating the correlations. Several of these methods address the treatment of missing data. See "Estimation Methods" on page 56.

Number of Components (Appears only when Sparse is the Estimation Method.) The number of components to be estimated. See "Sparse" on page 59

Estimation Methods

Use the estimation method that addresses your specific needs. Methods are available to handle missing values, outliers, wide data, and sparse data.

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 the Modeling Utilities chapter in the *Predictive and Specialized Modeling* book for details.

Default

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

- **Row-wise** is used for data tables with no missing values.
- **Pairwise** is used in the following circumstances:
  - the data table has more than 10 columns or more than 5,000 rows and has missing values
  - the data table has more columns than rows and has missing values
- **REML** is used otherwise.

REML

REML (restricted maximum likelihood) estimates are less biased than the ML (maximum likelihood) estimation method. 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.

REML uses all of your data, even if missing values are present, and is most useful for smaller datasets. Because of the bias-correction factor, this method is slow if your dataset is large and there are many missing data values. If there are no missing cells in the data, then the REML estimate is equivalent to the sample covariance matrix.

Note: If you select REML and your data table contains more columns than rows, JMP switches the Estimation Method. If there are no missing values, the Estimation Method switches to Row-wise. If there are missing values, then the Estimation Method switches to Pairwise.

ML

The maximum likelihood estimation method (ML) is useful for large data tables with missing cells. The ML estimates are similar to the REML estimates, but the ML estimates are generated faster. Observations with missing values are not excluded. For small data tables, REML is preferred over ML because REML's variance and covariance estimates are less biased.

Note: If you select ML and your data table contains more columns than rows, JMP switches the Estimation Method. If there are no missing values, the Estimation Method switches to Row-wise. If there are missing values, then the Estimation Method switches to Pairwise.

Robust

Robust estimation is useful for data tables that might have outliers. For statistical details, see "Robust" on page 47.

Note: If you select Robust and your data table contains more columns than rows, JMP switches the Estimation Method. If there are no missing values, the Estimation Method switches to Row-wise. If there are missing values, then the Estimation Method switches to Pairwise.

Row-wise

Row-wise estimation does not use observations containing missing cells. This method is useful in the following situations:

- Checking compatibility with JMP versions earlier than JMP 8. Row-wise estimation was the only estimation method available prior to JMP 8.
- Excluding any observations that have missing data.

Pairwise

Pairwise estimation performs correlations for all rows for each pair of columns with nonmissing values.

Wide

Note: The Wide method extracts components based on the standardized data. The On Covariance and On Unscaled options are not available.

The Wide method is useful when you have a very large number of columns in your data. It uses a computationally efficient algorithm that avoids calculating the covariance matrix. The algorithm is based on the singular value decomposition. For additional background, see "Wide Linear Methods and the Singular Value Decomposition" on page 226 in the "Statistical Details" appendix.

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 given as follows:

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

Using the singular value decomposition, $\mathbf{X}_s$ is written as $\mathbf{U}Diag(\mathbf{\Lambda})\mathbf{V}'$. 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 $\mathbf{X}_s\mathbf{V}$.

Note: If there are missing values and you select the Wide method, then the rows that contain missing values are deleted and the Wide method is applied to the remaining rows.

Note: When you select the Default estimation method and enter more than 500 variables as Y, Columns, a JMP Alert recommends that you switch to the Wide estimation method. 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. Click **Continue** to use the method you originally selected.

JMP PRO Sparse

Note: The Sparse method extracts components based only on the standardized data. The On Covariance and On Unscaled options are not available.

The Sparse method is useful when your data are sparse, meaning that they contain many zeros. It can also reduce computational time when there are a large number of columns in the data. 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 58. 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).

When you select Sparse as the estimation method in the launch window, the Number of Components option appears. By default, the Number of Components is 2. Typically, the Number of Components is much smaller than the dimension of your data.

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.

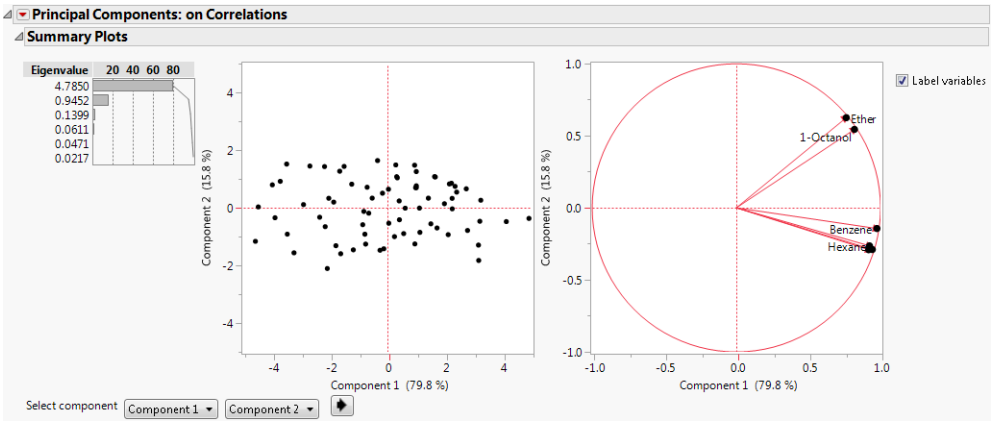
The initial Principal Components report is for an analysis on Correlations. It summarizes the variation of the specified *Y* variables with principal components. See Figure 4.4. You can switch to an analysis based on the covariance matrix or unscaled data by selecting the Principal Components option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

Figure 4.5 Correlations

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

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

Figure 4.6 Covariance Matrix

Covariance Matrix

	1-Octanol	Ether	Chloroform	Benzene	Carbon Tetrachloride	Hexane
1-Octanol	0.67553	0.76870	0.57055	0.72776	0.64207	0.69393
Ether	0.76870	1.00217	0.59845	0.79922	0.68356	0.76796
Chloroform	0.57055	0.59845	1.34948	1.34496	1.35876	1.41442
Benzene	0.72776	0.79922	1.34496	1.51350	1.49577	1.56172
Carbon Tetrachloride	0.64207	0.68356	1.35876	1.49577	1.61295	1.68446
Hexane	0.69393	0.76796	1.41442	1.56172	1.68446	1.94995

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 4.9) are given for each eigenvalue (Jackson, 2003).

Figure 4.7 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.

Figure 4.8 Eigenvectors

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
1-Octanol	0.37441	0.55987	-0.11070	-0.65842	0.31660	0.01874
Ether	0.34834	0.64314	0.11973	0.62764	-0.20890	0.11456
Chloroform	0.41940	-0.29864	-0.64850	0.30599	0.43061	0.18793
Benzene	0.44561	-0.14756	-0.21904	-0.09455	-0.49849	-0.68865
Carbon Tetrachloride	0.43102	-0.29736	0.18487	-0.24135	-0.45965	0.64968
Hexane	0.42217	-0.27117	0.68608	0.10831	0.45926	-0.23426

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 4.9 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 scaling of the loadings 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.

Figure 4.10 Loading Matrix

Loading Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
1-Octanol	0.81902	0.54432	-0.04140	-0.16274	0.06871	0.00276
Ether	0.76198	0.62528	0.04477	0.15513	-0.04534	0.01688
Chloroform	0.91742	-0.29035	-0.24252	0.07563	0.09345	0.02770
Benzene	0.97476	-0.14346	-0.08191	-0.02337	-0.10819	-0.10149
Carbon Tetrachloride	0.94285	-0.28910	0.06913	-0.05965	-0.09976	0.09575
Hexane	0.92348	-0.26364	0.25657	0.02677	0.09967	-0.03452

Note: 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.

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 4.11 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

Tip: Use the sliders to dim loadings whose absolute values fall below your selected value and to set the degree of transparency for the loadings.

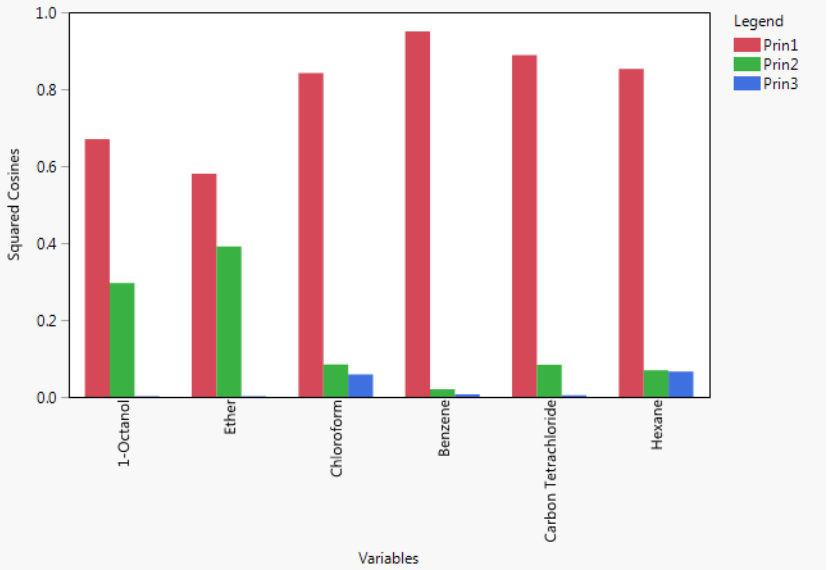
Squared Cosines of Variables Shows or hides a table that contains the squared cosines of variables. The sum of the squared cosine values across principal components is equal to one for each variable. The squared cosines enable you to see how well the variables are represented by the principal components. 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.

Figure 4.12 Squared Cosines

Squared Cosines of Variables

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
1-Octanol	0.67079	0.29628	0.00171	0.02648	0.00472	0.00001
Ether	0.58061	0.39098	0.00200	0.02407	0.00206	0.00029
Chloroform	0.84166	0.08430	0.05881	0.00572	0.00873	0.00077
Benzene	0.95016	0.02058	0.00671	0.00055	0.01170	0.01030
Carbon Tetrachloride	0.88897	0.08358	0.00478	0.00356	0.00995	0.00917
Hexane	0.85282	0.06951	0.06583	0.00072	0.00993	0.00119

Plot of Squared Cosines of Variables



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.

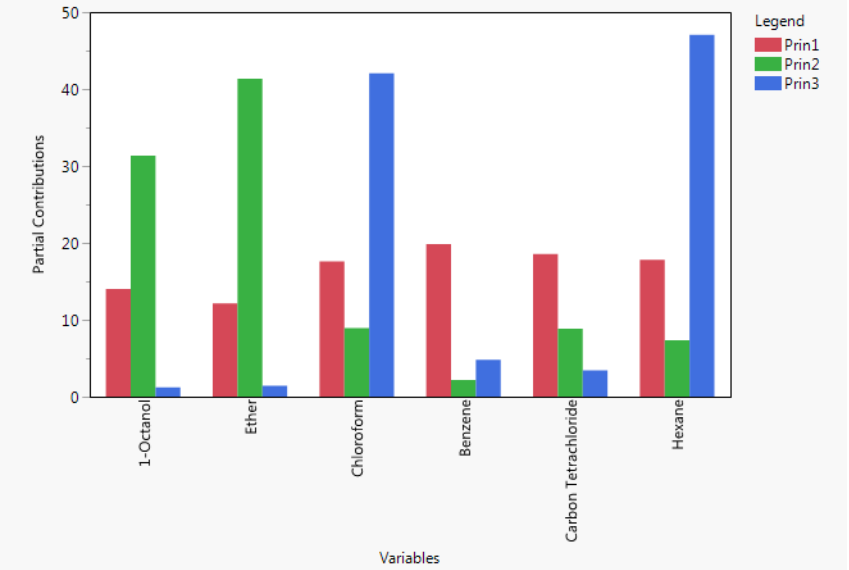
Partial Contribution of Variables Shows or hides a table that contains the partial contributions of variables. The partial contributions enable you to see the percentage that each variable contributes to each principal component. This option also shows a plot of the partial contributions for the first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

Figure 4.13 Partial Contribution of Variables

Partial Contribution of Variables

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
1-Octanol	14.01861	31.34495	1.22537	43.35231	10.02364	0.03513
Ether	12.13395	41.36336	1.43348	39.39286	4.36399	1.31235
Chloroform	17.58956	8.91883	42.05468	9.36274	18.54254	3.53165
Benzene	19.85699	2.17734	4.79789	0.89396	24.84960	47.42422
Carbon Tetrachloride	18.57813	8.84209	3.41761	5.82509	21.12809	42.20899
Hexane	17.82277	7.35342	47.07097	1.17304	21.09215	5.48766

Plot of Partial Contributions of Variables



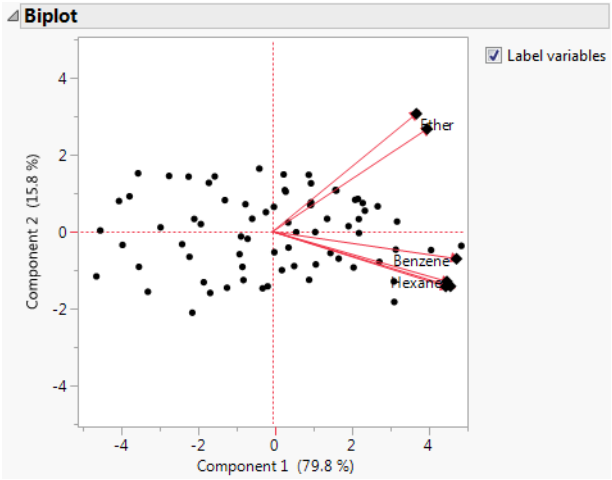
Note: If the Sparse estimation method is used and the number of components selected is less than three, only the specified number of components are displayed in the plot.

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 59.

Tip: Select the tips of arrows in the loading plot to select the corresponding columns in the data table. Hold down Control and click on 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.

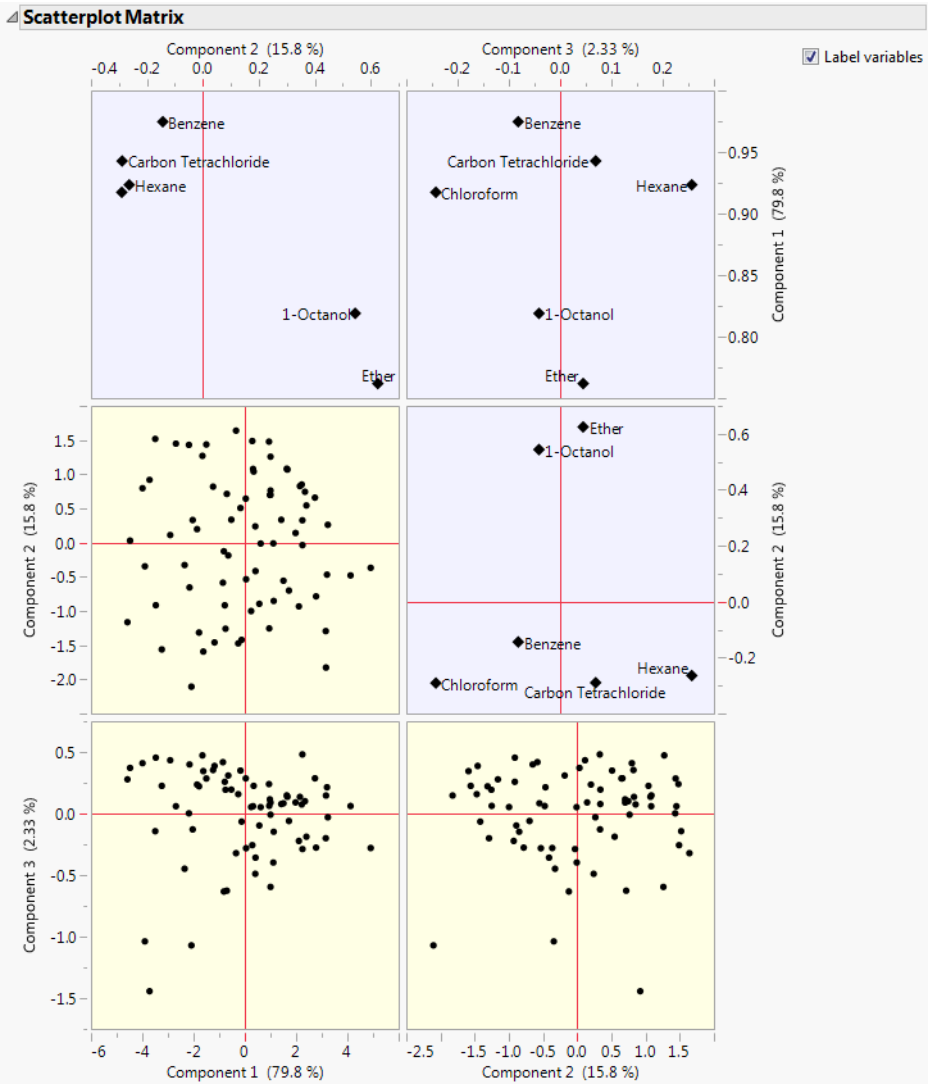
Figure 4.14 Biplot



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.

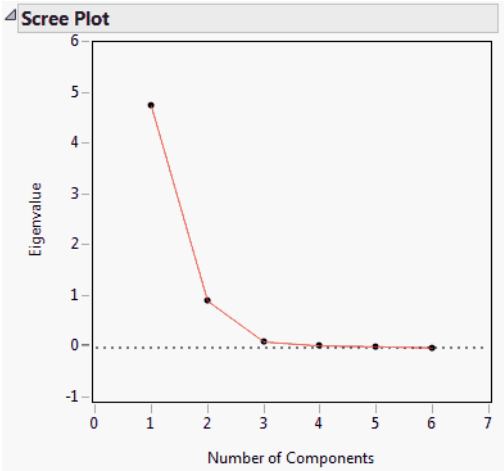
Figure 4.15 Scatterplot Matrix



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.

Figure 4.16 Scree Plot



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).

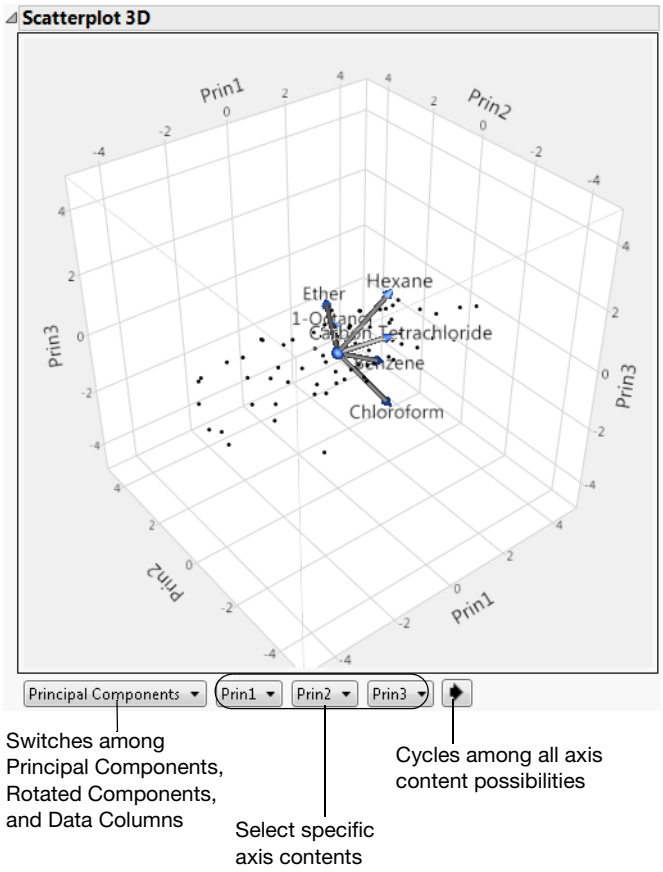
Loading Plot Shows or hides a matrix of two-dimensional representations of factor loadings for the specified number of components. The loading plot labels variables if the number of variables is 30 or fewer. If there are more than 30 variables, the labels are off by default. This information is shown 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. Hold down Ctrl and click on 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.

