



**Version 14**

# **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. 2018. *JMP® 14 Multivariate Methods*. Cary, NC: SAS Institute Inc.

## **JMP® 14 Multivariate Methods**

Copyright © 2018, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-63526-517-0 (Hardcopy)

ISBN 978-1-63526-518-7 (EPUB)

ISBN 978-1-63526-519-4 (MOBI)

ISBN 978-1-63526-520-0 (Web PDF)

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

**For a hard-copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

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

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

March 2018

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

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

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

### **Technology License Notices**

- Scintilla - Copyright © 1998-2017 by Neil Hodgson <neilh@scintilla.org>.

All Rights Reserved.

Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

- 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](http://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](http://www.mathworks.com/patents). MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See [www.mathworks.com/trademarks](http://www.mathworks.com/trademarks) for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.
- libopc is Copyright © 2011, Florian Reuter. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.

- Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- libxml2 - Except where otherwise noted in the source code (e.g. the files hash.c, list.c and the trio files, which are covered by a similar license but with different Copyright notices) all the files are:

Copyright © 1998 - 2003 Daniel Veillard. All Rights Reserved.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the “Software”), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

- Regarding the decompression algorithm used for UNIX files:

Copyright © 1985, 1986, 1992, 1993

The Regents of the University of California. All rights reserved.

THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
  3. Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.
- Snowball - Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.
3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE

DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

## **Get the Most from JMP<sup>®</sup>**

Whether you are a first-time or a long-time user, there is always something to learn about JMP.

Visit [JMP.com](http://www.jmp.com) to find the following:

- live and recorded webcasts about how to get started with JMP
- video demos and webcasts of new features and advanced techniques
- details on registering for JMP training
- schedules for seminars being held in your area
- success stories showing how others use JMP
- a blog with tips, tricks, and stories from JMP staff
- a forum to discuss JMP with other users

**<http://www.jmp.com/getstarted/>**



# Contents

## Multivariate Methods

---

<b>1</b>	<b>Learn about JMP</b>	17
	<b>Documentation and Additional Resources</b>	
	Formatting Conventions	19
	JMP Documentation	20
	JMP Documentation Library	20
	JMP Help	26
	Additional Resources for Learning JMP	26
	Tutorials	26
	Sample Data Tables	27
	Learn about Statistical and JSL Terms	27
	Learn JMP Tips and Tricks	27
	Tooltips	27
	JMP User Community	28
	JMP Books by Users	28
	The JMP Starter Window	28
	Technical Support	29
<b>2</b>	<b>Introduction to Multivariate Analysis</b>	31
	<b>Overview of Multivariate Techniques</b>	
<b>3</b>	<b>Correlations and Multivariate Techniques</b>	33
	<b>Explore the Multidimensional Behavior of Variables</b>	
	Launch the Multivariate Platform	35
	The Multivariate Report	37
	Multivariate Platform Options	38
	Nonparametric Correlations	43
	Scatterplot Matrix	44
	Outlier Analysis	46
	Item Reliability	48
	Impute Missing Data	48
	Example of Item Reliability	49
	Computations and Statistical Details	50
	Estimation Methods	50

Pearson Product-Moment Correlation .....	51
Nonparametric Measures of Association .....	51
Inverse Correlation Matrix .....	53
Distance Measures .....	53
Cronbach's $\alpha$ .....	55

## 4 **Principal Components**..... 57

### **Reduce the Dimensionality of Your Data**

Overview of Principal Component Analysis Platform .....	59
Example of Principal Component Analysis .....	59
Launch the Principal Components Platform .....	60
Principal Components Report .....	63
Principal Components Report Options .....	64
Outlier Analysis .....	74
Statistical Details for the Principal Components Analysis Platform .....	75
Estimation Methods .....	75
DModX Calculation .....	77
Calculations for Outlier Analysis .....	77

## 5 **Discriminant Analysis**..... 79

### **Predict Classifications Based on Continuous Variables**

Overview of the Discriminant Analysis Platform .....	81
Example of Discriminant Analysis .....	81
Discriminant Launch Window .....	82
Stepwise Variable Selection .....	84
Discriminant Methods .....	88
Shrink Covariances .....	91
The Discriminant Analysis Report .....	91
Principal Components .....	92
Canonical Plot and Canonical Structure .....	93
Discriminant Scores .....	97
Score Summaries .....	98
Discriminant Analysis Options .....	100
Score Options .....	102
Canonical Options .....	104
Example of a Canonical 3D Plot .....	108
Specify Priors .....	109
Consider New Levels .....	109
Save Discrim Matrices .....	110
Scatterplot Matrix .....	110
Validation in JMP and JMP Pro .....	111

Technical Details for the Discriminant Analysis Platform	112
Description of the Wide Linear Algorithm	112
Saved Formulas	113
Multivariate Tests	120
Approximate F-Tests	121
Between Groups Covariance Matrix	121
<b>6 Partial Least Squares Models</b>	<b>123</b>
<b>Develop Models Using Correlations between Ys and Xs</b>	
Overview of the Partial Least Squares Platform	125
Example of Partial Least Squares	126
Launch the Partial Least Squares Platform	129
Centering and Scaling	132
Standardize X	132
Model Launch Control Panel	133
Partial Least Squares Report	134
Model Comparison Summary	134
<Cross Validation Method> and Method = <Method Specification>	135
Model Fit Report	140
Partial Least Squares Options	141
Model Fit Options	141
Variable Importance Plot	143
VIP vs Coefficients Plots	144
Save Columns	145
Statistical Details for the Partial Least Squares Platform	147
Partial Least Squares	147
van der Voet $T^2$	148
$T^2$ Plot	149
Confidence Ellipses for X Score Scatterplot Matrix	150
Standard Error of Prediction and Confidence Limits	150
Standardized Scores and Loadings	151
PLS Discriminant Analysis (PLS-DA)	152
<b>7 Multiple Correspondence Analysis</b>	<b>153</b>
<b>Identify Associations between Levels of Categorical Variables</b>	
Example of Multiple Correspondence Analysis	155
Launch the Multiple Correspondence Analysis Platform	156
The Multiple Correspondence Analysis Report	158
Multiple Correspondence Analysis Platform Options	159
Correspondence Analysis Options	160
Show Plot	161

Show Detail .....	161
Show Adjusted Inertia .....	162
Show Coordinates .....	162
Show Summary Statistics .....	163
Show Partial Contributions to Inertia .....	163
Show Squared Cosines .....	164
Cross Table .....	164
Cross Table of Supplementary Rows .....	165
Additional Examples of the Multiple Correspondence Analysis Platform .....	166
Example Using a Supplementary Variable .....	166
Example Using a Supplementary ID .....	167
Statistical Details for the Multiple Correspondence Analysis Platform .....	168
Details Report .....	169
Adjusted Inertia .....	169
Summary Statistics .....	170
Partial Contributions to Inertia .....	170

## 8 Factor Analysis .....

<b>Identify Latent Variables in Your Data</b> .....	171
Overview of the Factor Analysis Platform .....	173
Example of the Factor Analysis Platform .....	174
Launch the Factor Analysis Platform .....	176
The Factor Analysis Report .....	177
Model Launch .....	178
Rotation Methods .....	180
Factor Analysis Platform Options .....	181
Factor Analysis Model Fit Options .....	182

## 9 Multidimensional Scaling .....

<b>Visualize Proximities among a Set of Objects</b> .....	187
Overview of the Multidimensional Scaling Platform .....	189
Example of Multidimensional Scaling .....	189
Launch the Multidimensional Scaling Platform .....	191
The Multidimensional Scaling Report .....	193
Multidimensional Scaling Plot .....	193
Shepard Diagram .....	193
Fit Details .....	194
Multidimensional Scaling Platform Options .....	194
Waern Links .....	195
Additional Example of the Multidimensional Scaling Platform .....	196
Statistical Details for the Multidimensional Scaling Platform .....	198

Stress .....	198
Transformations .....	199
Attributes List Format .....	199
<b>10 Item Analysis</b> .....	<b>201</b>
<b>Analyze Test Results by Item and Subject</b> .....	
Example of Item Analysis .....	203
Launch the Item Analysis Platform .....	206
The Item Analysis Report .....	207
Characteristic Curves .....	207
Information Plot .....	208
Dual Plot .....	209
Parameter Estimates .....	209
Item Analysis Platform Options .....	210
Statistical Details for the Item Analysis Platform .....	210
Item Response Curves .....	211
Item Response Curve Models .....	212
IRT Model Assumptions .....	214
Fitting the IRT Model .....	214
Ability Formula .....	216
<b>11 Hierarchical Cluster</b> .....	<b>217</b>
<b>Group Observations Using a Tree of Clusters</b> .....	
Overview of the Hierarchical Clustering Platform .....	219
Overview of Platforms for Clustering Observations .....	219
Example of Clustering .....	221
Launch the Hierarchical Cluster Platform .....	223
Clustering Method .....	225
Method for Distance Calculation .....	225
Data Structure .....	226
Transformations to Y, Columns Variables .....	228
Hierarchical Cluster Report .....	229
Dendrogram Report .....	230
Illustration of Dendrogram and Distance Graph .....	230
Clustering History Report .....	232
Hierarchical Cluster Options .....	232
Additional Examples of the Hierarchical Clustering Platform .....	236
Example of a Distance Matrix .....	236
Example of Wafer Defect Classification Using Spatial Measures .....	238
Statistical Details for the Hierarchical Clustering Platform .....	240
Spatial Measures .....	241

Distance Method Formulas .....	242
<b>12 K Means Cluster</b> .....	245
<b>Group Observations Using Distances</b>	
Overview of the K Means Cluster Platform .....	247
Overview of Platforms for Clustering Observations .....	247
Example of K Means Cluster .....	249
Launch the K Means Cluster Platform .....	252
Iterative Clustering Report .....	253
Iterative Clustering Control Panel .....	254
K Means NCluster=<k> Report .....	255
Cluster Comparison Report .....	255
K Means NCluster=<k> Report .....	255
K Means NCluster=<k> Report Options .....	256
Self Organizing Map .....	257
Self Organizing Map Control Panel .....	258
Self Organizing Map Report .....	259
Description of SOM Algorithm .....	259
Additional Example of K Means Cluster Platform .....	260
Example of a Self-Organizing Map .....	260
<b>13 Normal Mixtures</b> .....	263
<b>Group Observations Using Probabilities</b>	
Overview of the Normal Mixtures Clustering Platform .....	265
Overview of Platforms for Clustering Observations .....	265
Example of Normal Mixtures Clustering .....	267
Launch the Normal Mixtures Clustering Platform .....	268
Model Based Clustering Report .....	269
Model Based Clustering Control Panel .....	270
Normal Mixtures NCluster=<k> Report .....	271
Cluster Comparison Report .....	271
Normal Mixtures NCluster=<k> Report .....	272
Normal Mixtures NCluster=<k> Report Options .....	272
Statistical Details for the Normal Mixtures Clustering Platform .....	274
<b>14 Latent Class Analysis</b> .....	275
<b>Group Observations of Categorical Variables</b>	
Overview of the Latent Class Analysis Platform .....	277
Example of Latent Class Analysis .....	277
Launch the Latent Class Analysis Platform .....	280
The Latent Class Analysis Report .....	282

Latent Class Analysis Platform Options .....	284
Latent Class Analysis Options .....	285
Latent Class Model Options .....	285
Additional Example of the Latent Class Analysis Platform .....	286
Plot Probabilities of Cluster Membership .....	286
Statistical Details for the Latent Class Analysis Platform .....	287
Latent Class Model Fit .....	288
Maximum Number of Clusters .....	289
<b>15 Cluster Variables</b> .....	291
<b>Group Similar Variables into Representative Groups</b>	
Overview of the Cluster Variables Platform .....	293
Example of the Cluster Variables Platform .....	293
Launch the Cluster Variables Platform .....	295
The Cluster Variables Report .....	295
Color Map on Correlations .....	296
Cluster Summary .....	296
Cluster Members .....	297
Standardized Components .....	297
Cluster Variables Platform Options .....	297
Additional Examples of the Cluster Variables Platform .....	298
Example of Color Map on Correlations .....	298
Example of Cluster Variables Platform for Dimension Reduction .....	300
Statistical Details for the Cluster Variables Platform .....	303
Variable Clustering Algorithm .....	303
<b>A Statistical Details</b> .....	305
<b>Multivariate Methods</b>	
Wide Linear Methods and the Singular Value Decomposition .....	307
The Singular Value Decomposition .....	307
The SVD and the Covariance Matrix .....	308
The SVD and the Inverse Covariance Matrix .....	308
Calculating the SVD .....	309
<b>B References</b> .....	311
<b>Index</b> .....	315
<b>Multivariate Methods</b>	





# Chapter 1

## **Learn about JMP**

### **Documentation and Additional Resources**

---

This chapter includes details about JMP documentation, such as book conventions, descriptions of each JMP book, the Help system, and where to find other support.


## Contents

Formatting Conventions .....	19
JMP Documentation .....	20
JMP Documentation Library .....	20
JMP Help .....	26
Additional Resources for Learning JMP .....	26
Tutorials .....	26
Sample Data Tables .....	27
Learn about Statistical and JSL Terms .....	27
Learn JMP Tips and Tricks .....	27
Tooltips .....	27
JMP User Community .....	28
JMP Books by Users .....	28
The JMP Starter Window .....	28
Technical Support .....	29

---

# 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
- Features that are for JMP Pro only are noted with the JMP Pro icon . For an overview of JMP Pro features, visit <https://www.jmp.com/software/pro/>.

---

**Note:** Special information and limitations appear within a Note.

---

---

**Tip:** Helpful information appears within a Tip.

---

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

Document Title	Document Purpose	Document Content
<i>Basic Analysis</i>	Perform basic analysis using this document.	<p>Describes these Analyze menu platforms:</p> <ul style="list-style-type: none"> <li>• Distribution</li> <li>• Fit Y by X</li> <li>• Tabulate</li> <li>• Text Explorer</li> </ul> <p>Covers how to perform bivariate, one-way ANOVA, and contingency analyses through Analyze &gt; Fit Y by X. How to approximate sampling distributions using bootstrapping and how to perform parametric resampling with the Simulate platform are also included.</p>
<i>Essential Graphing</i>	Find the ideal graph for your data.	<p>Describes these Graph menu platforms:</p> <ul style="list-style-type: none"> <li>• Graph Builder</li> <li>• Scatterplot 3D</li> <li>• Contour Plot</li> <li>• Bubble Plot</li> <li>• Parallel Plot</li> <li>• Cell Plot</li> <li>• Scatterplot Matrix</li> <li>• Ternary Plot</li> <li>• Treemap</li> <li>• Chart</li> <li>• Overlay Plot</li> </ul> <p>The book also covers how to create background and custom maps.</p>
<i>Profilers</i>	Learn how to use interactive profiling tools, which enable you to view cross-sections of any response surface.	Covers all profilers listed in the Graph menu. Analyzing noise factors is included along with running simulations using random inputs.

Document Title	Document Purpose	Document Content
<i>Design of Experiments Guide</i>	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.
<i>Fitting Linear Models</i>	Learn about Fit Model platform and many of its personalities.	<div>Describes these personalities, all available within the Analyze menu Fit Model platform:</div> <ul style="list-style-type: none"><li>• Standard Least Squares</li><li>• Stepwise</li><li>• Generalized Regression</li><li>• Mixed Model</li><li>• MANOVA</li><li>• Loglinear Variance</li><li>• Nominal Logistic</li><li>• Ordinal Logistic</li><li>• Generalized Linear Model</li></ul>

Document Title	Document Purpose	Document Content
<i>Predictive and Specialized Modeling</i>	Learn about additional modeling techniques.	<p>Describes these Analyze &gt; Predictive Modeling menu platforms:</p> <ul style="list-style-type: none"> <li>• Modeling Utilities</li> <li>• Neural</li> <li>• Partition</li> <li>• Bootstrap Forest</li> <li>• Boosted Tree</li> <li>• K Nearest Neighbors</li> <li>• Naive Bayes</li> <li>• Model Comparison</li> <li>• Formula Depot</li> </ul> <p>Describes these Analyze &gt; Specialized Modeling menu platforms:</p> <ul style="list-style-type: none"> <li>• Fit Curve</li> <li>• Nonlinear</li> <li>• Functional Data Explorer</li> <li>• Gaussian Process</li> <li>• Time Series</li> <li>• Matched Pairs</li> </ul> <p>Describes these Analyze &gt; Screening menu platforms:</p> <ul style="list-style-type: none"> <li>• Response Screening</li> <li>• Process Screening</li> <li>• Predictor Screening</li> <li>• Association Analysis</li> <li>• Process History Explorer</li> </ul> <p>The platforms in the Analyze &gt; Specialized Modeling &gt; Specialized DOE Models menu are described in <i>Design of Experiments Guide</i>.</p>

Document Title	Document Purpose	Document Content
<i>Multivariate Methods</i>	Read about techniques for analyzing several variables simultaneously.	<p>Describes these Analyze &gt; Multivariate Methods menu platforms:</p> <ul style="list-style-type: none"> <li>• Multivariate</li> <li>• Principal Components</li> <li>• Discriminant</li> <li>• Partial Least Squares</li> <li>• Multiple Correspondence Analysis</li> <li>• Factor Analysis</li> <li>• Multidimensional Scaling</li> <li>• Item Analysis</li> </ul> <p>Describes these Analyze &gt; Clustering menu platforms:</p> <ul style="list-style-type: none"> <li>• Hierarchical Cluster</li> <li>• K Means Cluster</li> <li>• Normal Mixtures</li> <li>• Latent Class Analysis</li> <li>• Cluster Variables</li> </ul>
<i>Quality and Process Methods</i>	Read about tools for evaluating and improving processes.	<p>Describes these Analyze &gt; Quality and Process menu platforms:</p> <ul style="list-style-type: none"> <li>• Control Chart Builder and individual control charts</li> <li>• Measurement Systems Analysis</li> <li>• Variability / Attribute Gauge Charts</li> <li>• Process Capability</li> <li>• Pareto Plot</li> <li>• Diagram</li> <li>• Manage Spec Limits</li> </ul>



Document Title	Document Purpose	Document Content
<i>Reliability and Survival Methods</i>	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: <ul style="list-style-type: none"> <li>• Life Distribution</li> <li>• Fit Life by X</li> <li>• Cumulative Damage</li> <li>• Recurrence Analysis</li> <li>• Degradation and Destructive Degradation</li> <li>• Reliability Forecast</li> <li>• Reliability Growth</li> <li>• Reliability Block Diagram</li> <li>• Repairable Systems Simulation</li> <li>• Survival</li> <li>• Fit Parametric Survival</li> <li>• Fit Proportional Hazards</li> </ul>
<i>Consumer Research</i>	Learn about methods for studying consumer preferences and using that insight to create better products and services.	Describes these Analyze > Consumer Research menu platforms: <ul style="list-style-type: none"> <li>• Categorical</li> <li>• Choice</li> <li>• MaxDiff</li> <li>• Uplift</li> <li>• Multiple Factor Analysis</li> </ul>
<i>Scripting Guide</i>	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.
<i>JSL Syntax Reference</i>	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 <https://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”](#)
- [“Sample Data Tables”](#)
- [“Learn about Statistical and JSL Terms”](#)
- [“Learn JMP Tips and Tricks”](#)
- [“Tooltips”](#)
- [“JMP User Community”](#)
- [“JMP Books by Users”](#)
- [“The JMP Starter Window”](#)

## 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\14\Samples\Data

On Macintosh: \Library\Application Support\JMP\14\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.

## 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/>.

## JMP Books by Users

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

[https://www.jmp.com/en\\_us/software/books.html](https://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 <https://www.jmp.com/support>, including the technical support phone number.



## 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 Multiple Correspondence Analysis (MCA) platform takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA is frequently used in the social sciences particularly in France and Japan. It can be used in survey analysis to identify question agreement. For more information, see [Chapter 7, “Multiple Correspondence Analysis”](#).
- The Factor Analysis platform enables you to construct factors from a larger set of observed variables. These factors are expressed as linear combinations of a subset of the observed variables. Factor analysis enables you to explore the number of factors that are explained by a set of measured, observed variables, and the strength of the relationship between factors and variables. For more information, see [Chapter 8, “Factor Analysis”](#).
- The Multidimensional Scaling (MDS) platform enables you to create a visual representation of the pattern of proximities (similarities, dissimilarities, or distances) among a set of objects. For more information, see [Chapter 9, “Multidimensional Scaling”](#).
- The Item Analysis platform enables you to fit item response theory models. The Item Response Theory (IRT) method is used for the analysis and scoring of measurement instruments such as tests and questionnaires. IRT uses a system of models to relate a trait or ability to an individual’s probability of endorsing or correctly responding to an item.

IRT can be used to study standardized tests, cognitive development, and consumer preferences. For more information, see [Chapter 10, “Item Analysis”](#).

- The Hierarchical Cluster platform groups rows together that share similar values across a number of variables. It is a useful exploratory technique to help you understand the clumping structure of your data. See [Chapter 11, “Hierarchical Cluster”](#).
- The KMeans Clustering platform groups observations that share similar values across a number of variables. See [Chapter 12, “K Means Cluster”](#).
- The Normal Mixtures platform enables you to cluster observations when your data come from overlapping normal distributions. See [Chapter 13, “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 14, “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 15, “Cluster Variables”](#).



## 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



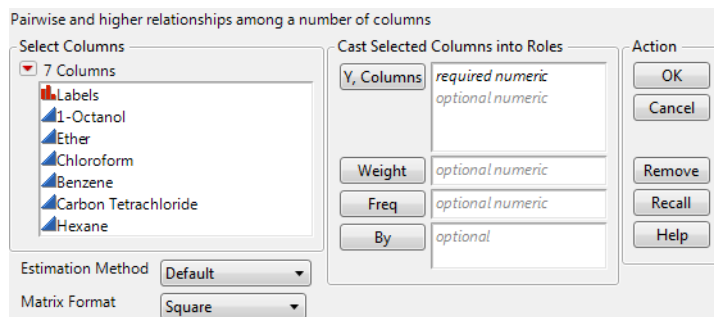
**Contents**

Launch the Multivariate Platform.....	35
The Multivariate Report.....	37
Multivariate Platform Options .....	38
Nonparametric Correlations .....	43
Scatterplot Matrix .....	44
Outlier Analysis.....	46
Item Reliability .....	48
Impute Missing Data .....	48
Example of Item Reliability .....	49
Computations and Statistical Details.....	50
Estimation Methods .....	50
Pearson Product-Moment Correlation.....	51
Nonparametric Measures of Association.....	51
Inverse Correlation Matrix.....	53
Distance Measures .....	53
Cronbach's $\alpha$ .....	55

## 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** Specifies the method for calculating the correlations. **REML** and **Pairwise** are the methods used most frequently. Several of these methods address the treatment of missing data. You can also estimate missing values by using the estimated covariance matrix, and then using the **Impute Missing Data** command. See [“Impute Missing Data”](#) on page 48.

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

- **Row-wise** estimation is used for data tables with no missing values.
- **Pairwise** estimation is used for data tables with missing values and either more than 10 columns, more than 5,000 rows, or more columns than rows.
- **REML** estimation is used otherwise.

---

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

---

**REML** Restricted maximum likelihood (REML) estimation uses all of the data, even if missing values are present. Due to a bias-correction factor, this method is slow if the dataset is large and there are many missing values. Therefore, REML is most useful for smaller datasets. If there are no missing cells in the data, then the REML and ML estimates are equivalent and equal to the sample covariance matrix. If there are missing cells, REML's variance and covariance estimates are less biased than the estimates from ML estimation. For more information, see [“REML”](#) on page 50.

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

**Robust** Robust estimation uses all of the data, even if missing values are present. This method down-weights extreme values and is therefore useful for data tables that might have outliers. For statistical details, see [“Robust”](#) on page 50.

**Row-wise** Row-wise estimation does not use observations with missing values, so rows that contain missing cells are deleted before the method is applied. This method is useful for excluding any observations that have missing data. Row-wise estimation was the only estimation method available prior to JMP 8, so it can also be used to check compatibility with JMP versions prior to JMP 8.

**Pair-wise** Pair-wise estimation uses all of the data, even if missing values are present. This estimation method performs correlations for all rows for each pair of columns with nonmissing values. It is most useful when a data table has missing values and either more columns than rows, more than 10 columns, or more than 5,000 rows.

---

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

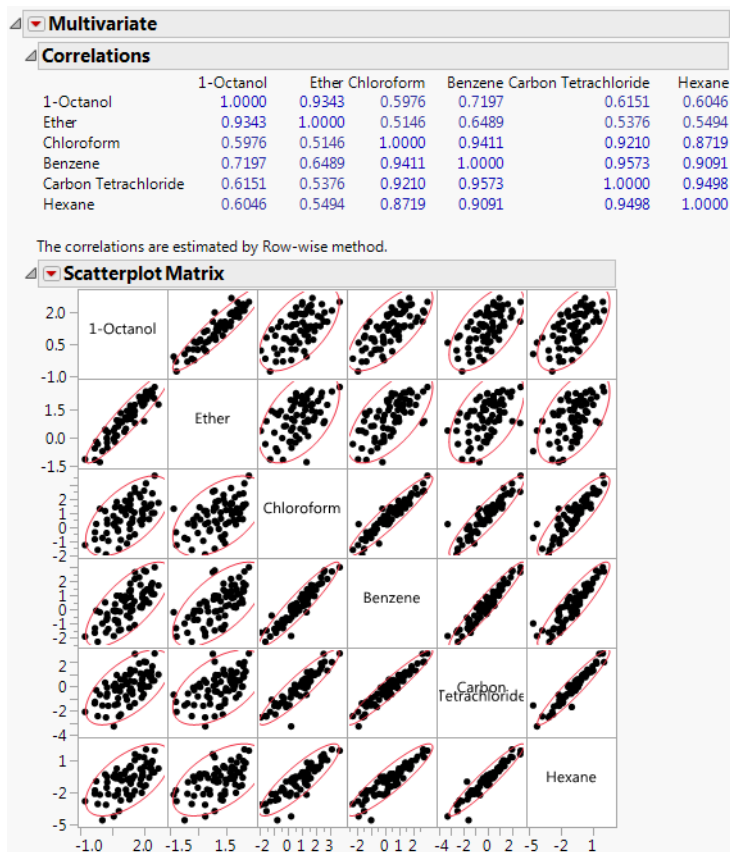
---

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

## The Multivariate Report

The default multivariate report shows the standard correlation matrix, the scatterplot matrix, and a note about which method was used to estimate the correlations. In some cases, information that explains why the given method was used is also displayed. The platform menu lists additional correlation options and other techniques for looking at multiple variables. See [“Multivariate Platform Options”](#) on page 38.

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	<p>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 (<i>Y</i>) variables. This option is on by default. See <a href="#">“Pearson Product-Moment Correlation”</a> on page 51.</p> <p>This correlation matrix is calculated by the method that you select in the launch window.</p>
Correlation Probability	<p>Shows the Correlation Probability report, which is a matrix of <i>p</i>-values. Each <i>p</i>-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.</p>
CI of Correlation	<p>Shows the two-tailed confidence intervals of the correlations. This option is off by default.</p> <p>The default confidence coefficient is 95%. Use the <b>Set <math>\alpha</math> Level</b> option to change the confidence coefficient.</p>

<b>Inverse Correlations</b>	<p>Shows or hides the inverse correlation matrix (Inverse Corr table). This option is off by default.</p> <p>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 <math>1/(1 - R^2)</math> 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.</p> <p>For statistical details about inverse correlations, see the <a href="#">“Inverse Correlation Matrix”</a> on page 53.</p>
<b>Partial Correlations</b>	<p>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.</p> <p>This table is the negative of the inverse correlation matrix, scaled to unit diagonal.</p>
<b>Covariance Matrix</b>	<p>Shows or hides the covariance matrix which measures the degree to which a pair of variables change together. This option is off by default.</p>
<b>Pairwise Correlations</b>	<p>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.</p> <p>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.</p>

---

**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.

---



<b>Simple Statistics</b>	<p>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.</p> <p><b>Univariate Simple Statistics</b> Shows statistics that are calculated on each column, regardless of values in other columns. These values match those produced by the Distribution platform.</p> <p><b>Multivariate Simple Statistics</b> Shows statistics that correspond to the estimation method selected in the launch window and the presence or absence of missing data. If there are no missing observations, this option is available only for the <b>Robust</b> method. If there are missing observations, this option is available for all estimation methods except for <b>Pairwise</b>. The option is never available for <b>Pairwise</b>. For the <b>REML</b>, <b>ML</b>, or <b>Robust</b> methods, the mean vector and covariance matrix are estimated by the selected method. For the <b>Row-wise</b> method, all rows with at least one missing value are excluded from the calculation of means and variances.</p> <p>These options are off by default.</p>
<b>Nonparametric Correlations</b>	<p>This menu contains three nonparametric measures: Spearman's Rho, Kendall's Tau, and Hoeffding's D. These options are off by default.</p> <p>For details, see <a href="#">“Nonparametric Correlations”</a> on page 43.</p>
<b>Set <math>\alpha</math> Level</b>	<p>You can specify any alpha value for the correlation confidence intervals.</p> <p>Four alpha values are listed: <b>0.01</b>, <b>0.05</b>, <b>0.10</b>, and <b>0.50</b>. Select <b>Other</b> to enter any other value.</p>
<b>Scatterplot Matrix</b>	<p>Shows or hides a scatterplot matrix of each pair of response variables. This option is on by default.</p> <p>For details, see <a href="#">“Scatterplot Matrix”</a> on page 44.</p>

Color Maps	<p>The <b>Color Map</b> menu contains three types of color maps.</p> <p><b>Color Map On Correlations</b> Produces a cell plot that shows the correlations among variables on a scale from red (+1) to blue (-1).</p> <p><b>Color Map On p-values</b> Produces a cell plot that shows the significance of the correlations on a scale from <math>p = 0</math> (red) to <math>p = 1</math> (blue).</p> <p><b>Cluster the Correlations</b> 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.</p> <p>These options are off by default.</p>
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	<p>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 <b>None</b> to remove the report. This option is off by default.</p> <p><i>Principal components</i> 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 <a href="#">“Principal Components”</a> on page 57.</p>
Outlier Analysis	<p>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 <math>T^2</math> statistic.</p> <p>For details, see <a href="#">“Outlier Analysis”</a> on page 46.</p>

<b>Item Reliability</b>	<p>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 <math>\alpha</math> or standardized <math>\alpha</math>. These options are off by default.</p> <p>For details, see <a href="#">“Item Reliability”</a> on page 48.</p>
<b>Impute Missing Data</b>	<p>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.</p> <p>For details, see <a href="#">“Impute Missing Data”</a> on page 48.</p>
<b>Save Imputed Formula</b>	<p>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_&lt;Column Name&gt;.</p>

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.

---

---

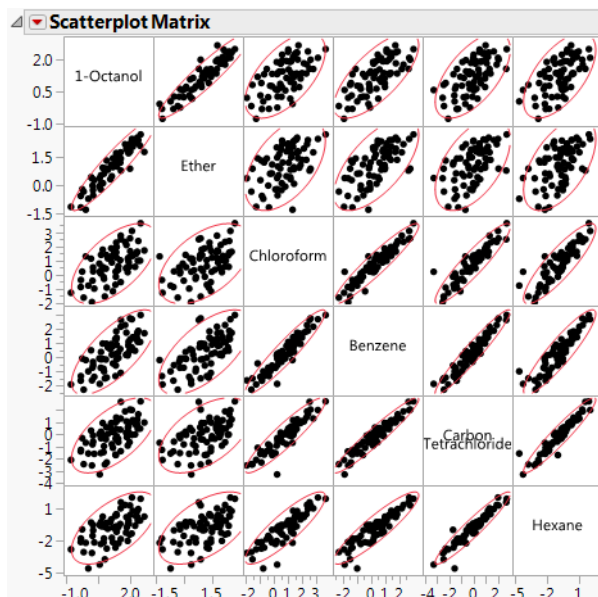
**Note:** When a Weight variable is specified, missing and zero-valued weights are excluded from the nonparametric correlation calculations. All other weight values are treated as 1.

---

For statistical details about these three methods, see the [“Nonparametric Measures of Association”](#) on page 51.

## 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  $\alpha$ -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*)

<b>Fit Line</b>	Shows or hides the regression line and 95% level confidence curves for the fitted regression line.
<b>Density Ellipses</b>	Shows or hides the 95% density ellipses in the scatterplots. Use the <b>Ellipse <math>\alpha</math></b> menu to change the $\alpha$ -level.
<b>Shaded Ellipses</b>	Colors each ellipse. Use the <b>Ellipses Transparency</b> and <b>Ellipse Color</b> menus to change the transparency and color.
<b>Show Correlations</b>	Shows or hides the correlation of each histogram in the upper left corner of each scatterplot.
<b>Show Histogram</b>	Shows either horizontal or vertical histograms in the label cells. Once histograms have been added, select <b>Show Counts</b> to label each bar of the histogram with its count. Select <b>Horizontal</b> or <b>Vertical</b> to either change the orientation of the histograms or remove the histograms.
<b>Ellipse <math>\alpha</math></b>	Sets the $\alpha$ -level used for the ellipses. Select one of the standard $\alpha$ -levels in the menu, or select <b>Other</b> to enter a different one.
<b>Ellipses Transparency</b>	Sets the transparency of the ellipses if they are colored. Select one of the default levels, or select <b>Other</b> to enter a different one. The default value is 0.2.
<b>Ellipse Color</b>	Sets the color of the ellipses if they are colored. Select one of the colors in the palette, or select <b>Other</b> to use another color. The default value is red.
<b>Nonpar Density</b>	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

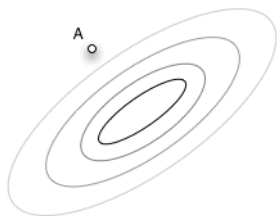
The **Outlier Analysis** menu contains options that show or hide plots that measure distance in the multivariate sense using one of these methods:

- 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 53 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 54 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 54 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  $\alpha$  (Cronbach 1951) is one measure of reliability. Two primary applications for Cronbach's  $\alpha$  are industrial instrument reliability and questionnaire analysis.

Cronbach's  $\alpha$  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  $\alpha$  can be requested if the items have variances that vary widely.

---

**Note:** Cronbach's  $\alpha$  is not related to a significance level  $\alpha$ . 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  $\alpha$  value. If  $\alpha$  increases when you exclude a variable (item), that variable is not highly correlated with the other variables. If the  $\alpha$  decreases, you can conclude that the variable is correlated with the other items in the scale. Nunnally (1978) suggests a Cronbach's  $\alpha$  of 0.7 as a rule-of-thumb acceptable level of agreement.

See "[Cronbach's  \$\alpha\$](#) " on page 55 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.