Figure 4.17 Scatterplot 3D Score Plot



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 length of the ray corresponds to the eigenvalue or variance of the principal component.

Display Options 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.

Arrow Lines Shows or hides arrow lines for the analysis variables in the all plots that can display arrows.

Show Supplementary Variable (Available only if you specify a supplementary variable.) Shows or hides the arrow lines for the supplementary variables in the biplot and loading plots.

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 in the *Consumer Research* book for details.

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 213 for details.

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

Figure 4.18 Cluster Summary

Variable Clustering

Cluster Summary

Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained	.2 .4 .6 .8
1	6	Benzene	0.798	0.798	

Proportion of variation explained by clustering: 0.798

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 with a specified number of principal components to new columns in the data table.

Save DModX Saves the observation distance to the principal components model (DModX) to a new column in the data table.DModX is defined as follows:

$$\text{DModX}=\frac{\sqrt{\sum_{k} e_{ik}}}{K-A}$$

DModX is calculated based on the residuals, e_{ik}, the number of variables, K, and the number of principal components, A. Larger DModX values indicate mild to moderate outliers in the data.

Save Individual Squared Cosines Saves the individual squared cosines to new columns in the data table.

Save Individual Partial Contributions Saves the individual partial contributions 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.

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 the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

See the JMP Reports chapter in the *Using JMP* book 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.

Chapter 5

Discriminant Analysis

Predict Classifications Based on Continuous Variables

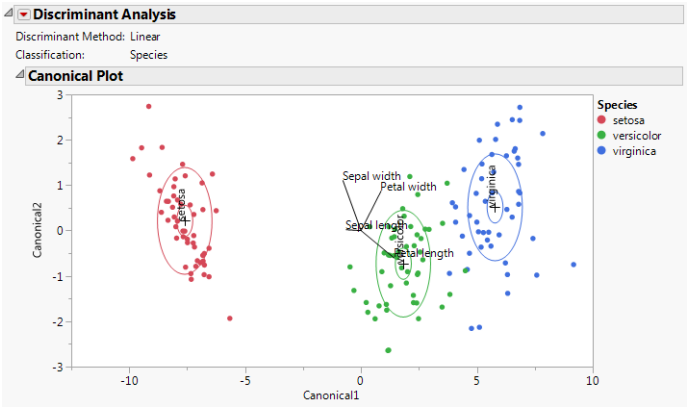
Discriminant analysis predicts membership in a group or category based on observed values of several continuous variables. Specifically, discriminant analysis predicts a classification (*X*) variable (categorical) based on known continuous responses (*Y*). The data for a discriminant analysis consist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

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:

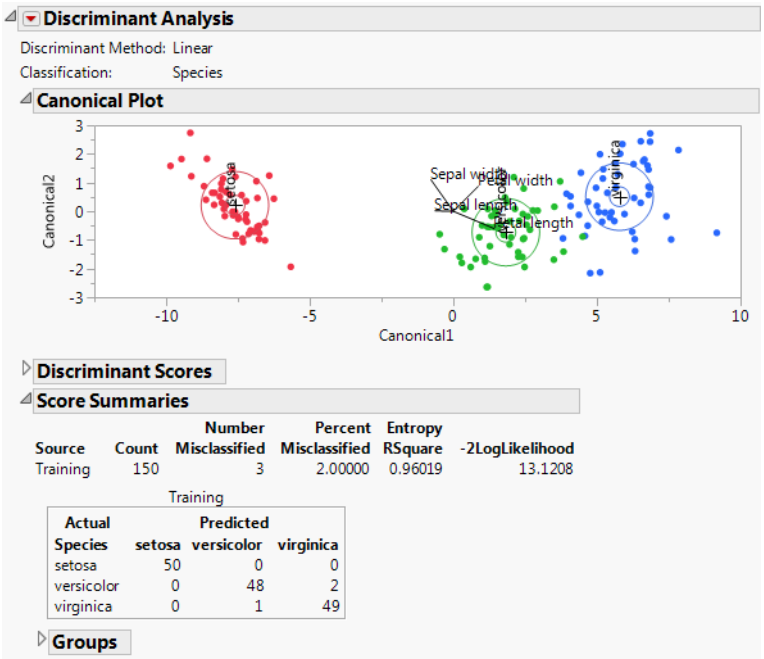
- Linear—Assumes that the within-group covariance matrices are equal. The covariate means for the groups defined by *X* are assumed to differ.
- Quadratic—Assumes that the within-group covariance matrices differ. This requires estimating more parameters than does the Linear method. If group sample sizes 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

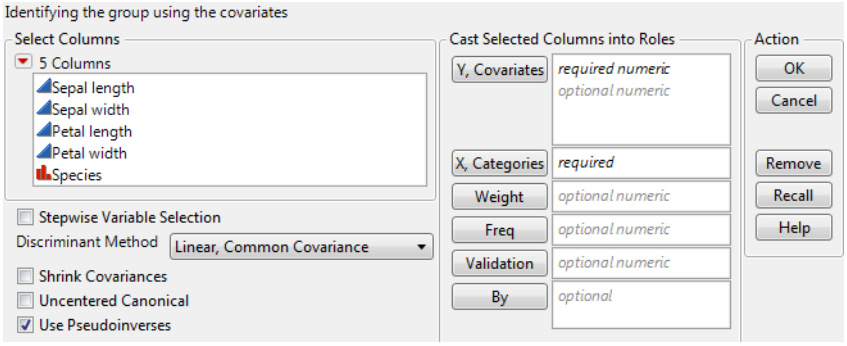


Because there are three classes for Species, there are two canonical variables. In the Canonical Plot, each observation is plotted against the two canonical coordinates. The plot shows that these two coordinates separate the three species. Since there was no validation set, the Score Summaries report 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.

Discriminant Launch Window

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

Figure 5.3 Discriminant Launch Window for Iris.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 105.

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

X, Categories A column containing 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.

JMP PRO **Validation** A numeric column containing two or 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 there are more than three values, all but the smallest three are ignored.

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 Modeling Utilities chapter in the *Predictive and Specialized Modeling* book.

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. For details, see "Stepwise Variable Selection" on page 79.

If you have specified a validation set, statistics that have been calculated for the validation set appear.

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 83.

Shrink Covariances Shrinks the off-diagonal entries of the pooled within-group covariance matrix and the within-group covariance matrices. See "Shrink Covariances" on page 86.

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

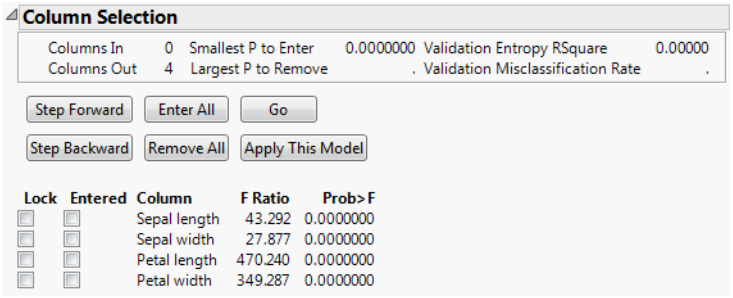
Use Pseudoinverses Uses Moore-Penrose pseudoinverses in the analysis when the covariance matrix is singular. The resulting scores involve all covariates. If left unchecked, the analysis drops covariates that are linear combinations of covariates that precede them in the list of **Y, Covariates**.

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. 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 details about how these are updated, see "Updating the F Ratio and Prob>F" on page 80.

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



Note: The Go button only appears when you use excluded rows for validation in JMP or a validation column in JMP Pro.

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 Number of columns currently selected for entry into the discriminant model.

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

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

Largest P to Remove 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 94. 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 Covariate of interest.

F Ratio *F* ratio for a test for the group variable obtained using an analysis of covariance model. For details, see "Updating the F Ratio and Prob>F" on page 80.

Prob > F *p*-value for a test for the group variable obtained using an analysis of covariance model. For details, see "Updating the F Ratio and Prob>F" on page 80.

Stepwise Example

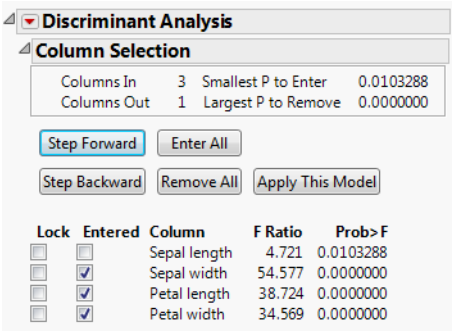
For an illustration of how to use Stepwise, consider the Iris.jmp sample data table.

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

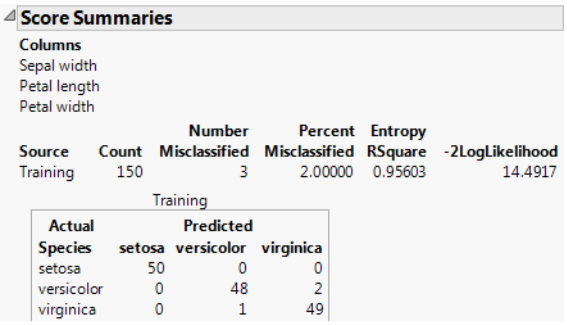


8. Click **Apply This Model**.

   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



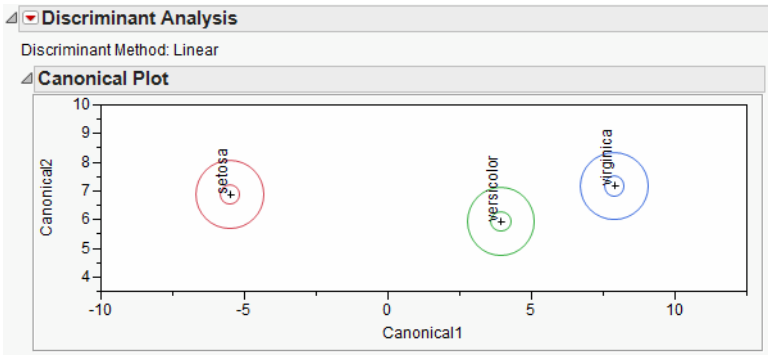
Discriminant Methods

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.

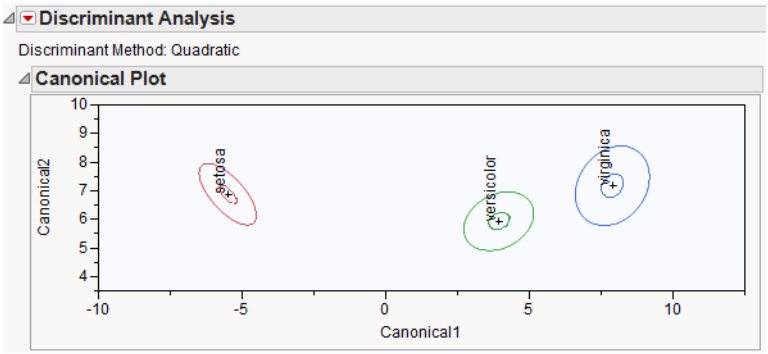
Note: When you enter more than 500 covariates, a JMP Alert recommends that you switch to the Wide Linear method. This is because computation time can be considerable when you use the other methods with a large number of columns. Click **Wide Linear, Many Columns** to switch to the Wide Linear method.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

Figure 5.7 Linear, Quadratic, and Regularized Discriminant Analysis

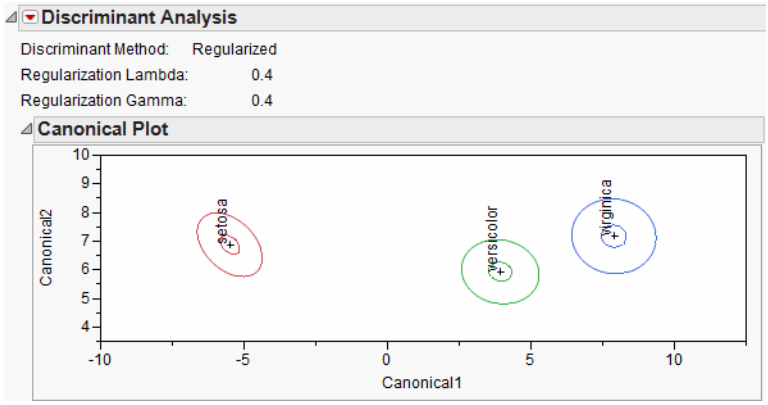
Linear



Quadratic



Regularized (λ=0.4, γ=0.4)



The Linear, Quadratic, and Regularized methods are illustrated in Figure 5.7. The methods are described here briefly. For technical details, see "Saved Formulas" on page 106.

Linear, Common Covariance Performs linear discriminant analysis. This method assumes that the within-group covariance matrices are equal. See"Linear Discriminant Method" on page 107.

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 108.

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 85 and "Regularized Discriminant Method" on page 109.

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 106.

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.

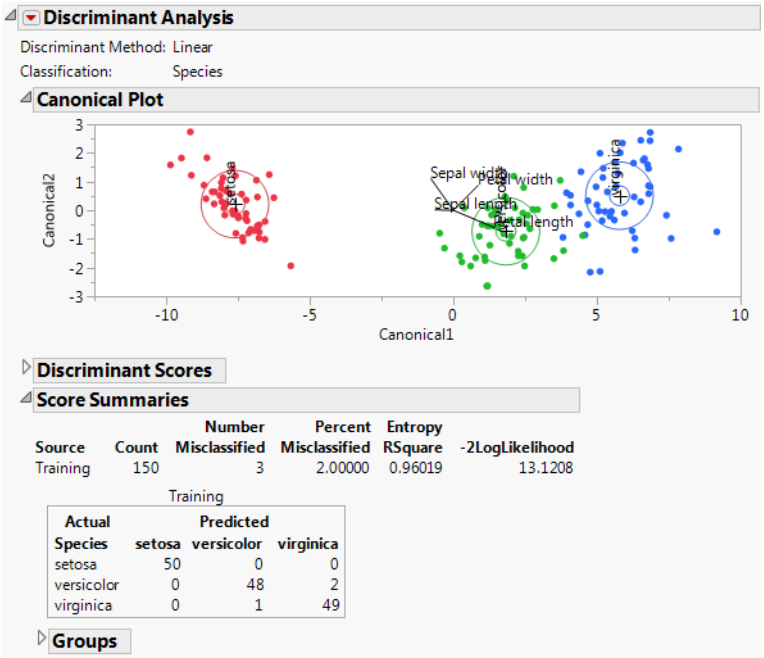
If you select the Shrink Covariances option with the Linear discriminant method in the launch window, this provides a shrinkage of the covariance matrices that is equivalent to the shrinkage provided by the Regularized discriminant method with appropriate Lambda and Gamma values. When you select the 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.

The 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 red triangle menu. The results in the report update to reflect the selected method.

Figure 5.8 Example of a Discriminant Analysis Report



The default Discriminant Analysis report is shown in Figure 5.8 and contains the following sections:

- When you select the Wide Linear discriminant method, a Principal Components report appears. See "Principal Components" on page 87.
- 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 88.
- The Discriminant Scores report provides details about how each observation is classified. See "Discriminant Scores" on page 92.
- The Score Summaries report provides an overview of how well observations are classified. See "Discriminant Scores" on page 92.

Principal Components

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

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

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 $\mathbf{X_s}$, namely $(\mathbf{X_s}'\mathbf{X_s})/(n - p)$, 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 $\mathbf{X_s}'\mathbf{X_s}$.

Singular Value The singular values of $\mathbf{X_s}$ 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

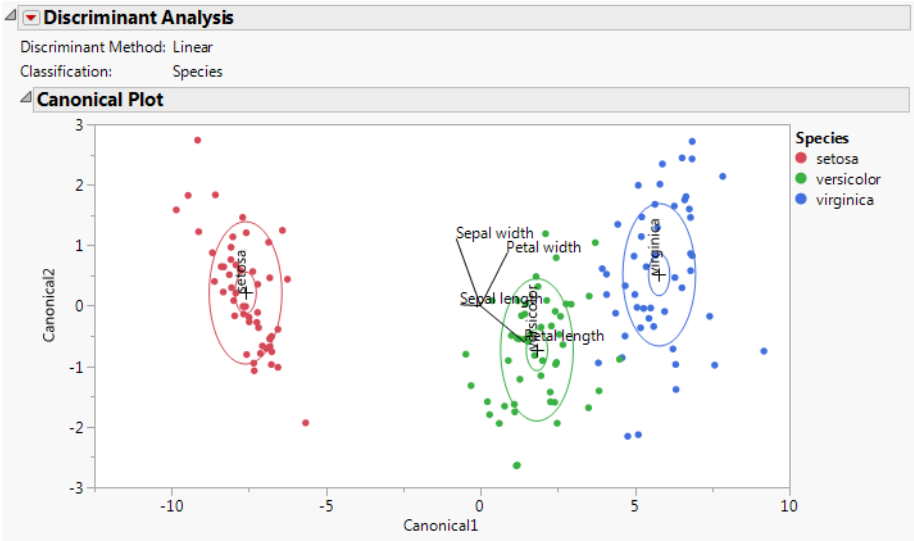
Each of the levels of the X, Categories column defines an indicator variable. A canonical correlation is performed between the set of indicator variables representing the categories and the covariates. Linear combinations of the covariates, called *canonical variables*, are derived. These canonical variables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



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 88.) 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 a covariate's canonical weight, 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 red triangle menu. At the bottom of the Canonical Details report, click Standardized Scoring Coefficients. See "Standardized Scoring Coefficients" on page 101 for details.

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 red triangle menu.
- Show or hide the rays by selecting **Canonical Options > Show Biplot Rays** from the 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 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 red triangle menu.
- Color code the points to match the ellipses by selecting **Canonical Options > Color Points** from the 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:

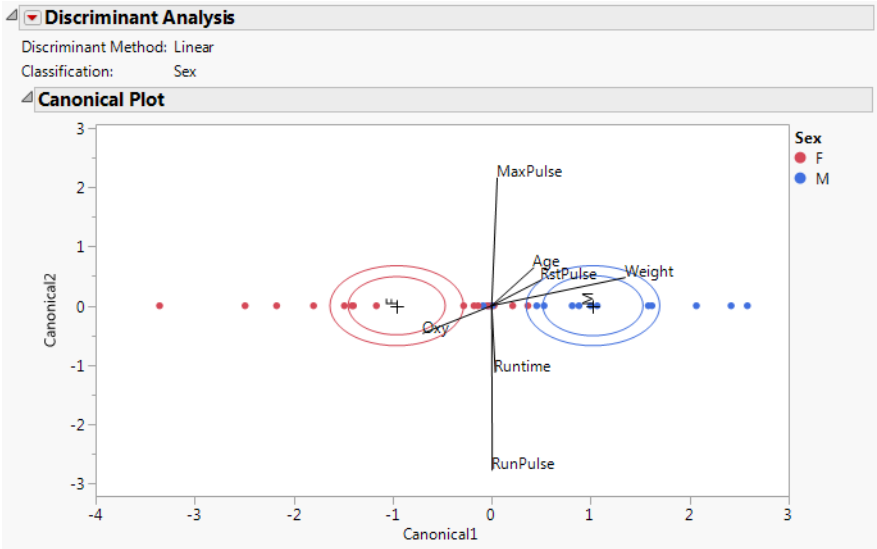
- Petal length is positively associated with Canonical1 and negatively associated with Canonical2. It carries more weight in defining Canonical1 than Canonical2.
- Petal width is positively associated with both Canonical1 and Canonical2. It carries about the same weight in defining both canonical variates.
- 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 Figure 5.10 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. For details, see "Score Options" on page 96.

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	68	3	4.41176	0.94993	7.45767
Validation	38	1	2.63158	0.93027	
Test	44	0	0.00000	0.99384	

Training

Actual Species	Predicted setosa	Predicted versicolor	Predicted virginica
setosa	20	0	0
versicolor	0	22	2
virginica	0	1	23

Validation

Actual Species	Predicted setosa	Predicted versicolor	Predicted virginica
setosa	12	0	0
versicolor	0	12	1
virginica	0	0	13

Test

Actual Species	Predicted setosa	Predicted versicolor	Predicted virginica
setosa	18	0	0
versicolor	0	13	0
virginica	0	0	13

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. See 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.

For details, see "Entropy RSquare" on page 94.

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. Larger values indicate better fit. Provided for the training set only. For more details, see the *Fitting Linear Models* book.

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. For more information, see "Validation in JMP and JMP Pro" on page 105.

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 following commands are available from the Discriminant Analysis red triangle menu.

Stepwise Variable Selection Selects or deselects the Stepwise control panel. See "Stepwise Variable Selection" on page 79.

Discriminant Method Selects the discriminant method. See "Discriminant Methods" on page 83.

Discriminant Scores Shows or hides the Discriminant Scores portion of the report.

Score Options Provides several options connected with the scoring of the observations. In particular, you can save the scoring formulas. See "Score Options" on page 96.

Canonical Plot Shows or hides the Canonical Plot. See "Canonical Plot and Canonical Structure" on page 88.

Canonical Options Provides options that affect the Canonical Plot. See "Canonical Options" on page 98.

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 102.

Specify Priors Enables you to specify prior probabilities for each level of the *X* variable. See "Specify Priors" on page 103.

Consider New Levels Used when you have some points that might not fit any known group, but instead might be from an unscored new group. For details, see "Consider New Levels" on page 103.

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 a Group Means report that provides the means 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 104.

Scatterplot Matrix Opens a separate window containing 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 104.

Not available for the Wide Linear discriminant method.

See the JMP Reports chapter in the *Using JMP* book 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, showing actual by predicted counts, in the Score Summaries report. By default, the Score Summaries report

shows a confusion 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. For more information, see "Validation in JMP and JMP Pro" on page 105.

Show Distances to Each Group Adds a report called Squared Distances to Each Group that shows each observation's squared Mahalanobis distance to each group mean.

Show Probabilities to Each Group Adds a report called Probabilities to Each Group that shows the probability that an observation belongs to each of the groups defined by the categorical X.

ROC Curve Appends a Receiver Operating Characteristic curve to the Score Summaries report. For details about the ROC Curve, see the Partition chapter in the *Predictive and Specialized Modeling* book.

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. The default is to define as uncertain any row whose probability differs from 0 or 1 by more than 0.1. 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. For details, see "Saved Formulas" on page 106.

- The distance formulas are SqDist[0] 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 details about the Response Probability column property, see the *Using JMP* book.
- 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 111.

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 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. (Available only in JMP Standard.)

JMP PRO **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 the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

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 Canonical 3D Plot.

Show Means CL Ellipses Shows or hides 95% confidence ellipses for the means on the canonical variables, assuming normality. Shows or hides 95% confidence ellipsoids in the Canonical 3D Plot.

Show Normal 50% Contours Shows or hides an ellipse or an ellipsoid that denotes a 50% contour for each group. In the Canonical Plot, each ellipse depicts a region in the space of the first two canonical variables that contains approximately 50% of the observations, assuming normality. In the Canonical 3D Plot, each ellipsoid depicts a region in the space of the first three canonical variables that contains approximately 50% of the observations, assuming normality.

Show Biplot Rays Shows or hides the biplot rays in the Canonical Plot and in 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 99.

Show Canonical Structure Shows or hides Canonical Structures report. See "Show Canonical Structure" on page 101. 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 within 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*

▷ Within Matrix
▷ Between Matrix
▷ Scoring Coefficients
▷ Standardized Scoring Coefficients

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 113 and "Approximate F-Tests" on page 114 in the "Discriminant Analysis" appendix.

Approx. F *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 114 in the "Discriminant Analysis" appendix.

NumDF Numerator degrees of freedom for the corresponding test.

DenDF Denominator degrees of freedom for the corresponding test.

Prob>F *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 115.

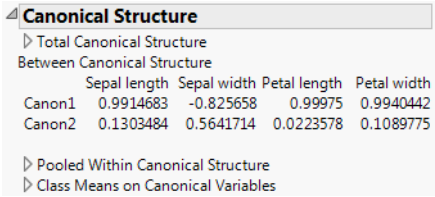
Scoring Coefficients Coefficients used to compute canonical scores in terms of the raw data. These are the coefficients used for the option **Canonical Options > Save Canonical Scores**. For details about how these are computed, see "The CANDISC Procedure" in SAS Institute Inc. (2011).

Standardized Scoring Coefficients Coefficients used to compute canonical scores in terms of the standardized data. Often called *canonical weights*. For details about how these are computed, see "The CANDISC Procedure" in SAS Institute Inc. (2011).

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


Canonical Structure

▷ Total 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

▷ Pooled Within Canonical Structure

▷ Class Means on Canonical Variables

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

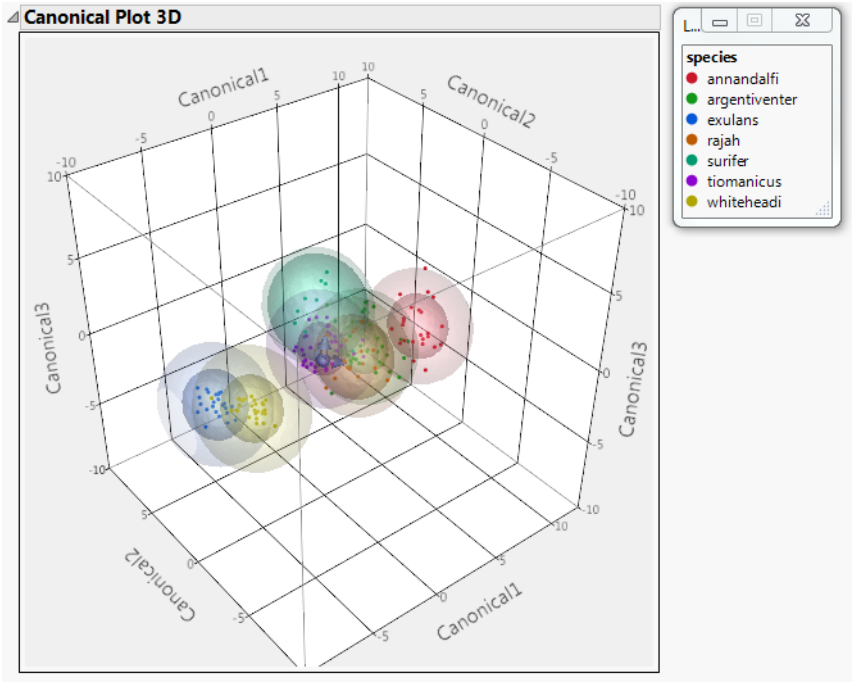
1. Select **Help > Sample Data Library** and open Owl Diet.jmp.
2. Select rows 180 through 294.

   These are the rows for which species is missing. You will 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. Select **Canonical 3D Plot** from the Discriminant Analysis red triangle menu.

Tip: Click on 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.

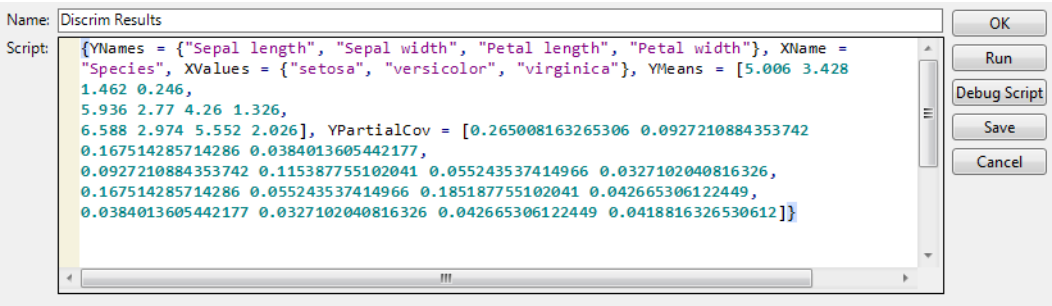
Save Discrim Matrices

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 red triangle menu, the script **Discrim Results** is saved to the data table. The script is shown in Figure 5.16.

Figure 5.16 Discrim Results Table Script for Iris.jmp



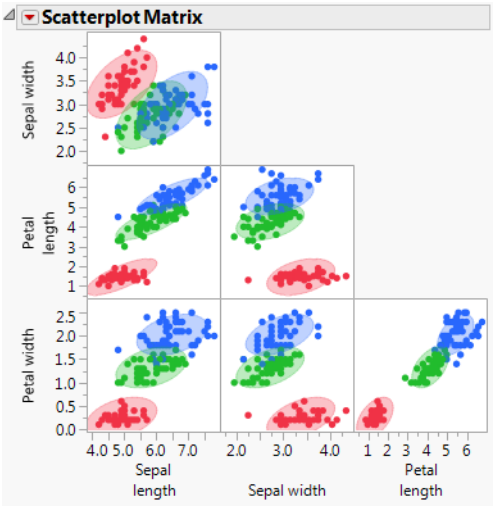
Note: In a script, you can send the scripting command Get Discrim Matrices to the Discriminant platform object. This obtains the same values as Save Discrim Matrices, but does not store them in the data table.

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 report's red triangle menu are described in the *Essential Graphing* book.

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. For details, see "Statistics" on page 80 and "Entropy RSquare for Validation and Test Sets" on page 94.
- 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.

Technical Details

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)
$t = 1, ..., T$	subscript to distinguish groups defined by X

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

n_t	number of observations in group t
$n = n_1 + n_2 + ... + n_T$	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 $\mathbf{y}$ in group t is estimated as follows:

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

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 given 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 $\mathbf{y}$ 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 109.

See Table 5.2 for notation. The Mahalanobis distance from an observation $\mathbf{y}$ to group t is defined as follows:

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

The likelihood for an observation $\mathbf{y}$ in group t is estimated as follows:

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

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

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

An observation $\mathbf{y}$ 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*>]	$d_t^2 + \log\|\mathbf{S}_t\| - 2\log(q_t)$
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:

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

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

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

The likelihood for an observation $\mathbf{y}$ in group t is estimated as follows:

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

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

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

An observation $\mathbf{y}$ 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>]	$d_t^2 + \log\left\|\mathbf{\Sigma}_t\right\| - 2\log(q_t)$
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.

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 steps in the Wide Linear calculation are as follows:

1. Compute the T by p matrix $\mathbf{M}$ of within-group sample means. The $(t,j)^{\text{th}}$ entry of $\mathbf{M}$, 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 as follows:
   - 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 $\mathbf{R}$. The matrix $\mathbf{R}$ 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 $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\mathbf{D}$ is a diagonal matrix with positive entries (the singular values) on the diagonal. See "The Singular Value Decomposition" on page 226 in the "Statistical Details" appendix.

   Then $\mathbf{R}$ 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 $\mathbf{R}$ 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 $\mathbf{D}$.

   If $\mathbf{R}$ is not of full rank, define a pseudo-inverse for $\mathbf{R}$ as follows:

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

Then define the inverse square root of $\mathbf{R}$ as follows:

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

9. If $\mathbf{R}$ 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 107. 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 $\mathbf{y}$, 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 113:

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

Saved Formulas

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

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 112.
Prob[<group *t*>]	$p(t\|\mathbf{y})$, given in "Linear Discriminant Method" on page 107
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 $\mathbf{M}'$ 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

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

- Hotelling-Lawley Trace

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

- Roy's Max Root

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

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

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

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

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 $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$, where the **L** matrix identifies elements of $\mathbf{X}'\mathbf{X}$ 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 115, 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

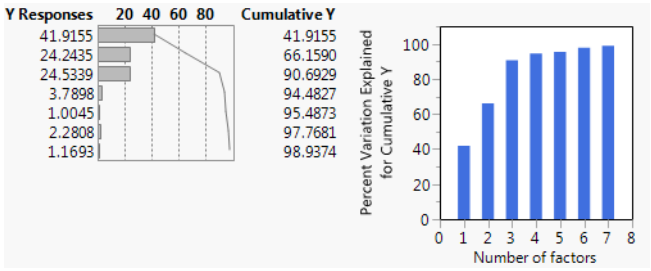
Develop Models Using Correlations between Ys and Xs

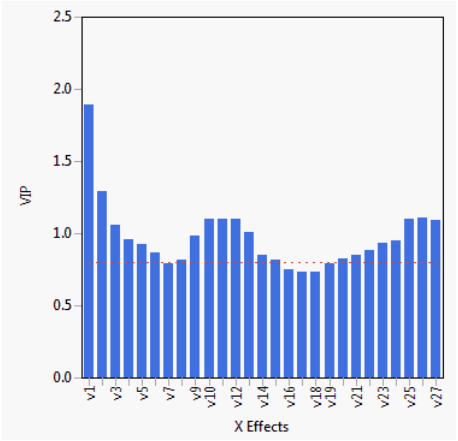
The Partial Least Squares (PLS) platform fits linear models based on factors, namely, linear combinations of the explanatory variables (*X*s). These factors are obtained in a way that attempts to maximize the covariance between the *X*s and the response or responses (*Y*s). PLS exploits the correlations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

JMP PRO JMP Pro provides additional functionality, allowing you to conduct PLS Discriminant Analysis (PLS-DA), include a variety of model effects, utilize 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





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 (1995), 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 NIPALS, see Wold, H., 1980; for 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 validation and cross validation methods.
- Impute missing data.
- Obtain bootstrap estimates of the distributions of various statistics. Right-click in the report of interest. For more details, see the *Basic Analysis* book.

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**.

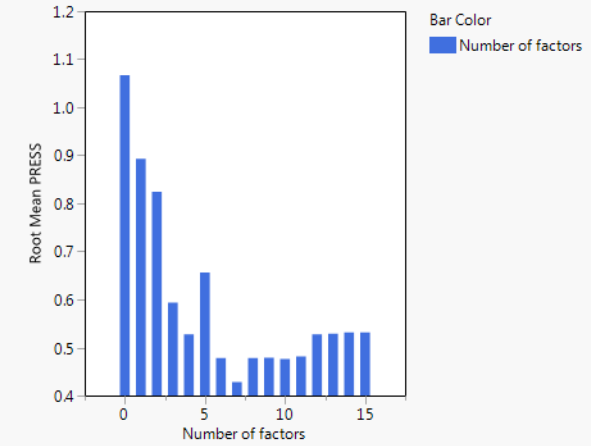
   A portion of the report appears in Figure 6.2. Since the van der Voet test is a randomization test, your Prob > van der Voet T^2 values can differ slightly from those in Figure 6.2.

Figure 6.2 Partial Least Squares Report

Leave-One-Out Cross Validation With Method=NIPALS

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	<.0001*	-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	0.0050*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0100*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0310*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3070	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	0.4050	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2520	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822873	0.1170	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.1970	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2230	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1540	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1540	0.816103	1.000000	0.000000	1.000000	0.000000	1.000000

Root Mean PRESS Plot

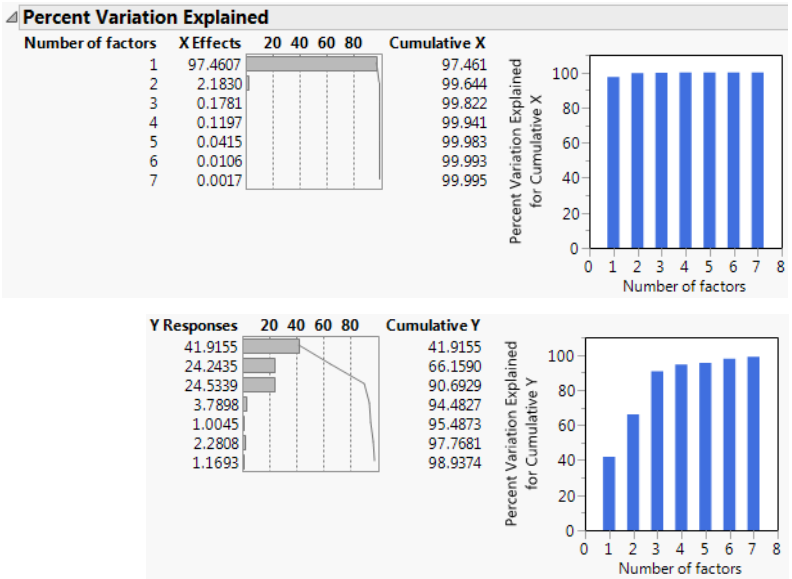


Note: The minimum root mean PRESS is 0.42814 and the minimizing number of factors is 7.

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** 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, 2011 and 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.

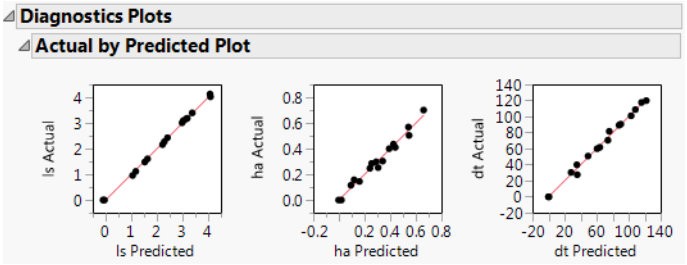
Figure 6.3 Seven Extracted Factors



8. Select **Diagnostics Plots** from the NIPALS Fit with 7 Factors red triangle menu.

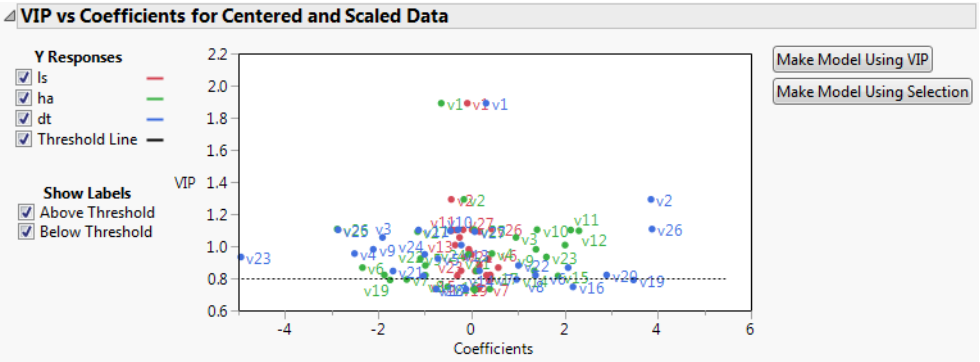
   This gives a report showing actual by predicted plots and three reports showing various residual plots. The Actual by Predicted Plot in Figure 6.4 shows the degree to which predicted compound amounts agree with actual amounts.

Figure 6.4 Diagnostics Plots



9. Select **VIP vs Coefficients Plot** from the NIPALS Fit with 7 Factors red triangle menu.

Figure 6.5 VIP vs Coefficients Plot



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**.
- JMP PRO 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.