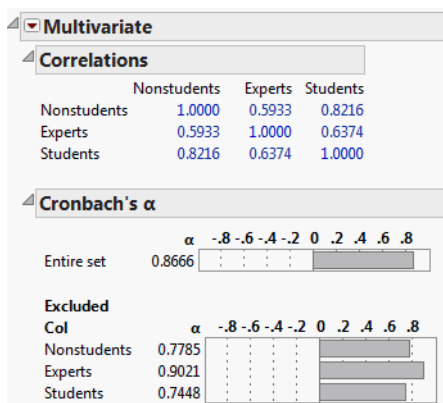
## Example of Item Reliability

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, non-students, and experts) ranked the items according to perceived level of danger. Note that Nuclear power is rated as very dangerous (1) by both students and non-students, but is ranked low (20) by experts. On the other hand, motorcycles are ranked either fifth or sixth by all three judging groups.

You can use Cronbach's  $\alpha$  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  $\alpha$** .
6. (Optional) From the red triangle menu for Multivariate, select **Scatterplot Matrix** to hide that plot.

Figure 3.6 Cronbach's  $\alpha$  Report



The Cronbach's  $\alpha$  results in Figure 3.6 show an overall  $\alpha$  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  $\alpha$  scores of 0.7785 and 0.7448, respectively.

## Computations and Statistical Details

- “Estimation Methods”
- “Pearson Product-Moment Correlation”
- “Nonparametric Measures of Association”
- “Inverse Correlation Matrix”
- “Distance Measures”
- “Cronbach’s  $\alpha$ ”

## Estimation Methods

### REML

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

### 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:

$$w_i = 1.0 \text{ if } Q < K \text{ and } w_i = K/Q \text{ otherwise,}$$

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  $i^{\text{th}}$  observation,  $\mu$  = 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 $\rho$ (rho) Coefficients

Spearman's  $\rho$  correlation coefficient is computed on the ranks of the data using the formula for the Pearson's correlation previously described.

### Kendall's $\tau_b$ Coefficients

Kendall's  $\tau_b$  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  $\tau_b$  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  $\text{sgn}(z)$  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  $i$ th group of tied  $x$  (respectively  $y$ ) values.
- The  $n$  is the number of observations.
- Kendall's  $\tau_b$  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  $\tau_b$ .

## 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:

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

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

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

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  $i$ th 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  $i$ th 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

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

where  $R_i^2$  is the coefficient of variation from the model regressing the  $i^{\text{th}}$  explanatory variable on the other explanatory variables. Thus, a large  $r^{ii}$  indicates that the  $i^{\text{th}}$  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

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

where:

$Y_i$  is the data for the  $i^{\text{th}}$  row

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

$S$  is the estimated covariance matrix for the data

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

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

where:

$n$  = number of observations

$p$  = number of variables (columns)

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

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

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

where:

$n$  = number of observations

$p$  = number of variables (columns)

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

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

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

### $T^2$ Distance Measures

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

The UCL on the  $T^2$  distance is:

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

where

$n$  = number of observations

$p$  = number of variables (columns)

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

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 $\alpha$

Cronbach's  $\alpha$  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

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

$r$  = the average correlation between items

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





# Chapter 4

## Principal Components

### Reduce the Dimensionality of Your Data

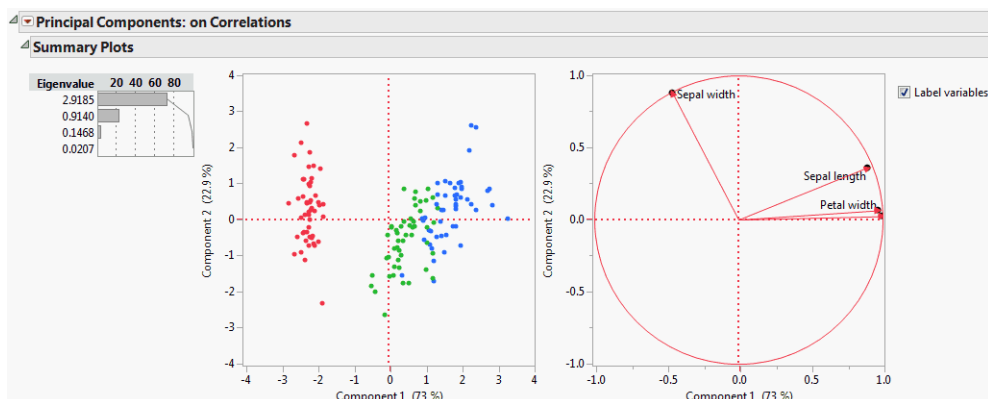
The purpose of principal component analysis is to derive a small number of independent linear combinations (*principal components*) of a set of measured variables that capture as much of the variability in the original variables as possible. Principal component analysis is a dimension-reduction technique,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

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

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

The Principal Components platform also supports factor analysis. JMP offers several types of orthogonal and oblique factor analysis-style rotations to help interpret the extracted components. For factor analysis, see the “[Factor Analysis](#)” chapter on page 171.

**Figure 4.1** Example of Principal Components



**Contents**

Overview of Principal Component Analysis Platform .....	59
Example of Principal Component Analysis .....	59
Launch the Principal Components Platform.....	60
Missing Data .....	63
Principal Components Report.....	63
Principal Components Report Options .....	64
Outlier Analysis.....	74
Statistical Details for the Principal Components Analysis Platform .....	75
Estimation Methods .....	75
DModX Calculation .....	77
Calculations for Outlier Analysis .....	77

---

## Overview of Principal Component Analysis Platform

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

If you want to see the arrangement of points across many correlated variables, you can use principal component analysis to show the most prominent directions of the high-dimensional data. Using principal component analysis reduces the dimensionality of a set of data.

Principal components is a way to picture the structure of the data as completely as possible by using as few variables as possible.

For  $p$  variables,  $p$  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 on page 171.

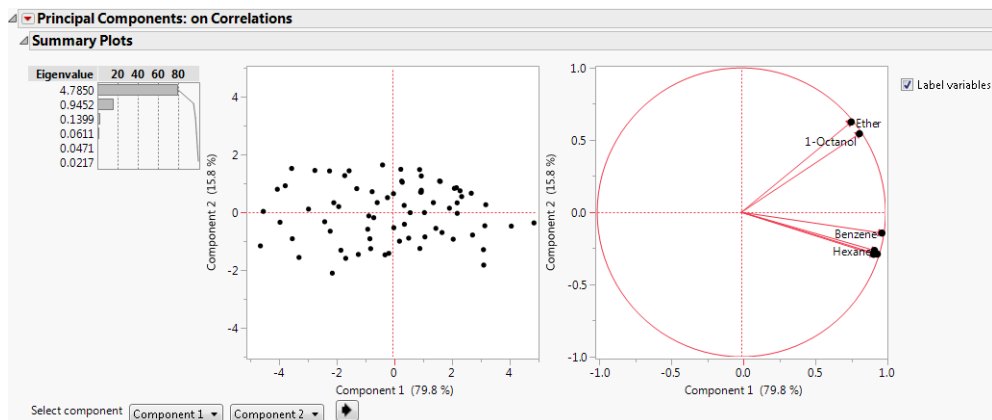
---

## Example of Principal Component Analysis

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

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

**Figure 4.2** Principal Components on Correlations Report



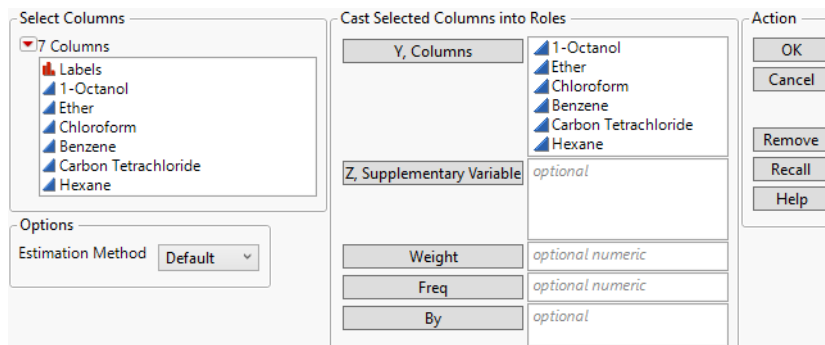
The report gives the eigenvalues and a bar chart of the percent of the variation accounted for by each principal component. In this example, the first principal component accounts for almost 80% of the variation in the data. Together, the first two principal components account for almost all of the variation in the data (95.6%). There is a Score Plot and a Loadings Plot as well. See [“Principal Components Report”](#) on page 63 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 59 uses all of the continuous variables from the Solubility.jmp sample data table.

**Figure 4.3** Principal Components Launch Window



**Y, Columns** The variables to analyze for components.

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

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

---

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

---

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

---

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

---

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

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

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

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

**REML** Restricted maximum likelihood (REML) estimation uses all of the data, even if missing values are present. Due to a bias-correction factor, this method is slow if the dataset is large and there are many missing values. Therefore, REML is most useful for smaller datasets. If there are no missing cells in the data, then the REML and ML estimates are equivalent and equal to the sample covariance matrix. If there are missing cells, REML's variance and covariance estimates are less biased than the estimates from ML estimation. For more information, see "[REML](#)" on page 75.

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

**Robust** Robust estimation uses all of the data, even if missing values are present. This method down-weights extreme values and is therefore useful for data tables that might have outliers. For statistical details, see “[Robust](#)” on page 50 in the “Correlations and Multivariate Techniques” chapter.

**Row-wise** Row-wise estimation does not use observations with missing values, so rows that contain missing cells are deleted before the method is applied. This method is useful for excluding any observations that have missing data. Row-wise estimation was the only estimation method available prior to JMP 8, so it can also be used to check compatibility with JMP versions prior to JMP 8.

**Pair-wise** Pair-wise estimation uses all of the data, even if missing values are present. This estimation method performs correlations for all rows for each pair of columns with nonmissing values. It is most useful when a data table has missing values and either more columns than rows, more than 10 columns, or more than 5,000 rows.

**Wide** Wide estimation does not use observations with missing values, so rows that contain missing cells are deleted before the method is applied. This estimation method uses an algorithm based on the full singular value decomposition. The algorithm avoids calculating the covariance matrix and is therefore computationally efficient. It is useful when you have a very large number of columns in your data. For additional information, see “[Wide](#)” on page 76.

**JMP PRO Sparse** Sparse estimation uses all of the data, even if missing values are present. This estimation method uses an algorithm based on the partial singular value decomposition, which computes only the first specified number of singular values and singular value vectors. The algorithm avoids calculating the covariance matrix, as well as unnecessary principal components and is therefore computationally efficient. It is useful when your data are sparse, meaning they contain many zeros, or when there are a large number of columns in the data. For additional information, see “[Sparse](#)” on page 76.

---

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

---

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

## Missing Data

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

- Use the Impute Missing Data option found under **Multivariate Methods > Multivariate**. See [“Impute Missing Data”](#) on page 48 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.

---

## 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.)

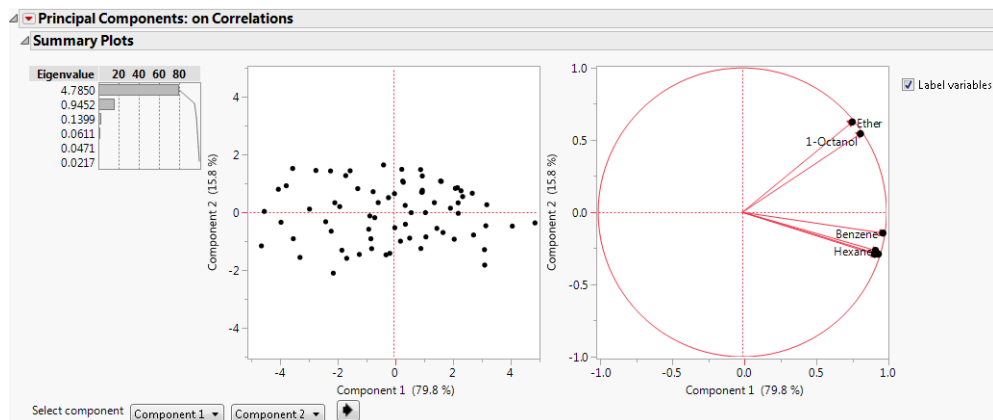
If you select the Wide method, the Wide Principal Components report appears. If you select the Sparse method, the Sparse Principal Components report appears.

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.

---

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

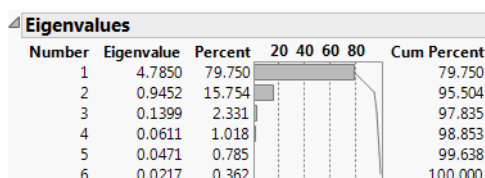
**Eigenvalues** Lists the eigenvalue that corresponds to each principal component in order from largest to smallest. The eigenvalues represent a partition of the total variation in the multivariate sample.

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

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

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

**Figure 4.5** Eigenvalues



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

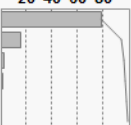
---

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

---

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

Figure 4.6 Bartlett Test

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

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

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

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

---


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


---

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

Figure 4.7 Formatted Loading Matrix

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

Suppress Absolute Loading Value Less Than  

Dim Text  

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

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

**Squared Cosines of Variables** Shows or hides a table that contains the squared cosines of variables. The sum of the squared cosine values across principal components is equal to one for each variable. The squared cosines enable you to see how well the variables are represented by the principal components. You can also determine how many principal components are necessary to represent certain variables. This option also shows a plot of the squared cosines for the first three principal components.

---

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

---

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

---

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

---

**Summary Plots** Shows or hides the summary information produced in the default report. This summary information includes a plot of the eigenvalues, a score plot, and a loading plot. By default, the report shows the score and loading plots for the first two principal

components. There are options in the report to specify which principal components to plot. See [“Principal Components Report”](#) on page 63.

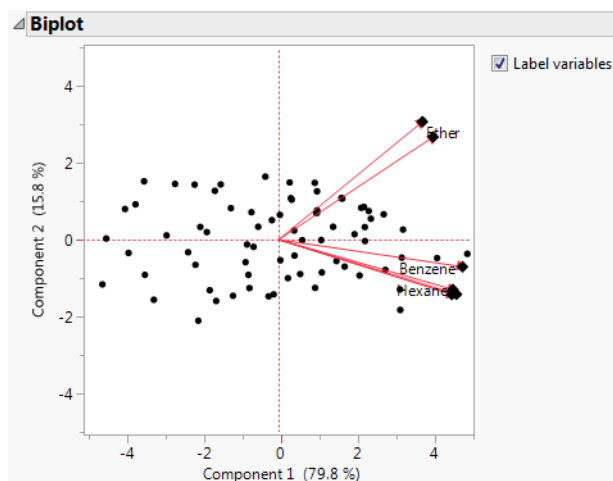
---

**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.

---

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

**Figure 4.8** Biplot



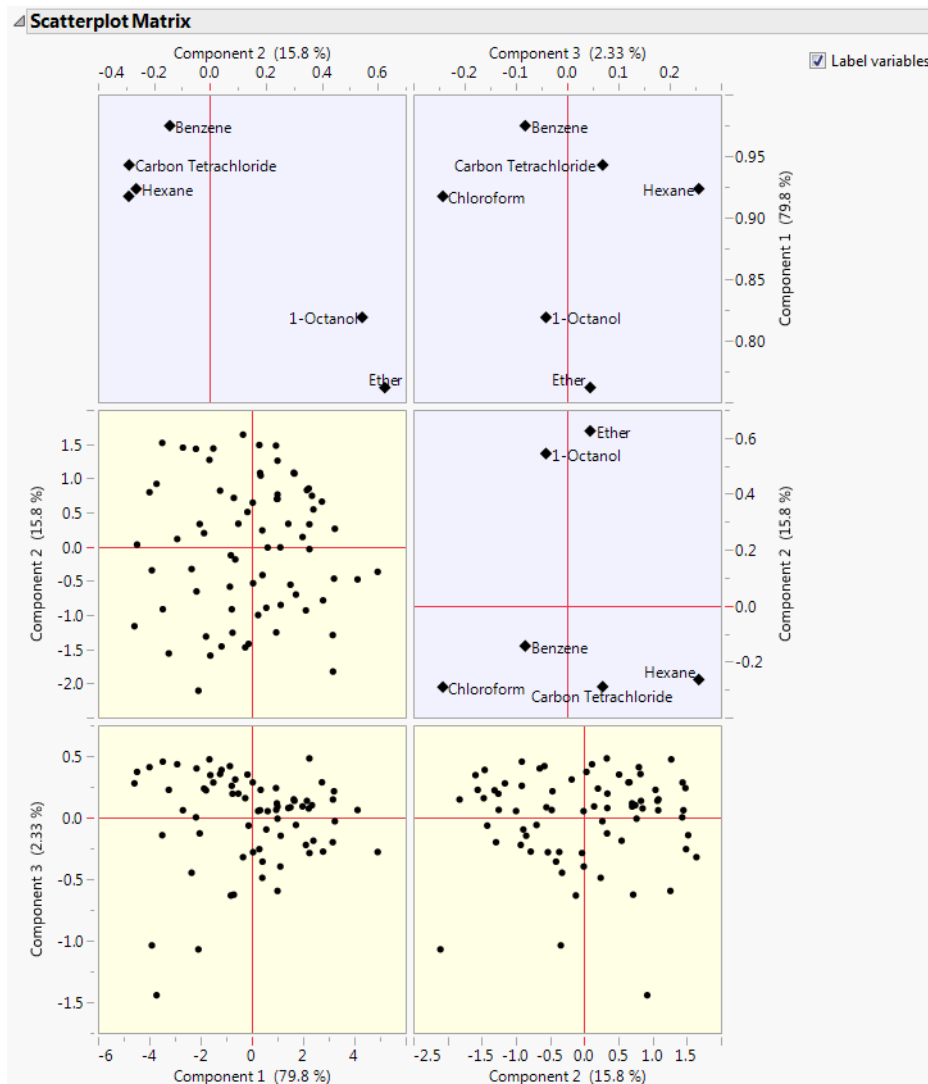
---

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

---

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

Figure 4.9 Scatterplot Matrix



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

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

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

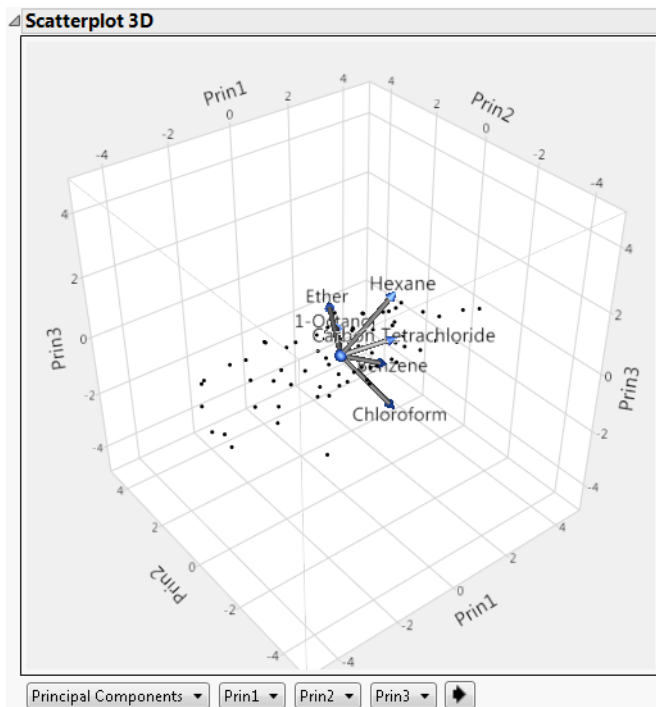
**Loading Plot** Shows or hides a matrix of two-dimensional representations of factor loadings for the specified number of components. The loading plot labels variables if the number of variables is 30 or fewer. If there are more than 30 variables, the labels are off by default. This information is shown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

**Tip:** Select the tips of arrows in the loading plot to select the corresponding columns in the data table. 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 4.10** Scatterplot 3D Score Plot



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

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

**Cycle Button**  Cycles through all axis content possibilities.

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

### Display Options

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

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

**Outlier Analysis** Shows or hides the Outlier Analysis report, which enables you to detect outliers in the data through  $T^2$  and contribution statistics. For details, see [“Outlier Analysis”](#) on page 74.

**Factor Analysis** (Not available for the Wide or Sparse estimation methods.) Performs factor analysis-style rotations of the principal components, or factor analysis. See the [“Factor Analysis”](#) chapter on page 171.

**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 291 for details.

---

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

---

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

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

- For the on Correlations option, the  $i^{\text{th}}$  principal component is a linear combination of the centered and scaled observations using the entries of the  $i^{\text{th}}$  eigenvector as coefficients.
- For the on Covariances options, the  $i^{\text{th}}$  principal component is a linear combination of the centered observations using the entries of the  $i^{\text{th}}$  eigenvector as coefficients.
- For the on Unscaled option, the  $i^{\text{th}}$  principal component is a linear combination of the raw observations using the entries of the  $i^{\text{th}}$  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. Larger DModX values indicate mild to moderate outliers in the data. For more information, see [“DModX Calculation”](#) on page 77.

**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.

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

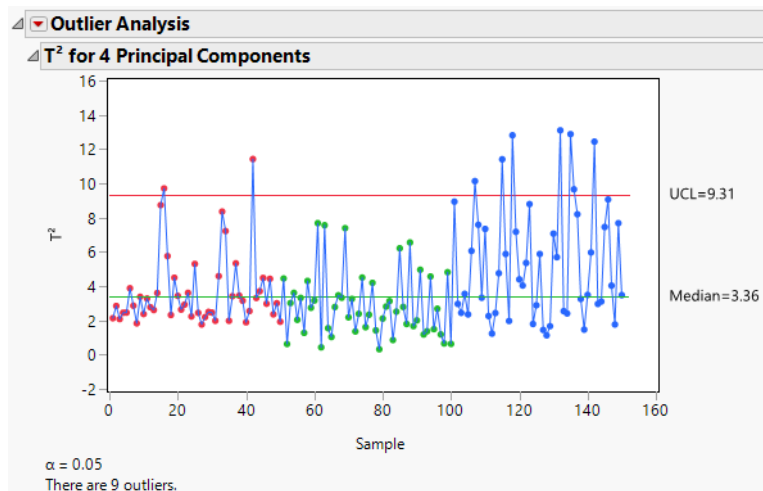
**Redo** Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

**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.

## Outlier Analysis

Figure 4.11 Outlier Analysis Report



By default, the Outlier Analysis report displays the  $T^2$  for <A> Principal Components plot. The plot shows the  $T^2$  value for each observation and horizontal lines at the Median and Upper Control Limit (UCL). For details on how the  $T^2$  values, median, and UCL are calculated, see [“Calculations for Outlier Analysis”](#) on page 77.

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

### Outlier Analysis Report Options

**T<sup>2</sup> Plot** Shows or hides the  $T^2$  plot. On by default.

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

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

**Contribution Plots with Selected Samples** (Available only if one or more points is selected in the  $T^2$  plot.) Shows a report with a  $T^2$  contribution plot for each selected sample. A  $T^2$  contribution plot shows the contribution of each variable to the sample's  $T^2$

statistic. For details on how the contributions are calculated, see [“Calculations for Outlier Analysis”](#) on page 77. Use contribution plots to investigate outliers. The variables with the largest positive or negative contributions are those that contribute most to a sample having a large  $T^2$  value. To remove the contribution plots, select Remove Plot from the contribution plot report red triangle menu.

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

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

**Set  $\alpha$  level** Enables you to specify the  $\alpha$  level.

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

**Save Contributions** Saves the  $T^2$  Contributions to new columns in the data table. There is one column for each principal component.

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

---

## Statistical Details for the Principal Components Analysis Platform

- [“Estimation Methods”](#)
- [“DModX Calculation”](#)
- [“Calculations for Outlier Analysis”](#)

### Estimation Methods

#### REML

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

with an unstructured covariance matrix. See the documentation for SAS PROC MIXED about REML estimation of mixed models.

## Wide

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

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

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

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

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

Using the singular value decomposition,  $\mathbf{X}_s$  is written as  $\mathbf{U} \text{Diag}(\Lambda) \mathbf{V}'$ . This representation is used to obtain the eigenvectors and eigenvalues of  $\mathbf{X}_s' \mathbf{X}_s$ . The principal components, or scores, are given by  $\mathbf{X}_s \mathbf{V}$ . For additional background information, see [“Wide Linear Methods and the Singular Value Decomposition”](#) on page 307 in the “Statistical Details” appendix.

## Sparse

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

Consider the same notation and standardization of  $\mathbf{X}$  that is described in [“Wide”](#) on page 76. The correlation matrix of  $\mathbf{X}$  is represented by the covariance matrix of  $\mathbf{X}_s$ :

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

The Sparse method differs from the Wide method in the calculation of the singular value decomposition. The Wide method performs a full singular value decomposition. However, the Sparse method uses an algorithm that computes only the first specified number of singular values and singular vectors in the singular value decomposition. Therefore, only the first

specified number of eigenvalues and principal components are returned. For more information about the algorithm, see Baglama and Reichel (2005).

## DModX Calculation

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

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

where

$e_{ik}$  = the residuals from the model

$K$  = the number of variables

$A$  = the number of principal components

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

## Calculations for Outlier Analysis

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

$n$  = number of observations

$A$  = number of principal components

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

### $T^2$ Statistic

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

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

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

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

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

where

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

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

### Contribution Statistic

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

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

---

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

---

# Chapter 5

## Discriminant Analysis

### Predict Classifications Based on Continuous Variables

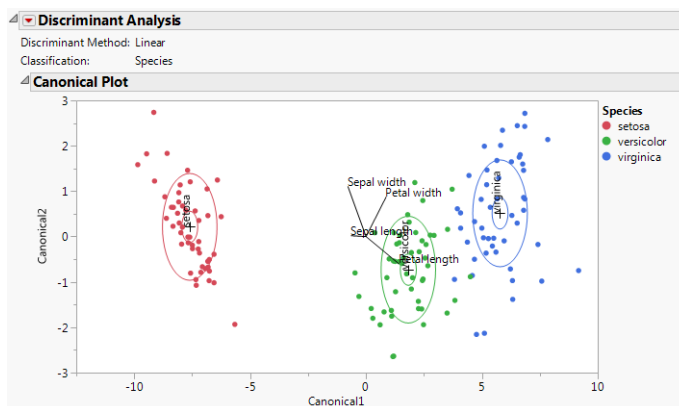
Discriminant analysis predicts membership in a group or category based on observed values of several continuous variables. Specifically, discriminant analysis predicts a classification (X) variable (categorical) based on known continuous responses (Y). The data for a discriminant analysis consist of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

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

Features of the Discriminant platform include the following:

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

**Figure 5.1** Canonical Plot



## Contents

Overview of the Discriminant Analysis Platform .....	81
Example of Discriminant Analysis .....	81
Discriminant Launch Window .....	82
Stepwise Variable Selection .....	84
Discriminant Methods .....	88
Shrink Covariances .....	91
The Discriminant Analysis Report .....	91
Principal Components .....	92
Canonical Plot and Canonical Structure .....	93
Discriminant Scores .....	97
Score Summaries .....	98
Discriminant Analysis Options .....	100
Score Options .....	102
Canonical Options .....	104
Example of a Canonical 3D Plot .....	108
Specify Priors .....	109
Consider New Levels .....	109
Save Discrim Matrices .....	110
Scatterplot Matrix .....	110
Validation in JMP and JMP Pro .....	111
Technical Details for the Discriminant Analysis Platform .....	112
Description of the Wide Linear Algorithm .....	112
Saved Formulas .....	113
Multivariate Tests .....	120
Approximate F-Tests .....	121
Between Groups Covariance Matrix .....	121



---

## Overview of the Discriminant Analysis Platform

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

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

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

Fitting methods include the following:

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

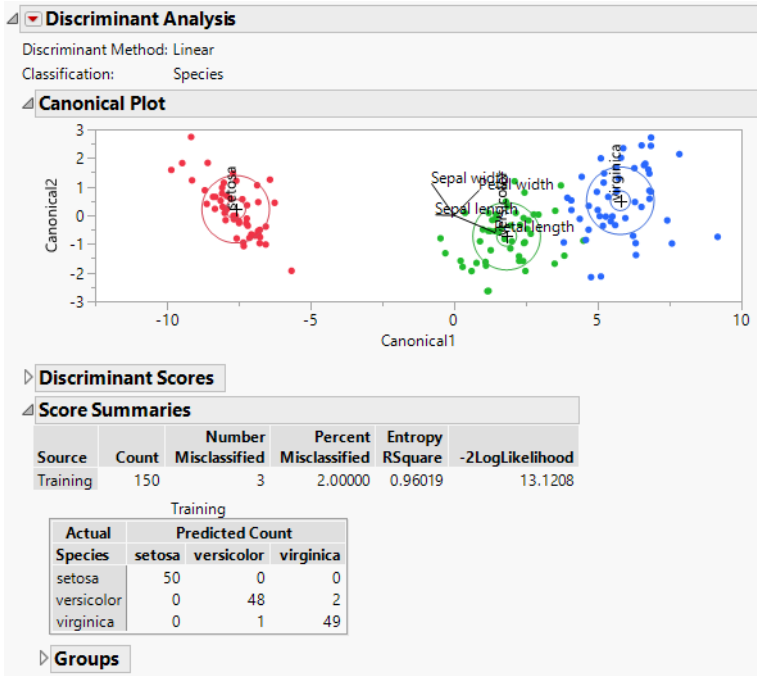
---

## Example of Discriminant Analysis

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

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

Figure 5.2 Discriminant Analysis Report Window



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 111.

**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 84.

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 88.

**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 91.

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

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

**Cross Validate by Excluded Rows** Excluded rows will form a validation set for which statistics of fit are calculated.

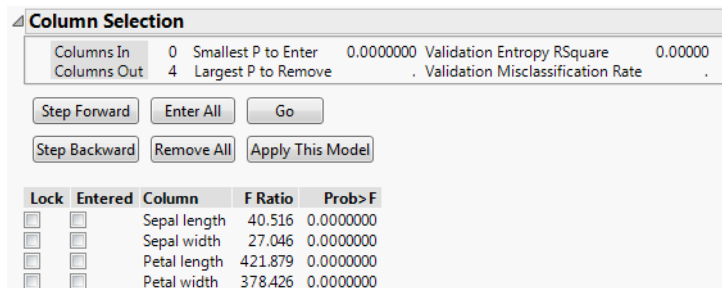
## Stepwise Variable Selection

---

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

---

If you select the Stepwise Variable Selection option in the launch window, the Discriminant Analysis report opens, showing the Column Selection panel. 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 85.

**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 99. 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 85.

**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 85.

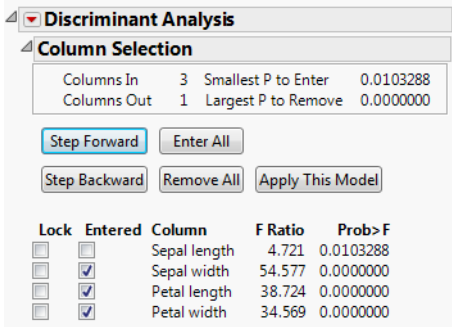
## Stepwise Example

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

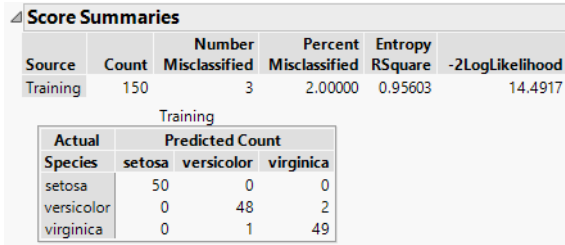


8. Click **Apply This Model**.

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

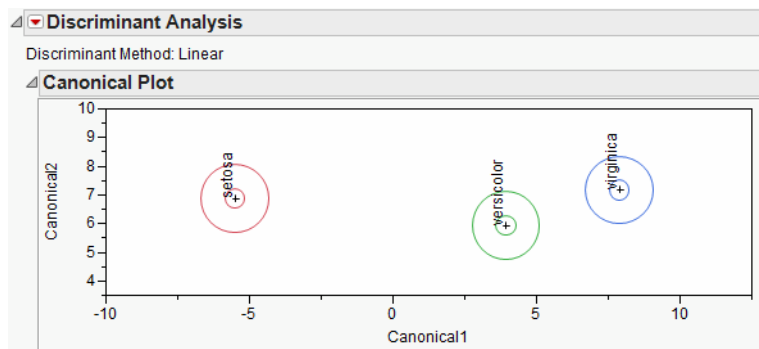
JMP offers these methods for conducting Discriminant Analysis: Linear, Quadratic, Regularized, and Wide Linear. The first three methods differ in terms of the underlying model. The Wide Linear method is an efficient way to fit a Linear model when the number of covariates is large.

**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.

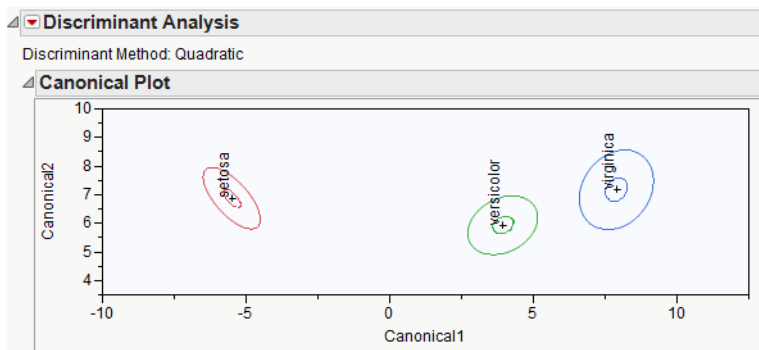


**Figure 5.7** Linear, Quadratic, and Regularized Discriminant Analysis

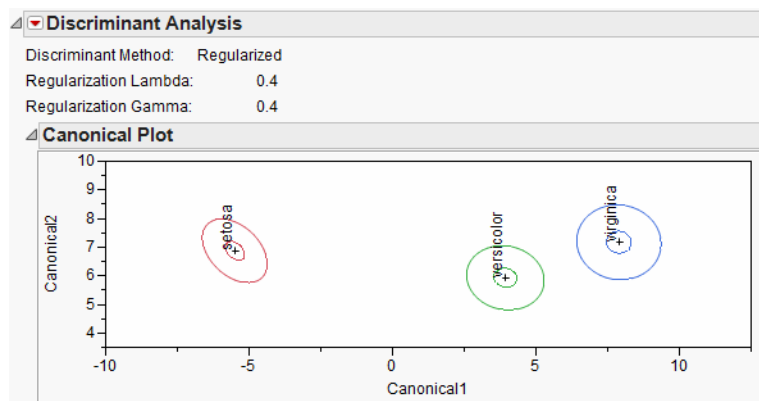
Linear



Quadratic



Regularized  
( $\lambda=0.4$ ,  $\gamma=0.4$ )



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 113.