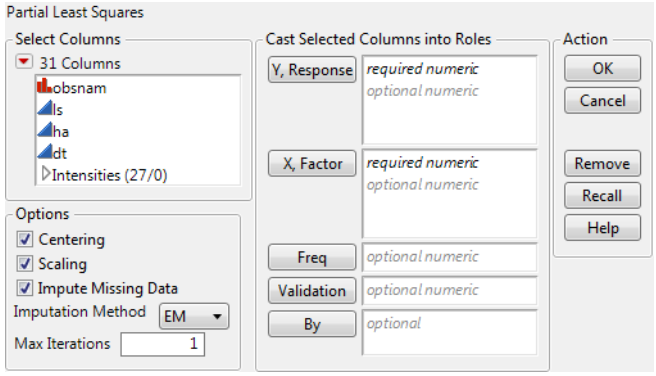
Some features on the Fit Model launch window are not applicable for the Partial Least Squares personality:

- 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, Scheffé Cubic, and Radial.

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



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.

JMP PRO In JMP Pro, you can enter nominal response columns in the Fit Model launch window to conduct PLS-DA. For details, see "PLS Discriminant Analysis (PLS-DA)" on page 144.

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

JMP PRO 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.

JMP PRO **Validation** Enter an optional validation column. A validation column must contain only consecutive integer values. Note the following:

- 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 126.

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, see *Basic Analysis*.

By Enter a column that creates separate reports for each level of the variable.

Centering Centers all Y variables and model effects by subtracting the mean from each column. See "Centering and Scaling" on page 125.

Scaling Scales all Y variables and model effects by dividing each column by its standard deviation. See "Centering and Scaling" on page 125.

JMP PRO **Standardize X** (Fit Model launch window only) Select this option to center and scale 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 125.

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 details 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 125.

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

JMP PRO 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.

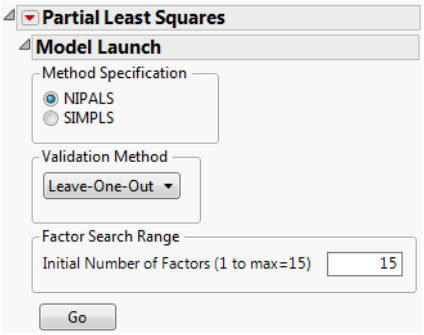
If the Standardize X option is not selected, and **Centering** and **Scaling** are both selected, then the term that is entered into the model is calculated as follows:

$$\frac{X1 \times X2 - mean(X1 \times X2)}{std(X1 \times X2)}$$

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 6.7 Partial Least Squares Model Launch Control Panel



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" on page 139 for details 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.

JMP PRO **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.

JMP PRO **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 Specify 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.

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.

Partial Least Squares Report

The first time you click **Go** in the Model Launch control panel (Figure 6.7), 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 Method> and Method = <Method Specification>
- 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 6.8 Model Comparison Summary

Model Comparison Summary

Method	Number of rows	Number of factors	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
NIPALS	16	6	99.993471	97.768092	22
NIPALS	16	7	99.995152	98.937438	22

In Figure 6.8, models for 7 and then 6 factors have been fit. The report includes the following summary information:

Method Shows the analysis method that you specified in the Model Launch control panel.

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, 1995 and Eriksson et al., 2006).

<Cross Validation Method> and Method = <Method Specification>

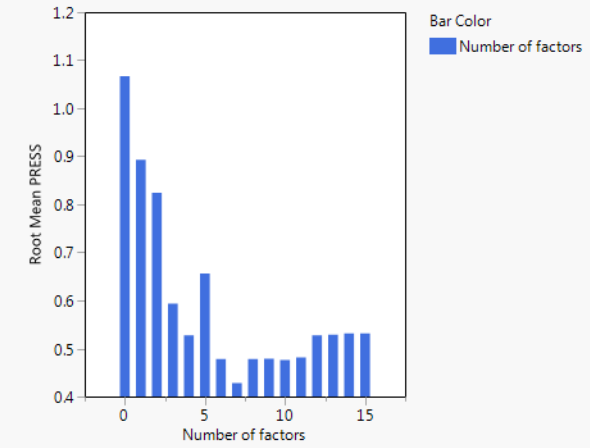
This report appears only when a form of cross validation is selected as a Validation Method in the Model Launch control panel. It shows summary statistics for models fit, using from 0 to the maximum number of extracted factors, as specified in the Model Launch control panel. The report also provides a plot of Root Mean PRESS values. See "Root Mean PRESS Plot" on page 131. An optimum number of factors is identified using the minimum Root Mean PRESS statistic.

Figure 6.9 Cross Validation Report

Leave-One-Out Cross Validation With Method=NIPALS

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	<.0001*	-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	0.0050*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0100*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0310*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3070	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	0.4050	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2520	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822873	0.1170	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.1970	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2230	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1540	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1540	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 Number of factors used in fitting the model.

Root Mean PRESS The square root of the average of the PRESS values across all responses. For details, see "Root Mean PRESS" on page 131.

van der Voet T^2 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. For more details, see "van der Voet T^2" on page 140.

Prob > van der Voet T^2 *p*-value for the van der Voet T^2 test. For more details, see "van der Voet T^2" on page 140.

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:

$$1 - PRESS/SSY$$

For details see "Calculation of Q^2" on page 131.

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:

$$1 - \prod_{i=1}^{f} (PRESS_i/SSY_i)$$

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 and R^2Y When Validation Is Used" on page 132.

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 and R^2Y When Validation Is Used" on page 132.

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

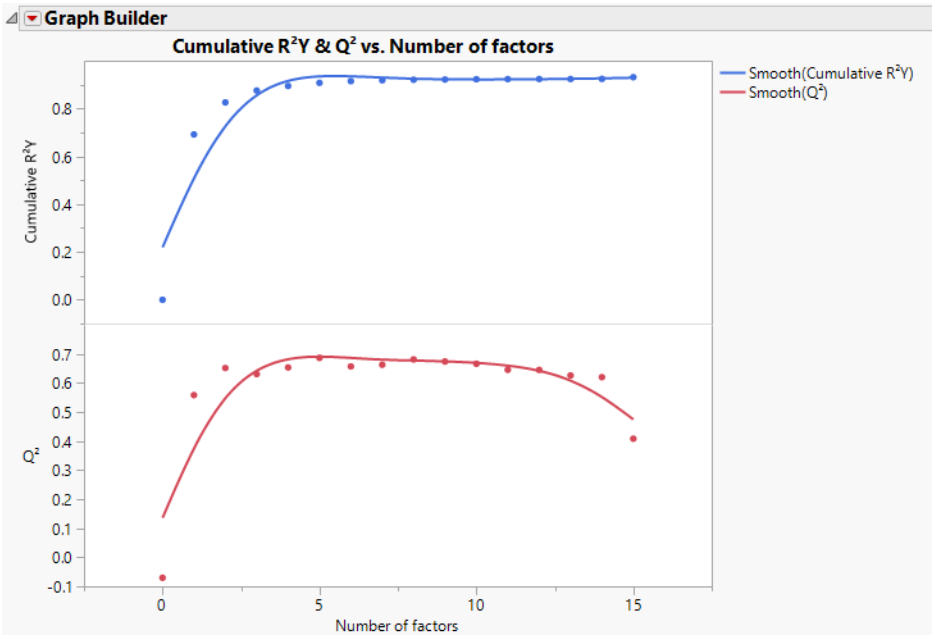
The statistics Q^2 and Cumulative R^2Y both measure the predictive ability of the model, but in different ways.

- 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 6.10 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 will explain a large portion of the variation in Y without overfitting the data.

Figure 6.10 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. See Figure 6.9.

Root Mean PRESS

For a specified number of factors, *a*, Root Mean PRESS is calculated as follows:

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 calculated as follows. 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.

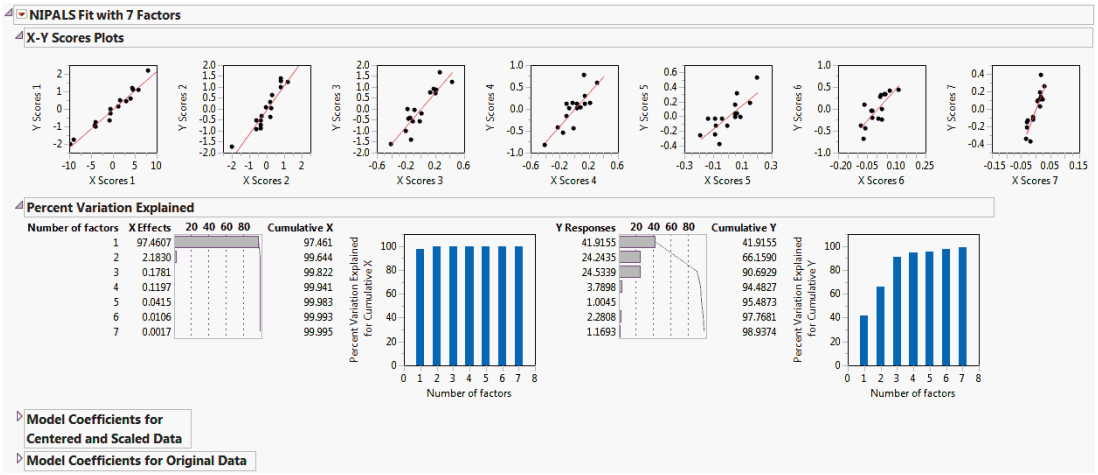
KFold R^2X is the average of the Percent Variation Explained for X Effects for the K models constructed by leaving out each fold.

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 and gives the number of extracted factors.

Figure 6.11 Model Fit Report



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.

Partial Least Squares Options

The Partial Least Squares red triangle menu contains the following options:

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.

See the JMP Reports chapter in the *Using JMP* book 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.

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 135.

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 136.

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. These should be used only for outlier detection.

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 141.

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 details about the computation of T^2 and the control limit, see "T^2 Plot" on page 141.

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 137.

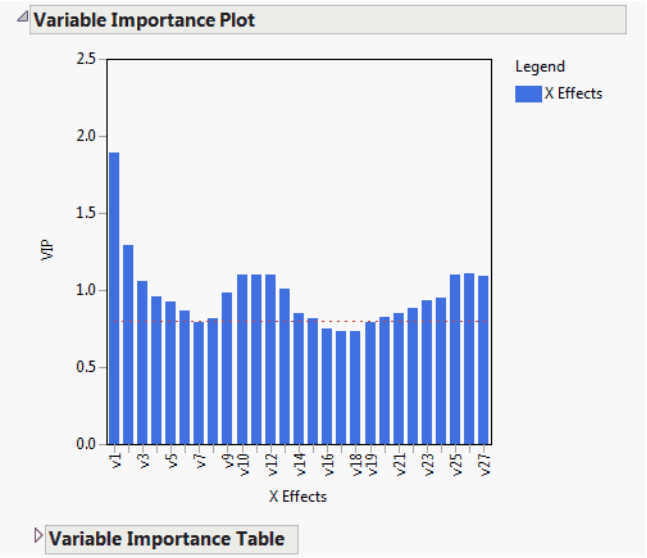
Remove Fit Removes the model report from the main platform report.

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 136.

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,1995). 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 6.12 Variable Importance Plot



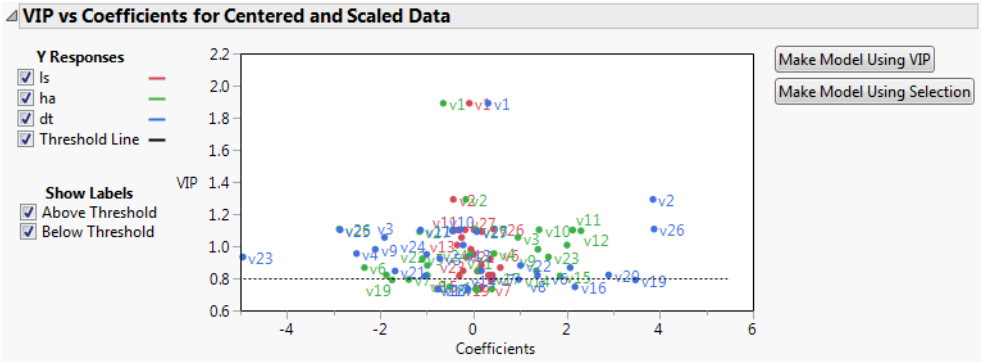
VIP vs Coefficients Plots

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 on the role button and the selected columns are populated.

Figure 6.13 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 its 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. For details, see "Standard Error of Prediction and Confidence Limits" on page 142.

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. For details, see "Standard Error of Prediction and Confidence Limits" on page 142.

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. For details, see "Standard Error of Prediction and Confidence Limits" on page 142.

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 139.

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 143.

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 143.

Save T Square Saves the T^2 values to the data table. These are the values 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.

JMP PRO **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.

JMP PRO **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.

JMP PRO **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 the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

JMP PRO **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 the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

Statistical Details

This section provides details about some of the methods used in the Partial Least Squares platform. For additional details, see Hoskuldsson (1988), Garthwaite (1994), or Cox and Gaudard (2013).

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 **t** is called a *score vector* and **w** is its associated *weight vector*. The PLS method predicts both $\mathbf{X}_0$ and $\mathbf{Y}_0$ by regression on **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 **p** and **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:

$$D_{i,jk} = R^2_{i,jk} - R^2_{opt,jk}$$

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:

$$C_i = \mathbf{d}_{i,.}{}' \mathbf{S}_{\mathbf{i}}^{-1} \mathbf{d}_{i,.}$$

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:

$$T_i^2 = (n-1) \sum_{j=1}^{p} \left(t_{ij}^2 / \sum_{k=1}^{n} t_{kj}^2 \right)$$

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, Young, and Mason, 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 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, Young, and Mason, 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:

$$SE(\bar{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + \mathbf{t}_0(\mathbf{T'T})^{-1}\mathbf{t}_0{}'\right)}$$

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:

$$SE(\hat{Y}_{x_0}) = s\sqrt{\left(\frac{1}{n} + 1 + \mathbf{t}_0(\mathbf{T'T})^{-1}\mathbf{t}_0{}'\right)}$$

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:

$$\overline{Y}_{x_0} \pm t_{0.975,\,df} SE(\hat{Y}_{x_0})$$

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 Ys.

Chapter 7

Hierarchical Cluster

Group Observations Using a Tree of Clusters

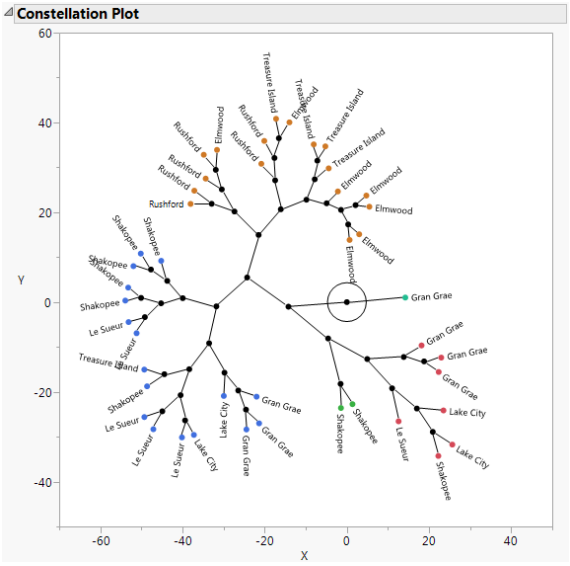
Clustering is a multivariate technique that groups together observations that share similar values across a number of variables. Use it to understand the clumping structure of your data.

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 7.1 Example of a Constellation Plot



Hierarchical Cluster Overview

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 146.

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, defining distances as follows:

- 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 n-dimensional space, but rather they 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 213.

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. A separate Robust Normal Mixtures option is an alternative to the Normal Mixture with uniform outlier cluster.

  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 7.1 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

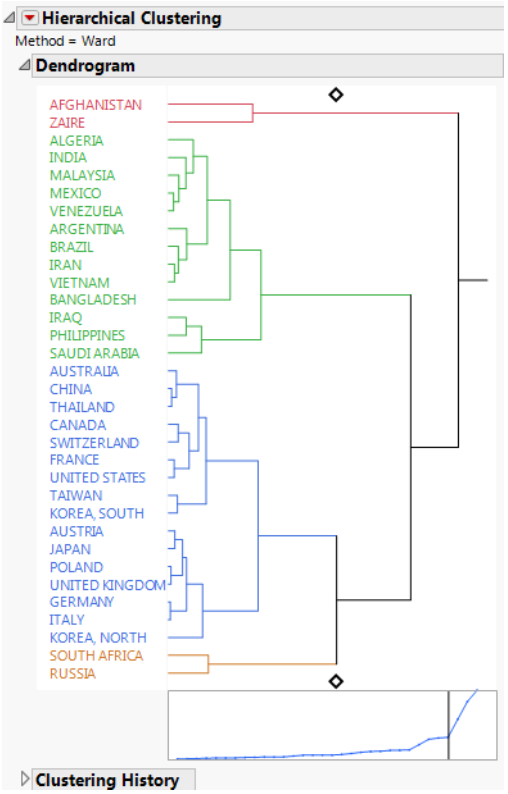
Example of Clustering

In this example, we group together countries by their 1976 crude birth and death rates per 100,000 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 red triangle next to Hierarchical Clustering and select **Color Clusters**.

Figure 7.2 Hierarchical Clustering Report



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.

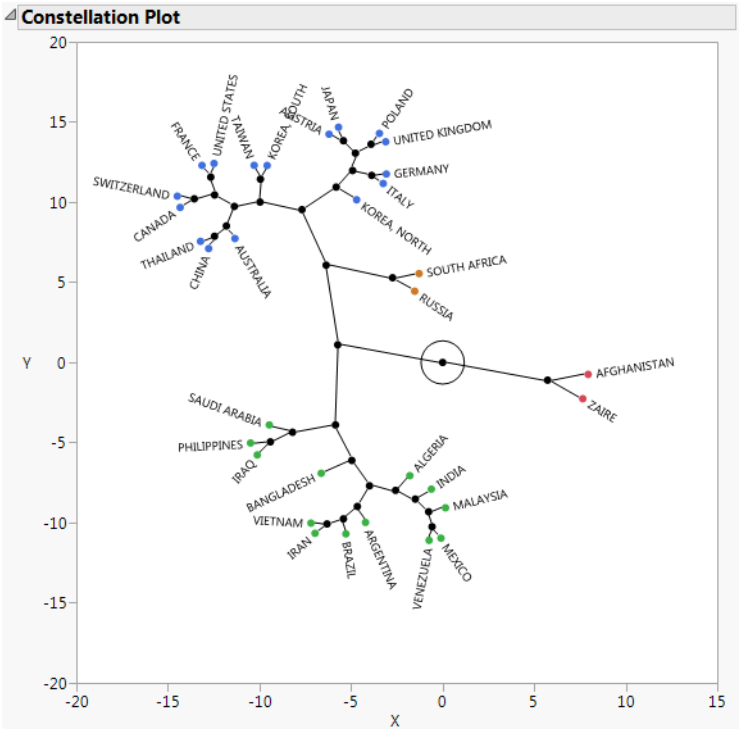
In the dendrogram, the relative distances between clusters are given by the horizontal distances between vertical lines that join the clusters. For example, Afghanistan and Zaire differ more than Malaysia differs from the cluster consisting of Mexico and Venezuela.

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 on 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 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 red triangle next to Hierarchical Clustering and select **Constellation Plot**.