**Linear, Common Covariance** Performs linear discriminant analysis. This method assumes that the within-group covariance matrices are equal. See [“Linear Discriminant Method”](#) on page 114.

**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 115.

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 90 and [“Regularized Discriminant Method”](#) on page 116.

**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 112.

---

**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).

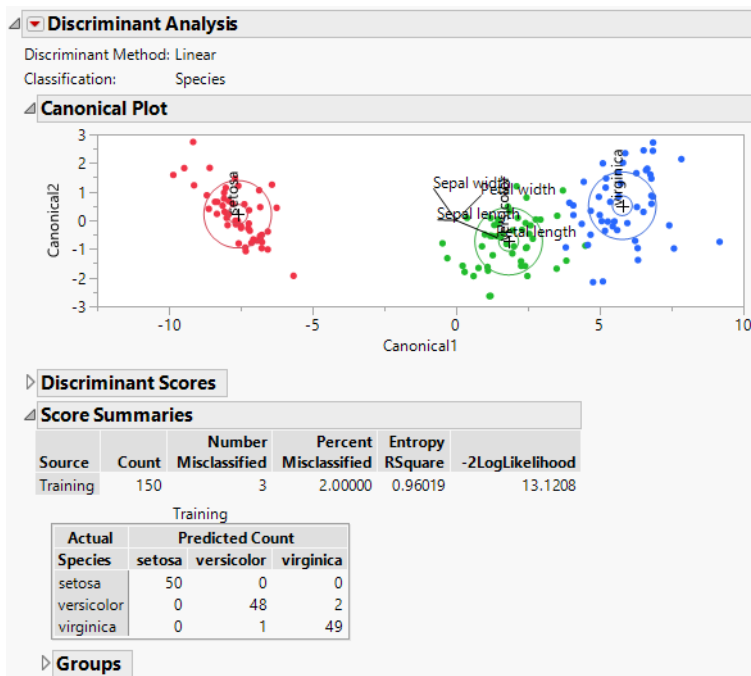
If you select the Shrink Covariances option with the Linear discriminant method in the launch window, this provides a shrinkage of the covariance matrices that is equivalent to the shrinkage provided by the Regularized discriminant method with appropriate Lambda and Gamma values. When you select the 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 92.
- 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 93.
- The Discriminant Scores report provides details about how each observation is classified. See [“Discriminant Scores”](#) on page 97.
- The Score Summaries report provides an overview of how well observations are classified. See [“Discriminant Scores”](#) on page 97.

## 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

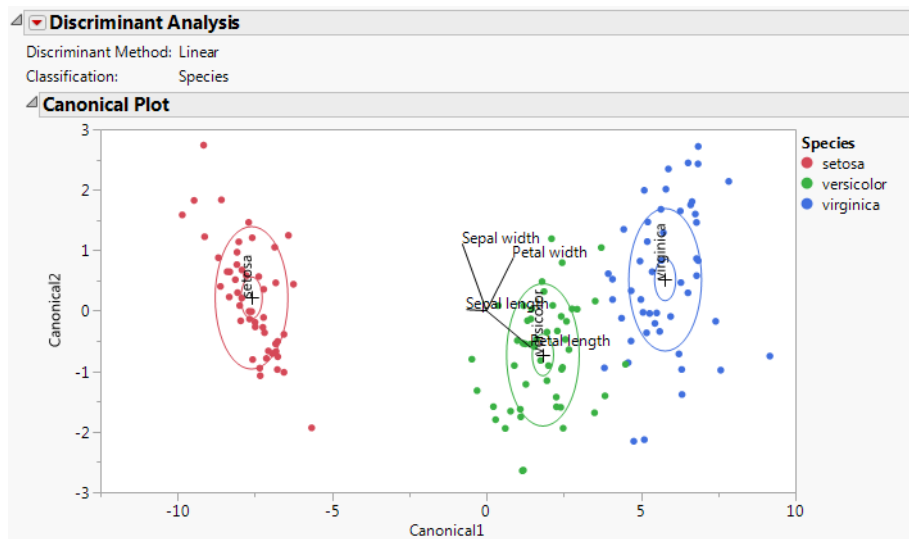
Each of the levels of the X, Categories column defines an indicator variable. A canonical correlation is performed between the set of indicator variables representing the categories and the covariates. Linear combinations of the covariates, called *canonical variables*, are derived. These canonical variables 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 93.) 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 107 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:

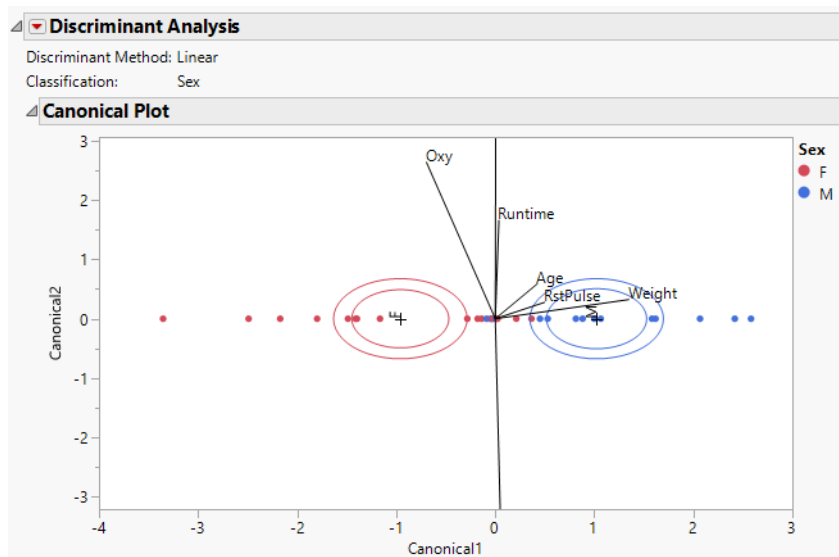
- Petal length is positively associated with Canonical1 and negatively associated with Canonical2. It carries more weight in defining Canonical1 than Canonical2.
- Petal width is positively associated with both Canonical1 and Canonical2. It carries about the same weight in defining both canonical variates.
- 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 102.

---

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

---

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

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

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

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

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

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

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

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

**Figure 5.11** Show Interesting Rows Only

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

\*\*) indicates misclassified

## Score Summaries

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

**Figure 5.12** Score Summaries for Iris.jmp

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

Training				Validation				Test			
Actual Species	setosa	versicolor	virginica	Actual Species	setosa	versicolor	virginica	Actual Species	setosa	versicolor	virginica
setosa	18	0	0	setosa	16	0	0	setosa	16	0	0
versicolor	0	22	1	versicolor	0	15	2	versicolor	0	10	0
virginica	0	0	24	virginica	0	1	15	virginica	0	0	10

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 99.

---

**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 111.

## 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 84.

**Discriminant Method** Selects the discriminant method. See [“Discriminant Methods”](#) on page 88.

**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 102.

**Canonical Plot** Shows or hides the Canonical Plot. See [“Canonical Plot and Canonical Structure”](#) on page 93.

**Canonical Options** Provides options that affect the Canonical Plot. See [“Canonical Options”](#) on page 104.

**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 108.

**Specify Priors** Enables you to specify prior probabilities for each level of the X variable. See [“Specify Priors”](#) on page 109.

**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 109.

**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 110.

**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 110.

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 111.

**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 113.

- 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 117.

---

**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.





## 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  $(1 - \text{Canonical Corr}^2)$  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 120 and ["Approximate F-Tests"](#) on page 121 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 121 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,  $S_B$ . See “[Between Groups Covariance Matrix](#)” on page 121.

**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 chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. [2017a](#)).

**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 chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. [2017a](#)).

## 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

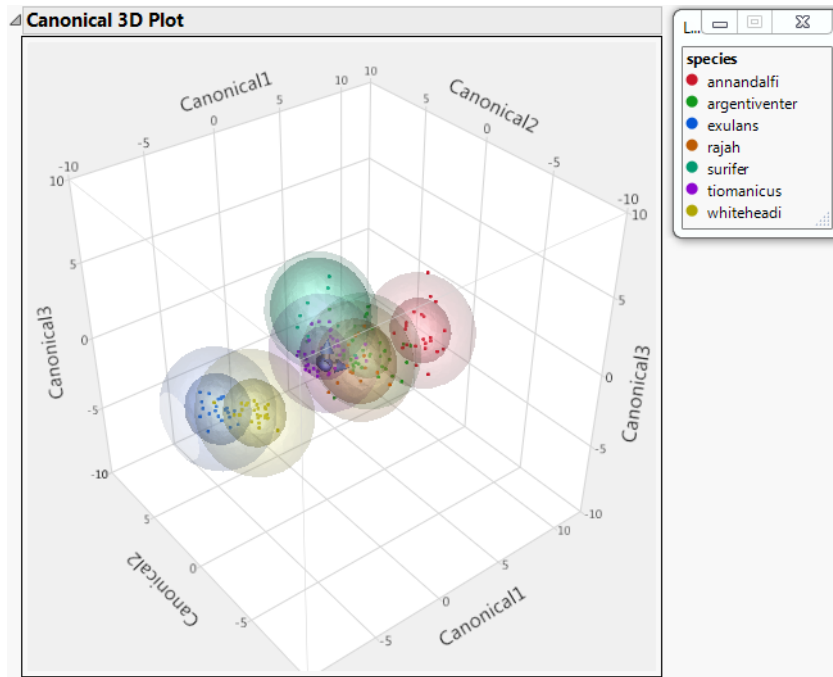
1. Select **Help > Sample Data Library** and open Owl Diet.jmp.
2. Select rows 180 through 294.  
These are the rows for which species is missing. You will hide and exclude these rows.
3. Select **Rows > Hide and Exclude**.
4. Select **Rows > Color or Mark by Column**.
5. Select species.
6. From the 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.

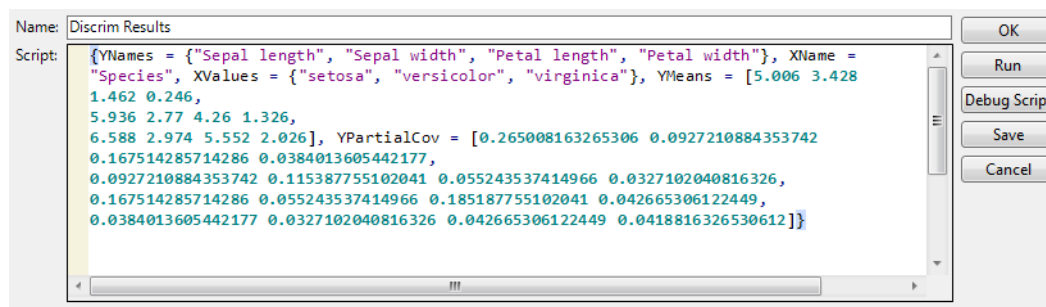
## Save Discrim Matrices

**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`



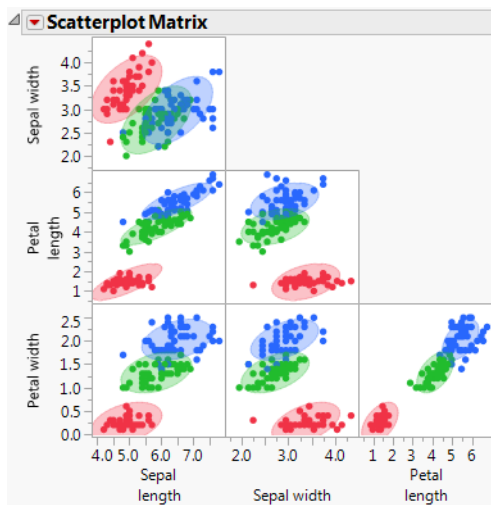
**Note:** In a script, you can send the scripting command `Get Discrim Matrices` to the Discriminant platform object. This obtains the same values as **Save Discrim Matrices**, but does not store them in the data table.

## 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 85 and [“Entropy RSquare for Validation and Test Sets”](#) on page 100.
- 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 for the Discriminant Analysis Platform

- [“Description of the Wide Linear Algorithm”](#)
- [“Saved Formulas”](#)
- [“Multivariate Tests”](#)
- [“Approximate F-Tests”](#)
- [“Between Groups Covariance Matrix”](#)

### Description of the Wide Linear Algorithm

Wide Linear discriminant analysis is performed as follows:

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



## Saved Formulas

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

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

**Table 5.2** Notation for Formulas Given by Save Formulas Options

$p$	number of covariates
$T$	total number of groups (levels of $X$ )
$t = 1, \dots, T$	subscript to distinguish groups defined by $X$
$n_t$	number of observations in group $t$
$n = n_1 + n_2 + \dots + n_T$	total number of observations
$\mathbf{y}$	$p$ by 1 vector of covariates for an observation
$\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \dots, y_{ipt})$	$i^{\text{th}}$ 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 $t$
$p(t \mathbf{y})$	posterior probability that $\mathbf{y}$ belongs to group $t$

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

$ A $	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  $S_p$ . 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:

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

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

The posterior probability of membership in group  $t$  is 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, $t = 1, \dots, T$

## 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 116.

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, $t = 1, \dots, T$

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

## Regularized Discriminant Method

Regularized discriminant analysis allows for two parameters:  $\lambda$  and  $\gamma$ .

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

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

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

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

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

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

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

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

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

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

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

**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  $j^{\text{th}}$  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  $j^{\text{th}}$  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  $\mathbf{Y}_s$ :

$$\mathbf{Y}_s = \mathbf{U} \mathbf{D} \mathbf{V}'$$

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 307 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:

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

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 114. 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 119:

$$\begin{aligned}
 d_t^2 &= (\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{S}_p^{-1} (\mathbf{y} - \bar{\mathbf{y}}_t) \\
 &= (\mathbf{y} - \bar{\mathbf{y}}_t)' \mathbf{T}_s \mathbf{T}_s' (\mathbf{y} - \bar{\mathbf{y}}_t) \\
 &= ((\mathbf{y} - \bar{\mathbf{y}}) - (\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' \mathbf{T}_s \mathbf{T}_s' ((\mathbf{y} - \bar{\mathbf{y}}) - (\bar{\mathbf{y}}_t - \bar{\mathbf{y}})) \\
 &= (\mathbf{T}_s' (\mathbf{y} - \bar{\mathbf{y}}))' (\mathbf{T}_s' (\mathbf{y} - \bar{\mathbf{y}})) - 2(\mathbf{T}_s' (\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' (\mathbf{T}_s' (\mathbf{y} - \bar{\mathbf{y}})) + (\mathbf{T}_s' (\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' (\mathbf{T}_s' (\bar{\mathbf{y}}_t - \bar{\mathbf{y}})) \\
 &= SqDist[0] - 2(\mathbf{T}_s' (\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' Discrim\ Prin\ Comp + (\mathbf{T}_s' (\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))' (\mathbf{T}_s' (\bar{\mathbf{y}}_t - \bar{\mathbf{y}}))
 \end{aligned}$$

### 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 <a href="#">“Mahalanobis Distance”</a> on page 119.
Prob[<group $t$ >]	$p(t \mathbf{y})$ , given in <a href="#">“Linear Discriminant Method”</a> on page 114
Pred <X>	$t$ for which $p(t \mathbf{y})$ is maximum, $t = 1, \dots, T$

## Multivariate Tests

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

Test statistics in the multivariate results tables are functions of the eigenvalues  $\lambda$  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  $\mathbf{E}$  and  $\mathbf{H}$  matrices are premultiplied by  $\mathbf{M}'$  and postmultiplied by  $\mathbf{M}$ .

---

- 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

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

$\mathbf{E}$  and  $\mathbf{H}$  are defined as follows:

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

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

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

The whole model  $\mathbf{L}$  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.  $\mathbf{L}$  matrices for effects are subsets of rows from the whole model  $\mathbf{L}$  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  $\mathbf{L}$  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 121, 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:

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



## Partial Least Squares Models

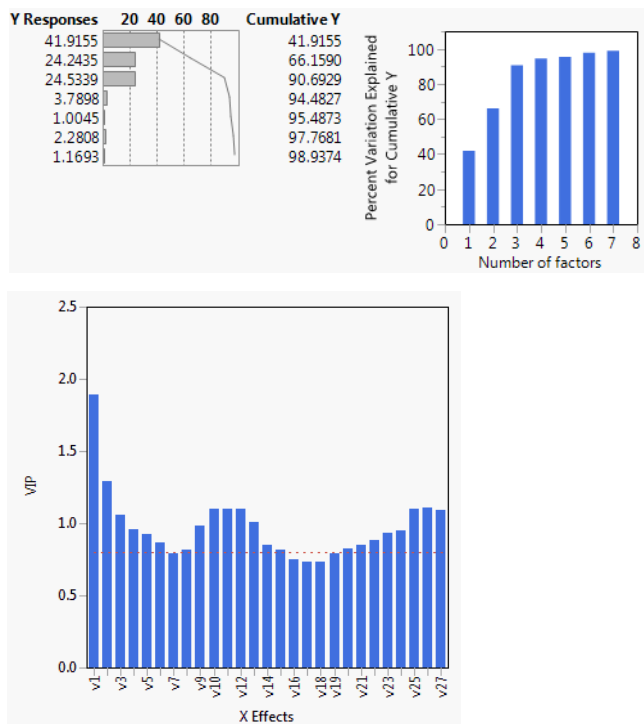
### Develop Models Using Correlations between Ys and Xs

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



## Contents

Overview of the Partial Least Squares Platform	125
Example of Partial Least Squares	126
Launch the Partial Least Squares Platform	129
Centering and Scaling	132
Standardize X	132
Model Launch Control Panel	133
Partial Least Squares Report	134
Model Comparison Summary	134
<Cross Validation Method> and Method = <Method Specification>	135
Model Fit Report	140
Partial Least Squares Options	141
Model Fit Options	141
Variable Importance Plot	143
VIP vs Coefficients Plots	144
Save Columns	145
Statistical Details for the Partial Least Squares Platform	147
Partial Least Squares	147
van der Voet $T^2$	148
$T^2_{\text{Plot}}$	149
Confidence Ellipses for X Score Scatterplot Matrix	150
Standard Error of Prediction and Confidence Limits	150
Standardized Scores and Loadings	151
PLS Discriminant Analysis (PLS-DA)	152

---


## Overview of the Partial Least Squares Platform

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

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


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

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

 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.
-  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).

---

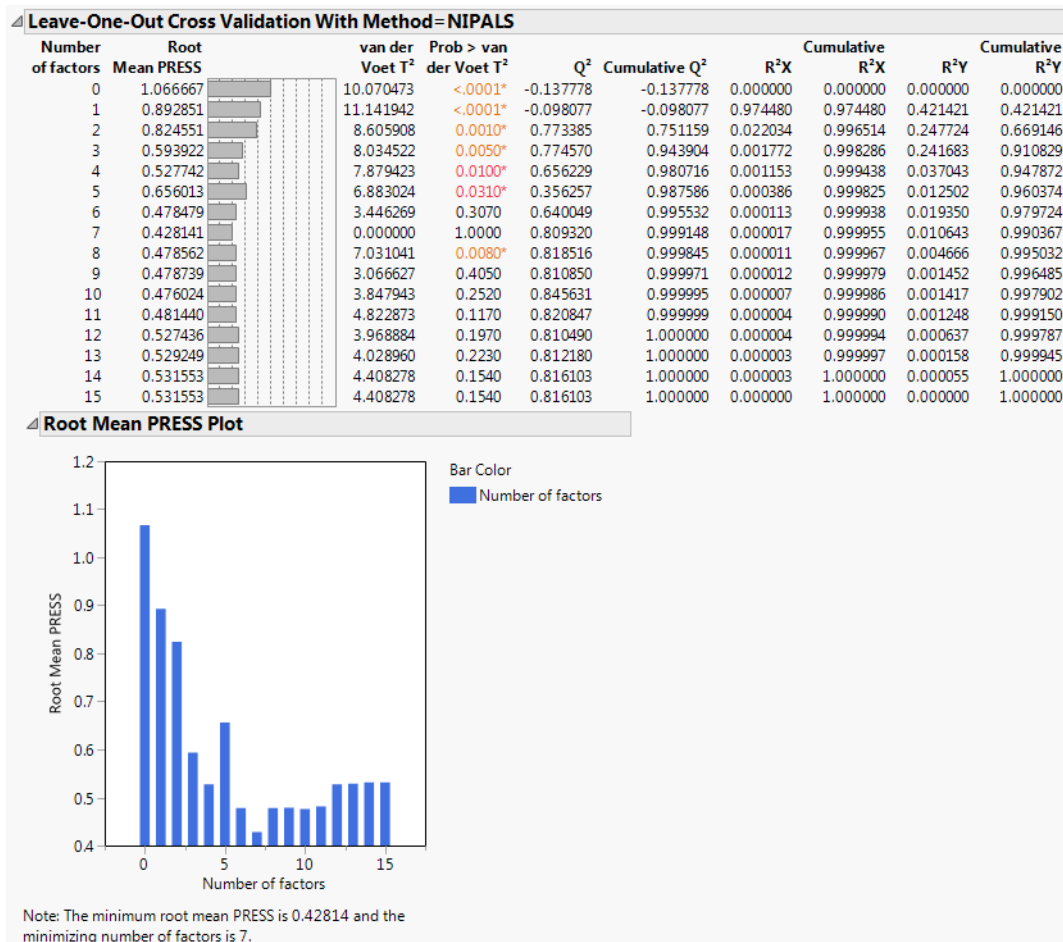
2. Select **Analyze > Multivariate Methods > Partial Least Squares**.
3. Assign ls, ha, and dt to the **Y, Response** role.
4. Assign Intensities, which contains the 27 intensity variables v1 through v27, to the **X, Factor** role.
5. Click **OK**.

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.

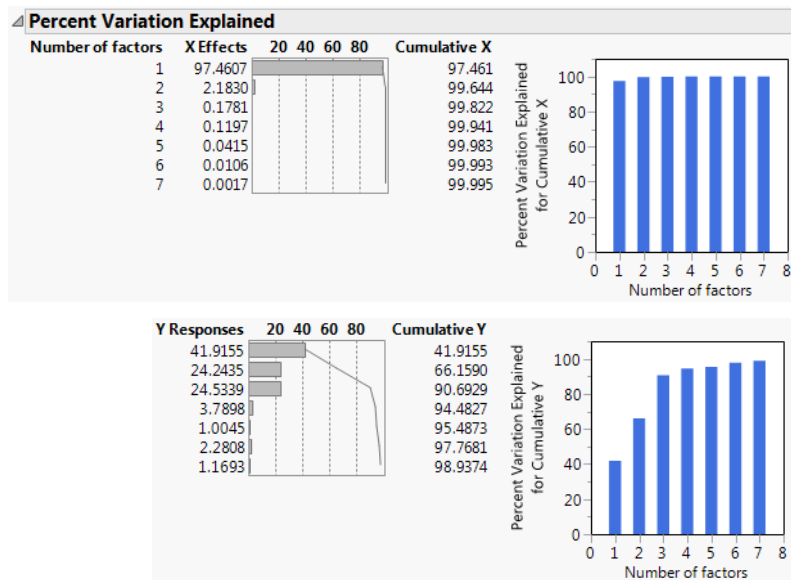
Figure 6.2 Partial Least Squares Report



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 2017d; 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.

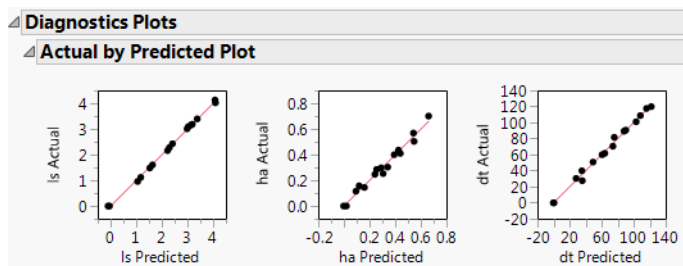
**Figure 6.3** Seven Extracted Factors



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.

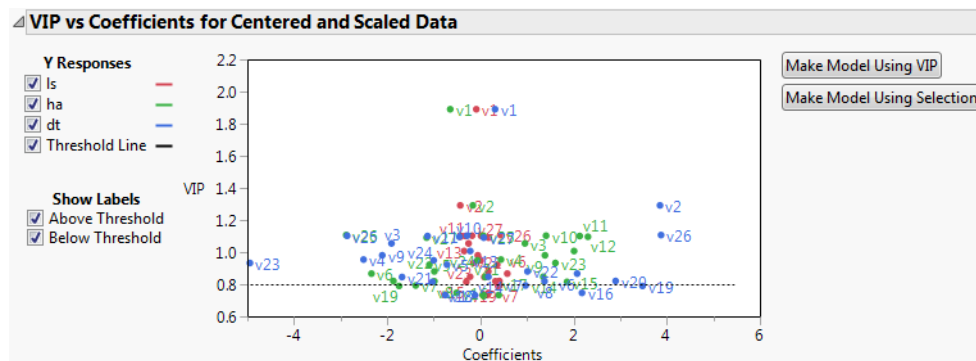
**Figure 6.4** Diagnostics Plots



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 Ys or Xs. Conduct PLS-DA by entering categorical Ys.
  - 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.

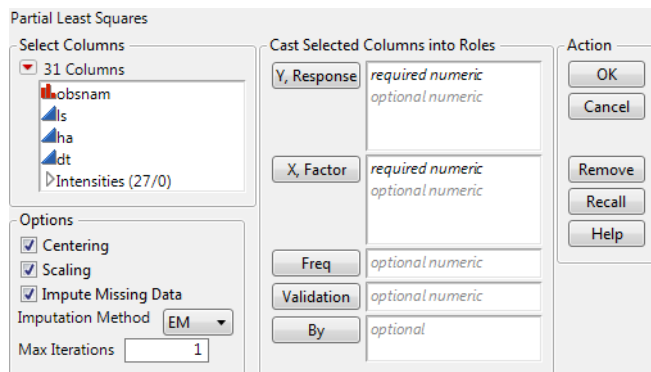
Some features on the Fit Model launch window are not applicable for the Partial Least Squares personality:

- 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 152.

**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 133.

---

**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 132.

**Scaling** Scales all Y variables and model effects by dividing each column by its standard deviation. See “[Centering and Scaling](#)” on page 132.

**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 132.

**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 133.

## 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:

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

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.

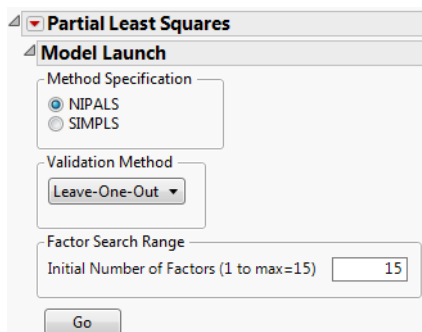
If the Standardize X option is not selected, and **Centering** and **Scaling** are both selected, then the term that is entered into the model is calculated as follows:

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

## 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 for the Partial Least Squares Platform”](#) on page 147 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.

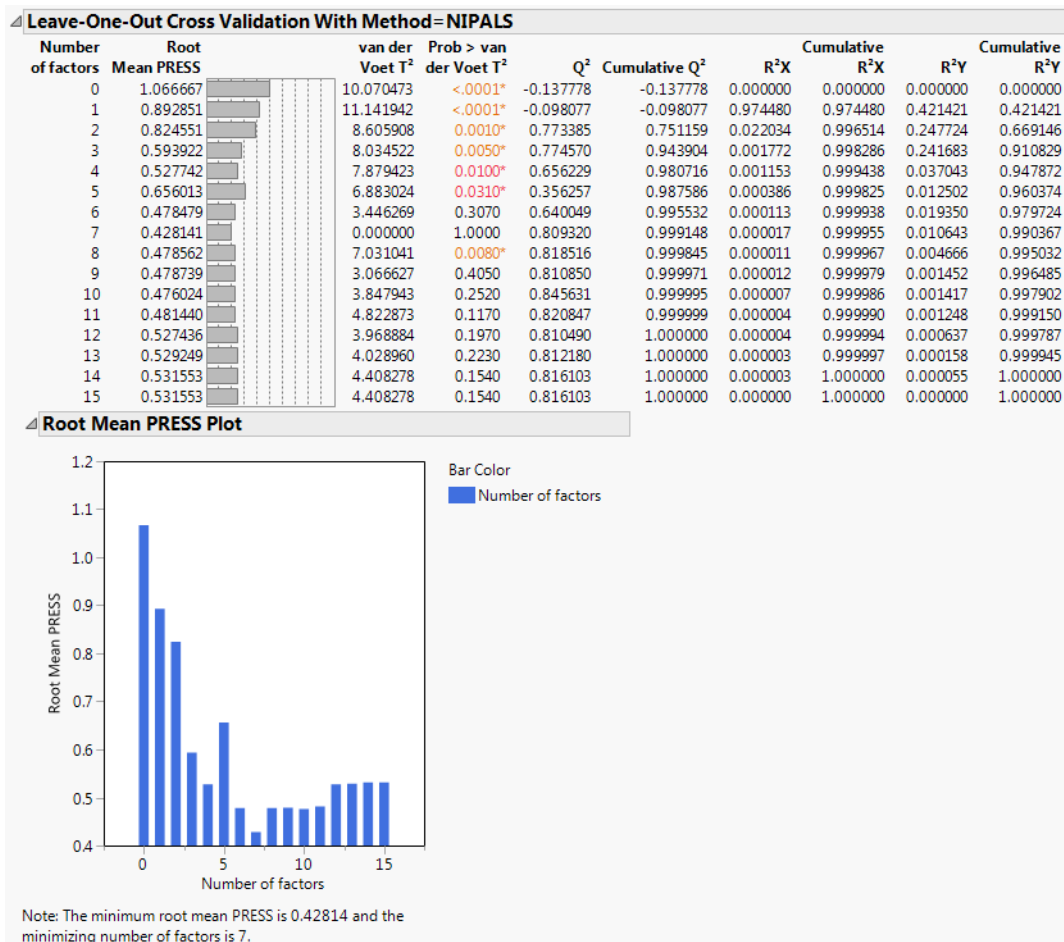
**Percent Variation Explained for Cumulative Y** Shows the percent of variation in Y that is explained by the model.

**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 [1994](#); 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 138. An optimum number of factors is identified using the minimum Root Mean PRESS statistic.

Figure 6.9 Cross Validation Report



**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 139.



**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 148.

**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 148.

**$Q^2$**  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 139.

**Cumulative  $Q^2$**  Indicator of the predictive ability of models with the given number of factors or fewer. For a given number of factors,  $f$ , Cumulative  $Q^2$  is defined as follows:

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

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 140.

**Cumulative  $R^2X$**  Percent of  $X$  variation explained by the model with the given number of factors. This is the sum of the  $R^2X$  values for  $i = 1$  to the given number of factors.

**$R^2Y$**  Percent of  $Y$  variation explained by the specified factor. A component with a large  $R^2Y$  explains a large amount of the variation in the  $Y$  variables. See [“Calculation of  \$R^{2X}\$  and  \$R^{2Y}\$  When Validation Is Used”](#) on page 140.

**Cumulative  $R^2Y$**  Percent of  $Y$  variation explained by the model with the given number of factors. This is the Sum of the  $R^2Y$  values for  $i = 1$  to the given number of factors.

### Interpretation of $Q^2$ and Cumulative $R^2Y$

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

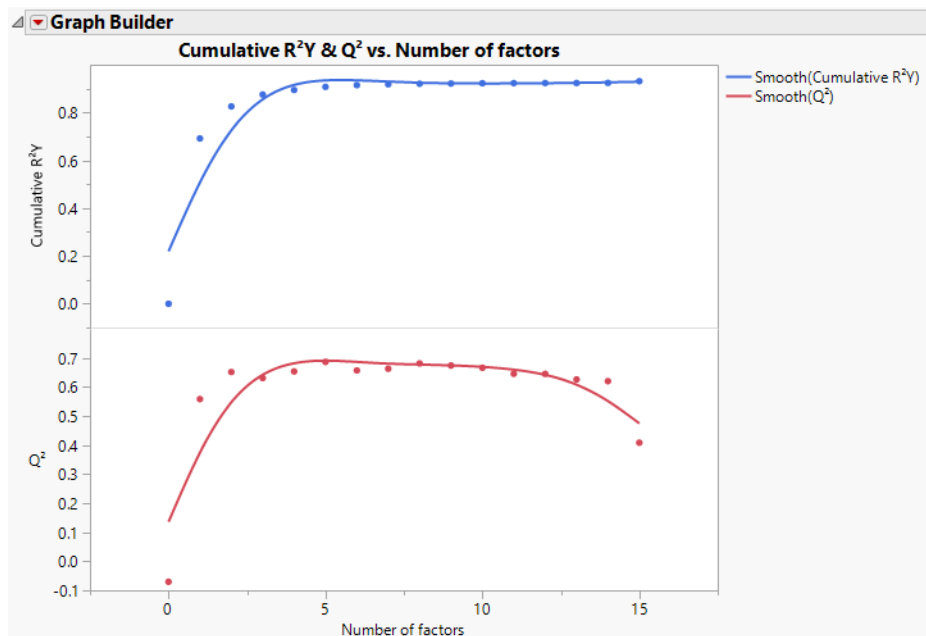
- Cumulative  $R^2Y$  increases as the number of factors increases. This is because, as factors are added to the model, more variation is explained.

- $Q^2$  tends to increase and then decrease, or at least discontinue increasing, as the number of factors increases. This is because, as more factors are added, the model becomes tuned to the training set and does not generalize well to new data, causing the PRESS statistic to decrease.

Analysis of  $Q^2$  and Cumulative  $R^2Y$  provides an alternative to using the van der Voet test for determining how many factors to include in your model. Select a number of factors for which  $Q^2$  is large and has not started decreasing. You also want Cumulative  $R^2Y$  to be large.

Figure 6.10 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  $Y$ s 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<sup>2</sup>X and R<sup>2</sup>Y When Validation Is Used

The statistics R<sup>2</sup>X and R<sup>2</sup>Y in the Cross Validation report are computed in the following ways, depending on the selected Validation Method:

**Note:** For all of these computations, R<sup>2</sup>Y is calculated analogously.

**Leave-One-Out** R<sup>2</sup>X is the average of the Percent Variation Explained for X Effects for the models constructed by leaving out one observation at a time.

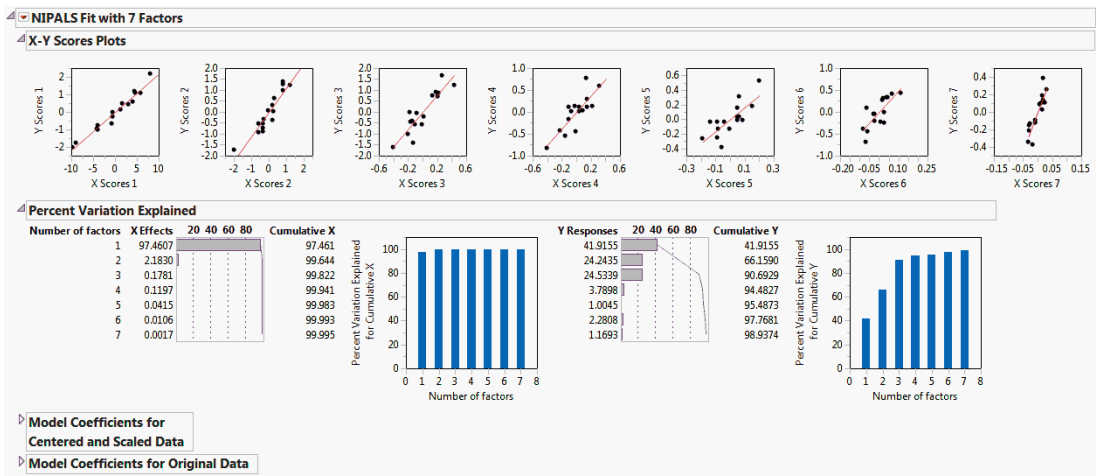
**KFold** R<sup>2</sup>X is the average of the Percent Variation Explained for X Effects for the K models constructed by leaving out each fold.

**Holdback or Validation Set** R<sup>2</sup>X 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 Xs and Ys.

**Variable Importance Plot** Plots the VIP values for each X variable. VIP scores appear in the Variable Importance Table. See [“Variable Importance Plot”](#) on page 143.

**VIP vs Coefficients Plots** Plots the VIP statistics against the model coefficients. You can show only those points corresponding to your selected Ys. Additional labeling options are provided. There are plots for both the centered and scaled data and the original data. See [“VIP vs Coefficients Plots”](#) on page 144.

**Set VIP Threshold** Sets the threshold level for the Variable Importance Plot, Variance Importance Table, and the VIP vs Coefficients Plots.

**Coefficient Plots** Plots the model coefficients for each response across the X variables. You can show only those points corresponding to your selected Ys. 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 Xs and Ys.

**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 150.

**Distance Plots** Shows plots of the following:

- the distance from each observation to the X model
- the distance from each observation to the Y model

- a scatterplot of distances to both the X and Y models

In a good model, both X and Y distances are small, so the points are close to the origin (0,0). Use the plots to look for outliers relative to either X or Y. If a group of points clusters together, then they might have a common feature and could be analyzed separately. When a validation set or a validation and test set are in use, separate reports are provided for these sets and for the training set.

**T Square Plot** Shows a plot of  $T^2$  statistics for each observation, along with a control limit. An observation's  $T^2$  statistic is calculated based on that observation's scores on the extracted factors. For details about the computation of  $T^2$  and the control limit, see "[T<sup>2</sup> Plot](#)" on page 149.

**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.

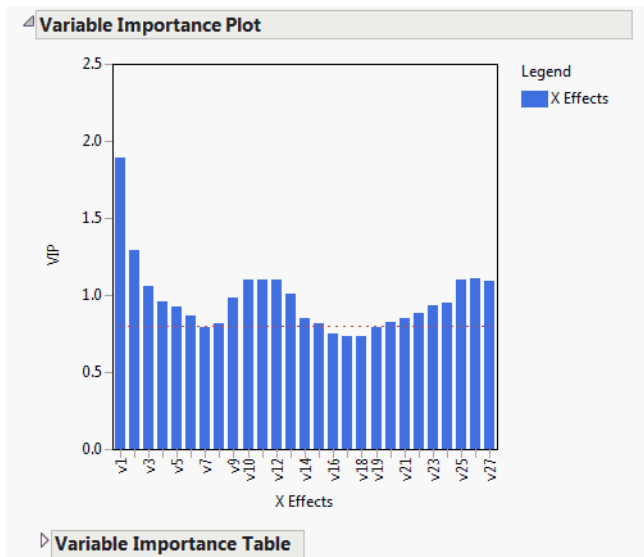
**Save Columns** Includes options for saving various formulas and results. See "[Save Columns](#)" on page 145.

**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 Ys and the variables whose VIPs exceed the specified threshold entered as Xs. Performs the same function as the button in the VIP vs Coefficients for Centered and Scaled Data report. See "[VIP vs Coefficients Plots](#)" on page 144.

## 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 (Wold1994). 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:

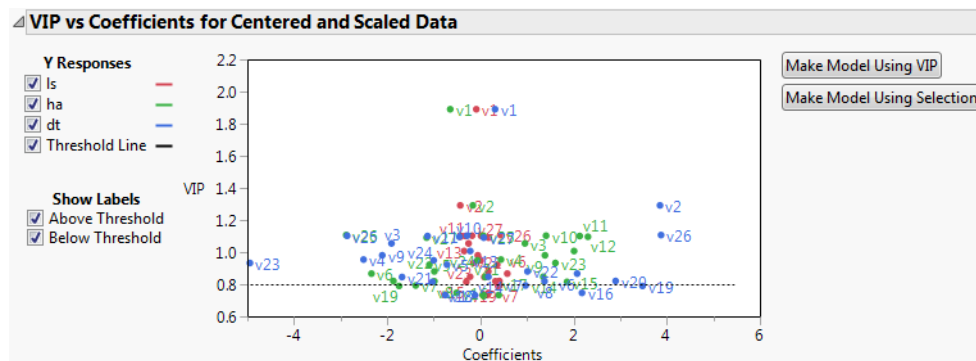
**Make Model Using VIP** Opens and populates a launch window with the appropriate responses entered as Ys and the variables whose VIPs exceed the specified threshold entered as Xs.

**Make Model Using Selection** Enables you to select Xs directly in the plot and then enters the Ys and only the selected Xs 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 150.

**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 150.

**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 150.

**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 147.

**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 151.

**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 151.

**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 for the Partial Least Squares Platform

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”
- “van der Voet  $T^2$ ”
- “ $T^2$  Plot”
- “Confidence Ellipses for X Score Scatterplot Matrix”
- “Standard Error of Prediction and Confidence Limits”
- “Standardized Scores and Loadings”
- “PLS Discriminant Analysis (PLS-DA)”

## Partial Least Squares

Partial least squares fits linear models based on linear combinations, called factors, of the explanatory variables (Xs). These factors are obtained in a way that attempts to maximize the covariance between the Xs and the response or responses (Ys). In this way, PLS exploits the correlations between the Xs and the Ys to reveal underlying latent structures. The factors address the combined goals of explaining response variation and predictor variation. Partial least squares is particularly useful when you have more X variables than observations or when the X variables are highly correlated.

### NIPALS

The NIPALS method works by extracting one factor at a time. Let  $\mathbf{X} = \mathbf{X}_0$  be the centered and scaled matrix of predictors and  $\mathbf{Y} = \mathbf{Y}_0$  the centered and scaled matrix of response values. The PLS method starts with a linear combination  $\mathbf{t} = \mathbf{X}_0\mathbf{w}$  of the predictors, where  $\mathbf{t}$  is called a *score*

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

$$\hat{\mathbf{X}}_0 = \mathbf{t}\mathbf{p}', \text{ where } \mathbf{p}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{X}_0$$

$$\hat{\mathbf{Y}}_0 = \mathbf{t}\mathbf{c}', \text{ where } \mathbf{c}' = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{Y}_0$$

The vectors  $\mathbf{p}$  and  $\mathbf{c}$  are called the *X-* and *Y-loadings*, respectively.

The specific linear combination  $\mathbf{t} = \mathbf{X}_0\mathbf{w}$  is the one that has maximum covariance  $\mathbf{t}'\mathbf{u}$  with some response linear combination  $\mathbf{u} = \mathbf{Y}_0\mathbf{q}$ . Another characterization is that the *X-* and *Y-*weights,  $\mathbf{w}$  and  $\mathbf{q}$ , 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 *Xs* and *Ys*, subject to the requirement that the *X-scores* are orthogonal. Unlike NIPALS, where the matrices  $\mathbf{X}_0$  and  $\mathbf{Y}_0$  are deflated, SIMPLS deflates the cross-product matrix,  $\mathbf{X}_0'\mathbf{Y}_0$ .

In the case of a single *Y* variable, these two algorithms are equivalent. However, for multivariate *Y*, the models differ. SIMPLS was suggested by De Jong (1993).

## van der Voet $T^2$

The van der Voet  $T^2$  test helps determine whether a model with a specified number of extracted factors differs significantly from a proposed optimum model. The test is a randomization test based on the null hypothesis that the squared residuals for both models have the same distribution. Intuitively, one can think of the null hypothesis as stating that both models have the same predictive ability.

To obtain the van der Voet  $T^2$  statistic given in the Cross Validation report, the calculation below is performed on each validation set. In the case of a single validation set, the result is the reported value. In the case of Leave-One-Out and KFold validation, the results for each validation set are averaged.

Denote by  $R_{i,jk}$  the  $j$ th predicted residual for response  $k$  for the model with  $i$  extracted factors. Denote by  $R_{opt,jk}$  is the corresponding quantity for the model based on the proposed optimum number of factors,  $opt$ . The test statistic is based on the following differences:

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

Suppose that there are  $K$  responses. Consider the following notation:

$$\mathbf{d}_{i,j} = (D_{i,j1}, D_{i,j2}, \dots, D_{i,jK})'$$

$$\mathbf{d}_{i,\cdot} = \sum_j \mathbf{d}_{i,j}$$

$$\mathbf{S}_i = \sum_j \mathbf{d}_{i,j} \mathbf{d}_{i,j}'$$

The van der Voet statistic for  $i$  extracted factors is defined as follows:

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

The significance level is obtained by comparing  $C_i$  with the distribution of values that results from randomly exchanging  $R_{i,jk}^2$  and  $R_{opt,jk}^2$ . 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:

$$((n-1)^2/n)*\text{BetaQuantile}(0.95, p/2, (n-p-1)/2)$$

where  $p$  = number of extracted factors, and  $n$  = number of observations used to train the model. If validation is not used,  $n$  = total number of observations. See Tracy et al. (1992).

## Confidence Ellipses for X Score Scatterplot Matrix

The Score Scatterplot Matrices option adds 95% confidence ellipses to the X Score scatterplots. The X scores are uncorrelated because both the NIPALS and SIMPLE algorithms produce orthogonal score vectors. The ellipses assume that each pair of X scores follows a bivariate normal distribution with zero correlation.

Consider a scatterplot for score  $i$  on the vertical axis and score  $j$  on the horizontal axis. The coordinates of the top, bottom, left, and right extremes of the ellipse are as follows:

- the top and bottom extremes are  $\pm \sqrt{\text{var}(\text{score } i)} * z$
- the left and right extremes are  $\pm \sqrt{\text{var}(\text{score } j)} * z$

where  $z = ((n-1)*(n-1)/n)*\text{BetaQuantile}(0.95, 1, (n-3)/2)$ . For background on the  $z$  value, see Tracy et al. (1992).

## Standard Error of Prediction and Confidence Limits

Let  $\mathbf{X}$  denote the matrix of predictors and  $\mathbf{Y}$  the matrix of response values, which might be centered and scaled based on your selections in the launch window. Assume that the components of  $\mathbf{Y}$  are independent and normally distributed with a common variance  $\sigma^2$ .

Hoskuldsson (1988) observes that the PLS model for  $\mathbf{Y}$  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  $\mathbf{Y}$ s. 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  $\mathbf{T}$  and consider a new observation on  $\mathbf{X}$ ,  $\mathbf{x}_0$ . The predictive model for  $\mathbf{Y}$  is obtained by regressing  $\mathbf{Y}$  on  $\mathbf{T}$ . 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  $\sigma^2$ .

### Standard Error of Prediction Formula

The standard error of the predicted mean at  $\mathbf{x}_0$  is estimated by the following:

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

### Mean Confidence Limit Formula

Let  $t_{0.975, df}$  denote the 0.975 quantile of a  $t$  distribution with degrees of freedom  $df = n - a - 1$  if the data are centered and  $df = n - a$  if the data are not centered.

The 95% confidence interval for the mean is computed as follows:

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

### Indiv Confidence Limit Formula

The standard error of a predicted individual response at  $\mathbf{x}_0$  is estimated by the following:

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

Let  $t_{0.975, df}$  denote the 0.975 quantile of a  $t$  distribution with degrees of freedom  $df = n - a - 1$  if the data are centered and  $df = n - a$  if the data are not centered.

The 95% prediction interval for an individual response is computed as follows:

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

## Standardized Scores and Loadings

Consider the following notation:

- $n_{tr}$  is the number of observations in the training set
- $m$  is the number of effects in  $X$
- $k$  is the number of responses in  $Y$
- $VarX_i$  is the percent variation in  $X$  explained by the  $i^{\text{th}}$  factor
- $VarY_i$  is the percent variation in  $Y$  explained by the  $i^{\text{th}}$  factor
- $\mathbf{XScore}_i$  is the vector of  $X$  scores for the  $i^{\text{th}}$  factor
- $\mathbf{YScore}_i$  is the vector of  $Y$  scores for the  $i^{\text{th}}$  factor
- $\mathbf{XLoad}_i$  is the vector of  $X$  loadings for the  $i^{\text{th}}$  factor
- $\mathbf{YLoad}_i$  is the vector of  $Y$  loadings for the  $i^{\text{th}}$  factor

## Standardized Scores

The vector of  $i^{\text{th}}$  Standardized X Scores is defined as follows:

$$\frac{\mathbf{XScore}_i}{(n_{tr} - 1) \sqrt{m \text{Var} X_i / n_{tr}}}$$

The vector of  $i^{\text{th}}$  Standardized Y Scores is defined as follows:

$$\frac{\mathbf{YScore}_i}{(n_{tr} - 1) \sqrt{k \text{Var} Y_i / n_{tr}}}$$

## Standardized Loadings

The vector of  $i^{\text{th}}$  Standardized X Loadings is defined as follows:

$$\mathbf{XLoad}_i \sqrt{m \text{Var} X_i}$$

The vector of  $i^{\text{th}}$  Standardized Y Loadings is defined as follows:

$$\mathbf{YLoad}_i \sqrt{k \text{Var} Y_i}$$

## 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

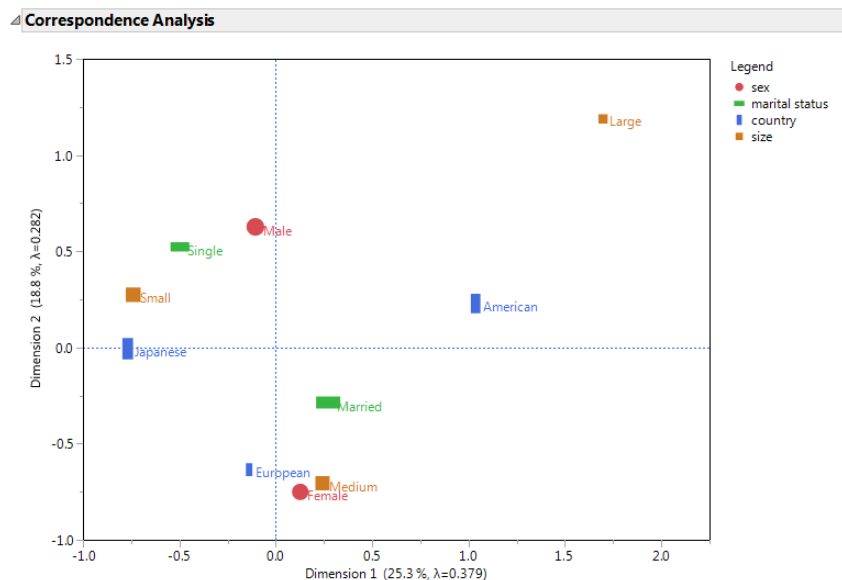
## Multiple Correspondence Analysis

### Identify Associations between Levels of Categorical Variables

Multiple Correspondence Analysis (MCA) takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA extends correspondence analysis from two variables to many. It can be thought of as analogous to principal component analysis for quantitative variables. Similar to other multivariate methods, it is a dimension reducing method; it represents the data as points in 2- or 3-dimensional space.

Multiple correspondence analysis is frequently used in the social sciences particularly in France and Japan. It can be used in survey analysis to identify question agreement. It is also used in consumer research to identify potential markets for products. Microarray studies in genetics also use MCA to identify potential relationships between genes.

**Figure 7.1** Multiple Correspondence Analysis



**Contents**

Example of Multiple Correspondence Analysis .....	155
Launch the Multiple Correspondence Analysis Platform .....	156
The Multiple Correspondence Analysis Report .....	158
Multiple Correspondence Analysis Platform Options .....	159
Correspondence Analysis Options .....	160
Show Plot .....	161
Show Detail .....	161
Show Adjusted Inertia .....	162
Show Coordinates .....	162
Show Summary Statistics .....	163
Show Partial Contributions to Inertia .....	163
Show Squared Cosines .....	164
Cross Table .....	164
Cross Table of Supplementary Rows .....	165
Cross Table of Supplementary Columns .....	165
Additional Examples of the Multiple Correspondence Analysis Platform .....	166
Example Using a Supplementary Variable .....	166
Example Using a Supplementary ID .....	167
Statistical Details for the Multiple Correspondence Analysis Platform .....	168
Details Report .....	169
Adjusted Inertia .....	169
Summary Statistics .....	170
Partial Contributions to Inertia .....	170

## Example of Multiple Correspondence Analysis

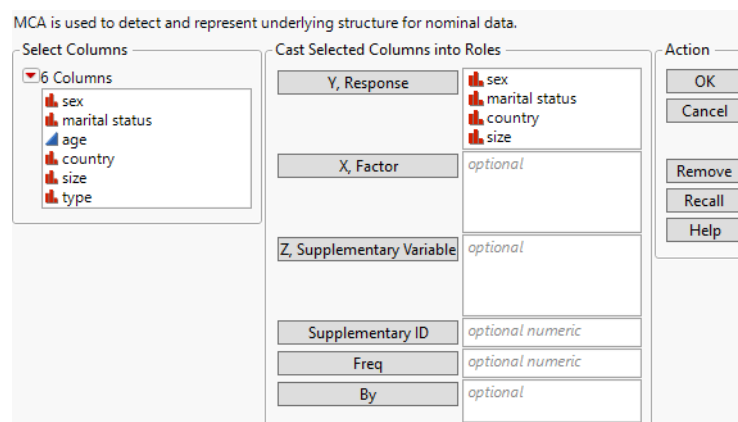
This example uses the Car Poll.jmp sample data table, which contains data collected from car polls. The data include aspects about the individuals polled, such as sex, marital status, and age. The data also include aspects about the car that they own, such as the country of origin, the size, and the type. You want to explore relationships between sex, marital status, country and size of car to identify consumer preferences.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.
3. Select sex, marital status, country, and size and click **Y, Response**.

In MCA, usually all columns are considered responses rather than some being responses and others explanatory.

4. Click **OK**.

**Figure 7.2** Completed Multiple Correspondence Analysis Launch Window

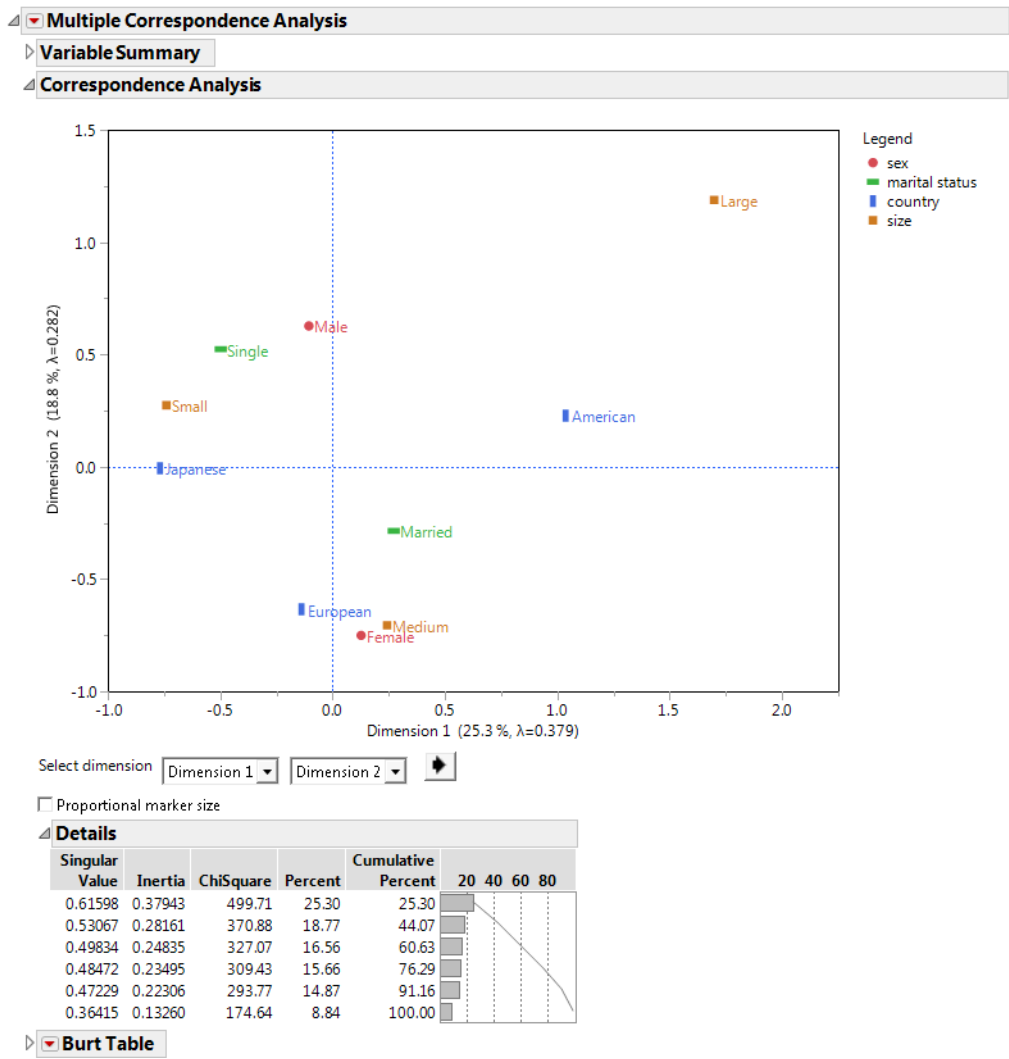


The Multiple Correspondence Analysis report is shown in Figure 7.3. Note that some of the outlines are closed.

The Variable Summary report provides a concise view of the analysis completed.

The Correspondence Analysis report shows the cloud of categories of the four variables as projected onto the two principal axes. From this cloud, you can see that Americans have a strong association with the large car size while Japanese are highly associated with the small car size. Also, males are strongly associated with the small car type and females are associated with the medium car size. This information could be used in market research to identify target audiences for advertisements.

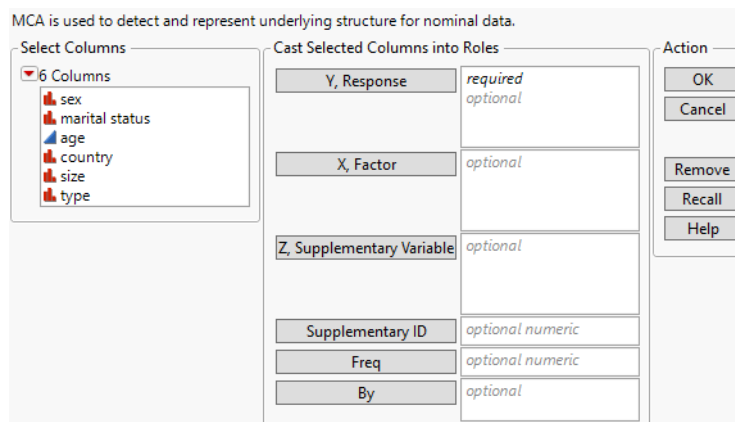
Figure 7.3 Multiple Correspondence Analysis Report



# Launch the Multiple Correspondence Analysis Platform

Launch the Multiple Correspondence Analysis platform by selecting **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.

**Figure 7.4** Multiple Correspondence Analysis Launch Window



**Y, Response** Assigns the categorical columns to be analyzed. In MCA, you are generally interested in the associations between variables, but there are not explicit “explanatory” and “response” variables.

**X, Factor** Assigns the categorical columns to be used as factor, or explanatory, variables.

**Z, Supplementary Variable** Assigns the columns to be used as supplementary variables. These variables are those you are interested in identifying associations with but not include in the calculations.

**Supplementary ID** Assigns the column that identifies rows to be used as supplementary. A supplementary ID column usually has 1s and 0s. The rows associated with ID 0 are treated as supplementary rows. The Supplementary ID column is ignored if there are levels of the X or Y variables present in the supplementary rows that are not present in the non-supplementary rows.

---

**Note:** Only one of the Supplementary ID and Z, Supplementary Variable roles can be specified at one time.

---

**Freq** Assigns a frequency variable to this role. This is useful if your data are summarized.

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

---

**Note:** The Multiple Correspondence Analysis platform handles missing values differently than many other JMP platforms. The analysis uses all nonmissing pairs of cells in a row. It does not remove entire rows from the computation.

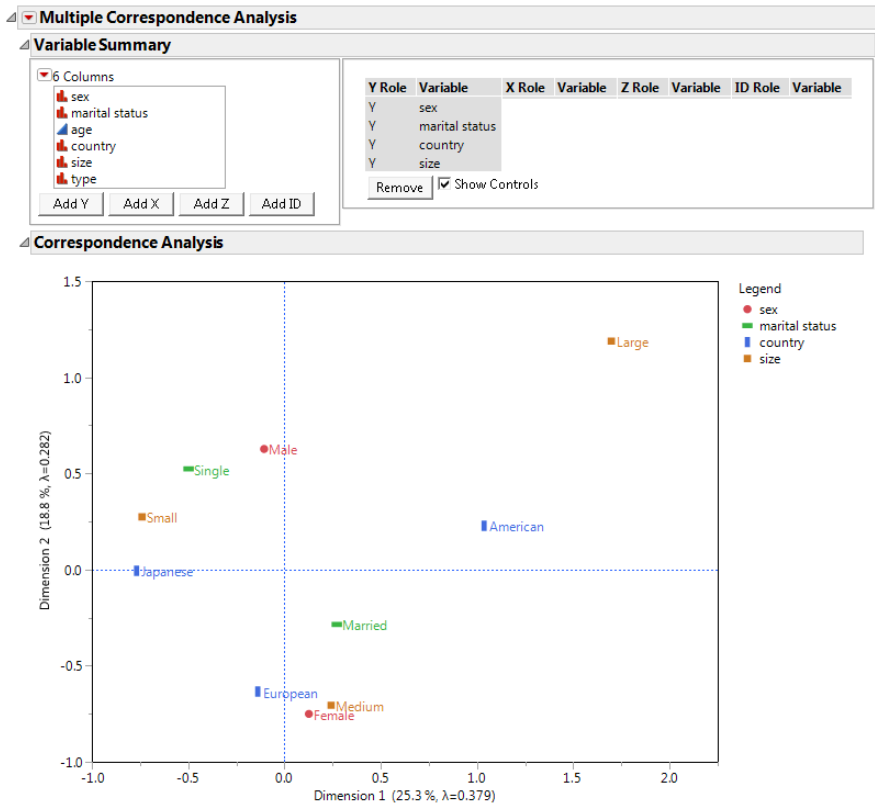
---

# The Multiple Correspondence Analysis Report

The initial Multiple Correspondence Analysis report shows the variable summary, correspondence analysis plot, and details of the dimensions of the data in order of importance. From the plot of the cloud of categories or individuals, you can identify associations that exist within the data. The details provide information about whether the two dimensions shown in the plot are sufficient to understand the relationships within the table.

The Variable Summary shows the columns used in the analysis and the roles that you selected in the launch window. If you select the Show Controls check box, a list of the columns in the data table appears to the left. You can change the columns in the analysis either by selecting a column and clicking Add Y, Add X, Add Z, or Add ID. Or you can drag the column to the header in the variable summary table. This enables you to modify the analysis without returning to the launch window.

Figure 7.5 Multiple Correspondence Analysis Report with Show Controls Selected



---

## Multiple Correspondence Analysis Platform Options

The Multiple Correspondence Analysis red triangle menu options give you the ability to customize reports according to your needs. The reports available are determined by the type of analysis that you conduct.

**Correspondence Analysis** Provides correspondence analysis reports. These reports give the plots, details, coordinates, and summary statistics. See [“Additional Examples of the Multiple Correspondence Analysis Platform”](#) on page 166.

**Cross Table** Provides the Burt table or contingency table as appropriate for variable roles selected. See [“Cross Table”](#) on page 164.

**Cross Table of Supplementary Rows** (Available only when a supplementary variable is specified.) Provides a contingency table of the supplementary variable(s) versus the response variable(s). This table appears by default only if a supplementary variable has been specified in the launch window.

**Cross Table of Supplementary Columns** (Available only when both X and Z variables are specified.) Provides a contingency table of the X, Factor variable(s) versus the supplementary variable(s). This table appears by default only if a factor variable and a supplementary variable have been specified in the launch window.

**Mosaic Plot** Displays a mosaic bar chart for each nominal or ordinal response variable. A mosaic plot is a stacked bar chart where each segment is proportional to its group's frequency count. This option is available if only one Y and only one X variable are selected.

**Tests for Independence** Provides the tests for independence whether there is association between the row and column variables. There are two versions of this test, the Pearson form and the Likelihood Ratio form, both with chi-square statistics. This option is available only when there is one Y variable and one X variable.

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.

## Correspondence Analysis Options

The reports available under Correspondence Analysis are determined by the type of analysis that you conduct. Several of these reports are shown by default.

**Show Plot** Shows the two-dimensional cloud of categories in the plane described by the first two principal axes. This plot appears by default. The plot uses isometric scaling.

**Show Detail** Provides the details of the analysis including the singular values, inertias, ChiSquare statistics, percent, and cumulative percent. This report appears by default. See [“Show Detail”](#) on page 161.

**Show Adjusted Inertia** Provides reports of the Benzecri and Greenacre adjusted inertia. See Benzécri (1979) and Greenacre (1984). This option is not available when there are one or more X variables. See [“Show Adjusted Inertia”](#) on page 162.

**Show Coordinates** Provides a report of up to the first three principal coordinates for the categories in the analysis, as appropriate. See [“Show Coordinates”](#) on page 162.

**Show Summary Statistics** Provides a report of the summary statistics, Quality, Mass, and Inertia, for each category in the analysis. See [“Show Summary Statistics”](#) on page 163.

**Show Partial Contributions to Inertia** Provides a report of the contribution of each category to the inertia for each of up to the first three dimensions. See [“Show Partial Contributions to Inertia”](#) on page 163.

**Show Squared Cosines** Provides a report of the squared cosines of each category for each of up to the first three dimensions. The report includes a bar chart that shows, for each level of each Y variable, the squared cosine values for each of up to the first three dimensions. See [“Show Squared Cosines”](#) on page 164.

**Cochran's Q Test** (Available only when all of the Y variables have the same set of only two levels and the X variable has a unique value for each row.) Provides Cochran's Q statistic, which tests that the marginal probability of a specific response is unchanged across the Y variables. Cochran's Q statistic is a generalization of McNemar's statistic for more than two response variables. See Agresti (2013).

**3D Correspondence Analysis** Shows the three-dimensional cloud of categories of the Y, X, and Z variables in the space described by the first three principal axes. This option is not available if there are less than three dimensions.



**Save Coordinates** Saves the principal coordinates to one or more JMP data tables. Column coordinates, row coordinates, supplementary column coordinates, and supplementary row coordinates are saved to separate JMP data tables. You can choose how many columns to save.

**Save Coordinate Formula** Saves formula columns to the data table for the principal coordinates in multiple dimensions. The value for each observation is the average of the coordinates for the Y variables scaled by the singular value for each dimension. You can choose how many columns to save.

## Show Plot

The plot displays a projection of the cloud of categories or individuals onto the plane described by the first two principal axes. The plot uses isometric scaling. You can toggle the dimensions shown in the plot using the Select Dimension controls below the plot. The first control defines the horizontal axis of the plot, and the second control defines the vertical axis of the plot. Click the arrow button to cycle through the dimensions shown in the plot. Below the Select Dimension controls, you can specify if the size of the points in the plot should be proportional to the count of observations corresponding to each point.

---

**Note:** Selecting a point in the correspondence analysis plot also selects the corresponding rows in other tables in the report window. However, rows in the data table are not selected. To select all of the points in the plot associated with a particular variable, select the name of the variable in the plot legend.

---

## Show Detail

Shows the table of singular values.

**Singular Value** Shows the singular values in the singular value decomposition of the contingency table or Burt table. For the formula, see [“Details Report”](#) on page 169.

**Inertia** Lists the square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

**ChiSquare** Lists the portion of the overall Chi-square for the Burt or contingency table represented by the dimension.

**Percent** Portion of inertia with respect to the total inertia.

**Cumulative Percent** Shows the cumulative portion of inertia. If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

## Show Adjusted Inertia

The principal inertias of a Burt table in MCA are the eigenvalues. The problem with these inertias is that they provide a pessimistic indication of fit. Benzécri proposed an inertia adjustment. Greenacre argued that the Benzécri adjustment overestimates the quality of fit and proposed an alternate adjustment. Both adjustments are calculated for your reference. See [“Adjusted Inertia”](#) on page 169.

**Inertia** Lists the square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

**Adjusted Inertia** Lists the adjusted inertia according to either the Benzécri or Greenacre adjustment.

**Percent** Portion of adjusted inertia with respect to the total inertia.

**Cumulative Percent** Shows the cumulative portion of adjusted inertia. If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

## Show Coordinates

Shows the Column Coordinates table or the Row and Column Coordinates tables.

**X** Lists the columns specified as X, Factor variables.

**Y** Lists the columns specified as Y, Response variables.

**Z** Lists the columns specified as Z, Supplementary Variables.

**Category** Lists the levels of the X, Y, or Z variables.

**Dimension 1, Dimension 2, Dimension 3** For each level or each response, lists its coordinate along the indicated principal axis. By default, the tables show coordinates for up to the first three dimensions. Coordinate columns for additional dimensions are hidden. To show these optional columns, right-click in a table and select the dimension columns from the **Columns** submenu.

---

**Note:** If there are columns specified as X, Factor variables, the Coordinates report displays tables of both X and Y with the same report headings. If a Z, Supplementary Variable is specified, the coordinates are listed below the X and Y coordinates as applicable.