Figure 7.3 Constellation Plot

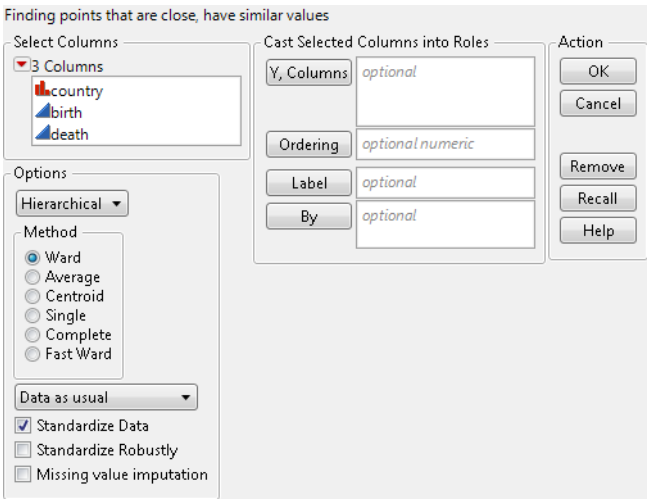


This constellation plot arranges the countries as endpoints and each cluster join as a new point. The lines represent membership in a cluster. The length of a line between cluster joins approximates the distance between the clusters that were joined. The constellation plot indicates that the cluster 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**. The Clustering launch window for the Birth Death Subset.jmp data table is shown in shown in Figure 7.4.

Figure 7.4 Hierarchical Cluster Launch Dialog



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.

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

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 171.

Normal Mixtures See the "Normal Mixtures" chapter on page 187.

Robust Normal Mixtures . See "Normal Mixtures NCluster=<k> Report" on page 195 in the "Normal Mixtures" chapter.

Self Organizing Map See"Self Organizing Map" on page 183 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 168.

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 (1976), 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 160.

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 162.

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 155.

– 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 164.

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 usual**, 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, and 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. For more details, see the Modeling Utilities chapter in the *Predictive and Specialized Modeling* book.

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 which spatial components to add measures for circle, pie, and streak spatial measures to aid in clustering defect patterns. This is a specialty method and is applicable in only very specific settings. See "Spatial Measures" on page 166 and "Example of Wafer Defect Classification Using Spatial Measures" on page 164.

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

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 158.

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 on any cluster stem to select all the members of the cluster in the dendrogram and in the data table.

Distance Graph

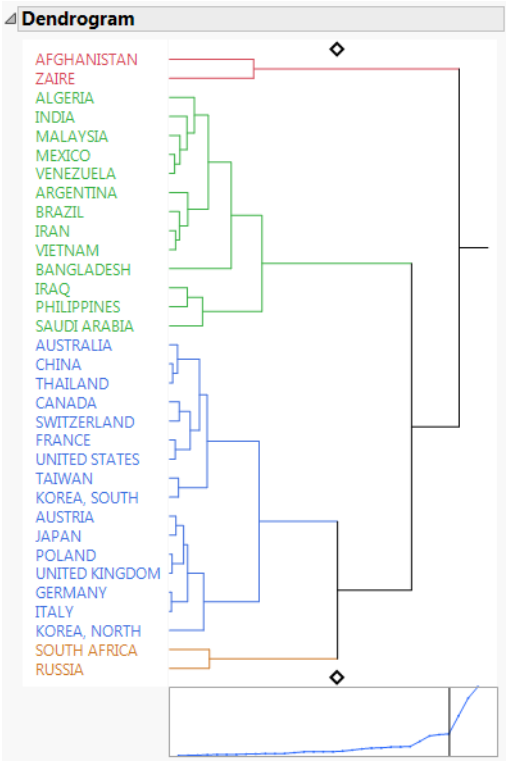
The Distance Graph is the plot that appears beneath the dendrogram. This graph has a point for each step where two clusters are joined into a single cluster. The horizontal coordinates represent the numbers of clusters, which decrease from left to right. The vertical coordinate of the point is the distance 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 148.

Figure 7.5 Dendrogram Report for Birth Death Subset.jmp



In Figure 7.5, 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 151.

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 includes 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 (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 Arranges the individuals as endpoints and each cluster join as a new point, with lines drawn that represent membership. The longer lines represent greater distances between clusters. To turn off the displayed labels, 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 as follows:

– 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 160.

Cluster Treatment Comparisons (Available only if you hold Shift and click the red triangle.) Select a response column and a two-level treatment column. Creates a Hierarchically Clustered Differences report.

See the JMP Reports chapter in the *Using JMP* book 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

The section contains two examples:

Example of a Distance Matrix

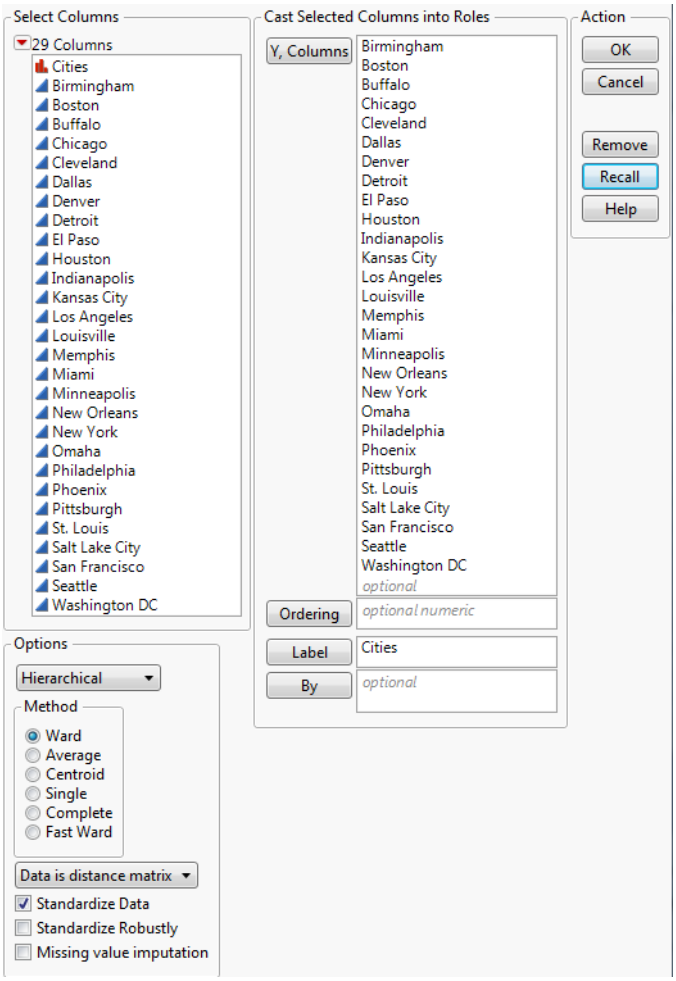
The proper data table structure for a distance matrix consists of the following:

- 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 Cites and click **Label**.
5. Select all remaining columns and click **Y, Columns**.

Figure 7.6 Completed Distance Matrix Launch Window



6. Click **OK**.
7. Click the red triangle next to Hierarchical Clustering and select **Color Clusters**.

Figure 7.7 Dendrogram Report for Flight Distances

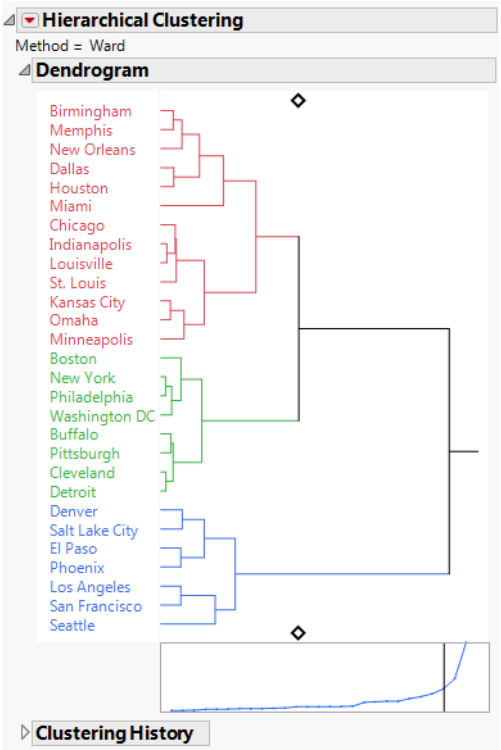


Figure 7.7 shows the Dendrogram report for the flight distances. The placement of the diamonds indicates that the model has grouped the cities into three clusters, which are color-coded on the dendrogram. For more details about how to interpret the report, see "Dendrogram Report" on page 155.

Example of Wafer Defect Classification Using Spatial Measures

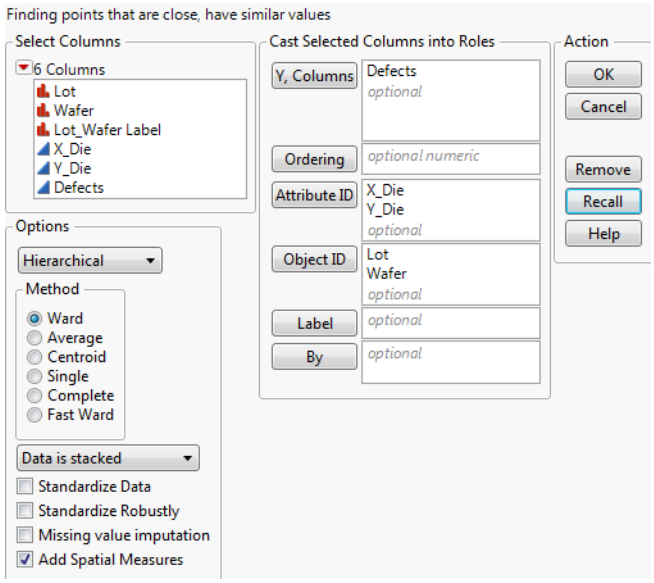
A specialty clustering option called Spatial Measures is available in the Hierarchical Cluster platform. In this example, you use this option to cluster wafers. For details about the option, see "Spatial Measures" on page 166.

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**.

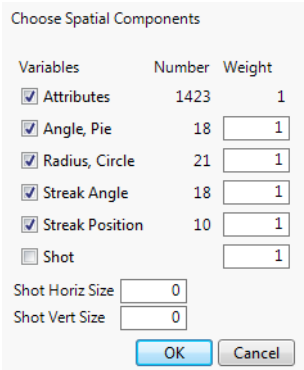
6. Select Lot and Wafer and click **Object ID**.
7. Select **Add Spatial Measures** from the list of options in the lower left corner.

Figure 7.8 Completed Clustering Launch Window



8. Click **OK**.

Figure 7.9 Spatial Components Window



Because Defects is measured at 1423 locations, there are 1423 Attributes variables.

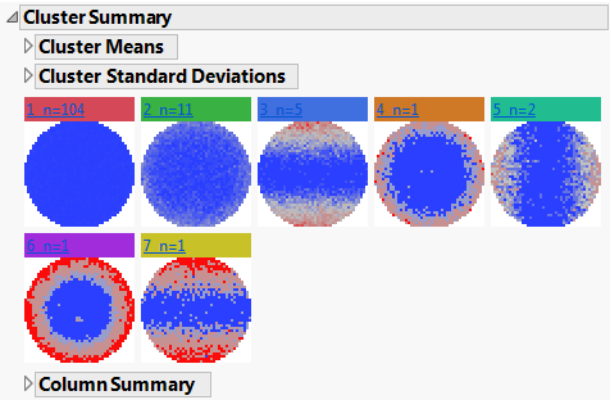
9. Click **OK** to accept the selections in the Spatial window.

   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 red triangle next to Hierarchical Clustering and select **Number of Clusters**.
12. Enter 7 and click **OK**.
13. Click the red triangle next to Hierarchical Clustering and select **Cluster Summary**.

Figure 7.10 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. For more details, see "Spatial Measures" on page 166.

Statistical Details

The following sections provide statistical details for the Hierarchical Clustering platform.

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 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 164.

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 add and weight variables that reflect spatial patterns to your cluster analysis.

Variables The types of variables that are constructed and used in the cluster analysis. The variables are constructed using the response, Y.

Attributes The value of the Y variable calculated for each location, 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.

Shot The variables that identify which rectangle an object is in, where you specify the numbers 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 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. (How is the weight used in the clustering algorithm?)

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 151.

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,..., 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 $\left\| x_i - x_j \right\|^2$

Average Linkage 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 Distance for the centroid method of clustering is:

$$D_{KL} = \left\| \overline{x_K} - \overline{x_L} \right\|^2$$

Ward's Distance for Ward's method is:

$$D_{KL} = \frac{\left\| \overline{x_K} - \overline{x_L} \right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

Single Linkage Distance for the single linkage cluster method is:

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} d(x_i, x_j)$$

Complete Linkage Distance for the Complete linkage cluster method is:

$$D_{KL} = \max_{i \in C_K} \max_{j \in C_L} d(x_i, x_j)$$

Chapter 8

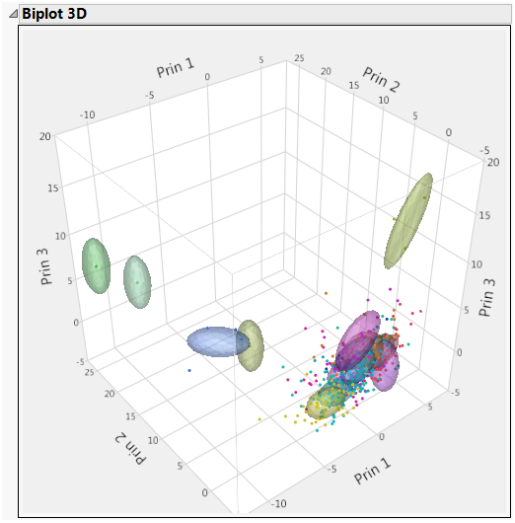
K Means Cluster

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 8.1 3D Biplot



K Means Cluster Platform Overview

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 172.

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 *n* 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 SAS Institute Inc. (2005) 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 *n*-dimensional space, but rather they 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 213.

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. A separate Robust Normal Mixtures option is an alternative to the Normal Mixture with uniform outlier cluster.

  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 8.1 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

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**.

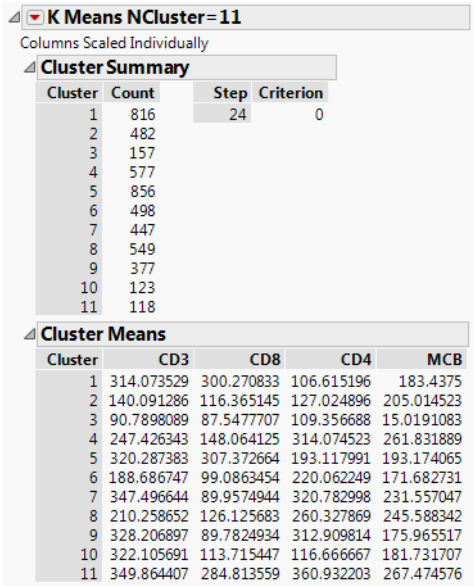
Figure 8.2 Cluster Comparison Report

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 8.3 K Means NCluster=11 Report



K Means NCluster=11

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	816	24	0
2	482		
3	157		
4	577		
5	856		
6	498		
7	447		
8	549		
9	377		
10	123		
11	118		

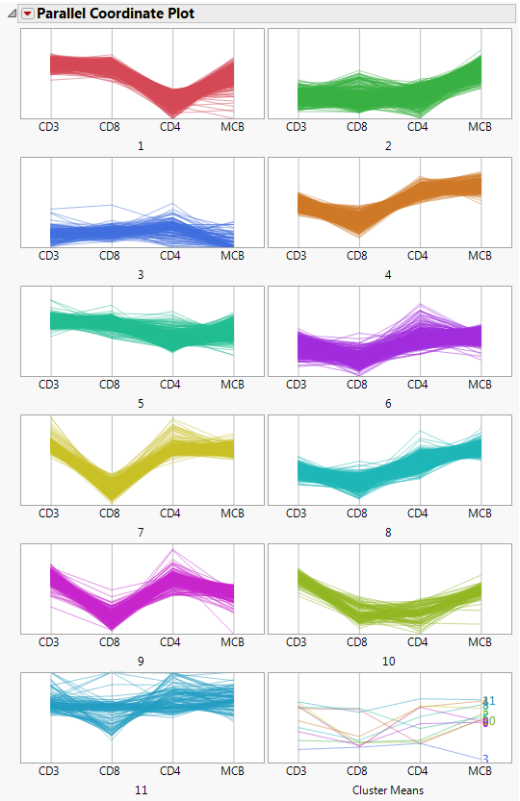
Cluster Means

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**.

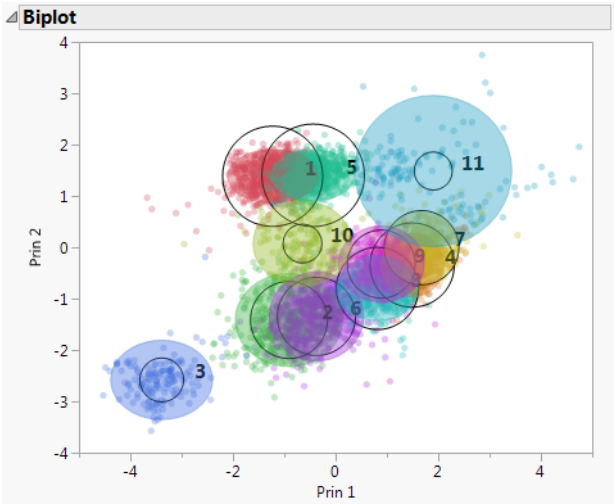
Figure 8.4 Parallel Coordinate Plots for Cytometry Data



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 8.5 Biplot for Cytometry Data

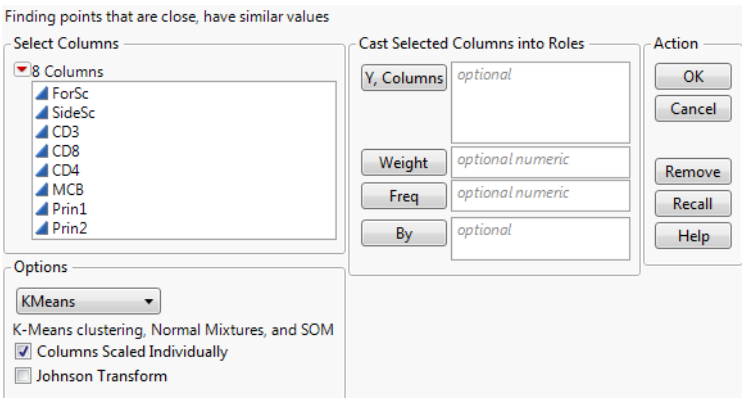


The clusters that appear to be most separated from the others based on their first two principal components are clusters 3, 10, and 11. This is supported by their parallel coordinate plots in Figure 8.4, which differ from the plots for the other clusters.

Launch the K Means Cluster Platform

Launch the K Means Cluster platform by selecting **Analyze > Clustering > K Means Cluster**. Figure 8.6 shows the Clustering launch window for Cytometry.jmp.

Figure 8.6 K Means Cluster Launch Window



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

KMeans is the default clustering method, but you have the option to select Hierarchical or Normal Mixtures. If you select KMeans or Normal Mixtures, when you click OK, a Control Panel appears. See "Iterative Clustering Control Panel" on page 179.

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.