---

## Show Summary Statistics

Shows the Summary Statistics for the Column Points table or the Summary Statistics for the Row and Column Points tables. The Y table gives Quality, Mass, and Inertia for each level of each response, called a *column point*. The X table gives Quality, Mass, and Inertia for each level of the X, Factor variables. See [“Summary Statistics”](#) on page 170.

**X** Lists the columns specified as X, Factor variables.

**Y** Lists the columns specified as Y, Response variables.

**Category** Lists the levels of the X or Y variables.

**Quality(dim=2)** Lists the quality of the representation of the level by the solution.

**Mass** Lists the row frequency for the level of the response divided by the total frequency. In the Burt table, this is the Total % for each row.

**Inertia** Lists the proportion of the total inertia accounted for by the level of the response. The inertia values sum to one across the levels and their responses.

---

**Note:** If there are columns specified as X, Factor variables, the Summary Statistics report displays tables of both X and Y with the same report headings.

---

## Show Partial Contributions to Inertia

Shows the Partial Contributions to Inertia for the Column Points table or the Partial Contributions to Inertia for the Row and Column Points tables. Also shows the Plot of Partial Contributions to Inertia for the Column Points. This is a bar chart that shows, for each level of each Y variable, its partial contributions to each of the dimensions shown in the table.

**X** Lists the columns specified as X, Factor variables.

**Y** Lists the columns specified as Y, Response variables.

**Category** Lists the levels of the X or Y variables.

**Dimension 1, Dimension 2, Dimension 3** Lists the contribution of the response or factor level to the inertia of the indicated dimension. By default, the tables show columns for up to the first three dimensions. Additional columns are hidden. To show these optional columns, right-click on a table and select the dimension columns from the **Columns** submenu.

Each level of each response contributes to the inertia of each dimension. The partial contributions within each dimension sum to 1.

---

**Note:** If there are columns specified as X, Factor variables, the Partial Contributions to Inertia report displays tables of both X and Y with the same report headings. See [“Partial Contributions to Inertia”](#) on page 170.

---

## Show Squared Cosines

Shows the Squared Cosines for the Column Points table or the Squared Cosines for the Row and Column Points. Also shows the Plot of Squared Cosines for the Column Points. This is a bar chart that shows, for each level of each Y variable, the squared cosine values for each of up to the first three dimensions shown.

**X** Lists the columns specified as X, Factor variables.

**Y** Lists the columns specified as Y, Response variables.

**Category** Lists the levels of the X or Y variables.

**Dimension 1, Dimension 2, Dimension 3** Lists the *quality* of the representation of the level by the indicated dimension. By default, the tables show results for up to the first three dimensions. Additional columns are hidden. To show these optional columns, right-click on a table and select the dimension columns from the **Columns** submenu.

The values indicate the quality of each point for the indicated dimension. The squared cosine can be interpreted as the squared correlation of the point with the dimension. The sum of the squared cosines of the first two dimensions equals  $\text{Quality}(\text{dim}=2)$  in the Summary Statistics report. See [“Summary Statistics”](#) on page 170.

---

**Note:** If there are columns specified as X, Factor variables, the Squared Cosines report displays tables of both X and Y with the same report headings.

---

## Cross Table

The Burt table is the basis of the multiple correspondence analysis. It is a partitioned symmetric table of all pairs of categorical variables. The diagonal partitions are diagonal matrices (a cross-table of a variable with itself). The off-diagonal partitions are ordinary contingency tables. When you select multiple Y, Response columns with no X, Factor columns, the Burt table is created. If you select any X, Factor columns, a traditional contingency table is created instead of a Burt table.

The red triangle menu for the Burt or contingency table contains options of statistics to display in the table.

**Count** Cell frequency, margin total frequencies, and grand total (total sample size). This appears by default.

**Total %** Percent of cell counts and margin totals to the grand total. This appears by default.

**Cell Chi Square** Chi-square values computed for each cell as  $(O - E)^2 / E$ .

**Col %** Percent of each cell count to its column total.

**Row %** Percent of each cell count to its row total.

**Expected** Expected frequency ( $E$ ) of each cell under the assumption of independence.  
Computed as the product of the corresponding row total and column total divided by the grand total.

**Deviation** Observed cell frequency ( $O$ ) minus the expected cell frequency ( $E$ ).

**Col Cum** Cumulative column total.

**Col Cum %** Cumulative column percentage.

**Row Cum** Cumulative row total.

**Row Cum %** Cumulative row percentage.

**Make Into Data Table** Creates one data table for each statistic shown in the table.

## Cross Table of Supplementary Rows

When a Z, Supplementary column is selected, a contingency table with the supplementary column levels as the rows and the response column levels as the columns is created. The red triangle menu contains the same options as the Burt Table.

## Cross Table of Supplementary Columns

When an X, Factor column and a Z, Supplementary column are selected, a contingency table with the X, Factor levels as rows and the Supplementary levels as columns is created. The red triangle menu contains the same options as the Burt Table.

---

## Additional Examples of the Multiple Correspondence Analysis Platform

- [“Example Using a Supplementary Variable”](#)
- [“Example Using a Supplementary ID”](#)

### Example Using a Supplementary Variable

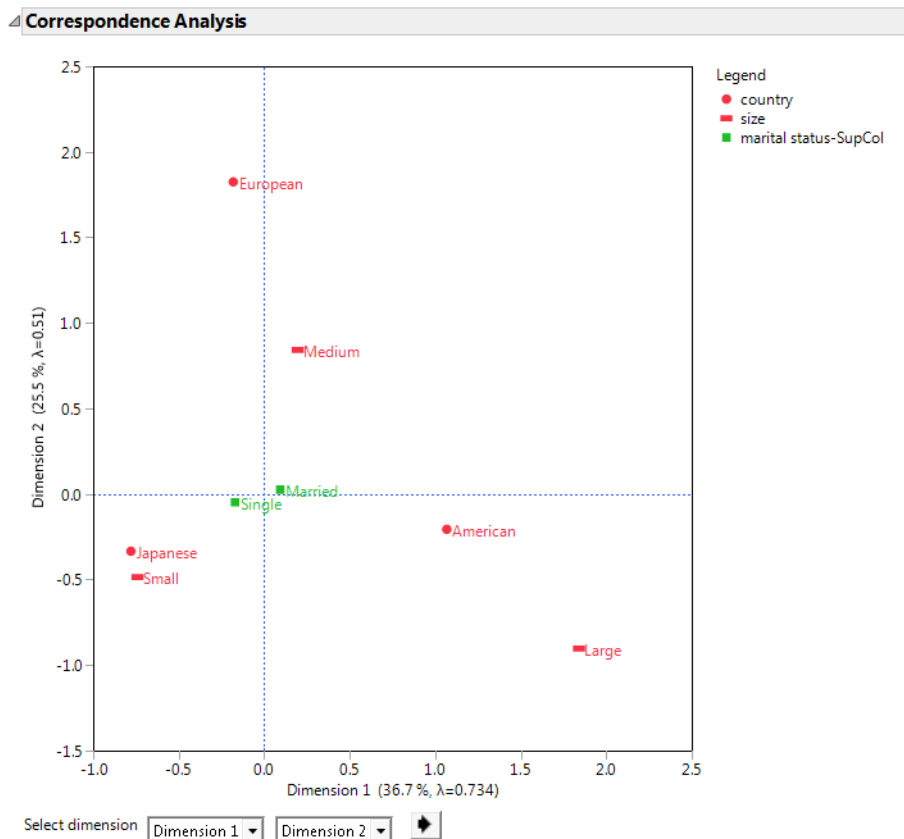
This example uses the Car Poll.jmp sample data table, which contains data collected from car polls. The data include aspects about the individuals polled, such as sex, marital status, and age. The data also include aspects about the car that they own, such as the country of origin, the size, and the type. You want to explore relationships between sex, country, and size of car to identify consumer preferences.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.
3. Select country and size and click **Y, Response**.
4. Select marital status and click **Z, Supplementary Variable**.
5. Click **OK**.

Unlike in the first example, this analysis does not use marital status in the calculations. Marital status is plotted after the calculations are complete.

You see from the plot strong relationships between Japanese and Small cars as well as American and Large cars. The two marital statuses are plotted in a different color. Single people seem to prefer smaller cars a bit more than married people.

**Figure 7.6** MCA with Supplementary Variable Report



## Example Using a Supplementary ID

The United States census allows for examining population growth over the last century. The US Regional Population.jmp sample data table contains populations of the 50 US states grouped into regions for each of the census years from 1920 to 2010. Alaska and Hawaii are treated as supplementary regions because they were not states during the entire time, and they are not part of the contiguous United States. You are interested in whether the population growth in these two states differs from the rest of the US.

1. Select **Help > Sample Data Library** and open US Regional Population.jmp.
2. Select **Analyze > Multivariate Methods > Multiple Correspondence Analysis**.
3. Select Year and click **Y, Response**.
4. Select Region and click **X, Factor**.
5. Select ID and click **Supplementary ID**.

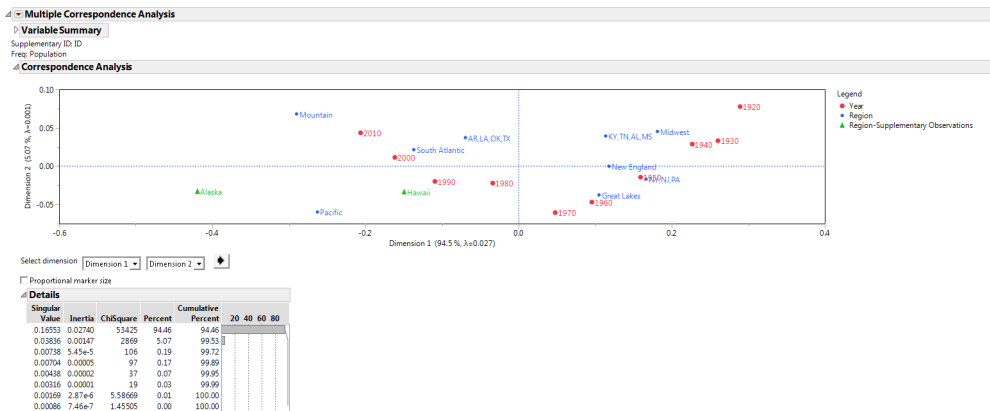
- 6. Select Population and click **Freq.**
- 7. Click **OK.**

The Details report shows that the association between years and regions is almost entirely explained by the first dimension. The plot shows that years are in the correct order on the first dimension. This ordering occurs naturally through the correspondence analysis; there is no information about the order provided to the analysis. The plot highlights the isometric scale used to plot the data.

Notice that the ordering of the regions reflects the population shift from the Midwest to the Northeast to the South and finally to the Mountain and West.

Alaska and Hawaii were not used in the computation of the analysis but are plotted based on the results. Their growth pattern is most similar to the Pacific states. Alaska’s growth is even more extreme than the Pacific region.

Figure 7.7 MCA with Supplementary ID Report



# Statistical Details for the Multiple Correspondence Analysis Platform

- “Details Report”
- “Adjusted Inertia”
- “Summary Statistics”
- “Partial Contributions to Inertia”



## Details Report

When a simple Correspondence Analysis is performed, the report lists the singular values of the following equation:

$$\mathbf{D}_r^{-0.5}(\mathbf{P} - r\mathbf{c}')\mathbf{D}_c^{-0.5} = \mathbf{U}\mathbf{D}_u\mathbf{V}'$$

where:

- $\mathbf{P}$  is the matrix of counts divided by the total frequency
- $r$  and  $c$  are the row and column sums of  $\mathbf{P}$
- the  $\mathbf{D}$  matrices are diagonal matrices of the values of  $r$  and  $c$

When Multiple Correspondence Analysis is performed, the singular value decomposition extends to:

$$\mathbf{D}^{-0.5}\left(\frac{\mathbf{C}}{Q^2n} - \mathbf{D}\mathbf{1}\mathbf{1}'\mathbf{D}\right)\mathbf{D}^{-0.5} = \mathbf{U}\mathbf{D}_u\mathbf{V}'$$

where:

$$\mathbf{D} = \left(\frac{1}{m}\right)\text{diag}(\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_Q)$$

- $\mathbf{C}$  is the Burt table.
- $Q$  is the number of categorical variables
- $n$  is the number of observations
- $\mathbf{1}$  is a column vector of ones

## Adjusted Inertia

The usual principal inertias of a Burt table constructed from  $m$  categorical variables in MCA are the eigenvalues  $u_k$  from  $\mathbf{D}_u^2$ . These inertias provide a pessimistic indication of fit. Benzécri (1979) proposed the following inertia adjustment; it is also described by Greenacre (1984, p. 145):

$$\left(\frac{m}{m-1}\right)^2 \times \left(u_k - \frac{1}{m}\right)^2 \text{ for } u_k > \frac{1}{m}$$

This adjustment computes the percent of adjusted inertia relative to the sum of the adjusted inertias for all inertias greater than  $1/m$ .

Greenacre (1984, p. 156) argues that the Benzécri adjustment overestimates the quality of fit. Greenacre proposes instead to compute the percentage of adjusted inertia relative to:

$$\frac{m}{m-1} \left( \text{trace}(\mathbf{D}_u^4) - \frac{n_c - m}{m^2} \right)$$

for all inertias greater than  $1/m$ , where  $\text{trace}(\mathbf{D}_u^4)$  is the sum of squared inertias and  $n_c$  is the total number of categories across the  $m$  variables.

## Summary Statistics

*Quality* is the ratio of the squared distance of a point from the origin in the space defined by the specified number of dimensions to the distance from the origin in the space with the maximum number of dimensions. For the Chi-Square metric, a point's quality in a given dimension can be obtained from the cosine that its vector makes with the vector that defines the dimension. Quality is also equal to the ratio of the sum of inertias in the specified dimensions to the sum of the inertias in all dimensions. Quality indicates how well the point is represented in the lower-dimensional space.

*Mass* is the proportion of row or column total frequency to the total frequency.

*Inertia* is analogous to variance in principal component analysis. The overall inertia is the total Pearson Chi-square for a two-way frequency table divided by the sum of all observations in the table.

*Relative inertia* is the proportion of the contribution of the point to the overall inertia. In the summary statistics table, the relative inertia is listed in the column labeled Inertia.

## Partial Contributions to Inertia

The contribution of a row or column to the inertia of a dimension is calculated as:

$$\text{contribution} = \frac{\text{mass} \times \text{coordinate}^2}{\text{dimensioninertia}}$$

# Chapter 8

## Factor Analysis

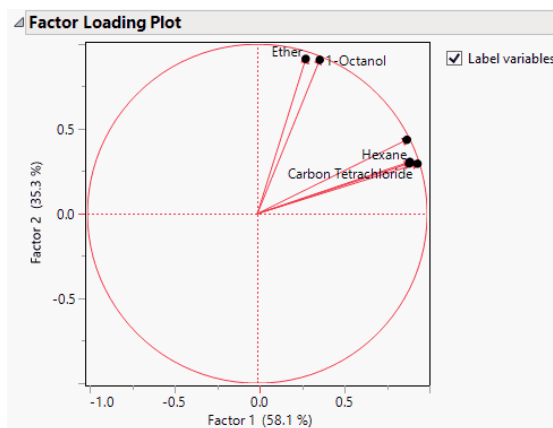
### Identify Latent Variables in Your Data

Factor analysis seeks to describe observed variables in terms of a smaller number of (unobservable) latent variables, or factors. Factor analysis is also known as common factor analysis and exploratory factor analysis. The factors are defined as linear combinations of the observed variables. They are constructed to explain variation that is *common* to the observed variables. A goal of factor analysis is to find a meaningful interpretation of the observed variables in terms of the unobserved factors. Another goal is to reduce the number of variables.

Factor analysis is used in many areas, with roots in psychology, sociology, and education. In these areas, factor analysis is used to understand how observed behavior can be interpreted in terms of underlying patterns and structures. For example, measures of participation in outdoor activities, hobbies, exercise, and travel, might all relate to a factor that can be described as “active versus inactive personality type”.

Use factor analysis when you need to explore or interpret underlying patterns and structure in your data. Also consider using it to summarize the information in your variables using a smaller number of latent variables.

**Figure 8.1** Rotated Factor Loading



**Contents**

Overview of the Factor Analysis Platform ..... 173

Example of the Factor Analysis Platform ..... 174

Launch the Factor Analysis Platform ..... 176

The Factor Analysis Report ..... 177

    Model Launch ..... 178

    Rotation Methods ..... 180

Factor Analysis Platform Options ..... 181

Factor Analysis Model Fit Options ..... 182

## Overview of the Factor Analysis Platform

Factor analysis models a set of observable variables in terms of a smaller number of unobservable factors. The factors are constructed to account for the correlation or covariance between the observed variables. Factor rotation is used to change the reference axes of the factors to increase their interpretability.

Consider a situation where you have ten observed variables,  $X_1, X_2, \dots, X_{10}$ . Suppose you want to model these ten variables in terms of two latent factors,  $F_1$  and  $F_2$ . For convenience, it is assumed that the factors are uncorrelated and that each has mean zero and variance one. The model that you want to derive is of the form:

$$X_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + \varepsilon_i$$

It follows that  $Var(X_i) = \beta_{i1}^2 + \beta_{i2}^2 + Var(\varepsilon_i)$ . The portion of the variance of  $X_i$  that is attributable to the factors, the common variance or *communality*, is  $\beta_{i1}^2 + \beta_{i2}^2$ . The remaining variance,  $Var(\varepsilon_i)$ , is the unique variance, and is considered to be a combination of specific and error variances that are unique to  $X_i$ .

The platform provides a scree plot for the eigenvalues of the correlation or covariance matrix. You can use this as a guide in determining the number of factors to extract. The platform's default number of factors is the number of eigenvalues that exceed one.

The platform provides two factoring methods for estimating the parameters of this model: Principal Axis and Maximum Likelihood. There are two Prior Communality options for estimating the proportion of variance contributed by common factors for each variable. These options impose assumptions on the diagonal of the correlation (or covariance) matrix. The Principal Components option treats the correlation matrix, which has ones on its diagonal (or the covariance matrix with variances on its diagonal), as the structure to be analyzed. The Common Factor Analysis option sets the diagonal entries to square multiple comparisons. These are values that reflect the proportion of the variation shared with other variables.

Factor rotation is used to support interpretability of the extracted factors. The Factor Analysis platform provides a variety of rotation methods that encompass both orthogonal and oblique rotations.

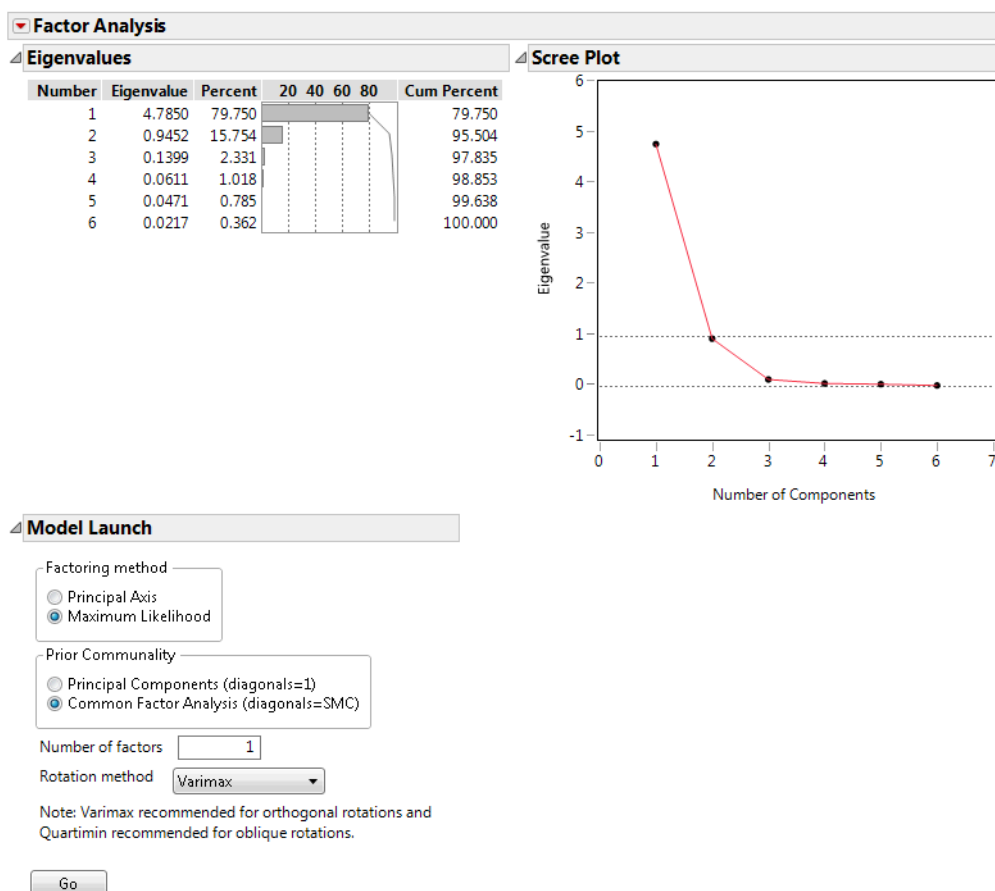
In contrast to factor analysis, which looks at common variance, principal component analysis accounts for the total variance of the observed variables. See the [“Principal Components”](#) chapter on page 57.

For more information about factor analysis, see Jöreskog (1977) or Cudeck and MacCallum (2007).

## Example of the Factor Analysis Platform

1. This example uses the Solubility.jmp sample data table to extract two factors explained by six solvents. Select **Help > Sample Data Library** and open Solubility.jmp.
2. Select **Analyze > Multivariate Methods > Factor Analysis**.
3. Select 1-Octanol through Hexane and click **Y, Columns**.
4. Click **OK**.

**Figure 8.2** Initial Factor Analysis Report

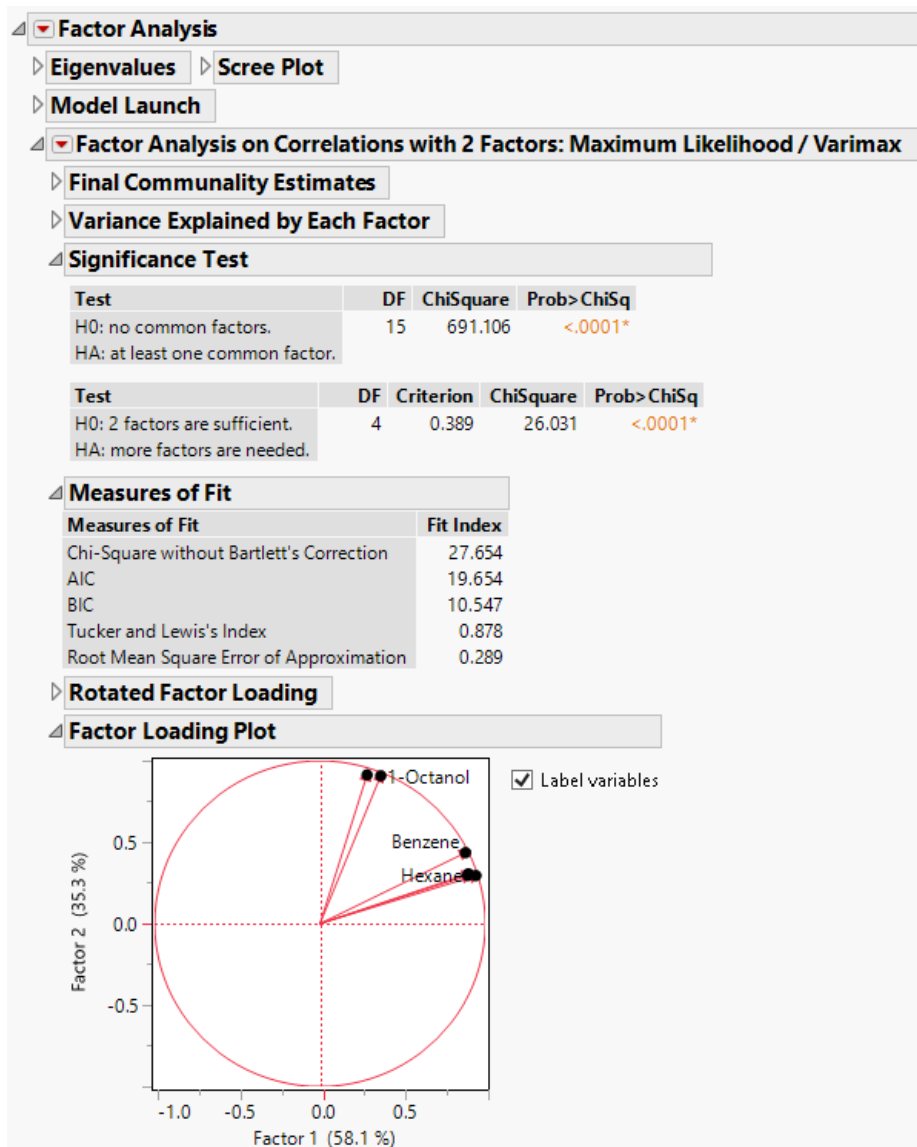


The elbow on the scree plot indicates that a model with two factors is appropriate.

5. In the Model Launch outline, make the following selections:
  - Factoring Method= **Maximum Likelihood**

- Prior Communality= **Common Factor Analysis**
  - Number of factors = 2
  - Rotation Method= **Varimax**
6. Click **Go**.

**Figure 8.3** Factor Analysis Report



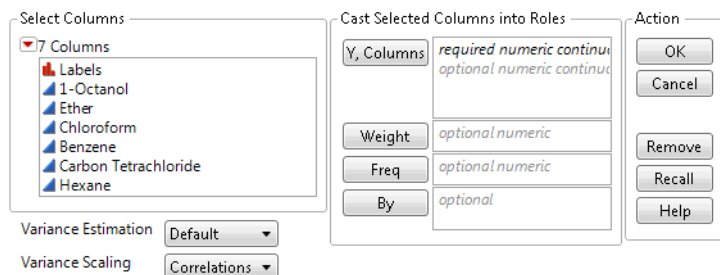
The report lists the communality estimates, variance estimates, significance tests, measures of fit, rotated factor loadings, and a factor loading plot. The Rotated Factor Loadings and Factor Loading Plot suggest that Factor 1 relates to the Carbon Tetrachloride-Chloroform-Benzene-Hexane cluster of variables, and that Factor 2 relates to the Ether-1-Octanol cluster of variables. See [“Factor Analysis Model Fit Options”](#) on page 182 for details of the information shown in the report.

**Tip:** Click on the points in the Factor Loading Plot to select and move the labels. Click in the lower right corner to increase the plot size to more easily view the labels.

## Launch the Factor Analysis Platform

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

**Figure 8.4** Factor Analysis Launch Window



**Y, Columns** The columns to be analyzed. These must have a Numeric data type.

**Weight** A column containing a weight for each observation in the data table. A row is included in the analyses only when its value is greater than zero.

**Freq** Assigns a frequency to each row in the analysis. This is useful when your data are summarized.

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

**Variance Estimation** Lists the methods for estimating the variance-covariance matrix for the analysis. For more information about the methods, see [“The Multivariate Report”](#) on page 37.

**Variance Scaling** Lists the scaling methods for performing the factor analysis.



**Correlations** Default method that enables analysis on correlations.

**Covariances** Enables the analysis on a weighted correlation matrix where the weights are the variables' variances.

**Unscaled.** Enables the analysis of variables that are already centered or scaled.

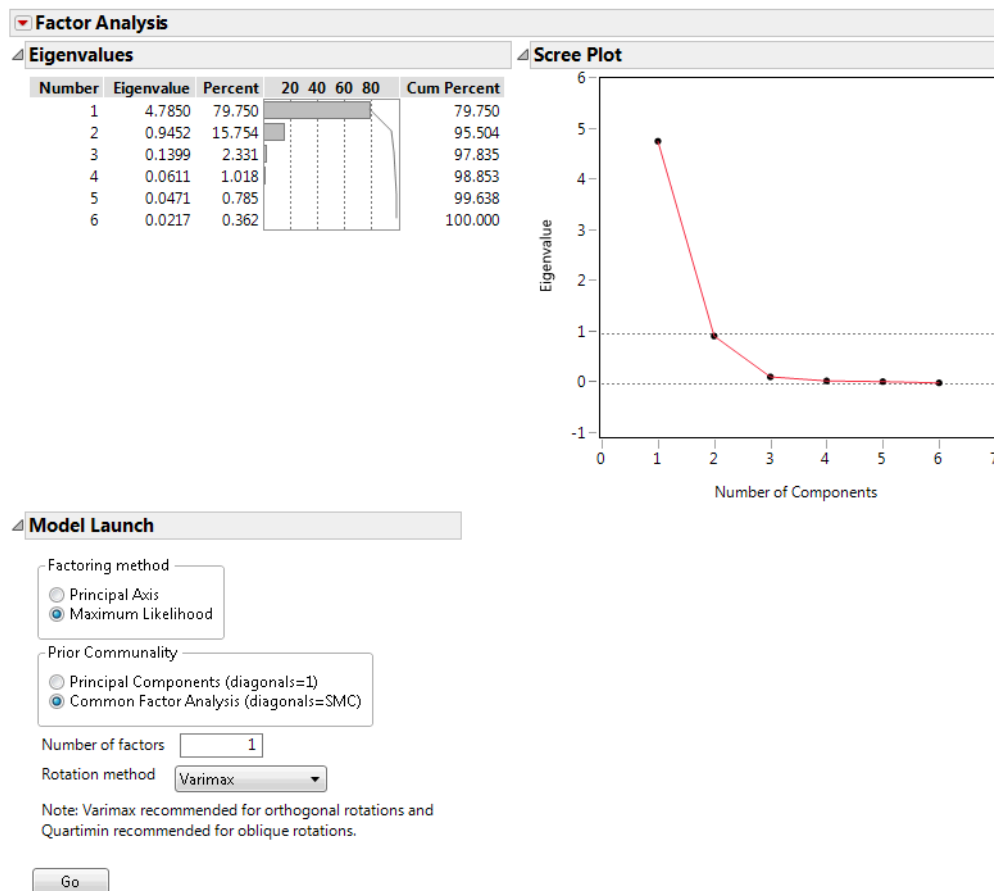
---

## The Factor Analysis Report

The initial Factor Analysis report shows Eigenvalues and the Scree Plot. The eigenvalues are obtained from a principal components analysis. The scree plot graphs these eigenvalues. The initial number of factors in the Model Launch equals the number of eigenvalues that exceed 1.0. You can change the number of factors to extract.

Use the scree plot as a guide for the number of factors. The number of eigenvalues before the scree plot levels out provides an upper bound on the number of factors. For this example, two components (factors) are suggested by the scree plot.

**Figure 8.5** Factor Analysis Report



The Eigenvalues table indicates that the first eigenvalue accounts for 79.75% of the variation and the second eigenvalue accounts for 15.75%. Therefore, the first two eigenvalues account for 95.50% of the total variation. The third eigenvalue explains only 2.33% of the variation, and the contributions from the remaining eigenvalues are negligible. Although the **Number of factors** box is initially set to 1, this analysis suggests that it is appropriate to extract 2 factors.

## Model Launch

Configure the Factor Analysis from the Model Launch control panel. Click **Go** to obtain the results of the factor analysis.

**Figure 8.6** Model Launch

**Model Launch**

Factoring method

☐ Principal Axis

☒ Maximum Likelihood

Prior Communality

☐ Principal Components (diagonals=1)

☒ Common Factor Analysis (diagonals=SMC)

Number of factors

Rotation method

Note: Varimax recommended for orthogonal rotations and Quartimin recommended for oblique rotations.

**Factoring method** Defines the method for extracting factors:

**Principal Axis** Performs eigenvalue decomposition on a reduced correlation or covariance matrix, where the diagonal of the matrix is replaced by an estimate of the communality of the variables. This method is a computationally efficient method, but it does not allow for hypothesis testing.

**Maximum Likelihood** Enables you to test hypotheses about the number of common factors and to obtain model fit statistics.

---

**Note:** The Maximum Likelihood method requires a positive definite correlation matrix. If your correlation matrix is not positive definite, select the Principal Axis method.

---

**Prior Communality** Defines the method for estimating the proportion of variance contributed by common factors for each variable.

**Principal Components (diagonals = 1)** Sets all communalities equal to 1, indicating that 100% of each variable's variance is explained by all of the factors.

---

**Tip:** Using this option with Factoring Method set to Principal Axis results in principal component analysis.

---

**Common Factor Analysis (diagonals = SMC)** Sets the communalities equal to squared multiple correlation (SMC) coefficients. For a given variable, the SMC is the RSquare for a regression of that variable on all other variables.

**Number of factors** Specifies the number of factors to extract from the analysis. The default is the number of eigenvalues that are greater than or equal to 1.0. You can set the number of factors to at least one and no more than the number of variables.

**Rotation method** Defines the method for rotation. The default is Varimax. See [“Rotation Methods”](#) on page 180 for a description of the available rotation methods.

## Rotation Methods

Rotations are used to change the reference axes of the factors to make the factors more interpretable. Rotations are applied to the factors extracted from the data. Rotation methods are based on various complexity or simplicity functions. For more information about rotations see the FACTOR Procedure chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017b), Browne (2001), or Frank and Todeschini (1994).

After the initial extraction, the factors are uncorrelated with each other. If the factors are rotated by an orthogonal transformation, the rotated factors are also uncorrelated. If the factors are rotated by an oblique transformation, the rotated factors become correlated. Oblique rotations often produce more interpretable factors than do orthogonal rotations. However, a consequence of correlated factors is that there is no single unambiguous measure of the importance of a factor in explaining a variable.

### Orthogonal Rotation Methods

**Varimax** Maximizes the sum of the variances of the squared loadings of a factor on all variables. This common method results in each variable having either a small or large loading on each factor. (Orthomax with  $\gamma = 1$ )

**Biquartimax** Maximizes the sum of the variances of the squared factor loadings across factors and variables. It is a mixture of the Varimax and Quartimax rotations. (Orthomax with  $\gamma = 0.5$ )

**Equamax** A weighted solution between the Varimax rotation and the Quartimax rotation. (Orthomax with  $\gamma = N/2$  where  $N$  = number of factors)

**Factorparsimax** A solution that aims to minimize the complexity of factors. This method might result in cross-loadings as variable complexity is not considered in the algorithm. (Orthomax with  $\gamma = N$  where  $N$  = number of factors)

**Orthomax** A general weighted rotation method where the weight is denoted by  $\gamma$ . Many specific orthogonal rotation methods are Orthomax rotations with a specific  $\gamma$ .