Johnson Transform Fits a Johnson family distribution to the Y variables. If Columns Scaled Individually is selected, a separate Johnson transformation is fit to each Y variable. If Columns Scaled Individually is not selected, a single Johnson transformation is fit to values for all the Y variables. Two of the Johnson families (Sb and Su) are considered and the fitting method uses maximum likelihood.See the Distributions chapter in the *Basic Analysis* book.

The Johnson transformations attempt to normalize the data, mitigating skewness, and pulling outliers in toward the center of the distribution.

Iterative Clustering Report

When you click OK in the launch window, the Iterative Clustering report window appears, showing:

- A Transformations report. (Shown only if Johnson Transformation option is selected.) The Transformations report contains a row for each variable entered in Y, Columns. For each variable, the best-fit Johnson distribution is reported from the Sb or Su family, as well as the estimated values for the parameters γ, δ, θ, and σ. For more information about the Johnson distributions, see the Distributions chapter in the *Basic Analysis* book.
- A Control Panel for fitting models. See "Iterative Clustering Control Panel" on page 179.

As you fit models, additional reports are added to the window. See "K Means NCluster=<k> Report" on page 181.

Iterative Clustering Options

Save Transformed (Available only if the Johnson Transformation option is selected.) Saves the transformed data as new formula columns in the data table.

See the JMP Reports chapter in the *Using JMP* book 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.

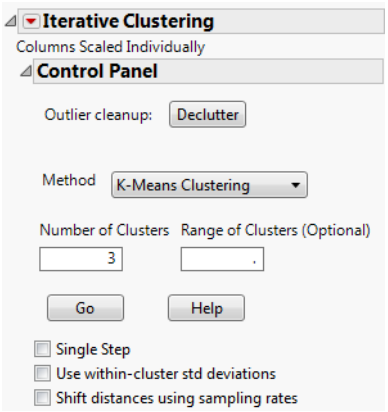
Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

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 8.7. You can interactively fit different numbers of clusters or you can specify a range using the Range of Clusters option.

Figure 8.7 Iterative Clustering Control Panel



The Control Panel has the following options:

Declutter Locates multivariate outliers. Enter the maximum number of neighbors to examine. When you click OK, plots appear giving distances between each point and that point's nearest neighbor, second nearest neighbor, up to the maximum nearest neighbor specified.

Caution: The total number of nearest neighbors specified is denoted using a small *k*. This number is not directly connected to the number of clusters, also sometimes denoted with a small *k*.

The following options appear beneath the nearest neighbor plots:

Reset Excluded Specifies that excluded rows be treated as excluded in future cluster fits within the given report. For example, if you exclude rows based on nearest neighbor plots or the scatterplot matrix, you must select this option to exclude these rows from cluster fits performed within the report using the Control Panel.

Scatterplot Matrix Creates a scatterplot matrix in a separate window that uses all of the Y variables.

Save NN Distances Saves the distances from each row to its k^{th} nearest neighbor as new columns in the data table.

Close Removes the Declutter report.

Method The following clustering methods are available:

KMeans Clustering Described in this chapter.

Normal Mixtures Described in the "Iterative Clustering Control Panel" on page 193 in the "Normal Mixtures" chapter.

Robust Normal Mixtures Described in the "Normal Mixtures NCluster=<k> Report" on page 195 in the "Normal Mixtures" chapter.

Self Organizing Map Described in "Self Organizing Map Control Panel" on page 183.

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 NCluster=<k> Report

When you click Go in the Control Panel, the following reports appear:

- A Cluster Comparison report. See "Cluster Comparison Report" on page 181.
- One or more K Means NCluster=<k> reports, where *k* is 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 8.2 and "K Means NCluster=11 Report" on page 175.

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 NCluster=<k> Report

The K Means NCluster=<k> 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 NCluster=<k> Report Options

Each K Means NCluster=<k> report contains the following options:

Biplot Shows a plot of the points and clusters in the first two principal components of the data. Circles are drawn around the cluster centers. The size of the circles is proportional to the count inside the cluster. The shaded area is the 50% density contour around the mean, and indicates where 50% of the observations in that cluster would fall (Mardia et al., 1980). Below the plot is an option to save the cluster colors to the data table. The eigenvalues are shown in decreasing order.

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.

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. For details, see the Parallel Plots chapter in the *Essential Graphing* book.

Scatterplot Matrix Creates a scatterplot matrix using all of the Y variables in a separate window.

Save Colors to Table Assigns colors that identify the clusters to the rows of the data table.

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 a scaled 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. The square root of the sum is given as the Distance. If Johnson Transform is selected, Distance is calculated in transformed units.

Save Cluster Formula Saves a formula column called Cluster Formula to the data table. This is the formula that identifies cluster membership for each.

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

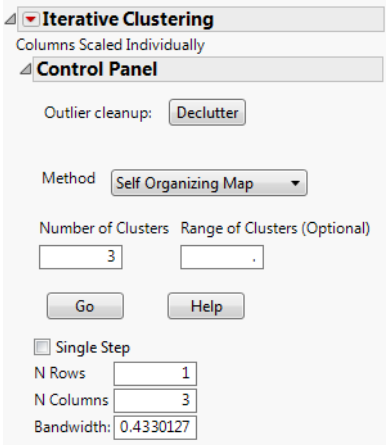
The *Self-Organizing Map* (SOM) technique was developed by Teuvo Kohonen (1989, 1990) and extended by other neural network enthusiasts and statisticians. The original SOM was cast as a learning process, like the original neural net algorithms, but the version implemented here is a variation on *k*-means 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 184.

Self Organizing Map Control Panel

Select the Self Organizing Map option from the Method list in the Iterative Clustering Control Panel (Figure 8.8).

Figure 8.8 Self Organizing Map Control Panel



Some of the options on the panel are described in "Iterative Clustering Control Panel" on page 179. The other options are described as follows:

Number of Clusters Not applicable for SOM.

Range of Clusters (Optional) Not applicable for SOM.

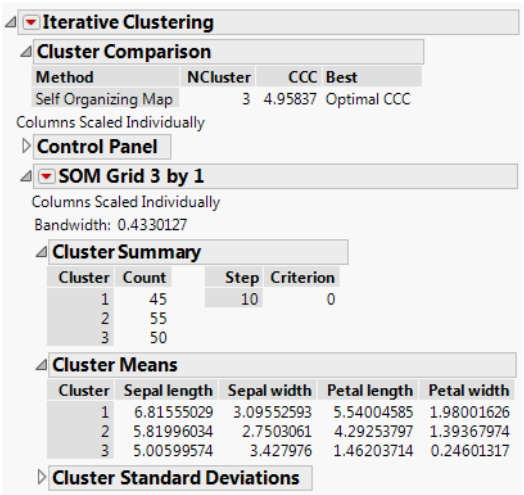
N Rows The number of rows in the cluster grid.

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 8.9 Self Organizing Map Report



Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
Self Organizing Map	3	4.95837	Optimal CCC

Columns Scaled Individually

Control Panel

SOM Grid 3 by 1

Columns Scaled Individually

Bandwidth: 0.4330127

Cluster Summary

Cluster	Count	Step	Criterion
1	45	10	0
2	55		
3	50		

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

Cluster Standard Deviations

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 NCluster=<k> Report" on page 181.

For details about the red-triangle options for Self Organizing Maps, see "K Means NCluster=<k> Report Options" on page 181.

Description of SOM Algorithm

The SOM implementation in JMP proceeds as follows:

- 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.

Chapter 9

Normal Mixtures

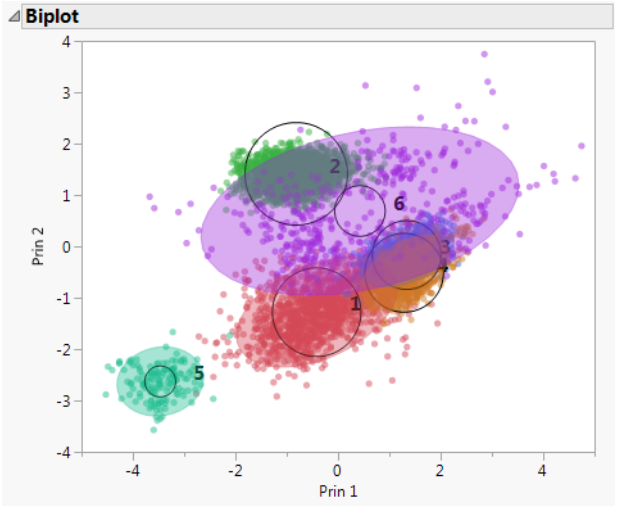
Group Observations Using Probabilities

Use Normal Mixtures for clustering when your data come from overlapping normal distributions. If you have multivariate outliers, use the Robust Normal Mixtures method. 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 9.1 Normal Mixtures Biplot



Normal Mixtures Clustering Platform Overview

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 188.

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 you can select the Robust Normal Mixtures method. The outlier cluster option assumes a uniform distribution and the Robust Normal Mixtures option uses a robust method for estimating the parameters. These methods are less sensitive to outliers than the Normal Mixtures method. For details, see "Outlier Cluster" on page 195 and "Additional Details for Robust Normal Mixtures" on page 199.

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 *n*-dimensional space, but rather they 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 213.

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. A separate Robust Normal Mixtures option is an alternative to the Normal Mixture with uniform outlier cluster.

  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 9.1 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

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 9.2 Normal Mixtures NCluster=6 Report

Iterative Clustering

Cluster Comparison

Method	NCluster	BIC	AICc	Best
Normal Mixtures	6	208033	207456	Smallest BIC Smallest AICc

Columns Scaled Individually

Control Panel

Normal Mixtures NCluster=6

Cluster Summary

Cluster	Count	Proportion
1	944	0.18462
2	1602	0.31663
3	1194	0.24344
4	147	0.02932
5	393	0.08556
6	720	0.14044

Cluster Means

Cluster	CD3	CD8	CD4	MCB
1	233.76885	140.185266	298.685852	256.78374
2	317.251078	306.005523	150.869012	189.618348
3	173.952765	109.762903	187.953147	195.550558
4	87.8647425	86.2327269	107.791878	10.1460043
5	336.442204	193.303126	238.447762	206.502571
6	338.616551	86.6610652	315.558822	208.347735

Cluster Standard Deviations

-LogLikelihood	BIC	AICc
103637.52	208033.07	207456.3

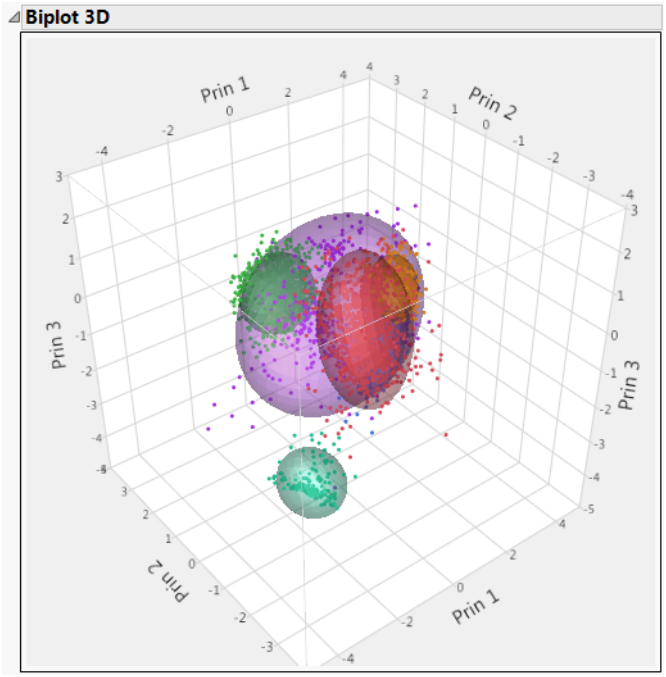
Correlations for Normal Mixtures

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 9.3 3D Biplot of Cytometry Data

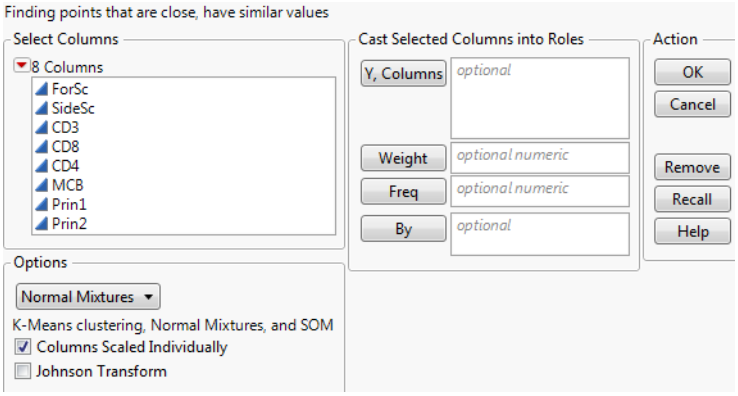


The plot shows contours for the normal densities that are fit to the clusters. Note that one cluster appears to be distinctly separated from the other clusters based on the first three principal components.

Launch the Normal Mixtures Clustering Platform

Launch the Normal Mixtures Clustering platform by selecting Analyze > Clustering > Normal Mixtures. The Clustering launch window in Figure 9.4 uses the Cytometry.jmp sample data table.

Figure 9.4 Normal Mixtures Launch Window



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.

Options

Normal Mixtures is the default clustering method, but you have the option to select Hierarchical or KMeans. If you select KMeans or Normal Mixtures, when you click OK, a Control Panel appears. See "Iterative Clustering Control Panel" on page 193.

Columns Scaled Individually 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.

Johnson Transform Fits a Johnson family distribution to each variable entered in Y, if Columns Scaled Individually is selected. Two of the Johnson families (Sb and Su) are considered and the fitting method uses maximum likelihood. See the Distributions chapter in the *Basic Analysis* book.

The Johnson transformations attempt to normalize the data, mitigating skewness and pulling outliers in toward the center of the distribution.

Note: If you select Johnson Transform but do not select Columns Scaled Individually, a single Johnson transformation is fit to values for all the variables entered in Y.

Iterative Clustering Report

When you click OK in the launch window, the Iterative Clustering report window opens, showing:

- A Transformations report. (Only shown if Johnson Transformation option is selected.) The Transformations report contains a row for each variable entered in Y, Columns. For each variable, the best-fit Johnson distribution from the Sb or Su family is reported, as well as the estimated values for the parameters γ, δ, θ, and σ. For more information about the Johnson distributions, see the Distributions chapter in the *Basic Analysis* book.
- A Control Panel for fitting models. See "Iterative Clustering Control Panel" on page 193. As you fit models, additional reports are added to the window. See "Normal Mixtures NCluster=<k> Report" on page 195.

Iterative Clustering Options

Save Transformed (Only available if the Johnson Transformation option is selected.) Saves the transformed data as new columns in the data table.

See the JMP Reports chapter in the *Using JMP* book 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.

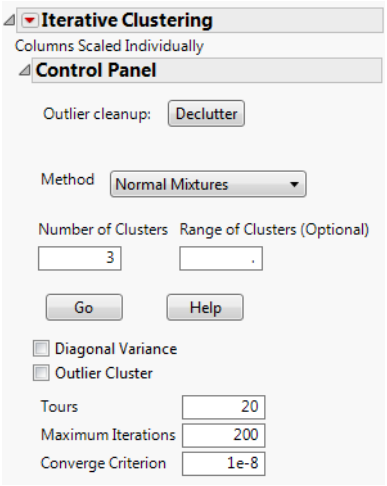
Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

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

Iterative 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 9.5. 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 9.5 Control Panel for Normal Mixtures Method



The Normal Mixtures Control Panel has these options:

Declutter Locates multivariate outliers. Enter the maximum number of neighbors to examine. When you click OK, plots appear giving distances between each point and that point's nearest neighbor, second nearest neighbor, up to the maximum nearest neighbor specified.

Caution: The total number of nearest neighbors specified is denoted using a small *k*. This number is not directly connected to the number of clusters, also sometimes denoted with a small *k*.

The following options appear beneath the nearest neighbor plots:

Reset Excluded Tells JMP to treat excluded rows as excluded in future cluster fits within the given report. For example, if you exclude rows based on nearest neighbor plots or the scatterplot matrix, you must select this option to exclude these rows from cluster fits performed within the report using the Control Panel.

Scatterplot Matrix Creates a scatterplot matrix in a separate window that uses all of the Y variables.

Save NN Distances Creates new columns in the data table that give the distances from each row to its k^{th} nearest neighbor.

Close Removes the Declutter report.

Method The available clustering methods:

KMeans Clustering Described in the "Iterative Clustering Control Panel" on page 179 in the "K Means Cluster" chapter.

Normal Mixtures Described in this chapter.

Robust Normal Mixtures Described in this chapter. See"Normal Mixtures NCluster=<k> Report" on page 195.

Self Organizing Map Described in "Self Organizing Map Control Panel" on page 183 in the "K Means Cluster" chapter.

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.

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 NCluster=<k> Report

When you click Go in the Control Panel, the following reports appear:

- A Cluster Comparison report. See "Cluster Comparison Report" on page 196
- One or more Normal Mixtures NCluster=<k> reports, where *k* is 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 inFigure 9.2 and "Normal Mixtures NCluster=6 Report" on page 190.

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. The Cluster Comparison report does not provide a fit statistic for Robust Normal Mixture fits.

Normal Mixtures NCluster=<k> Report

The Normal Mixtures NCluster=<k> 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 loglikelihood, BIC, and AICc.
- The Correlations for Normal Mixtures report gives the estimated correlation matrix for each cluster

Normal Mixtures NCluster=<k> Report Options

Biplot Shows a plot of the points and clusters in the first two principal components of the data. Circles are drawn around the cluster centers. The size of the circles is proportional to the count inside the cluster. The shaded area is the 50% density contour around the mean, and indicates where 50% of the observations in that cluster would fall (Mardia et al., 1980). Below the plot is an option to save the cluster colors to the data table. 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 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.

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.For details, see the Parallel Plots chapter in the *Essential Graphing* book.

Scatterplot Matrix Creates a scatterplot matrix in a separate window, using all the variables.

Save Colors to Table Assigns colors that identify the clusters to the rows of the data table.

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. (Not available for Robust Normal Mixtures.)

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 (Not available for Robust Normal Mixtures) 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.

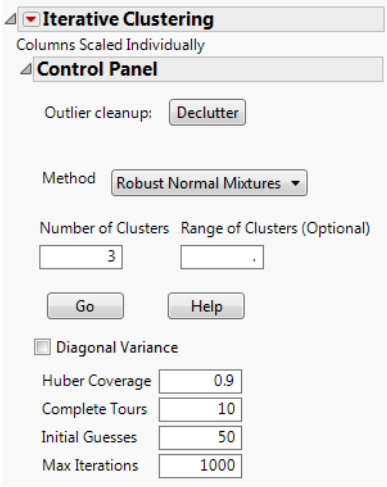
Robust Normal Mixtures

Because Normal Mixtures clustering is sensitive to outliers, an outlier-robust alternative called Robust Normal Mixtures is provided. This method uses a robust method of estimating the normal parameters. JMP computes the estimates via maximum likelihood with respect to a mixture of Huberized normal distributions. Huberized normal distributions are multivariate normal distributions that are constructed to be outlier resistant. For details, see "Additional Details for Robust Normal Mixtures" on page 199.

Robust Normal Mixtures Control Panel

Select the Robust Normal Mixtures method from the Method menu of the Iterative Clustering Control Panel (Figure 9.5).

Figure 9.6 Control Panel for Robust Normal Mixtures Method



Several of the options on the panel are described in "Iterative Clustering Control Panel" on page 193. The remaining options are described as follows:

Huber Coverage A number between 0 and 1. Robust Normal Mixtures protects against outliers by down-weighting them. Huber Coverage loosely reflects the proportion of the observations that are not considered outliers, and not down-weighted. Values closer to 1 result in a larger proportion of the data *not* being down-weighted. In other words, values closer to 1 protect only against the most extreme outliers. Values closer to 0 result in a larger proportion of the data being down-weighted, and might falsely consider less extreme data points to be outliers. For details, see "Additional Details for Robust Normal Mixtures" on page 199.

Complete Tours The number of independent restarts of estimation process. Each restart has a different starting value. Independent starts help guard against finding local solutions.

Initial Guesses Number of random starts within each tour. Random starting values for the parameters are used for each new start.

Max Iterations Maximum number of iterations during the convergence stage.

Robust Normal Mixture Reports

Robust Normal Mixture reports have the same structure and options as Normal Mixture reports. The reports have titles of the form Robust Normal Mixtures NCluster=<k>. See "Normal Mixtures NCluster=<k> Report" on page 195.

Note: The Cluster Comparison report does not provide a fit statistic for Robust Normal Mixture fits.

Statistical Details for the Normal Mixtures Method

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.

Additional Details for Robust Normal Mixtures

For the Robust Normal Mixtures method, JMP computes estimates via maximum likelihood with respect to a mixture of Huberized normal distributions. This is a class of modified normal distributions that was developed to be more outlier resistant than the normal distribution.

The Huberized Gaussian distribution has pdf $\Phi_k(x)$.

$$\Phi_k(x) = \frac{\exp(-\rho(x))}{c_k}$$

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} & \text{if } |x| \le k \\ k|x| - \dfrac{k^2}{2} & \text{if } |x| > k \end{cases}$$

$$c_k = \sqrt{2\pi}[\Phi(k) - \Phi(-k)] + 2\frac{\exp(-k^2/2)}{k}$$

So, in the limit as k becomes arbitrarily large, $\Phi_k(x)$ tends toward the normal PDF. As $k \to 0$, $\Phi_k(x)$ tends toward the exponential (Laplace) distribution.

The regularization parameter k is set so that P(Normal(x) < k) = Huber Coverage, where Normal(x) indicates a multivariate normal variate. Huber Coverage is a user field, which defaults to 0.90.

Chapter 10

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 10.1 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	0.28596	0.3764	0.6236	0.4163	0.5837	0.1555	0.2303	0.6142	0.0009	0.2838	0.7154	0.0187	0.9629	0.0185
Cluster 2	0.25892	0.4067	0.5933	0.6742	0.3258	0.0258	0.0562	0.9180	0.0030	0.2948	0.7022	0.6436	0.0359	0.3205
Cluster 3	0.20696	0.6794	0.3206	0.7524	0.2476	0.5625	0.1376	0.2999	0.0026	0.9918	0.0056	0.8010	0.1952	0.0038
Cluster 4	0.19717	0.4706	0.5294	0.9922	0.0078	0.8534	0.0827	0.0640	0.4958	0.2047	0.2996	0.7651	0.0020	0.2329
Cluster 5	0.05099	0.1840	0.8160	0.0329	0.9671	0.8476	0.1468	0.0056	0.7761	0.1176	0.1063	0.4364	0.0956	0.4680

Cluster	Overall	sex	marital status	country	size	type
Cluster 1	0.28596					
Cluster 2	0.25892					
Cluster 3	0.20696					
Cluster 4	0.19717					
Cluster 5	0.05099					

Latent Class Analysis Platform Overview

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 appetite 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 a subset of 12 of these 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.
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**.

The Fit Group outline contains three Latent Class Analysis reports. The reports fit models for three, four, and five clusters.

5. Click the red triangle next to Fit Group and select **Order by Goodness of Fit**.

   Because it has the smallest BIC value of the three models, the model with five clusters now appears first in the Fit Group report.

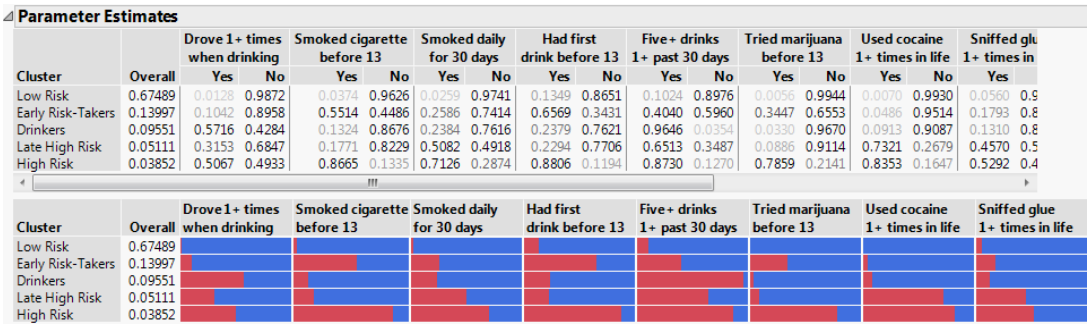
6. In the Latent Class Analysis for 5 Clusters report, examine the bar charts under Parameter Estimates. Note the following:
   - Cluster 1 has mostly No answers to all of the risk behaviors.
   - Cluster 2 has high numbers of Yes answers for the four risk behaviors before the age of 13.
   - Cluster 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.

7. Click the red triangle next to Latent Class Analysis 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.
8. Click **OK**.
9. Click **OK** in the JMP Alert that appears.

Note: The new cluster names are not saved to scripts.

Figure 10.2 Partial Parameter Estimates Report



Parameter Estimates

Cluster	Overall	Drove 1+ times when drinking		Smoked cigarette before 13		Smoked daily for 30 days		Had first drink before 13		Five+ drinks 1+ past 30 days		Tried marijuana before 13		Used cocaine 1+ times in life		Sniffed glu 1+ times in
		Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Low Risk	0.67489	0.0128	0.9872	0.0374	0.9626	0.0259	0.9741	0.1349	0.8651	0.1024	0.8976	0.0056	0.9944	0.0070	0.9930	0.0560
Early Risk-Takers	0.13997	0.1042	0.8958	0.5514	0.4486	0.2586	0.7414	0.6569	0.3431	0.4040	0.5960	0.3447	0.6553	0.0486	0.9514	0.1793
Drinkers	0.09551	0.5716	0.4284	0.1324	0.8676	0.2384	0.7616	0.2379	0.7621	0.9646	0.0354	0.0330	0.9670	0.0913	0.9087	0.1310
Late High Risk	0.05111	0.3153	0.6847	0.1771	0.8229	0.5082	0.4918	0.2294	0.7706	0.6513	0.3487	0.0886	0.9114	0.7321	0.2679	0.4570
High Risk	0.03852	0.5067	0.4933	0.8665	0.1335	0.7126	0.2874	0.8806	0.1194	0.8730	0.1270	0.7859	0.2141	0.8353	0.1647	0.5292

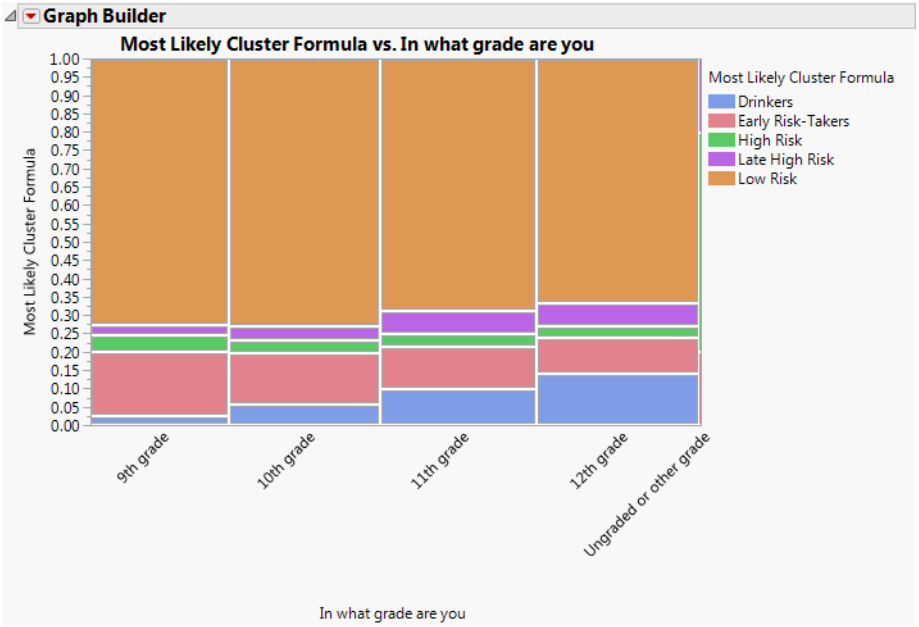
Cluster	Overall	Drove 1+ times when drinking	Smoked cigarette before 13	Smoked daily for 30 days	Had first drink before 13	Five+ drinks 1+ past 30 days	Tried marijuana before 13	Used cocaine 1+ times in life	Sniffed glue 1+ times in life
Low Risk	0.67489								
Early Risk-Takers	0.13997								
Drinkers	0.09551								
Late High Risk	0.05111								
High Risk	0.03852								

Figure 10.2 shows parameter estimates for the first 8 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".

10. Click the red triangle next to Latent Class Analysis for 5 Clusters and select **Save Mixture and Cluster Formulas**.
11. Select **Graph > Graph Builder**.
12. Enter In what grade are you as **X**.
13. Enter Most Likely Cluster Formula as **Y**.
14. Select the Mosaic element.
15. Click **Done**.

Figure 10.3 Mosaic Plot of Cluster Membership versus Grade Level

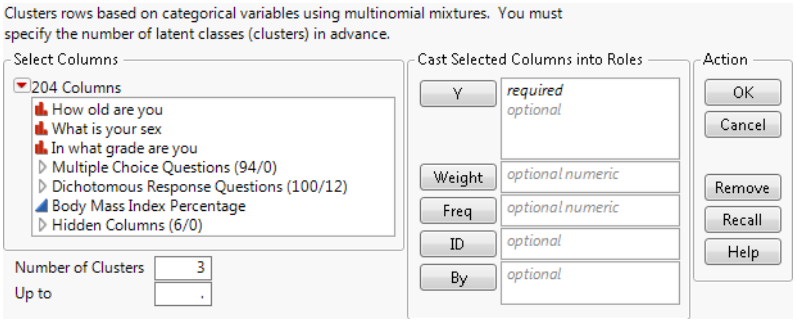


Observe that most of the respondents fall into the Low Risk cluster. The class labeled Drinkers includes more respondents as the grade level increases.

Launch the Latent Class Analysis Platform

Launch the Latent Class Analysis platform by selecting **Analyze > Clustering > Latent Class Analysis**.

Figure 10.4 Latent Class Analysis Launch Window



The Latent Class Analysis platform launch window contains the following options:

Y One or more categorical response columns that you want to analyze.

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 a Fit Group outline.

After you click **OK** on the launch window, the Latent Class Analysis report appears.

The Latent Class Analysis Report

By default, the Latent Class Analysis report contains a Fit Group report that contains Latent Class Analysis reports for each specified number of clusters. Options in the Fit Group report enable you to arrange the Latent Class Analysis reports in rows and to order the reports by goodness of fit.

The Latent Class Analysis also contains the following results and outlines:

BIC

The BIC value for the model with the specified number of clusters appears at the top of each Latent Class Analysis report. The BIC is a goodness of fit measure. Lower values of BIC indicate better fits.

Parameter Estimates

The Parameter Estimates report contains two tables. Each table contains rows corresponding to the model clusters. The first table gives the numerical results. The second table graphs the results with share charts.

The Overall column in both tables 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 211.)

The remaining columns in the first table are grouped with vertical dividers according to the Y columns specified in the Latent Class Analysis launch window. Each group of columns has a column for each level of the corresponding Y column. 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.)

The graph in the second table 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. The stacking of bars follows the order of appearance of the variables in the table of values.

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 first table shown in the Parameter Estimates report. 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. The statistics in each row of this table are obtained from a contingency table analysis of expected counts for cluster membership by levels 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 of the Y column.

For each response, the Pearson chi-square statistic, X^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:

$$\text{Effect Size} = \sqrt{\frac{X^2}{n}}$$

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. 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 in the *Consumer Research* book.

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

Fit Group Options

The Fit Group red triangle menu contains the following options:

Arrange in Rows Arranges the Latent Class Analysis reports horizontally in the Fit Group report.

Order by Goodness of Fit Rearranges the Latent Class Analysis reports in order of increasing BIC values. The report for the model with the smallest BIC value appears first in the Fit Group report.

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.

Color by Cluster Colors each row in the data table according to its most likely cluster. For an example, see "Additional Example: Plot Probabilities of Cluster Membership" on page 210.

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.

JMP PRO **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 the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

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.

See the JMP Reports chapter in the *Using JMP* book 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 Example: 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 Analysis for 3 Clusters and select **Color by Cluster**.
6. Click the red triangle next to Latent Class Analysis 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 10.5 Ternary Plot of Cluster Membership Probabilities

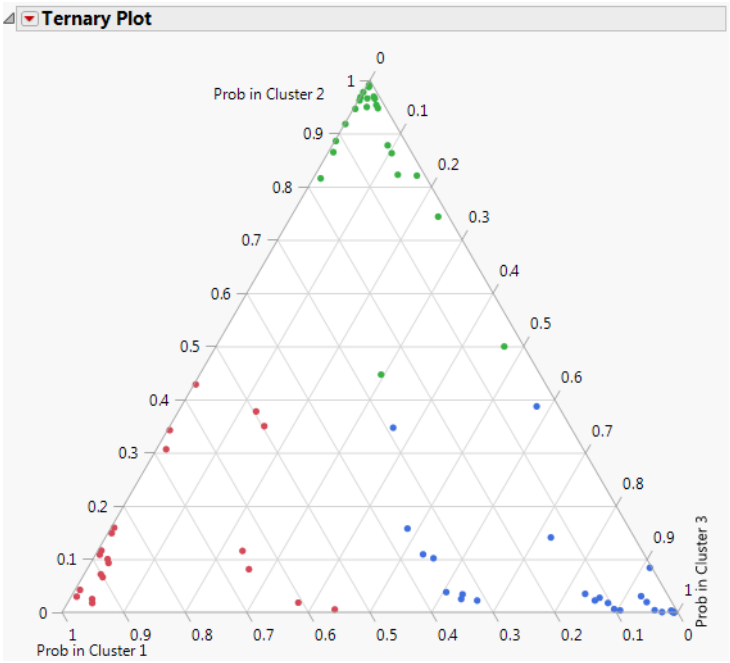


Figure 10.5 shows the ternary plot of cluster probabilities for each observation. Most of the cluster membership probabilities fall near the vertices, which indicates that they have high values for one cluster and lower values for the other two. However, there are some points in the middle of the plot, 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 no random seed was specified.

Statistical Details for the Latent Class Analysis Platform

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 (2002).

Note: The LCA algorithm that is used in the Text Explorer platform takes advantage of the specific structure 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 $j = 1, ..., J$ 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 $W = R_1*...*R_J$ 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 $\mathbf{y} = y_1, ..., y_j$. Define $\mathbf{Y}$ to be the W by J array of all the response patterns considered as row vectors. Each row $\mathbf{y}_w$ 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 $\mathbf{y}_w = y_1, ..., y_j$ is the sum of the conditional probabilities of observing that vector of responses for each of the C latent classes:

$$\Pr(\mathbf{y_w}) = \sum_{c=1}^{C} \gamma_c \prod_{j=1}^{J} \prod_{k=1}^{R_j} \rho_{j,k|c}^{I(y_j = r_{j,k})}$$

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.

Chapter 11

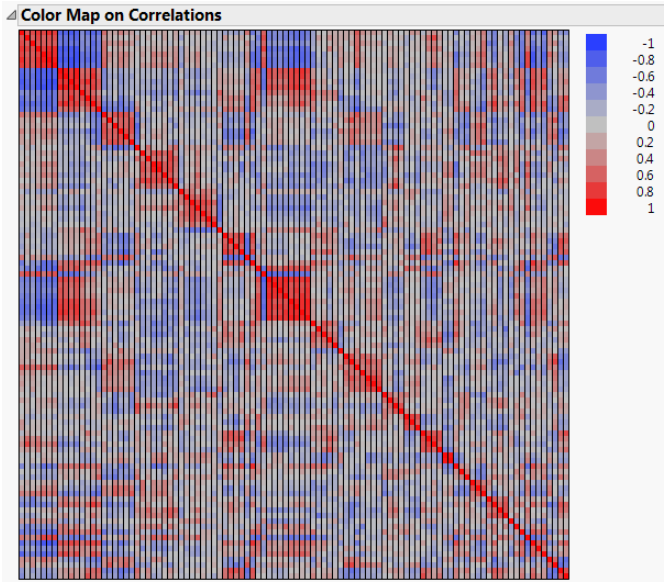
Cluster Variables

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 11.1 Example of Correlation Map for Variables



Cluster Variables Platform Overview

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 226 in the "Statistical Details" appendix.

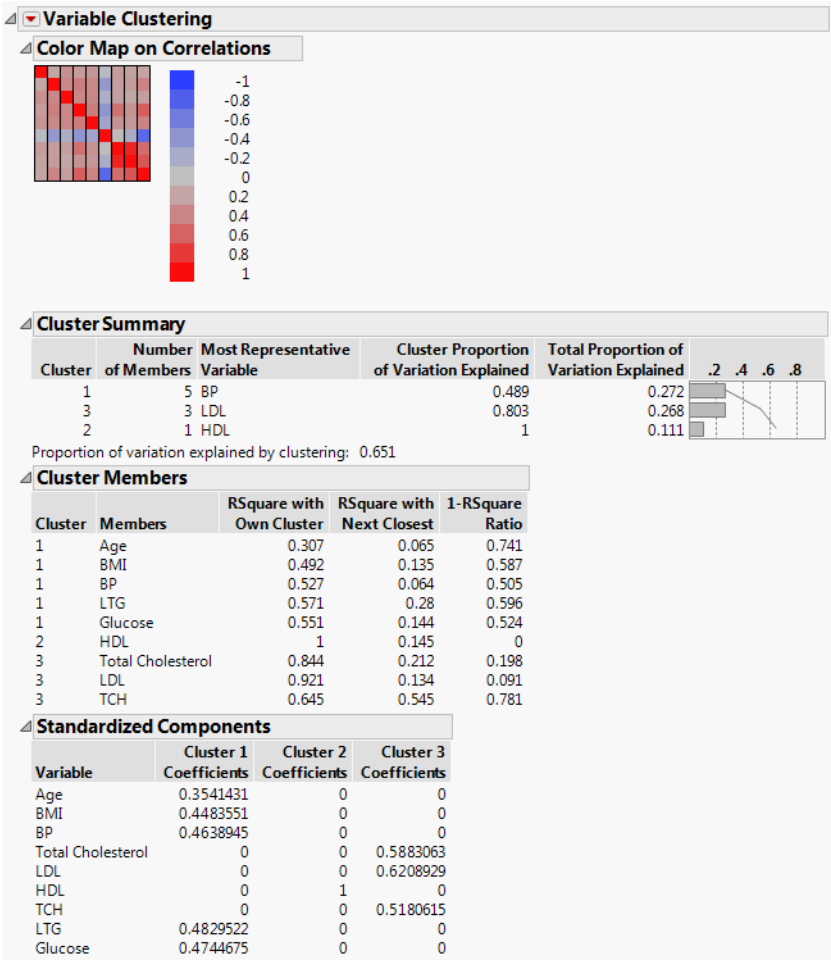
Example of the Cluster Variables Platform

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 11.2 Cluster Variables Report for Diabetes Data



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	5	BP	0.489	0.272
3	3	LDL	0.803	0.268
2	1	HDL	1	0.111

Proportion of variation explained by clustering: 0.651

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	Age	0.307	0.065	0.741
1	BMI	0.492	0.135	0.587
1	BP	0.527	0.064	0.505
1	LTG	0.571	0.28	0.596
1	Glucose	0.551	0.144	0.524
2	HDL	1	0.145	0
3	Total Cholesterol	0.844	0.212	0.198
3	LDL	0.921	0.134	0.091
3	TCH	0.645	0.545	0.781

Standardized Components

Variable	Cluster 1 Coefficients	Cluster 2 Coefficients	Cluster 3 Coefficients
Age	0.3541431	0	0
BMI	0.4483551	0	0
BP	0.4638945	0	0
Total Cholesterol	0	0	0.5883063
LDL	0	0	0.6208929
HDL	0	1	0
TCH	0	0	0.5180615
LTG	0.4829522	0	0
Glucose	0.4744675	0	0

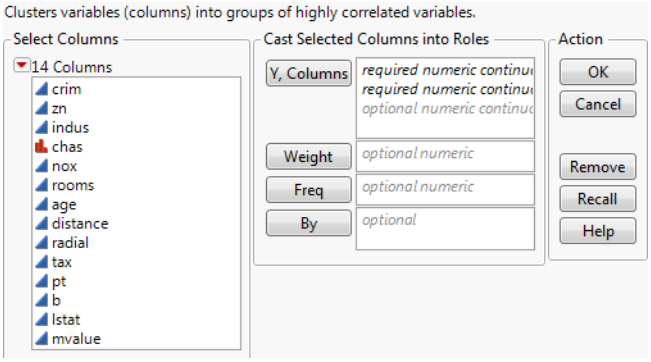
The Variable Clustering report is shown in Figure 11.2. The Cluster Summary report shows that the variables were grouped into three clusters:

- Cluster 1 consists of Age, BMI, BP, LTG, and Glucose, as shown in the Cluster Members report. The Cluster Summary report shows that BP is the most representative variable for Cluster 1 and that it explains 48.9% of the Cluster 1 variation.
- Cluster 2 consists of the single variable HDL.
- Cluster 3 consists of Total Cholesterol, LDL, and TCH. The Cluster Summary report shows that the most representative variable is LDL and it explains 80.4% of the Cluster 3 variation.