**Parsimax** Balances the variable and the factor complexity. (Orthomax with  $\gamma = (I(N-1))/(I+N-2)$  where  $I$  = the number of items and  $N$  = number of factors)

**Quartimax** Minimizes the number of factors needed to explain each variable. (Orthomax with  $\gamma = 1$ )

## Oblique Rotation Methods

**Biquartimin** A rotation to minimize the ratio of the covariances (Oblimin with  $\tau = 0.5$ ).

**Covarimin** Oblique Varimax rotation. (Oblimin with  $\tau = 1$ ).

**Obbiquartimax** Oblique Biquartimax rotation.

**Obequamax** Oblique Equamax rotation.

**Obfactorparsimax** Oblique factor Parsimax rotation.

**Oblimin** A general weighted oblique rotation method where the weight is denoted by  $\tau$ . Many specific oblique rotation methods are Oblimin rotations with a specific  $\tau$ .

**Obparsimax** Oblique Parsimax rotation

**Obquartimax** Oblique Quartimax rotation, equivalent to the Quartimin method.

**Obvarimax** Oblique Varimax rotation.

**Quartimin** Oblique Quartimin rotation, equivalent to oblique Quartimax (Oblimin with  $\tau = 0$ )

**Promax** A two step rotation in which Varimax is performed first and then the Procrustes rotation is used to attain simple structure. This is a computationally efficient method that is an alternative to Oblimin. Recommended for large data sets.

---

## Factor Analysis Platform Options

The Factor Analysis red triangle menu contains the following options:

**Eigenvalues** Shows or hides a table of the eigenvalues of the original correlation, covariance, or unscaled matrix. The table includes the percent of the total variance represented by each eigenvalue, a bar chart illustrating the percent contribution, and the cumulative percent contributed by each successive eigenvalue. The number of eigenvalues that are greater than or equal to 1.0 can be taken as a guideline of the number of factors for analysis.

**Scree Plot** Shows or hides a plot of the eigenvalues versus the number of components (or factors). The plot can be used as an additional guideline to determine the number of factors that contribute to the maximum amount of variance. The point at which the plotted line levels out can be used as the number of sufficient factors for analysis.

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.

---

## Factor Analysis Model Fit Options

The Factor Analysis Model Fit red triangle menu contains the following options:

**Prior Communality** (Available only for Common Factor Analysis.) Shows or hides an initial estimate of the communality for each variable. For a given variable, this estimate is the squared multiple correlation coefficient (SMC), or RSquare, for a regression of that variable on all other variables.

**Eigenvalues** (Available only for Common Factor Analysis.) Shows or hides the eigenvalues of the reduced correlation matrix and the percent of the common variance for which they account. The reduced correlation matrix is the correlation matrix with its diagonal entries replaced by the communality estimates. The eigenvalues indicate the common variance explained by the factors. The Cum Percent can exceed 100% because the reduced correlation matrix is not necessarily positive definite and can have negative eigenvalues.

Note the table indicates the number of factors retained for analysis.

**Unsorted and Unrotated Factor Loading** Shows or hides the factor loading matrix before sorting and rotation.

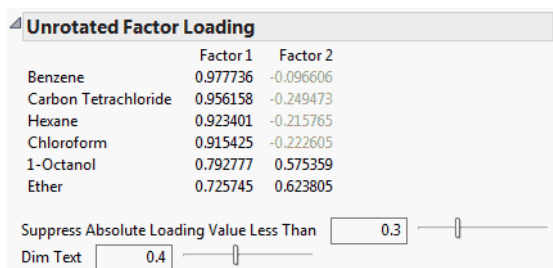
**Unrotated Factor Loading** Shows or hides the factor loading matrix before rotation. Factor loadings measure the influence of a common factor on a variable. Because the unrotated factors are orthogonal, the factor loading matrix is the matrix of correlations between the

variables and the factors. The closer the absolute value of a loading is to 1, the stronger the effect of the factor on the variable.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient for factor values less in absolute value than the specified **Suppress Absolute Loading Values Less Than** value.

**Figure 8.7** Unrotated Factor Loading with Dim Text Controls



The screenshot shows a software window titled "Unrotated Factor Loading". It contains a table with 6 rows of variables and 2 columns of factors. The values are as follows:

	Factor 1	Factor 2
Benzene	0.977736	-0.096606
Carbon Tetrachloride	0.956158	-0.249473
Hexane	0.923401	-0.215765
Chloroform	0.915425	-0.222605
1-Octanol	0.792777	0.575359
Ether	0.725745	0.623805

Below the table are two controls:

- "Suppress Absolute Loading Value Less Than" with a text box containing "0.3" and a slider.
- "Dim Text" with a text box containing "0.4" and a slider.

---

**Note:** The Unrotated Factor Loading matrix is ordered so that variables associated with the same factor appear next to each other.

---

**Rotation Matrix** Shows or hides the values used for rotating the factor loading plot and the factor loading matrix.

**Interfactor Correlations** (Available only for oblique rotations.) Shows or hides the matrix of correlations between factors.

**Target Matrix** (Available only for the Promax rotation.) Shows or hides the matrix to which the varimax factor pattern is rotated.

**Factor Structure** (Available only for oblique rotations.) Shows or hides the matrix of correlations between variables and common factors.

**Final Community Estimates** Shows or hides estimates of the communalities after the factor model has been fit. When the factors are orthogonal, the final communality estimate for a variable equals the sum of the squared loadings for that variable.

**Standard Score Coefficients** Shows or hides a table of the multipliers used to estimate factor scores when saving rotated factors to the source data table.

**Variance Explained by Each Factor** (Available only for orthogonal rotations.) Shows or hides the variance, percent, and cumulative percent, of common variance explained by each rotated factor.

**Variance Explained by Each Factor Ignoring Other Factors** (Available only for oblique rotations.) Shows or hides the variance and percent of common variance explained by each rotated factor regardless of other factors.

**Significance Test** (Available only for the Maximum Likelihood factoring method.) Provides the results of two Chi-square tests.

The first test is for  $H_0$ : No common factors. This null hypothesis indicates that none of the common factors explain the intercorrelations among the variables. This test is Bartlett's Test for Sphericity, whose null hypothesis is that the correlation matrix of the factors is an identity matrix (Bartlett, 1954).

The second test is for  $H_0$ :  $N$  factors are sufficient, where  $N$  is the specified number of factors. Rejection of this null hypothesis indicates that more factors might be required to explain the intercorrelations among the variables (Bartlett, 1954). The Criterion is the log likelihood objective function value.

**Measures of Fit** (Available only for the Maximum Likelihood factoring method.) Shows or hides measures of fit: Chi-Squared without Bartlett's Correction, AIC, BIC, Tucker-Lewis's Index, and the root mean square error of approximation.

**Measures of Factor Scores** Shows or hides measures of factor score determinacy. These measures are used to evaluate if the factor scores might be useful for secondary analyses.

**Unsorted and Rotated Factor Loading** Shows or hides the unsorted factor loading matrix after rotation.

**Rotated Factor Loading** Shows or hides the factor loading matrix after rotation. If the rotation is orthogonal, these values are the correlations between the variables and the rotated factors.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient. The lower the value, the more transparency the font for factor values less in absolute value than the specified **Suppress Absolute Loading Values Less Than** value.



**Figure 8.8** Rotated Factor Loading with Dim Text Controls

Rotated Factor Loading		
	Factor 1	Factor 2
Carbon Tetrachloride	0.9430402	0.2952119
Hexane	0.8973972	0.3064346
Chloroform	0.8942593	0.2964072
Benzene	0.8803182	0.4362790
Ether	0.2848126	0.9136307
1-Octanol	0.3673323	0.9080748

Suppress Absolute Loading Value Less Than

Dim Text

---

**Note:** The Rotated Factor Loading matrix is ordered so that variables associated with the same factor appear next to each other.

---

**Factor Loading Plot** shows or hides a plot of the rotated factor loadings. When more than 2 factors are modeled, the loading plot is a matrix of plots.

**Score Plot** Shows or hides a scatterplot of the estimated factor scores. When more than 2 factors are modeled, the score plot is a matrix of plots.

**Score Plot with Imputation** (Available only if there are missing values.) Shows or hides a scatterplot of the estimated factor scores with imputed values for missing values.

**Display Options** Enables you to show or hide arrows on the loading plots.

**Save Rotated Factors** Saves the rotated factor scores and formulas to the data table.

---

**Note:** The formula cannot evaluate rows with missing values.

---

**Remove Fit** Removes the fit model results from the Factor Analysis report. This option enables you to change the Model Launch configuration for a new report.



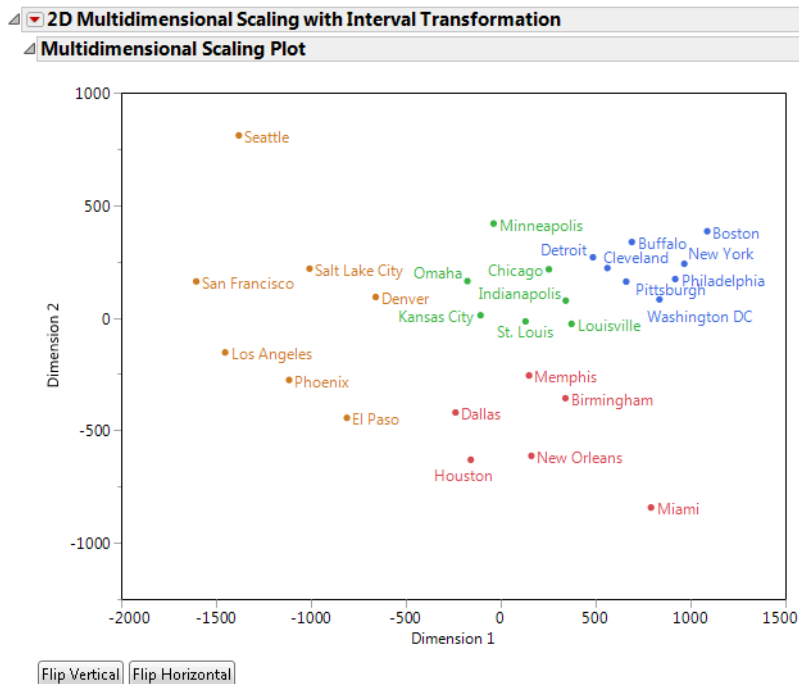
## Multidimensional Scaling

### Visualize Proximities among a Set of Objects

Multidimensional Scaling (MDS) is a technique that is used to create a visual representation of the pattern of proximities (similarities, dissimilarities, or distances) among a set of objects. For example, given a matrix of distances between cities, MDS can be used to generate a map of the cities in two dimensions.

Multidimensional Scaling is frequently used in consumer research where researchers have measures of perceptions about brands, tastes, or other product attributes. MDS is applicable to many other areas where one is interested in visualizing the proximity of objects based on a set of attributes or proximities.

**Figure 9.1** Multidimensional Scaling Example



**Contents**

- Overview of the Multidimensional Scaling Platform ..... 189
- Example of Multidimensional Scaling ..... 189
- Launch the Multidimensional Scaling Platform ..... 191
- The Multidimensional Scaling Report ..... 193
  - Multidimensional Scaling Plot ..... 193
  - Shepard Diagram ..... 193
  - Fit Details ..... 194
- Multidimensional Scaling Platform Options ..... 194
  - Waern Links ..... 195
- Additional Example of the Multidimensional Scaling Platform ..... 196
- Statistical Details for the Multidimensional Scaling Platform ..... 198
  - Stress ..... 198
  - Transformations ..... 199
  - Attributes List Format ..... 199

---

## Overview of the Multidimensional Scaling Platform

The Multidimensional Scaling platform generates a plot of proximities among a set of objects. This plot can be used to visually explore structure in a data set. MDS is a multivariate technique that is used to visualize the patterns of proximities (distances, similarities) among a set of objects in a small number of dimensions. MDS is applied to a distance matrix. The coordinates for the MDS plot are obtained by minimizing a stress function (the difference between the actual and predicted proximities).

The term distance can refer to a measure of physical distance, such as between cities. More often distance is a subjective assessment rather than a precise measurement. Proximities can measure perceived similarities between brands of a product, correlations of crime rates, or economic similarities for a sample of countries. Distance can also be called proximity or similarity (dissimilarity). If the data are given as an attribute list, then a distance matrix is first constructed from the correlation structure of the attribute list.

For more information about multidimensional scaling, see Borg and Groenen ([2005](#)) or Jackson ([2003](#)).

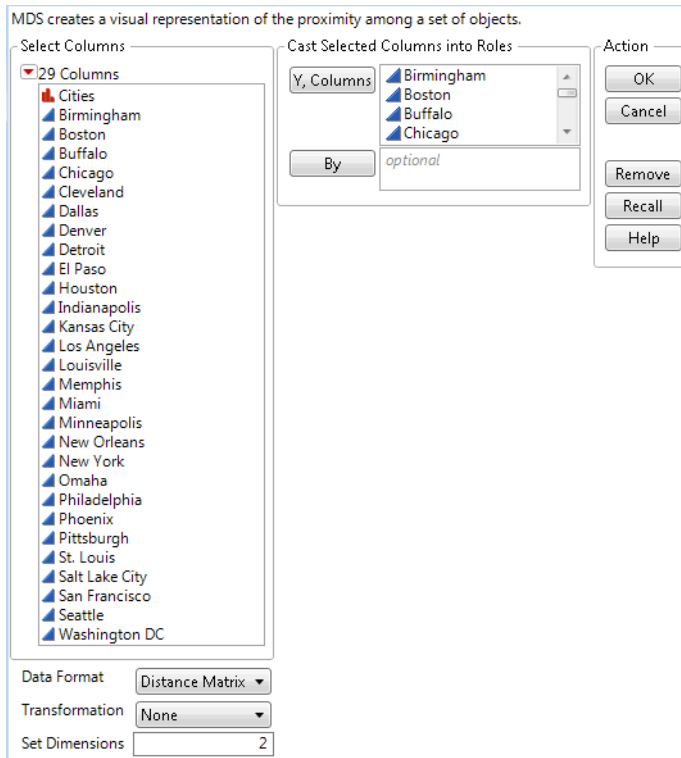
---

## Example of Multidimensional Scaling

This example uses the Flight Distances.jmp sample data table, which is a distance matrix of flight distances between 28 US cities. You can use MDS to construct a map of the cities in two dimensions that is based on the pairwise distances in the data table.

1. Select **Help > Sample Data Library** and open Flight Distances.jmp.
2. Select **Analyze > Multivariate Methods > Multidimensional Scaling**.
3. Select Birmingham through Washington DC and click **Y, Columns**.

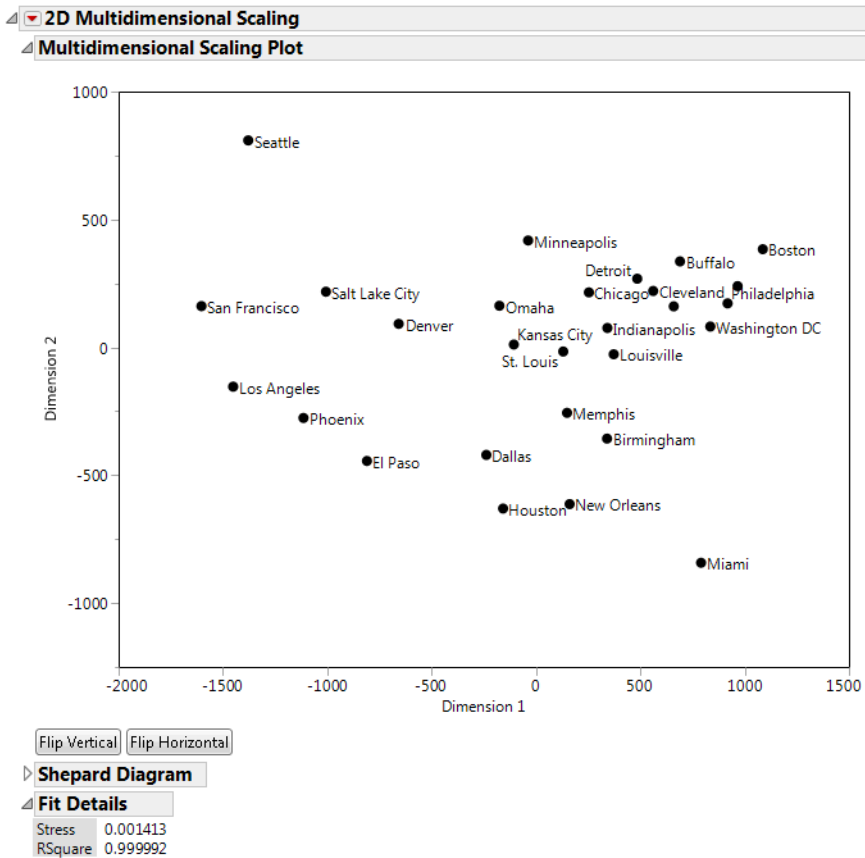
**Figure 9.2** Completed Multidimensional Scaling Launch Window



4. Click **OK**.
5. Select the Flight Distances data table.
6. Right-click on the column Cities and select **Label/Unlabel**.
7. Select **Rows > Row Selection> Select all Rows**.
8. Select **Rows > Label/Unlabel**.
9. Select the Multidimensional Scaling Plot.
10. Click on the **Flip Vertical** button.
11. Click on the **Flip Horizontal** button.

The Flip Vertical and Flip Horizontal buttons enable you to change the orientation of the MDS Plot. The MDS results are invariant to orientation. When the results have a known orientation, such as physical locations, then you might want to rotate or flip your plot.

**Figure 9.3** Multidimensional Scaling Plot



## Launch the Multidimensional Scaling Platform

Launch the Multidimensional Scaling Platform by selecting **Analyze > Multivariate Methods > Multidimensional Scaling**.

**Figure 9.4** Multidimensional Scaling Launch Window

MDS creates a visual representation of the proximity among a set of objects.

Select Columns

▼ 9 Columns

- MAMMAL
- Top incisors
- Bottom incisors
- Top canines
- Bottom canines
- Top premolars
- Bottom premolars
- Top molars
- Bottom molars

Cast Selected Columns into Roles

Y, Columns *required numeric continuous*  
*optional numeric continuous*

By *optional*

Action

OK

Cancel

Remove

Recall

Help

Data Format Distance Matrix ▼

Transformation None ▼

Set Dimensions 2

**Y, Columns** The columns to be analyzed. These must have a Numeric data type.

**By** A column or columns whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed using the other variables that you have specified. The results are presented in separate reports. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

---

**Note:** When using a distance matrix, the By variable requires a full matrix for each level of the By variable.

---

**Data Format** MDS supports two data formats:

**Distance Matrix** A full symmetric or lower triangular matrix where the number of rows equals the number of columns).The diagonal entries can either be zeros or missing.

**Attribute List** A set of columns that contain measures of a quality or characteristic of an object. The objects are typically named in a column. The object column is not used in the analysis but rather is used as a label for the data points on the MDS plot.

**Transformation** Supported transformations are Ratio, Interval, and Ordinal.

**None** No transformation used.

**Ratio** Data has an ordering from smallest to largest, the differences between values have meaning, and the scale has a true zero. Used to scale the MDS plot.

**Interval** Data has an ordering from smallest to largest and the differences between values have meaning. Used to scale and shift the MDS plot.

**Ordinal** Data has an ordering from smallest to largest. Used for ordinal data.



**Set Dimensions** The number of dimensions for the visual representation of the proximities among your objects. Typically, two or three dimensions are used. With greater than three dimensions, the visualization becomes complex.

---

**Note:** The dimension selected can be between 1 to  $n - 1$  where  $n$  = the number of objects, otherwise the dimension is set to 2.

---

---

## The Multidimensional Scaling Report

The initial Multidimensional Scaling report shows these reports: Multidimensional Scaling Plot, the Shepard Diagram, and the Fit Details. If you specify three or more dimensions for the fit in the launch window, then the Multidimensional Scaling Plot provides controls for selecting the dimensions that you view.

Objects that are close together on the MDS plot share similar attributes. Adding labels and colors to the plot can help in the identification of similar groups. The Shepard diagram and summary of fit statistics provide measures of how well the MDS plot represents the proximities of the objects.

### Multidimensional Scaling Plot

The MDS plot displays the multidimensional scaling in two dimensions. Below the plot are two buttons to flip the axis either in the vertical or horizontal direction. The MDS solution can be reflected, rotated, or translated without changing the inter-point proximities. The rotating or reflection of the axes is most common when working with geographical objects that have a known map orientation.

If more than two dimensions were used in the analysis, then you can toggle the dimensions shown in the plot using the **Select Dimension** controls below the plot. The first control defines the horizontal axis of the plot, and the second control defines the vertical axis of the plot.

### Shepard Diagram

The Shepard plot is a plot of the actual or transformed proximities versus the predicted proximities. The plot indicates how well the Multidimensional Scaling Plot reflects the actual proximities. The Shepard is analogous to an Actual by Predicted plot. Ideally the points fall on the  $Y = X$  line, which is shown in red.

## Fit Details

The Fit Details provides statistics that summarize how well the MDS proximities match the actual proximities as well as details about transformations when used.

**Stress** The value of the stress function (Stress1) that was minimized in the fitting procedure. Stress can be between 0 and 1 with lower values indicating a better fit.

**RSquare** The  $R^2$  value for linear fit of the actual or transformed proximities versus the predicted proximities.

**Slope** If a ratio or interval transformation was used, the slope for the transformation is provided. It is the slope of the linear regression of the actual against transformed proximities.

**Intercept** If an interval transformation was used, the intercept for the transformation is provided. It is the intercept of the linear regression of the actual against transformed proximities.

---

## Multidimensional Scaling Platform Options

The Multidimensional Scaling red triangle menu options give you the ability to customize reports according to your needs. The options available are determined by the type of data and the number of dimensions that you use for your analysis.

**MDS Plot** Shows or hides the MDS Plot.

**Diagnostics** Provides diagnostics for the MDS.

**Shepard Diagram** Shows a plot of actual proximity (or transformed proximity if a transformation is used) versus the predicted proximity. This report appears by default. See [“Shepard Diagram”](#).

**Waern Links** Displays the Waern links on the MDS plot. Controls for the portion (smallest or largest) are available when this option is selected. See [“Waern Links”](#).

**Show Coordinates** Provides a report of the solution coordinates. These are the coordinates of the points on the Multidimensional Scaling Plot. The report shows the coordinates of up to three dimensions. Right-click in the report and select **Columns** to add additional dimensions to the report. The maximum number of dimensions is the number of dimensions set in the launch window.

**Show Proximity** Provides a report of the proximities. The original and derived proximities (distances) are provided between each pair of objects. The pairs are identified in the From

and To object columns. If a transformation was used, the transformed proximities are also included in the table.

**Save Proximity** (Available only if Attribute List is the data format.) Saves the distance matrix to the data table.

**3D Plot** (Available only if three or more dimensions are specified for Set Dimensions in the launch window.) Shows a 3-D plot of the first three dimensions.

**Save Coordinates** Saves the solution coordinates to the data table in separate columns.

See 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.

## Waern Links

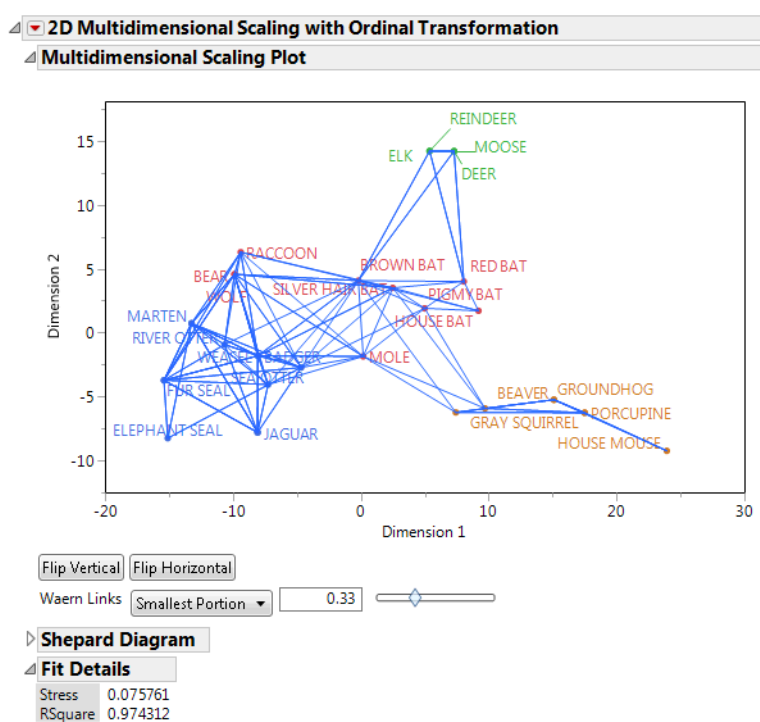
Waern links provide a visual check of the MDS results by comparing actual proximities to predicted proximities. The links join points on the Multidimensional Scaling Plot based on their actual proximities. The objects with the smallest (largest) proximities are connected. A typical scenario to consider is the smallest 33% of the proximities between objects. If the MDS Plot is a good representation of the proximities, then the links for the smallest actual proximities should connect the closest objects in the plot. If a link for a small proximity stretches across the plot connecting distant objects, then the MDS fit would be questioned.

### Waern Link Controls

There is a list from which you can choose to show the Smallest Portion or the Largest Portion of links on the plot. The portion of links shown is controlled by entering a value in the box or by using the slider. Figure 9.5 shows Waern links for the *Teeth.jmp* data table for the 33% smallest portion.

For more information about Waern links, see Waern ([1972](#)).

Figure 9.5 MDS Plot with Waern Links



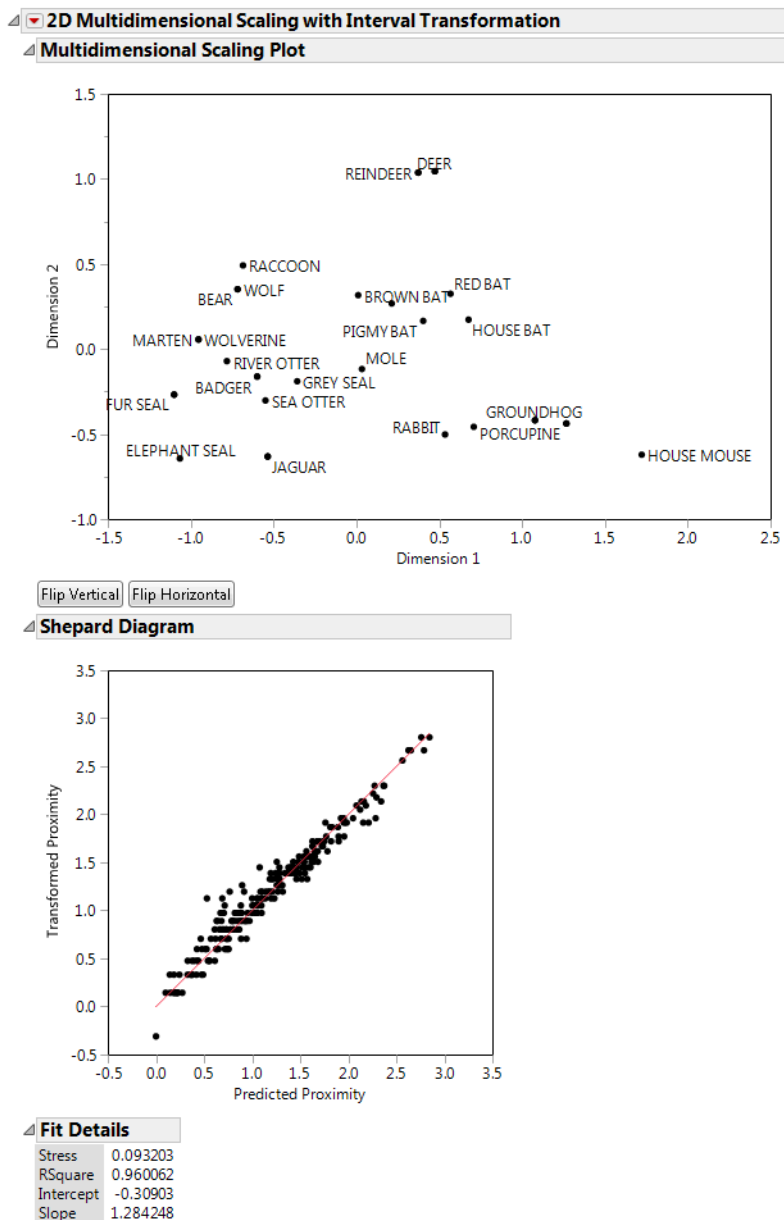
## Additional Example of the Multidimensional Scaling Platform

This example uses the Teeth.jmp sample data table, which is an attribute list of the counts of eight teeth types in 32 mammals. You can use MDS to explore the similarities of mammals based on their teeth. An interval transformation is used to illustrate the output from that transformation. The data do have an ordering that has a meaning (four teeth are twice as many as two teeth). One might explore other transformations such as the ordinal transformation.

1. Select **Help > Sample Data Library** and open Teeth.jmp.
2. Right-click on the column MAMMAL and select **Label/Unlabel**.
3. Select **Rows > Row Selection> Select all Rows**.
4. Select **Rows > Label/Unlabel**
5. Select **Analyze > Multivariate Methods > Multidimensional Scaling**.
6. Select Top incisors through Bottom molars and click **Y, Columns**.

7. Select **Data Format > Attribute List**.
8. Select **Transformation > Interval**.
9. Click **OK**.

**Figure 9.6** Multidimensional Scaling Report



The Shepard Diagram and the Fit Details indicate that the MDS Plot is a good representation of similarities of animals due to similarities in their teeth. The Stress statistics of 0.093 is low and the  $R^2$  fit of the transformed versus predicted proximities is high at 0.96. In addition, the Fit Details provides the intercept and slope for the transformation of the actual proximities.

---

## Statistical Details for the Multidimensional Scaling Platform

JMP uses a quasi Newton optimization method to minimize the Stress function to determine the MDS coordinates. This minimization leads to a set of coordinates in a predetermined number of dimensions that minimize the derived proximity measures for each pairwise set of the dimensions. When the data is ordinal, then monotonic regression is used. Otherwise, standard least squares regression is used.

- [“Stress”](#)
- [“Transformations”](#)
- [“Attributes List Format”](#)

### Stress

The following notation is used to define Stress:

- $i, j$  - indexes for the number of objects
- $d_{ij}$  - the derived distance between objects  $i$  and  $j$
- $\delta_{ij}$  - the observed relative distance between objects  $i$  and  $j$
- $f(\delta_{rs})$  - transformation function for the distance

The Stress function is given by

$$\text{Stress} = \left[ \frac{\sum_{i < j} [f(\delta_{ij}) - d_{ij}]^2}{\sum_{i < j} d_{ij}^2} \right]^{\frac{1}{2}}$$

This measure of stress is also known as Kruskal's Stress, Type I, or simply Stress1.

## Transformations

The section uses the notation described in “Stress” on page 198. Transformations are used to scale the actual proximities. Transformations would be considered to improve the MDS representation of the actual proximities by taking into account specific structures in the data. The parameters in the transformation functions become additional parameters in the minimization algorithm.

### Ratio Transformation

For ratio data:

$$f(\delta_{rs}) = b\delta_{rs}$$

### Interval Transformation

For interval data:

$$f(\delta_{rs}) = a + b\delta_{rs}$$

### Ordinal Transformation

For ordinal data the data is not transformed, rather the algorithm uses monotone regression rather than least squares regression.

## Attributes List Format

When the data table is in the attributes list format, it is converted to a distance matrix and then MDS is applied. The distance matrix is determined by the correlation structure of the data.

JMP generates the distance matrix, or dissimilarity matrix, using the following transformation:

$$\delta_{ij} = \frac{1 - r_{ij}}{2}$$

---

**Note:** For an advanced example of the MDS platform, see the San Francisco Crime Distances.jmp sample data table and the source script for that table. The script creates the distance matrix that is then used to explore the relationships between crime categories.

---





# Chapter 10

## Item Analysis

### Analyze Test Results by Item and Subject

---

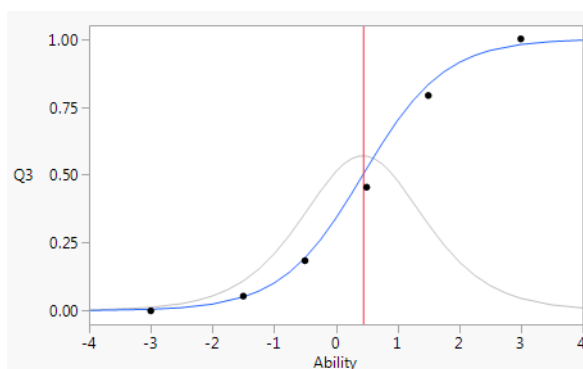
The Item Analysis platform enables you to fit *item response theory* models. The Item Response Theory (IRT) method is used for the analysis and scoring of measurement instruments such as tests and questionnaires. Item response theory uses a system of models to relate a trait or ability to an individual's probability of endorsing or correctly responding to an item. Frequently, the trait or ability of interest is not directly measurable and is therefore called *latent*. IRT can be used to study standardized tests, cognitive development, and consumer preferences. IRT is an alternative method to classical test theory (CTT) where the focus is on the total observed score rather than the item scores.

The Item Analysis platform implements the IRT method with the following outcomes:

- Measurement instruments are scored at the item level, providing insight into the contributions of each item on the latent response.
- Scores for both the responders and the items are produced on the same scale.
- Respondent and item scores are shown on a single plot.
- Item characteristic curves are shown. These curves can be used to explore the relationship between items and respondent's underlying trait or ability.

For more information about item response theory, see de Ayala (2009).