Launch the Cluster Variables Platform

Launch the Cluster Variables platform by selecting **Analyze > Clustering > Cluster Variables**. The Cluster Variables launch window for the Diabetes.jmp table is shown in Figure 11.3.

Figure 11.3 Cluster Variables Launch Dialog



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 217
- "Cluster Summary" on page 217
- "Cluster Members" on page 217
- "Standardized Components" on page 218

Tip: In any of the Cluster Variables reports, select rows in order to select the corresponding columns in the data table. Hold down 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 219.

Tip: Place the cursor 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 "Row-wise" on page 33 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 217.

Cluster Summary Shows or hides the Cluster Summary report. See "Cluster Summary" on page 217.

Cluster Members Shows or hides the Cluster Members report. See "Cluster Members" on page 217.

Cluster Components Shows or hides the Standardized Components report. See "Standardized Components" on page 218.

Save Cluster Components Saves columns called Cluster <i> Components to the data table. Each column 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 the JMP Reports chapter in the *Using JMP* book 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

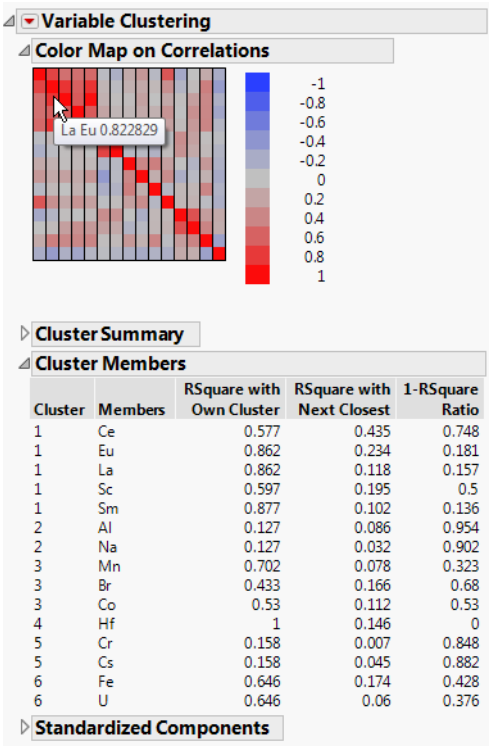
Example of Color Map on Correlations

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. Place your cursor over the cell in the third row and second 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 11.4 Color Map on Correlations for Cherts.jmp



Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	Ce	0.577	0.435	0.748
1	Eu	0.862	0.234	0.181
1	La	0.862	0.118	0.157
1	Sc	0.597	0.195	0.5
1	Sm	0.877	0.102	0.136
2	Al	0.127	0.086	0.954
2	Na	0.127	0.032	0.902
3	Mn	0.702	0.078	0.323
3	Br	0.433	0.166	0.68
3	Co	0.53	0.112	0.53
4	Hf	1	0.146	0
5	Cr	0.158	0.007	0.848
5	Cs	0.158	0.045	0.882
6	Fe	0.646	0.174	0.428
6	U	0.646	0.06	0.376

The Cluster Members report shows that there are five variables in Cluster 1. In the Color Map on Correlations, the five-by-five 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 Cluster 2 and Cluster 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. For more details, see "Color Map on Correlations" on page 217.

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.

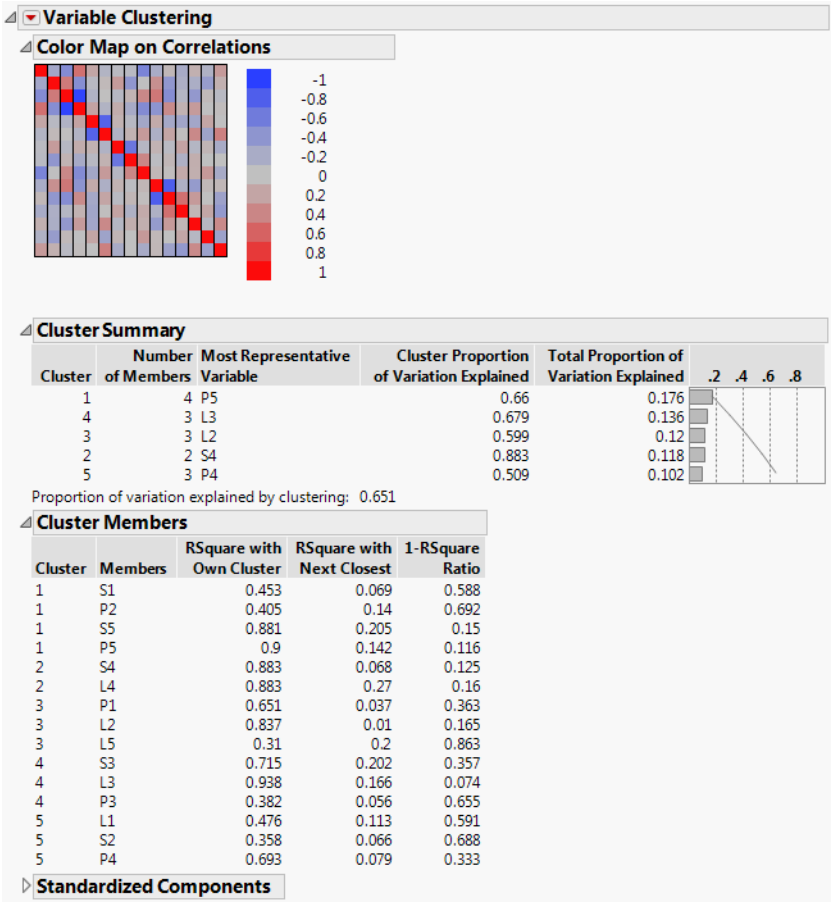
Cluster Variables

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**.

   Three formula columns are added to the data table.

Figure 11.5 Cluster Variables Report for Penta.jmp



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 red triangle next to Variable Clustering 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 11.6 Fit Least Squares Report for Model with All Continuous Predictors

Response log RAI

Effect Summary

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

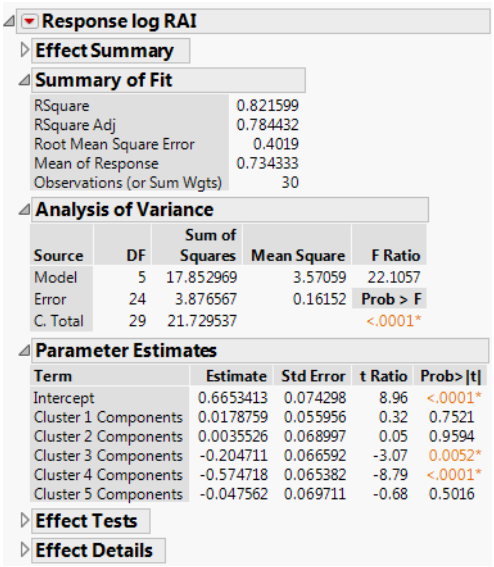
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.802632	0.924946	-0.87	0.4002
P5	2.0563803	1.651272	1.25	0.2335
S4	-0.062354	0.134935	-0.46	0.6511
L2	0.0860287	0.061206	1.41	0.1817
L3	0.3185383	0.080091	3.98	0.0014*
P4	0.4136598	0.394449	1.05	0.3121
S1	-0.09783	0.038948	-2.51	0.0249*
L1	0.032362	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

Effect Tests

Effect Details

6. In the Fit Model window, select all variables in the Construct Model Effects window and click **Remove**.
7. Select the five Cluster Component variables and click **Add**.
8. Click **Run**.

Figure 11.7 Fit Least Squares Report for Model with Cluster Components as Predictors



Response log RAI

Effect Summary

Summary of Fit

RSquare	0.821599
RSquare Adj	0.784432
Root Mean Square Error	0.4019
Mean of Response	0.734333
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	17.852969	3.57059	22.1057
Error	24	3.876567	0.16152	Prob > F
C. Total	29	21.729537		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.6653413	0.074298	8.96	<.0001*
Cluster 1 Components	0.0178759	0.055956	0.32	0.7521
Cluster 2 Components	0.0035526	0.068997	0.05	0.9594
Cluster 3 Components	-0.204711	0.066592	-3.07	0.0052*
Cluster 4 Components	-0.574718	0.065382	-8.79	<.0001*
Cluster 5 Components	-0.047562	0.069711	-0.68	0.5016

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 11.6).

Statistical Details for the Cluster Variables Platform

This section contains 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., 2011).

Note: The algorithm uses only observations for which there are no missing values for any variable in the Y, Columns list.

The iterative steps in the algorithm are as follows:

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 as follows:
   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. It also gives details on computations used in multivariate tests and exact and approximate *F*-statistics.

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 228.

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 $\mathbf{X}$ can be written as follows:

$$\mathbf{X} = \mathbf{U}Diag(\mathbf{\Lambda})\mathbf{V}'$$

Let r be the rank of $\mathbf{X}$. Denote the r by r identity matrix by $\mathbf{I}_r$.

The matrices $\mathbf{U}$, $Diag(\mathbf{\Lambda})$, and $\mathbf{V}$ have the following properties:

$\mathbf{U}$ is an n by r semi-orthogonal matrix with $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$

$\mathbf{V}$ is a p by r semi-orthogonal matrix with $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$

$Diag(\mathbf{\Lambda})$ is an r by r diagonal matrix with positive diagonal elements given by the column vector $\mathbf{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_r)'$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$.

The λ_i are the nonzero *singular values* of $\mathbf{X}$.

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 $\mathbf{V}$ are eigenvectors of $\mathbf{X}'\mathbf{X}$.

Note: There are various conventions in the literature regarding the dimensions of the matrices $\mathbf{U}$, $\mathbf{V}$, and the matrix containing the singular values. However, the differences have no practical impact on the decomposition up to the rank of $\mathbf{X}$.

For further details, 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 given 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 "Wide" on page 58 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:

$$\mathbf{S} = (\mathbf{U}Diag(\mathbf{\Lambda})\mathbf{V}')'(\mathbf{U}Diag(\mathbf{\Lambda})\mathbf{V}') = \mathbf{V}Diag(\mathbf{\Lambda})^2\mathbf{V}'$$

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:

$$\mathbf{S}^{-1} = (\mathbf{V}Diag(\mathbf{\Lambda})^2\mathbf{V}')^{-1} = \mathbf{V}Diag(\mathbf{\Lambda})^{-2}\mathbf{V}'$$

If $\mathbf{S}$ is not of full rank, then $Diag(\mathbf{\Lambda})^{-1}$ can be replaced with a generalized inverse, $Diag(\mathbf{\Lambda})^-$, where the diagonal elements of $Diag(\mathbf{\Lambda})$ are replaced by their reciprocals. This defines a generalize inverse of $\mathbf{S}$ as follows:

$$\mathbf{S}^- = \mathbf{V}(Diag(\mathbf{\Lambda})^+)^2\mathbf{V}'$$

This generalized inverse can be calculated using only the SVD.

To see the specific details behind the application of the SVD for wide linear discriminant analysis, see "Wide Linear Discriminant Method" on page 111 in the "Discriminant Analysis" chapter.

Calculating the SVD

In the Multivariate Methods platforms, JMP's calculation of the SVD of a matrix follows 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. (2002), *Categorical Data Analysis*, Second Edition, New York: John Wiley and Sons, Inc.

Baglama, J. and Reichel, L. (2005), "Augmented implicitly restarted Lanczos bidiagonalization methods," *SIAM Journal on Scientific Computing*, 27.1, 19-42.

Ballard, D.H., (1981), "Generalizing the Hough Transform to Detect Arbitrary Shapes," *Pattern Recognition*, 13:2, 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, 16 (Series B)*, 296–298.

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(1), 32–44.

Collins, L. and Lanza, S. (2010), *Latent Class and Latent Transition Analysis*, Hoboken NJ: John Wiley and 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.

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.

Dwass, M. (1955), "A Note on Simultaneous Confidence Intervals," *Annals of Mathematical Statistics* 26: 146–147.

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.

Farebrother, R.W. (1981), "Mechanical Representations of the L1 and L2 Estimation Problems," *Statistical Data Analysis*, 2nd Edition, Amsterdam, North Holland: edited by Y. Dodge.

Fieller, E.C. (1954), "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society, Series B*, 16, 175–185.

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.

Garthwaite, P. (1994), "An Interpretation of Partial Least Squares," *Journal of the American Statistical Association*, 89:425, 122–127.

Golub, G.H., 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* 2:2, 205–224.

Golub, G.H. and van der Vorst, H.A., (2000), "Eigenvalue Computation in the 20th Century," *Journal of Computational and Applied Mathematics* 123, 35-65.

Goodman, L.A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika* 61:2, 215–231.

Goodnight, J.H. (1978), "Tests of Hypotheses in Fixed Effects Linear Models," *SAS Technical Report R–101*, Cary NC: SAS Institute Inc, also in Communications in Statistics (1980), A9 167–180.

Goodnight, J.H. and W.R. Harvey (1978), "Least Square Means in the Fixed Effect General Linear Model," *SAS Technical Report R–103*, Cary NC: SAS Institute Inc.

Hand, D, Mannila, H, and Smyth, P. (2001), *Principles of Data Mining*, 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), "Consistence 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), *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, New York: Springer Science and Business Media.

Hocking, R.R. (1985), *The Analysis of Linear Models*, Monterey: Brooks–Cole.

Hoskuldsson, A. (1988), "PLS Regression Methods," *Journal of Chemometrics*, 2:3, 211–228.

Hoeffding, W (1948), "A Non-Parametric Test of Independence", Annals of Mathematical Statistics, 19, 546–557.

Huber, P.J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35:1, 73–101.

Huber, Peter J. (1973), "Robust Regression: Asymptotics, Conjecture, and Monte Carlo," *Annals of Statistics*, Volume 1, Number 5, 799–821.

Huber, P.J. and Ronchetti, E.M. (2009), *Robust Statistics,* Second Edition, Wiley.

Jackson, J. Edward (2003), *A User's Guide to Principal Components*, New Jersey: John Wiley and Sons.

Jardine, N. and Sibson, R. (1971), *Mathematical Taxonomy*, New York: John Wiley and Sons.

Kohonen, T. (1989), *Self-Organization and Associative Memory*, Springer Series in Information Sciences, Volume 8.

Kohonen, T. (1990), "The Self-Organizing Map," *Proceedings of the IEEE*, 78:9, 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.

Mardia, K., Kent, J., and Bibby, J. (1980), *Multivariate Analysis,* First Edition, New York: Academic Press.

Mason, R.L. and Young, J.C. (2002), *Multivariate Statistical Process Control with Industrial Applications,* Philadelphia: ASA-SIAM.

McLachlan, G.J. and Krishnan, T. (1997), *The EM Algorithm and Extensions,* New York: John Wiley and 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, Philip R.C., Taylor, Paul A., MacGregor, John F. (1996), "Missing Data Methods in PCA and PLS: Score calculations with incomplete observations," *Chemometrics and Intelligent Laboratory Systems*, 35, 45–65.

Press, W.H, Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1998), *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition, Cambridge, England: Cambridge University Press.

SAS Institute Inc. (1983), "SAS Technical Report A-108: Cubic Clustering Criterion," Cary, NC: SAS Institute Inc. Retrieved December 16, 2015 from https://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf.

SAS Institute Inc. (2011), *SAS/STAT 9.2 User's Guide*, "The VARCLUS Procedure," Cary, NC: SAS Institute Inc. Retrieved April 15, 2015 from http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#varclus_toc.htm.

SAS Institute Inc. (2011), *SAS/STAT 9.3 User's Guide*, "The PLS Procedure," Cary, NC: SAS Institute Inc. Retrieved April 15, 2015 from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_pls_sect006.htm.

SAS Institute Inc. (2011), *SAS/STAT 9.3 User's Guide*, "The CANDISC Procedure," Cary, NC: SAS Institute Inc. Retrieved April 15, 2015 from http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_candisc_sect004.htm.

SAS Institute Inc. (2005), *SAS/STAT 9.2 User's Guide,* "The FASTCLUS Procedure," Cary, NC: SAS Institute Inc. Retrieved June 21, 2016 from http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_fastclus_sect005.htm.

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, 1175–1189.

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.

Tobias, R.D. (1995), "An Introduction to Partial Least Squares Regression," *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Tracy, N.D., Young, J.C., 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.

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*, 21:2, 272-278.

Wold, (1980), "Soft Modelling: Intermediate between Traditional Model Building and Data Analysis," *Mathematical Statistics* (Banach Center Publications, Warsaw), 6, 333–346.

Wold, S. (1994), "PLS for Multivariate Linear Modeling", *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry*.

Wold, S., Sjostrom, M., and Eriksson, L. (2001), "PLS-Regression: A Basic Tool of Chemometrics," *Chemometrics and Intelligent Laboratory Systems*, 58:2, 109–130.

Wright, S.P. and R.G. O'Brien (1988), "Power Analysis in an Enhanced GLM Procedure: What it Might Look Like," *SUGI 1988, Proceedings of the Thirteenth Annual Conference*, 1097–1102, Cary NC: SAS Institute Inc.

Index

Multivariate Methods

Q

R

S

T

U

W-Z