**Figure 10.1** Item Analysis Characteristic Plot



**Contents**

Example of Item Analysis .....	203
Launch the Item Analysis Platform .....	206
Logistic 3PL Model Details .....	207
The Item Analysis Report.....	207
Characteristic Curves .....	207
Information Plot .....	208
Dual Plot.....	209
Parameter Estimates .....	209
Item Analysis Platform Options .....	210
Statistical Details for the Item Analysis Platform.....	210
Item Response Curves .....	211
Item Response Curve Models .....	212
IRT Model Assumptions .....	214
Fitting the IRT Model .....	214
Ability Formula .....	216

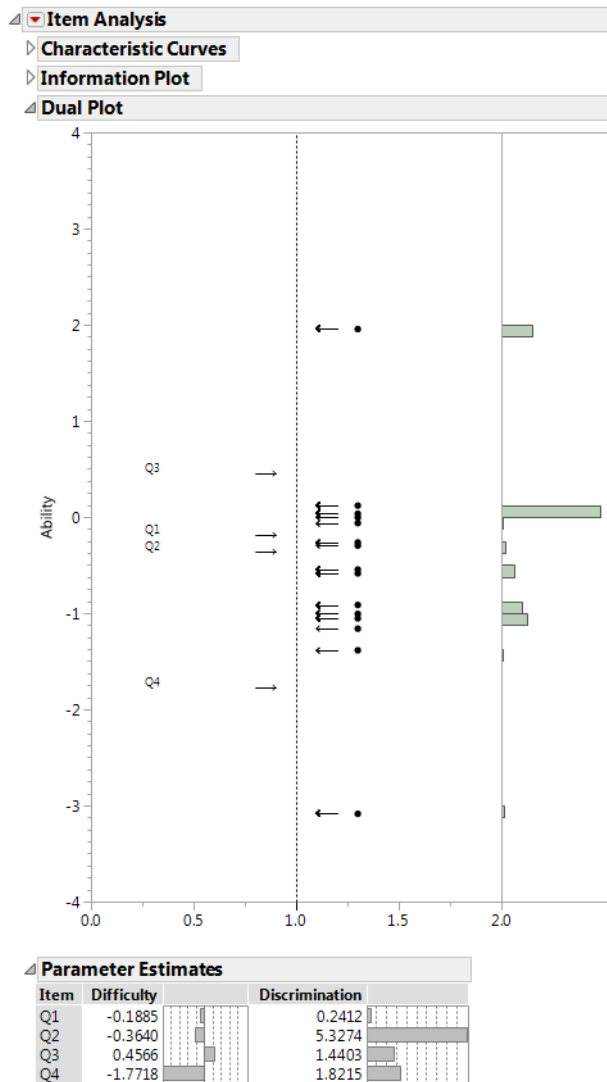
---

## Example of Item Analysis

This example uses the MathScienceTest.jmp sample data table, which is a subset of the data from the Third International Mathematics and Science Study (TIMSS) conducted in 1996. The data table contains scores (1 = correct or 0 = incorrect) for 1263 students on 14 questions. You examine the first four questions to understand the relationship between questions and respondent's mathematical ability. The questions on the test are the items that are used to measure the latent mathematical ability. Fit a 2PL model to this data.

1. Select **Help > Sample Data Library** and open MathScienceTest.jmp.
2. Select **Analyze > Multivariate Methods > Item Analysis**.
3. Select Q1 through Q4, click **Y**, **Test Items** and click **OK**.

Figure 10.2 Item Response Report

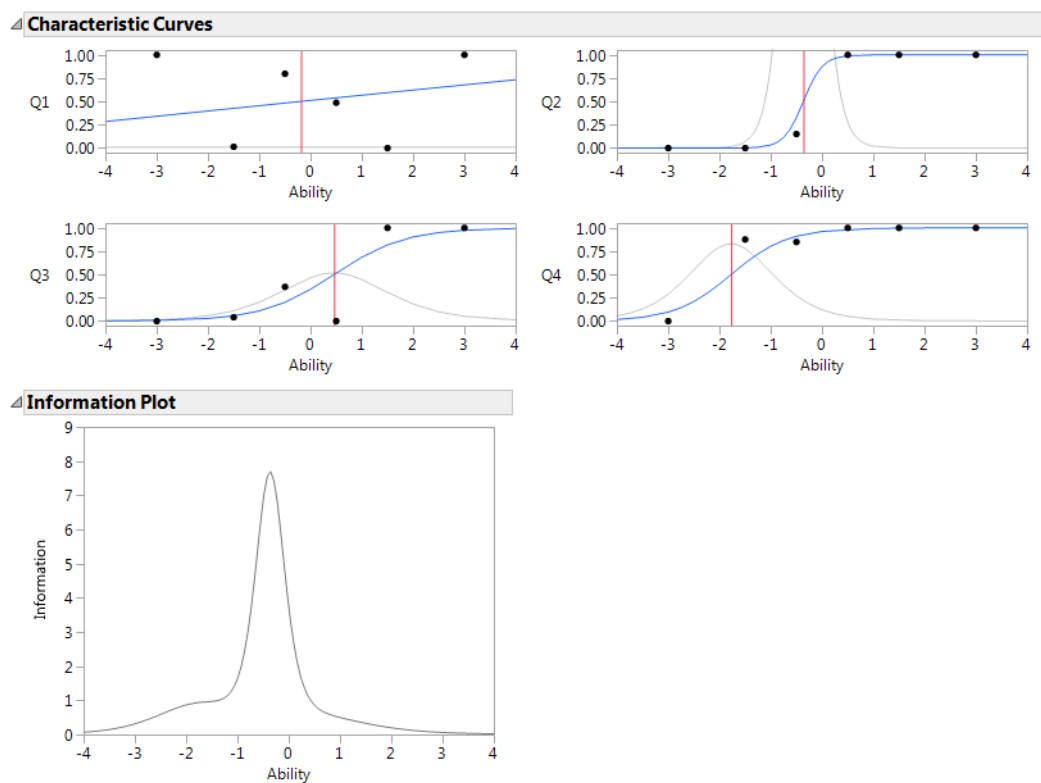


From the dual plot you note that Q4 is the easiest of the four questions to answer as it has the lowest difficulty score at -1.78. Q3 is the hardest with a difficulty score of 0.46. Most of the respondents fall in the middle to lower end of the ability scale as shown by the data points in the center part of the graph. In the histogram, you can see that roughly 40% of the respondents fall slightly above zero on the ability scale.

**Note:** Ability scores are not computed for individuals with all incorrect or all correct answers. For more information, see [“Fitting the IRT Model”](#) on page 214.

4. Click on the gray Characteristic Curves report disclosure icon to open.
5. Click the Item Analysis red triangle menu, and select **Number of Plots Across**.
6. Enter 2 and click **OK**.
7. Click on the gray Information Plot report disclosure icon to open.

**Figure 10.3** Item Response Example



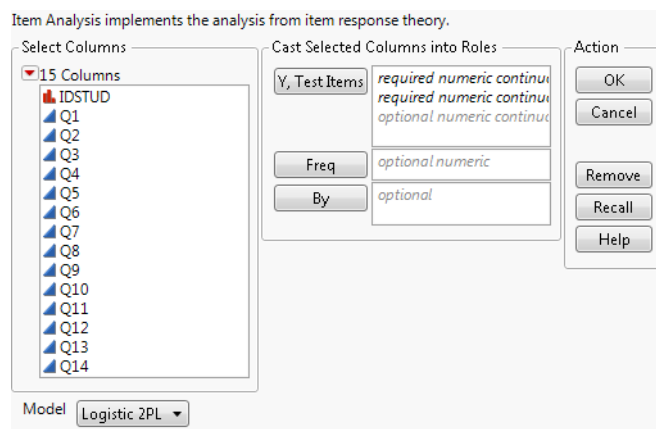
Q1 has a flat characteristic curve and a flat information curve. This suggests that Q1 does not provide much information to discriminate respondents' mathematical ability. The characteristic curve for Q2 is steep, which indicates that Q2 is useful for discriminating respondent ability. The vertical line in each plot is at the inflection point for the characteristic curve. This vertical line is the ability level at which the respondent has a 50% probability of answering the specified question correctly.

The information plot indicates that together the four questions analyzed provide the most information about ability levels between about -1 and 0. Including more questions of higher difficulty in the model could broaden the information curve.

## Launch the Item Analysis Platform

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

**Figure 10.4** Item Analysis Launch Window



**Y, Test Items** Assigns two or more columns to be analyzed. The columns must be numeric, continuous, and contain only 0s and 1s.

**Tip:** Use **Cols > Recode** if you need to recode your data to 0s and 1s. See the Enter and Edit Data chapter in the *Using JMP* book.

**Freq** Assigns a frequency variable to this role. This is useful if your data are summarized.

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

**Model** Specifies the desired model from the following options:

**Logistic 2PL** The 2-parameter logistic model.

**Logistic 3PL** The 3-parameter logistic model.

**Logistic 1PL** The 1-parameter logistic model with a Rasch parameterization.

## Logistic 3PL Model Details

If you select Logistic 3PL for Model, you are prompted to enter a penalty for the guessing parameters after you click OK. For a 3PL model, the default value of the penalty is zero. However, you can enter a non-zero penalty for the  $c$  parameters (the guessing for each item). This penalty is similar to the type of penalty parameter that you would use in ridge regression. The penalty is on the variance of the estimated guessing parameters. The use of the penalty has the following benefits:

- Stabilizes the estimation of model parameters.
- Speeds up computations.
- Reduces the variability of the guessing parameter across items at the expense of some bias.

Large values of the penalty force the guessing parameters to zero while smaller values help reduce the variability of the guessing parameter across items. A value of zero can be used for no penalty.

---

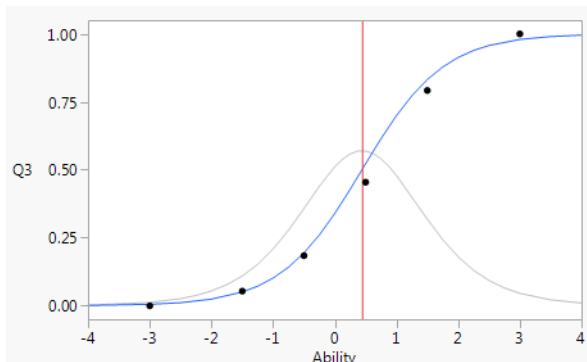
## The Item Analysis Report

- [“Characteristic Curves”](#)
- [“Information Plot”](#)
- [“Dual Plot”](#)
- [“Parameter Estimates”](#)

### Characteristic Curves

The Characteristic Curves contains an item characteristic curve (ICC) for each item that you specified in the launch window. The Characteristic Curves are initially closed.

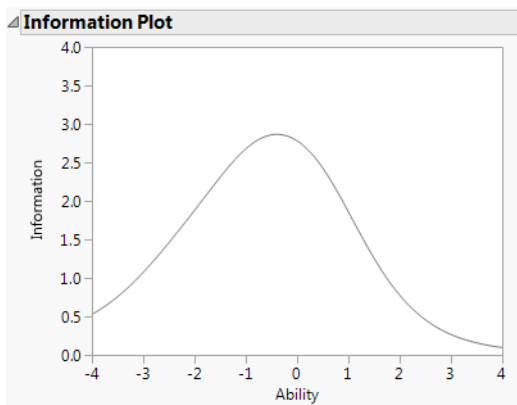
The item characteristic curve plots the probability of answering an item correctly versus ability. Ability is measured on a standardized scale, so a respondent with ability equal to 0 is a respondent of average ability. Data points for the observed probability of correct answers for fixed ability levels are plotted. Comparing the fitted characteristic curve to the data points provides a visual measure of goodness of fit of the model for each individual item. In addition, the characteristic plots have a background information curve and a vertical line at the characteristic curve inflection point. The background information curve is a plot of the slope of the item characteristic curve, which is maximized at the inflection point.

**Figure 10.5** Item Characteristic Curve

**Tip:** You can adjust the number of characteristic curves that appear in each row of the report using the Number of Plots Across option in the Item Analysis red triangle menu.

## Information Plot

The Information Plot report contains a plot of the overall information curve, which is constructed by summing the individual item information curves. The information plot provides insight into the appropriate ability levels that the test is able to measure. Figure 10.6 describes a test with items that are appropriate for assessing individuals with average to low levels of the ability more so than individuals with high levels of ability. This plot is initially closed.

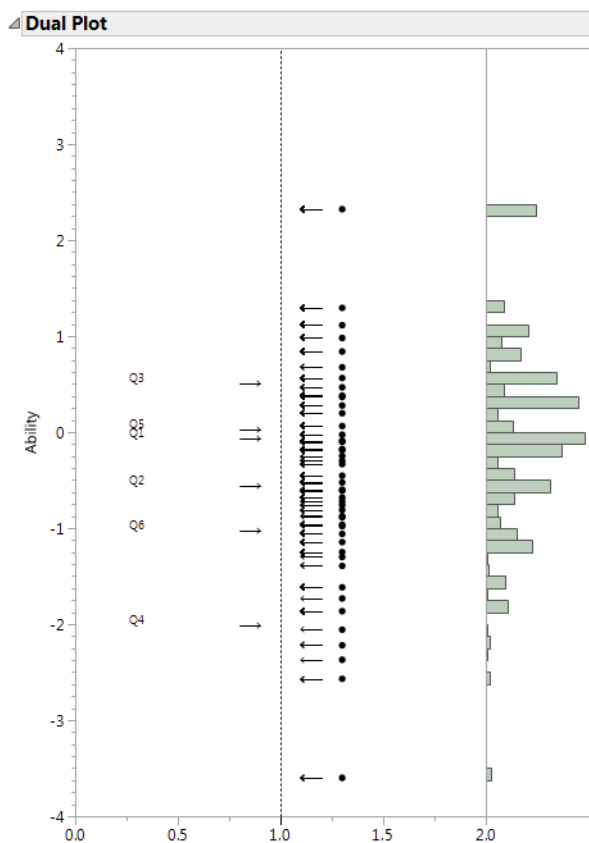
**Figure 10.6** Information Plot



## Dual Plot

The Dual Plot report contains a plot that shows item difficulty and subject ability in one plot. Difficulty and ability use a common standardized scale shown on the  $y$ -axis. The items are plotted by their difficulty on the left side of the plot. The subjects are plotted to the right with data points and a histogram. The dual plot enables you to relate the difficulty of each item to the ability of each respondent.

**Figure 10.7** Dual Plot



## Parameter Estimates

The Parameter Estimates report contains a table of estimated parameters for each item. The parameters provided depend on the model used in your analysis (1PL, 2PL, or 3PL).

**Item** The test item.

**Difficulty** The  $b$  parameter or the measure of the difficulty of the item. A histogram of the difficulty parameters is shown beside the difficulty estimates.

**Discrimination** (Available only for 2PL and 3PL models.) The  $a$  parameter or the measure of the item discrimination. A histogram of the discrimination parameters is shown beside the discrimination estimates.

**Lower Asymptote** (Available only for 3PL models.) The  $c$  parameter or a measure of guessing.

---

## Item Analysis Platform Options

**Number of Plots Across** Enables you to specify how many ICC plots to display in each row of plots in the Characteristic Curve report. The default is one ICC plot per row.

**Save Ability Formula** Saves the ability formula to a new column in the data table.

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.

---

## Statistical Details for the Item Analysis Platform

Item response theory (IRT) uses a series of equations to relate items to an unobserved (latent) trait or ability. Items, or questions, are indicators of an underlying latent construct that cannot be directly observed. At the time of data collection, both the subject abilities and the item characteristics are unknown.

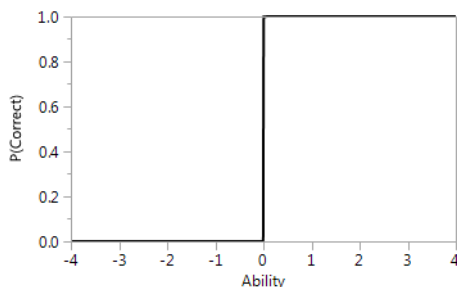
- [“Item Response Curves”](#)
- [“Item Response Curve Models”](#)
- [“IRT Model Assumptions”](#)

- “Fitting the IRT Model”
- “Ability Formula”

## Item Response Curves

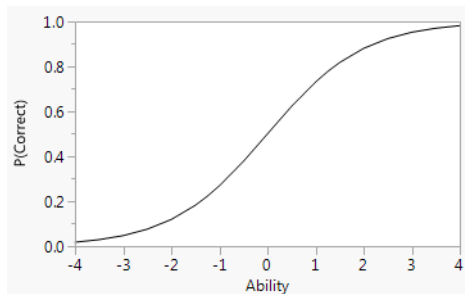
Item response curves (item characteristic curves) are used to describe the relationship between the ability, defined on an *ability* scale, and each item. An item response curve plots the probability of correctly answering an item against different levels of ability. An item with perfect discrimination has a 0% probability of correct answers for respondents with ability below a threshold and a 100% probability of a correct response for subjects with ability above the threshold (Figure 10.8).

**Figure 10.8** Characteristic Curve for an Item with Perfect Discrimination



A typical relationship between the probability of correctly answering an item and ability is an S-shaped function with lower and upper asymptotes. As a respondent's ability increases, their probability of correctly answering the item increases to 100%. The shape of the curve for a specific item is related to the difficulty and discriminatory properties of the item.

**Figure 10.9** Typical Item Response Curve



## Item Response Curve Models

One-, two-, and three-parameter logistic models can be used to model the item response curves. The three-parameter logistic (3PL) model is defined as follows.

$$P(\theta) = c + \frac{I - c}{1 + e^{-(a)(\theta - b)}}$$

- $P(\theta)$  is the probability of answering the item correctly for an ability level  $\theta$ . For more information about fitting the item response theory model, see [“Fitting the IRT Model”](#) on page 214.
- The  $a$  parameter defines the steepness of the curve at its inflection point. It provides an estimate of the discriminatory power of the item.
- The  $b$  parameter defines the location of the inflection point on the Ability axis. It provides an estimate of the difficulty of an item.
- The  $c$  parameter is the lower asymptote. It provides an estimate of the probability that an item is answered correctly by guessing.
- For the 2PL model, the  $c$  parameter is set to 0.

$$P(\theta) = \frac{I}{1 + e^{-(a)(\theta - b)}}$$

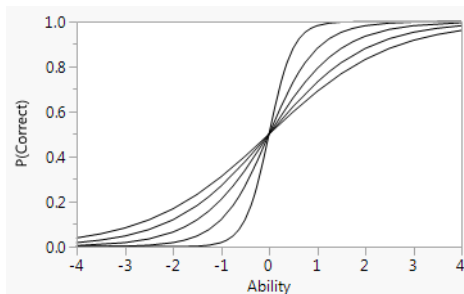
- For the 1PL model, the  $c$  parameter is set to 0 and the  $a$  parameter is set to 1. This parameterization is also known as the Rasch model (Rasch 1980).

$$P(\theta) = \frac{I}{1 + e^{-(\theta - b)}}$$

### The $a$ Parameter: Item Discrimination

In the 2PL and 3PL models, the  $a$  parameter, or the steepness of the curve at its inflection point, provides a measure of the discriminatory power of an item. The discriminatory power, or discrimination, of an item refers to how well an item can distinguish between respondents with low ability levels versus those with high ability levels. A steep item response curve indicates that the item has strong discrimination. Respondents with low ability levels have a low probability of a correct response to the item while respondents with high ability have a high probability of a correct response. Items whose curves are relatively flat have low discrimination. Items with low discrimination are candidates to be dropped from the measurement instrument.

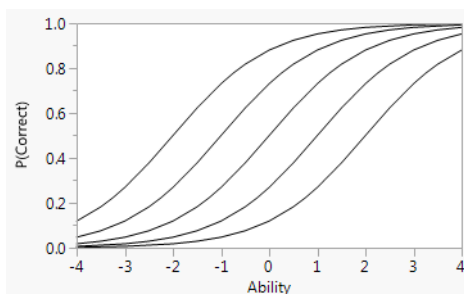
**Figure 10.10** Logistic Model for Several Values of  $a$



### The $b$ Parameter: Item Difficulty

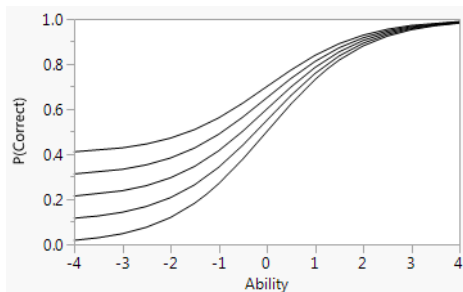
The  $b$  parameter, or the location of the inflection point with respect to ability, provides a measure of item difficulty. Item response curves with inflection points farther to the right on the ability scale are indicative of items that are more difficult to answer than items with inflection points to the left. In the 1PL and 2PL models, the  $b$  parameter provides an estimate of the ability level required for a 50% probability of correctly answering the item.

**Figure 10.11** Logistic Curve for Several Values of  $b$



### The $c$ Parameter: Guessing

In the 3PL model, the  $c$  parameter, or the lower asymptote of the item response curve, provides a measure of the guessing parameter. A nonzero lower asymptote represents the nonzero probability of a person with a very low ability level answering an item correctly.

**Figure 10.12** Logistic Model for Several Values of  $c$ 


## IRT Model Assumptions

The 2PL model is the default model in the Item Analysis platform. The 1PL model is appropriate when you can assume that all items have equal discriminating power. When this assumption is not appropriate, the 2PL or 3PL model should be used. The 2PL model has greater numerical stability than the 3PL model, especially for small data sets. Additionally, in the 2PL model,  $b$  can be interpreted as the ability level required for a 50% chance of a responder answering an item correctly.

The IRT model assumes that the underlying trait is unidimensional. That is, there is a single underlying latent construct. If there are several traits that have complex interactions with each other being measured, then a unidimensional model is not appropriate. The IRT model is appropriate for continuous latent variables. For a categorical latent variable, you should consider a latent class model. See the [“Latent Class Analysis”](#) chapter on page 275. IRT models are assumed to be item-invariant. Item-invariance means that  $P(\theta)$  is interpreted as the probability of a correct response for a set of individuals with ability level  $\theta$ . If a large group of individuals with equal ability levels answered the item,  $P(\theta)$  predicts the proportion who would answer the item correctly. This implies that IRT models would have the same parameters regardless of the group of subjects tested. Additionally, the IRT model assumes local independence, which means that once the latent construct has been accounted for, the items are independent of one another.

## Fitting the IRT Model

The IRT model is fit using Marginal Maximum Likelihood estimation (MMLE). MMLE is an alternative method to Joint Maximum Likelihood estimation (JLE). MMLE treats the subjects as random effects. The items and abilities are related as conditional probabilities as follows:

$$p(\mathbf{x}|\theta, \vartheta) = \prod_{j=1}^L p_j(\theta)^{x_j} (1 - p_j(\theta))^{1-x_j}$$

where  $p(\mathbf{x}|\theta, \vartheta)$  is the probability of a response vector  $\mathbf{x}$  given the subject ability  $\theta$  and the vector of item parameters  $\vartheta$ . The number of item parameters depends on the model used (1PL, 2PL, or 3PL).

MMLE integrates out the subject effects using Gaussian quadrature to obtain item parameter estimates. The probability of response vector  $\mathbf{x}$  is as follows:

$$p(\mathbf{x}) = \int_{-\infty}^{\infty} p(\mathbf{x}|\theta, \vartheta)g(\theta|\mathbf{v})d\theta$$

where  $g(\theta|\mathbf{v})$  is the distribution of the subjects and  $\mathbf{v}$  is a vector of the population location and scale parameters. The normal distribution with mean 0 and standard deviation 1 is used for  $g(\theta|\mathbf{v})$  in JMP.

---

**Note:** A missing value for a test question is treated as an incorrect response. Ability scores are not calculated for individuals with all incorrect or all correct answers. The patterns of the responses for these subjects are included in the model estimation.

---

The MMLE procedure for fitting the IRT model can be compared to fitting a random effects model in two stages. The ability parameters are treated as random effects with variance of 1. In the first step, these random effects are integrated out using Gaussian quadrature. The item parameters are treated as fixed effects that are estimated using ML from the marginal likelihood with the ability parameters integrated out. The ability parameters are in essence best linear unbiased predictions that are estimated using the full unintegrated (joint) likelihood, treating the item parameters as known and held fixed at the values obtained in the first stage.

There are  $2^L$  patterns of responses for  $L$  items. The ability level for each pattern can be calculated by finding the ability level with the highest probability for the response pattern by applying the following until  $\theta$  converges:

$$\theta_i^{t+1} = \theta_i^t - \frac{X_i - \sum_{j=1}^L p_{ij}(t)}{L - \sum_{j=1}^L p_{ij}(t)(1 - p_{ij}(t))}$$

where:

$\theta$  maximizes the likelihood of obtaining the response pattern

$t$  is the number of iterations

$L$  is the number of items

$X_i$  is the observed score

$p_{ij}$  is the probability of a correct response on the  $j^{\text{th}}$  item by the  $i^{\text{th}}$  person based on the item parameters.

## Ability Formula

The Save Ability Formula option from the Item Analysis red triangle menu saves the ability formula to a new column in the data table. This formula can be used to score additional subjects added to the data table or it can be copied to a new table to score a new set of subjects.

The function saved to the data table is called the IRT Ability function. The item parameter estimates are stored in a matrix in this function.



# Chapter 11

## Hierarchical Cluster Group Observations Using a Tree of Clusters

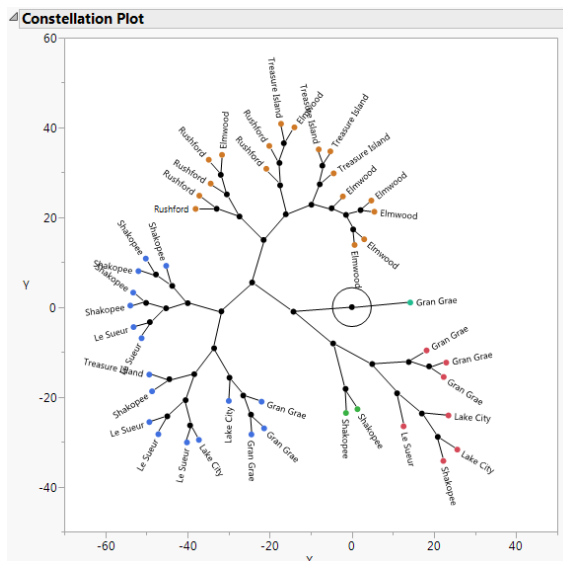
Clustering is a multivariate technique that groups together observations that share similar values across a number of variables. Use it to understand the clumping structure of your data.

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 11.1** Example of a Constellation Plot



**Contents**

Overview of the Hierarchical Clustering Platform .....	219
Overview of Platforms for Clustering Observations .....	219
Example of Clustering .....	221
Launch the Hierarchical Cluster Platform .....	223
Clustering Method .....	225
Method for Distance Calculation .....	225
Data Structure .....	226
Transformations to Y, Columns Variables .....	228
Hierarchical Cluster Report .....	229
Dendrogram Report .....	230
Illustration of Dendrogram and Distance Graph .....	230
Clustering History Report .....	232
Hierarchical Cluster Options .....	232
Additional Examples of the Hierarchical Clustering Platform .....	236
Example of a Distance Matrix .....	236
Example of Wafer Defect Classification Using Spatial Measures .....	238
Statistical Details for the Hierarchical Clustering Platform .....	240
Spatial Measures .....	241
Distance Method Formulas .....	242

---

## Overview of the Hierarchical Clustering Platform

Hierarchical Clustering is one of four platforms that JMP provides for clustering observations. For a comparison of all four methods, see [“Overview of Platforms for Clustering Observations”](#) on page 219.

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 291.

---

JMP provides four platforms that you can use to cluster observations:

- Hierarchical Cluster is useful for smaller tables with up to several tens of thousands of rows and allows character data. Hierarchical clustering combines rows in a hierarchical sequence that is portrayed as a tree. You can choose the number of clusters that is most appropriate for your data after the tree is built.
- K Means Cluster is appropriate for larger tables with up to millions of rows and allows only numerical data. You need to specify the number of clusters, *k*, in advance. The algorithm guesses at cluster seed points. It then conducts an iterative process of alternately assigning points to clusters and recalculating cluster centers.
- Normal Mixtures is appropriate when your data come from a mixture of multivariate normal distributions that might overlap and allows only numerical data. For situations where you have multivariate outliers, you can use an outlier cluster with an assumed uniform distribution.

You need to specify the number of clusters in advance. Maximum likelihood is used to estimate the mixture proportions and the means, standard deviations, and correlations jointly. Each point is assigned a probability of being in each group. The EM algorithm is used to obtain estimates.

- Latent Class Analysis is appropriate when most of your variables are categorical. You need to specify the number of clusters in advance. The algorithm fits a model that assumes a multinomial mixture distribution. A maximum likelihood estimate of cluster membership is calculated for each observation. An observation is classified into the cluster for which its probability of membership is the largest.

**Table 11.1** Summary of Clustering Methods

Method	Data Type or Modeling Type	Data Table Size	Specify Number of Clusters
Hierarchical Cluster	Any	With Fast Ward, up to 200,000 rows  With other methods, up to 5,000 rows	No
K Means Cluster	Numeric	Up to millions of rows	Yes
Normal Mixtures	Numeric	Any size	Yes
Latent Class Analysis	Nominal or Ordinal	Any size	Yes

Some of the clustering platforms have options to handle outliers in the data. However, if your data has outliers, it is best to explore them first prior to analyzing. This can be done using the

Explore Outliers Utility. For more information, see the Modeling Utilities chapter in the *Predictive and Specialized Modeling* book.

---

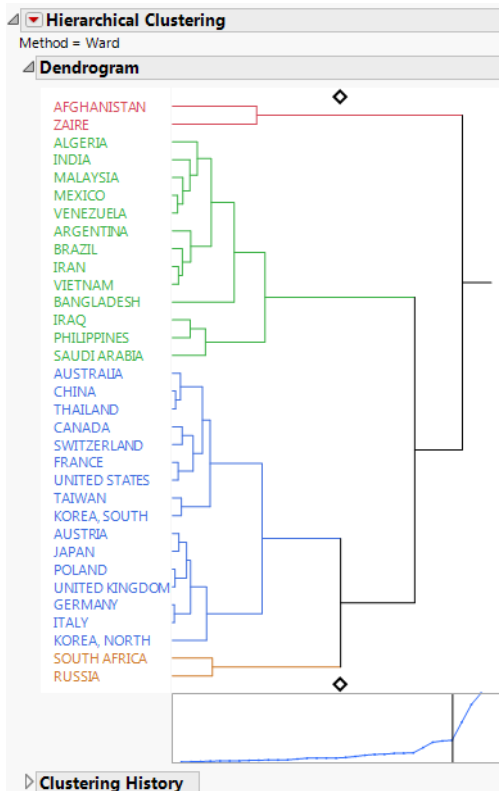
## Example of Clustering

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 11.2** 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.

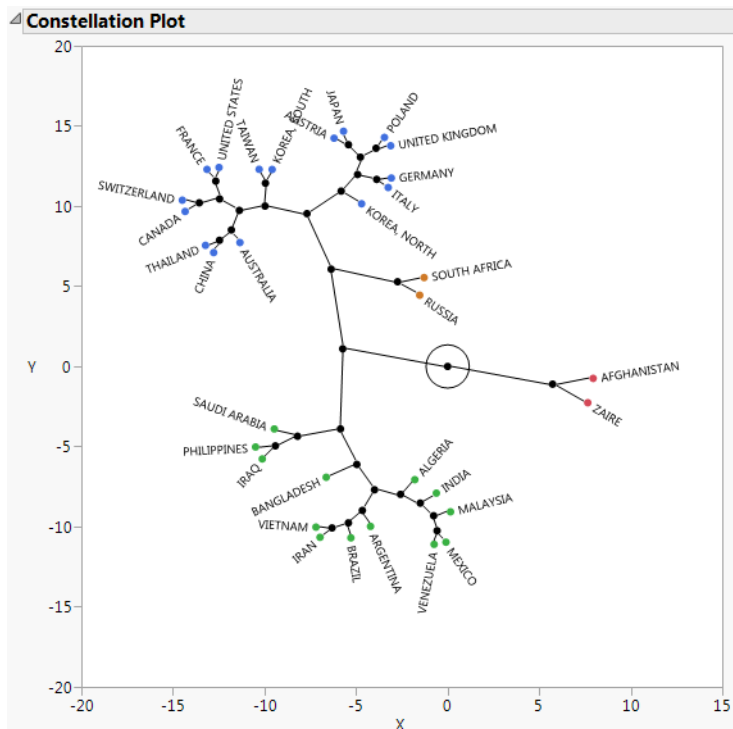
The plot that appears beneath the dendrogram has a point for each step where two clusters are joined into a single cluster. The horizontal coordinates represent the numbers of clusters and they decrease from left to right. The vertical coordinate of the point is the distance between the two clusters 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 11.3** 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 Figure 11.4.

**Figure 11.4** Hierarchical Cluster Launch Dialog

Finding points that are close, have similar values

Select Columns 3 Columns country birth death	Cast Selected Columns into Roles Y, Columns: optional Ordering: optional numeric Label: optional By: optional	Action OK Cancel Remove Recall Help
--	---	--

Options

Method: Hierarchical

Method:

- ☒ Ward
- ☐ Average
- ☐ Centroid
- ☐ Single
- ☐ Complete
- ☐ Fast Ward

Data as usual

☒ Standardize Data

☐ Standardize Robustly

☐ Missing value imputation

**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"
- "Method for Distance Calculation"
- "Data Structure"
- "Transformations to Y, Columns Variables"



## Clustering Method

Hierarchical is the default clustering method, but the dialog enables you to switch to KMeans or Normal Mixtures. If you select KMeans or Normal Mixtures, when you click OK, a Control Panel appears where you can select any of the following as Method:

**K-Means Clustering** See the [“K Means Cluster”](#) chapter on page 245.

**Normal Mixtures** See the [“Normal Mixtures”](#) chapter on page 263.

**Robust Normal Mixtures** See [“Normal Mixtures NCluster=<k> Report”](#) on page 271 in the [“Normal Mixtures”](#) chapter.

**Self Organizing Map** See [“Self Organizing Map”](#) on page 257 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 242.

**Ward** In Ward’s minimum variance method, the distance between two clusters is the ANOVA sum of squares between the two clusters summed over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give the proportions of variance (squared semipartial correlations).

Ward’s method joins clusters to maximize the likelihood at each level of the hierarchy under the assumptions of multivariate normal mixtures, spherical covariance matrices, and equal sampling probabilities.

Ward’s method tends to join clusters with a small number of observations and is strongly biased toward producing clusters with approximately the same number of observations. It is also very sensitive to outliers. See Milligan ([1980](#)).

**Average** The distance between two clusters is the average distance between pairs of observations. Average linkage tends to join clusters with small variances and is slightly biased toward producing clusters with the same variance. See Sokal and Michener ([1958](#)).

**Centroid** The distance between two clusters is defined as the squared Euclidean distance between their means. The centroid method is more robust to outliers than most other hierarchical methods but in other respects might not perform as well as Ward’s method or average linkage. See Milligan ([1980](#)).

**Single** The distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster. Single linkage has many desirable theoretical properties but has performed poorly in Monte Carlo studies. See Jardine and Sibson (1971), Fisher and Van Ness (1971), Hartigan (1981), and Milligan (1980). Single linkage was originated by Florek et al. (1951a, 1951b) and later reinvented by McQuitty (1957) and Sneath (1957).

By imposing no constraints on the shape of clusters, single linkage sacrifices performance in the recovery of compact clusters in return for the ability to detect elongated and irregular clusters. Single linkage tends to chop off the tails of distributions before separating the main clusters. See Hartigan (1981).

**Complete** The distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. Complete linkage is strongly biased toward producing clusters with approximately equal diameters and can be severely distorted by moderate outliers. See Milligan (1980).

**Fast Ward** Applies an algorithm that computes Ward's method more quickly for large numbers of rows. The computation time is shorter because this algorithm does not require the calculation of a distance matrix. It is used automatically whenever there are more than 2,000 rows.

## Data Structure

These options describe the form of the data that is used in calculating multivariate distances:

**Data as usual** Data that are rectangular with one row for each observation and one column for each variable.

**Data as summarized** Data that are summarized by the levels of one or more identifying columns. When you select this option, an Object ID text box appears in the launch window. Specify the identifying columns as the Object ID. The **Data as summarized** option calculates level means and treats these means as your input data.

**Data is distance matrix** Data that consist of distances between observations. For  $n$  observations, the distance table should have  $n$  rows and  $n + 1$  columns. One column (usually the first) must contain a unique identifier for each of the  $n$  observations. The remaining columns contain distances between that observation and the  $n$  observations. Note the following:

- The diagonal elements of the table should be zero or missing, because the distance between a point and itself is zero. Values that are not zero or missing are treated as zero, and a note appears in the report.

- The distance columns can be a symmetric square matrix, or they can be upper or lower triangular with missing entries in the lower or upper portion. If the distances are given as a square matrix, a warning appears in the report if the table is not symmetric.
- You can begin with a different data structure and then save a distance matrix. See [“Save Distance Matrix”](#) on page 234.

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 236.

**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 229.
- 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 238.

---

---

**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; 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 241 and [“Example of Wafer Defect Classification Using Spatial Measures”](#) on page 238.

---

## Hierarchical Cluster Report

The Hierarchical Cluster report displays the method used, a dendrogram, and the Clustering History table. If you assigned a column as a Label in the launch window, the column’s values identify each observation in the dendrogram.

- [“Dendrogram Report”](#)
- [“Illustration of Dendrogram and Distance Graph”](#)
- [“Clustering History Report”](#)

## Dendrogram Report

The dendrogram is a tree diagram that represents the agglomeration of observations into clusters. The dendrogram also gives information about the degree of dissimilarity of clusters.

The clustering process can be viewed by reading the dendrogram from left to right. Each step consists of combining the two *closest* clusters into a single cluster.

- The joining of clusters is indicated by horizontal lines that are connected by vertical lines.
- The horizontal position of the vertical line represents the distance between the two clusters that are most recently joined to form the specified number of clusters.

---

**Note:** When the number of observations is less than 256, the distances are proportional to the distances shown in the Distance Graph. Otherwise, Geometric Spacing is used. See [“Dendrogram Scale”](#) on page 233.

---

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

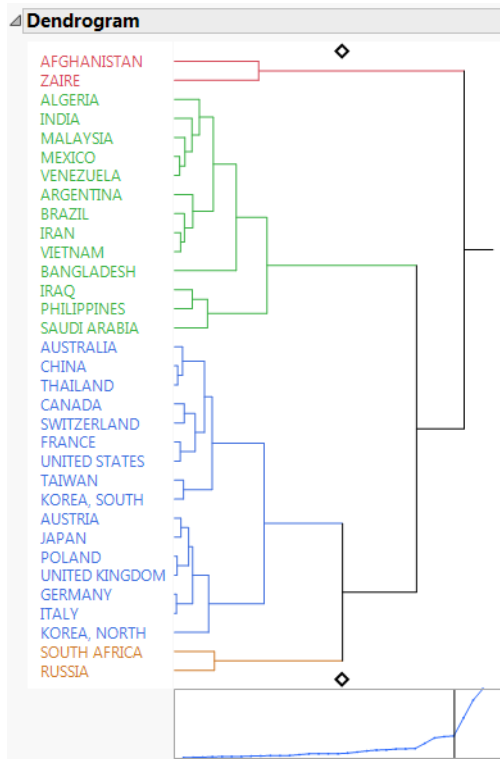
The Distance Graph is the plot that appears beneath the dendrogram. This graph has a point for each step where two clusters are joined into a single cluster. The horizontal coordinates represent the numbers of clusters, which decrease from left to right. The vertical coordinate of the point is the distance 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 221.

**Figure 11.5** Dendrogram Report for Birth Death Subset.jmp



In Figure 11.5, 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 225.

**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 Institute Inc. (1983). (Not available when **Data is distance matrix** is selected.)

**Show Dendrogram** Shows or hides the Dendrogram report.



**Dendrogram Scale** Contains the following options for scaling the dendrogram:

**Distance Scale** Shows the horizontal distances between any two join points as the distances between the two clusters joined at that point, based on the distance method specified on the launch window. The distance scale is the same scale as used in the Distance Graph and is the default scale for the dendrogram.

**Even Spacing** Shows the horizontal distances between any two join points as equal.

**Geometric Spacing** Increases the horizontal distances between join points as the number of clusters increases. This option is useful when there are many objects and you want the smaller clusters to be more visible than the larger clusters.

**Distance Graph** Shows or hides the distance plot beneath the dendrogram.

**Show NCluster Handle** Shows or hides the handles on the dendrogram used to manually change the number of clusters.

**Zoom to Selected Rows** Selects and enlarges a particular cluster after you select the cluster in the dendrogram. Alternatively, you can double-click the cluster to zoom in on it. Use Release Zoom to return to the original view.

**Release Zoom** Returns the dendrogram to the original view after zooming.

**Pivot on Selected Cluster** Reverses the order of the two sub-clusters of the currently selected cluster.

**Color Map** Gives the option to add a color map, or heat map, showing each Y, Column variable colored by value. Several color theme choices are available in a submenu.

**Two Way Clustering** Clusters by the variables specified in Y, Columns as well as rows. A color map is added with a dendrogram for the Y, Column variables at its base. Typically, for two-way clustering, your variables are measured on the same scale and you do not select Standardize Data. (Not available when **Data is stacked** is selected.)

**Positioning** Provides options for changing the positions of labels and other parts of the dendrogram.

**Legend** Shows or hides a legend for the colors used in color maps. This option is available only if a color map is enabled.

**More Color Map Columns** Adds a color map for specified columns. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.)

**Constellation Plot** Shows or hides an alternative way to present the information in the hierarchical clustering dendrogram. Each observation (row) is represented by an endpoint and each cluster join is represented by a new point. The lines that are drawn represent

cluster membership. The lengths of the lines represent the distance between clusters. Longer lines represent greater distances between clusters.

You can hover over the lines in the constellation plot to see their length. However, the length values are only meaningful with respect to each other. The axis scaling, orientation of points, and angles of the lines are arbitrary. They are determined such that the ends of the nodes are spaced out and the plot does not appear cluttered, which is important with larger data sets.

To turn off the labels on the endpoints, right-click inside the Constellation Plot and deselect **Show Labels**.

**Save Constellation Coordinates** Saves the coordinates of the constellation plot to the data table. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.)

**Save Clusters** Creates a data table column that contains the cluster number. If Add Spatial Measures is selected on the launch window, the cluster numbers are also saved to the Hough Data Table.

**Save Formula for Closest Cluster** Creates a data table column that contains a formula for the closest cluster. This option calculates the squared Euclidean distance to each cluster's centroid and selects the cluster that is closest. Note that this formula does not always reproduce the cluster assignment given by Hierarchical Clustering since the clusters are determined differently. However, the cluster assignment is very similar. (Not available when **Data as summarized**, **Data is distance matrix**, or **Data is stacked** is selected.)

**Save Display Order** Creates a data table column that contains the order in which the row appears in the dendrogram.

**Save Cluster Hierarchy** Creates a data table that contains the information needed to write a script for a custom dendrogram. For each cluster join, there are three rows: the first for the joiner, the second for the leader, and the third for the result, giving the cluster centers, size, and other information.

**Save Cluster Tree** Creates a new data table that contains information needed to compare cluster trees between JMP and SAS. For each cluster join, there is one row for each new cluster, with the cluster's size and other information.

**Save Distance Matrix** Creates a new data table that contains the distances between the observations.

**Save Cluster Means** Creates a new data table that contains the number of rows and the means of each column in each cluster.

**Cluster Summary** (Not available when **Data is distance matrix** is selected.) Displays the following information:

**Cluster Means** A table that gives, for each cluster, the number of observations (or Object IDs, if the data are stacked) and means for each variable.

**Cluster Standard Deviations** A table that gives, for each cluster, the number of observations (or Object IDs, if the data are stacked) and standard deviations for each variable.

**Cluster Means Plot** Either a parallel plot or a two-dimensional heat map of the cluster means.

The plot is a parallel plot unless **Data is stacked** is selected and there are two Attribute ID variables. For the parallel plot, the axis for each variable is scaled 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 235.

**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

- [“Example of a Distance Matrix”](#)
- [“Example of Wafer Defect Classification Using Spatial Measures”](#)

### Example of a Distance Matrix

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

- An identifier column (usually the first column) that has a Character data type.
- A set of  $n$  columns, where  $n$  is also the number of rows. These  $n$  columns define a symmetric matrix with zero or missing values on the diagonal.

Notice that the distance matrix in Flight Distances.jmp follows the preceding format.

1. Select **Help > Sample Data Library** and open Flight Distances.jmp.
2. Select **Analyze > Clustering > Hierarchical Cluster**.
3. In the list at the bottom left corner of the launch window, change **Data as usual** to **Data is distance matrix**.
4. Select Cities and click **Label**.
5. Select all remaining columns and click **Y, Columns**.

**Figure 11.6** Completed Distance Matrix Launch Window

**Select Columns**

▼ 29 Columns

- Cities
- Birmingham
- Boston
- Buffalo
- Chicago
- Cleveland
- Dallas
- Denver
- Detroit
- El Paso
- Houston
- Indianapolis
- Kansas City
- Los Angeles
- Louisville
- Memphis
- Miami
- Minneapolis
- New Orleans
- New York
- Omaha
- Philadelphia
- Phoenix
- Pittsburgh
- St. Louis
- Salt Lake City
- San Francisco
- Seattle
- Washington DC

**Options**

Method: Hierarchical

Method:

- ☒ Ward
- ☐ Average
- ☐ Centroid
- ☐ Single
- ☐ Complete
- ☐ Fast Ward

Data is distance matrix

☒ Standardize Data

☐ Standardize Robustly

☐ Missing value imputation

**Cast Selected Columns into Roles**

Y, Columns

- Birmingham
- Boston
- Buffalo
- Chicago
- Cleveland
- Dallas
- Denver
- Detroit
- El Paso
- Houston
- Indianapolis
- Kansas City
- Los Angeles
- Louisville
- Memphis
- Miami
- Minneapolis
- New Orleans
- New York
- Omaha
- Philadelphia
- Phoenix
- Pittsburgh
- St. Louis
- Salt Lake City
- San Francisco
- Seattle
- Washington DC

Ordering: optional numeric

Label: Cities

By: optional

**Action**

OK

Cancel

Remove

Recall

Help

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

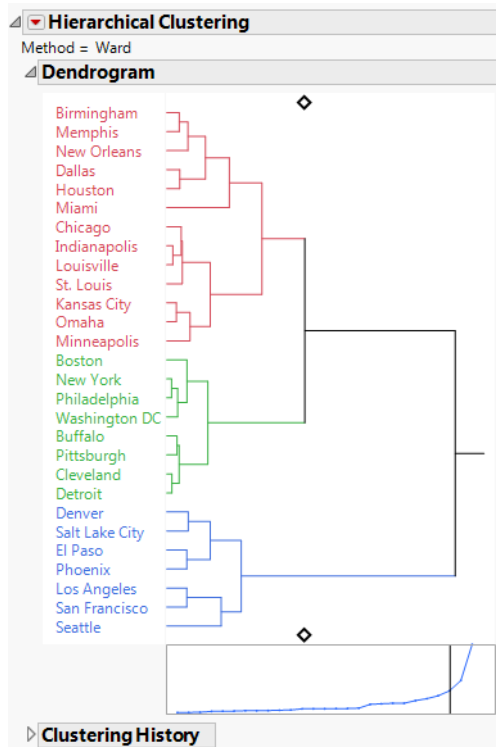
**Figure 11.7** Dendrogram Report for Flight Distances


Figure 11.7 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 230.

## 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 241.

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 11.8** Completed Clustering Launch Window

Finding points that are close, have similar values

Select Columns

▼ 6 Columns

- Lot
- Wafer
- Lot\_Wafer Label
- X\_Die
- Y\_Die
- Defects

Options

Hierarchical ▼

Method

- ☒ Ward
- ☐ Average
- ☐ Centroid
- ☐ Single
- ☐ Complete
- ☐ Fast Ward

Data is stacked ▼

- ☐ Standardize Data
- ☐ Standardize Robustly
- ☐ Missing value imputation
- ☒ Add Spatial Measures

Cast Selected Columns into Roles

Y, Columns: Defects  
optional

Ordering: optional numeric

Attribute ID: X\_Die  
Y\_Die  
optional

Object ID: Lot  
Wafer  
optional

Label: optional

By: optional

Action

OK

Cancel

Remove

Recall

Help

8. Click **OK**.

**Figure 11.9** Spatial Components Window

Choose Spatial Components

Variables	Number	Weight
<input checked="" type="checkbox"/> Attributes	1423	1
<input checked="" type="checkbox"/> Angle, Pie	18	<input type="text" value="1"/>
<input checked="" type="checkbox"/> Radius, Circle	21	<input type="text" value="1"/>
<input checked="" type="checkbox"/> Streak Angle	18	<input type="text" value="1"/>
<input checked="" type="checkbox"/> Streak Position	10	<input type="text" value="1"/>
<input type="checkbox"/> Position in Shot		<input type="text" value="1"/>
<input type="checkbox"/> Shot		<input type="text" value="1"/>

Shot Horiz Size:

Shot Vert Size:

OK Cancel

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

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

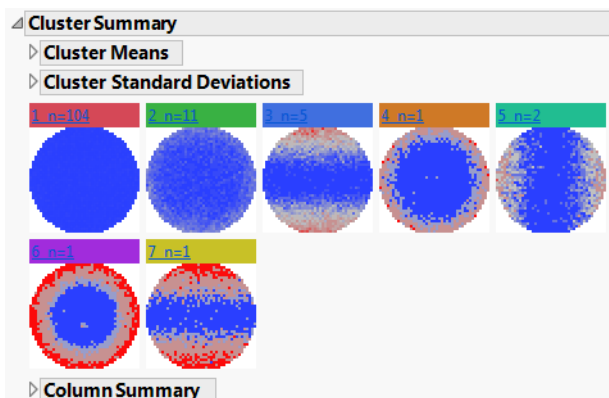
Two windows open: the Hierarchical Clustering report and the Wafer Stacked Defects Spatial data table.

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 11.10** 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 241.

---

## Statistical Details for the Hierarchical Clustering Platform

- [“Spatial Measures”](#)
- [“Distance Method Formulas”](#)



## Spatial Measures

To use the Add Spatial Measures option, your data must be stacked and contain two attribute columns that correspond to spatial coordinates. Some 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 238.

### 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:

$$\text{floor}[(h\text{Size}+h\text{ShotSize}-1)/h\text{ShotSize}] * \text{floor}[(v\text{Size}+v\text{ShotSize}-1)/v\text{ShotSize}]$$

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 225.

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  $i$ th observation

$C_K$  is the  $K$ th cluster, subset of  $\{1, 2, \dots, n\}$

$N_K$  is the number of observations in  $C_K$

$\bar{x}$  is the sample mean vector

$\bar{x}_K$  is the mean vector for cluster  $C_K$

$\|\mathbf{x}\|$  is the square root of the sum of the squares of the elements of  $\mathbf{x}$  (the Euclidean length of the vector  $\mathbf{x}$ )

$d(x_i, x_j)$  is  $\|x_i - x_j\|^2$

**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:

$$D_{KL} = \|\bar{x}_K - \bar{x}_L\|^2$$

**Ward's** Distance for Ward's method is:

$$D_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

**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 12

## K Means Cluster

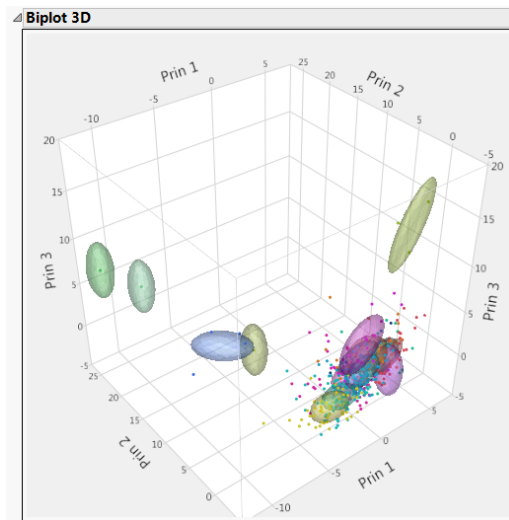
### Group Observations Using Distances

---

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 12.1** 3D Biplot



**Contents**

Overview of the K Means Cluster Platform .....	247
Overview of Platforms for Clustering Observations .....	247
Example of K Means Cluster .....	249
Launch the K Means Cluster Platform .....	252
Iterative Clustering Report .....	253
Iterative Clustering Options .....	253
Iterative Clustering Control Panel .....	254
K Means NCluster=<k> Report .....	255
Cluster Comparison Report .....	255
K Means NCluster=<k> Report .....	255
K Means NCluster=<k> Report Options .....	256
Self Organizing Map .....	257
Self Organizing Map Control Panel .....	258
Self Organizing Map Report .....	259
Description of SOM Algorithm .....	259
Additional Example of K Means Cluster Platform .....	260
Example of a Self-Organizing Map .....	260

---

## Overview of the K Means Cluster Platform

K Means Cluster is one of four platforms that JMP provides for clustering observations. For a comparison of all four methods, see [“Overview of Platforms for Clustering Observations”](#) on page 247.

The K Means Cluster platform forms a specified number of clusters using an iterative fitting process. The  $k$ -means algorithm first selects a set of  $k$  points called *cluster seeds* as an initial guess for the means of the clusters. Each observation is assigned to the nearest cluster seed to form a set of temporary clusters. The seeds are then replaced by the cluster means, the points are reassigned, and the process continues until no further changes occur in the clusters.

The  $k$ -means algorithm is a special case of the *EM algorithm*, where  $E$  stands for Expectation, and  $M$  stands for maximization. In the case of the  $k$ -means algorithm, the calculation of temporary cluster means represents the Expectation step, and the assignment of points to the closest clusters represents the Maximization step.

$K$ -Means clustering supports only numeric columns.  $K$ -Means clustering ignores modeling types (nominal and ordinal) and treats all numeric columns as continuous.

You must specify the number of clusters,  $k$ , or a range of values for  $k$ , in advance. However, you can compare the results of different values of  $k$  to select an optimal number of clusters for your data.

For background on  $K$ -Means clustering, see the FASTCLUS Procedure chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017c) 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 291.

---

JMP provides four platforms that you can use to cluster observations:

- Hierarchical Cluster is useful for smaller tables with up to several tens of thousands of rows and allows character data. Hierarchical clustering combines rows in a hierarchical sequence that is portrayed as a tree. You can choose the number of clusters that is most appropriate for your data after the tree is built.

- K Means Cluster is appropriate for larger tables with up to millions of rows and allows only numerical data. You need to specify the number of clusters,  $k$ , in advance. The algorithm guesses at cluster seed points. It then conducts an iterative process of alternately assigning points to clusters and recalculating cluster centers.
- Normal Mixtures is appropriate when your data come from a mixture of multivariate normal distributions that might overlap and allows only numerical data. For situations where you have multivariate outliers, you can use an outlier cluster with an assumed uniform distribution.

You need to specify the number of clusters in advance. Maximum likelihood is used to estimate the mixture proportions and the means, standard deviations, and correlations jointly. Each point is assigned a probability of being in each group. The EM algorithm is used to obtain estimates.

- Latent Class Analysis is appropriate when most of your variables are categorical. You need to specify the number of clusters in advance. The algorithm fits a model that assumes a multinomial mixture distribution. A maximum likelihood estimate of cluster membership is calculated for each observation. An observation is classified into the cluster for which its probability of membership is the largest.

**Table 12.1** Summary of Clustering Methods

Method	Data Type or Modeling Type	Data Table Size	Specify Number of Clusters
Hierarchical Cluster	Any	With Fast Ward, up to 200,000 rows  With other methods, up to 5,000 rows	No
K Means Cluster	Numeric	Up to millions of rows	Yes
Normal Mixtures	Numeric	Any size	Yes
Latent Class Analysis	Nominal or Ordinal	Any size	Yes

Some of the clustering platforms have options to handle outliers in the data. However, if your data has outliers, it is best to explore them first prior to analyzing. This can be done using the Explore Outliers Utility. For more information, see the Modeling Utilities chapter in the *Predictive and Specialized Modeling* book.



## 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 12.2** 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 12.3 K Means NCluster=11 Report

K Means NCluster=11

Columns Scaled Individually

Cluster Summary

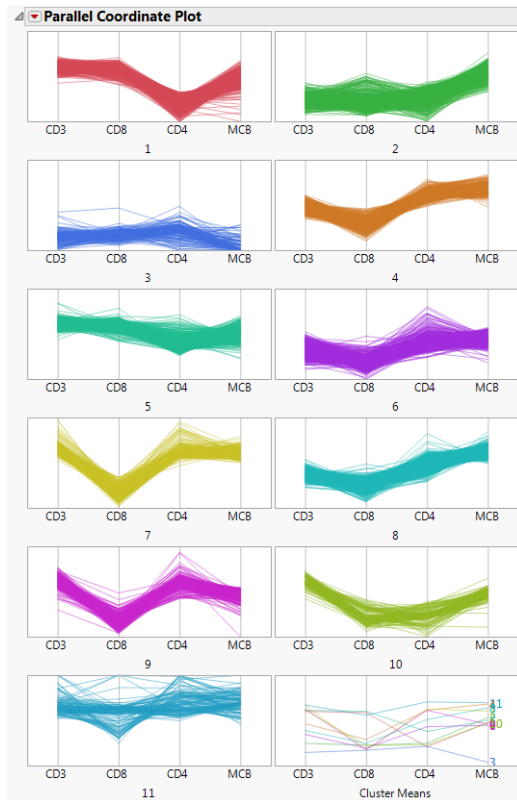
Cluster	Count	Step	Criterion
1	816	24	0
2	482		
3	157		
4	577		
5	856		
6	498		
7	447		
8	549		
9	377		
10	123		
11	118		

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.
- Click the K Means NCluster=11 red triangle and select **Parallel Coord Plots**.

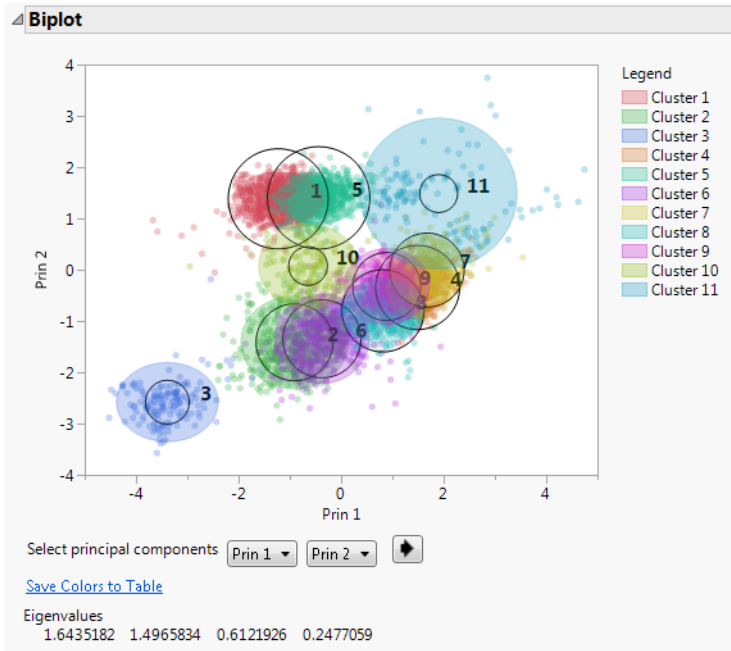
**Figure 12.4** Parallel Coordinate Plots for Cytometry Data



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 12.5 Biplot for Cytometry Data

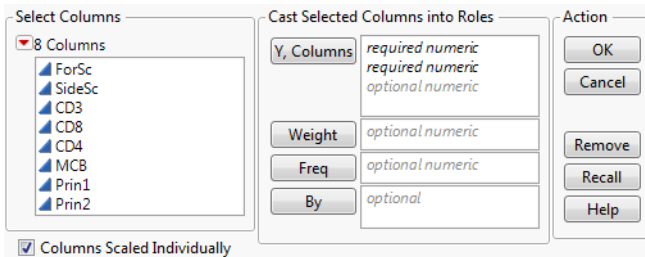


A legend that identifies the colors of the clusters is shown to the right of the plot. The clusters that appear to be most separated from the others based on their first two principal components are clusters 3, 10, and 11. This is supported by their parallel coordinate plots in Figure 12.4, which differ from the plots for the other clusters. Use the list below the plot to see the biplot for other combinations of principal components.

# Launch the K Means Cluster Platform

Launch the K Means Cluster platform by selecting **Analyze > Clustering > K Means Cluster**. Figure 12.6 shows the Clustering launch window for Cytometry.jmp.

Figure 12.6 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

**Columns Scaled Individually** Scales each column independently of the other columns. Use when variables do not share a common measurement scale, and you do not want one variable to dominate the clustering process. For example, one variable might have values that are between 0 and 1000, and another variable might have values between 0 and 10. In this situation, you can use the option so that the clustering process is not dominated by the first variable.

When you click OK, a Control Panel appears. See [“Iterative Clustering Control Panel”](#) on page 254.

---

## Iterative Clustering Report

When you click OK in the launch window, the Iterative Clustering report window appears, showing a Control Panel for fitting models. See [“Iterative Clustering Control Panel”](#) on page 254. As you fit models, additional reports are added to the window. See [“K Means NCluster=<k> Report”](#) on page 255.

### Iterative Clustering Options

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.

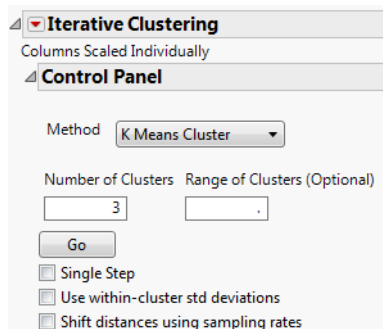
**Save Script** Contains options that enable you to save a script that reproduces the report to several destinations.

**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 12.7. You can iteratively fit different numbers of clusters or you can specify a range using the Range of Clusters option.

**Figure 12.7** Iterative Clustering Control Panel



The Control Panel has the following options:

**Method** The following clustering methods are available:

**KMeans Clustering** Described in this chapter.

**Self Organizing Map** Described in [“Self Organizing Map Control Panel”](#) on page 258.

**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 255.
- 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 12.2 and [“K Means NCluster=11 Report”](#) on page 250.

## 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, along with a legend identifying the cluster colors. Circles are drawn around the cluster centers and the size of the circles is proportional to the count inside the cluster. The shaded area is the density contour around the mean. By default, this area indicates where 90% of the observations in that cluster would fall (Mardia et al. 1980). Use the list below the plot to change the plot axes to other principal components. Alternatively, use the arrow button to cycle through all possible axes combinations. An option to save the cluster colors to the data table is also located below the plot. For details, see [“Save Colors to Table”](#) on page 256. The eigenvalues are shown in decreasing order.

---

**Note:** If Columns Scaled Individually is checked in the launch window, the biplot uses a correlation matrix. If Columns Scaled Individually is not checked, the biplot uses a covariance matrix.

---

**Biplot Options** Contains the following options for controlling the appearance of the Biplot:

**Show Biplot Rays** Shows the biplot rays. The labeled rays show the directions of the covariates in the subspace defined by the principal components. They represent the degree of association of each variable with each principal component.

**Biplot Ray Position** Enables you to specify the position and radius scaling of the biplot rays. By default, the rays emanate from the point (0,0). In the plot, you can drag the rays or use this option to specify coordinates. You can also adjust the scaling of the rays to make them more visible with the radius scaling option.

**Biplot Contour Density** Enables you to specify the confidence level for the density contours. The default confidence level is 90%.

**Mark Clusters** Assigns markers that identify the clusters to the rows of the data table.

**Biplot 3D** Shows a three-dimensional biplot of the data. Available only when there are three or more variables.

**Parallel Coord Plots** Creates a parallel coordinate plot for each cluster. The plot report has options for showing and hiding the data and means. For details, see the Parallel Plots chapter in the *Essential Graphing* book.

**Scatterplot Matrix** Creates a scatterplot matrix using all of the Y variables.

**Save Colors to Table** Assigns colors that identify the clusters to the rows of the data table. If there is a Biplot in the report window, the colors saved to the data table match the colors



of the clusters in the Biplot. If the colors are changed in the Biplot and the Save Colors To Table option is selected again, the colors in the table will update to match those in the Biplot.

**Save Clusters** Saves the following two columns to the data table:

- The **Cluster** column contains the number of the cluster to which the given row is assigned.
- (Not available for Self Organizing Maps.) The **Distance** column contains the squared Euclidean distance between the given observation and its cluster mean. For each variable, the difference between the observation's value and the cluster mean on that variable is divided by the overall standard deviation for the variable. These scaled differences are squared and summed across the variables.

**Save Cluster Distance** (Not available for Self Organizing Maps.) Saves a **Distance** column to the data table. This column is the same as the **Distance** column obtained from the **Save Clusters** option.

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

**Save Distance Formula** (Not available for Self Organizing Maps.) Saves a formula column called **Distance Formula** to the data table. This is the formula that calculates the distance to the assigned cluster.

**Save K Cluster Distances** (Not available for Self Organizing Maps.) Saves  $k$  columns containing the squared Euclidean distances to each cluster center.

**Save K Distance Formulas** (Not available for Self Organizing Maps.) Saves  $k$  columns containing the formulas for the squared Euclidean distances to each cluster center.

**Publish Cluster Formulas** Publishes to the Formula Depot the same scoring code used in the **Save Cluster Formula** option.

**Simulate Clusters** Creates a new data table containing simulated cluster observations on the  $Y$  variables, using the cluster means and standard deviations.

**Remove** Removes the clustering report.

---

## Self Organizing Map

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

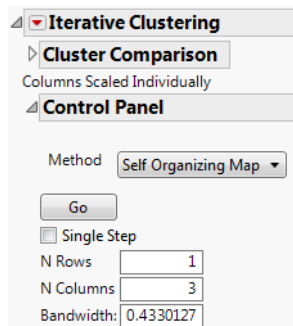
a variation on  $k$ -means clustering. In the SOM literature, this variation is called a *batch algorithm* using a *locally weighted linear smoother*.

The goal of a SOM is not only to form clusters in a particular layout on a cluster grid, such that points in clusters that are near each other in the SOM grid are also near each other in multivariate space. In classical  $k$ -means clustering, the structure of the clusters is arbitrary, but in SOMs the clusters have a grid structure. The grid structure helps interpret the clusters in two dimensions: clusters that are close are more similar than distant clusters. See [“Description of SOM Algorithm”](#) on page 259.

## Self Organizing Map Control Panel

Select the Self Organizing Map option from the Method list in the Iterative Clustering Control Panel (Figure 12.8).

**Figure 12.8** Self Organizing Map Control Panel



**Iterative Clustering**

Cluster Comparison

Columns Scaled Individually

**Control Panel**

Method: Self Organizing Map

Go

☐ Single Step

N Rows: 1

N Columns: 3

Bandwidth: 0.4330127

Some of the options on the panel are described in [“Iterative Clustering Control Panel”](#) on page 254. The other options are described as follows:

**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.



- The cluster assignment proceeds as with  $k$ -means. Each point is assigned to the cluster closest to it.
- The means are estimated for each cluster as in  $k$ -means. JMP then uses these means to set up a weighted regression with each variable as the response in the regression, and the SOM grid coordinates as the regressors. The weighting function uses a kernel function that gives large weight to the cluster whose center is being estimated. Smaller weights are given to clusters farther away from the cluster in the SOM grid. The new cluster means are the predicted values from this regression.
- These iterations proceed until the process has converged.

## Additional Example of K Means Cluster Platform

### Example of a Self-Organizing Map

This example uses the Iris.jmp sample data table, which includes measurements of sepal length, sepal width, petal length, and petal width for three species of irises.

1. Select **Help > Sample Data Library** and open Iris.jmp.
2. Select **Analyze > Clustering > K Means Cluster**.
3. Select Sepal length, Sepal width, Petal length, and Petal width and click **Y, Columns**.
4. Click **OK**.
5. Select **Self Organizing Map** from the Method menu on the Control Panel.
6. Set **N Rows** equal to 1 and **N Columns** equal to 2.
7. Click **Go**.
8. Open the Control Panel Report.
9. Set **N Rows** equals to 1 and **N Columns** equal to 3.
10. Click **Go**.
11. Open the Control Panel Report.
12. Set **N Rows** equal to 2 and **N Columns** equal to 2.
13. Click **Go**.

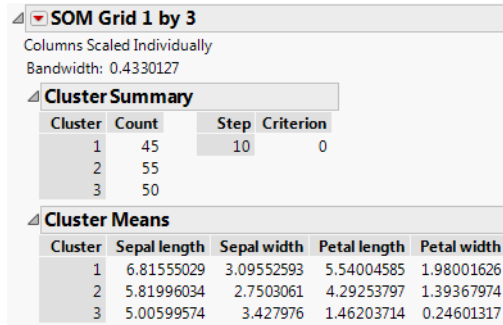
**Figure 12.10** SOM Cluster Comparison

Cluster Comparison			
Method	NCluster	N Rows	CCC Best
Self Organizing Map	2	1	3.35952
Self Organizing Map	3	1	4.95837 Optimal CCC
Self Organizing Map	4	2	3.64938

The Cluster Comparison report appears at the top of the report window. The best fit is determined by the highest CCC value. Notice the number of clusters that gives the largest CCC is 3, which is the number of species.

14. Scroll to the SOM Grid 1 by 3 report. We can see the classification was not perfect; each cluster should represent each species, with 50 rows for each.

**Figure 12.11** Self-Organizing Map Report for Iris.jmp


 The image shows a screenshot of the 'SOM Grid 1 by 3' report window in JMP. The window has a title bar with a red triangle icon and the text 'SOM Grid 1 by 3'. Below the title bar, it says 'Columns Scaled Individually' and 'Bandwidth: 0.4330127'. There are two expandable sections: 'Cluster Summary' and 'Cluster Means'. The 'Cluster Summary' section is expanded, showing a table with columns 'Cluster', 'Count', 'Step', and 'Criterion'. The 'Cluster Means' section is also expanded, showing a table with columns 'Cluster', 'Sepal length', 'Sepal width', 'Petal length', and 'Petal width'.

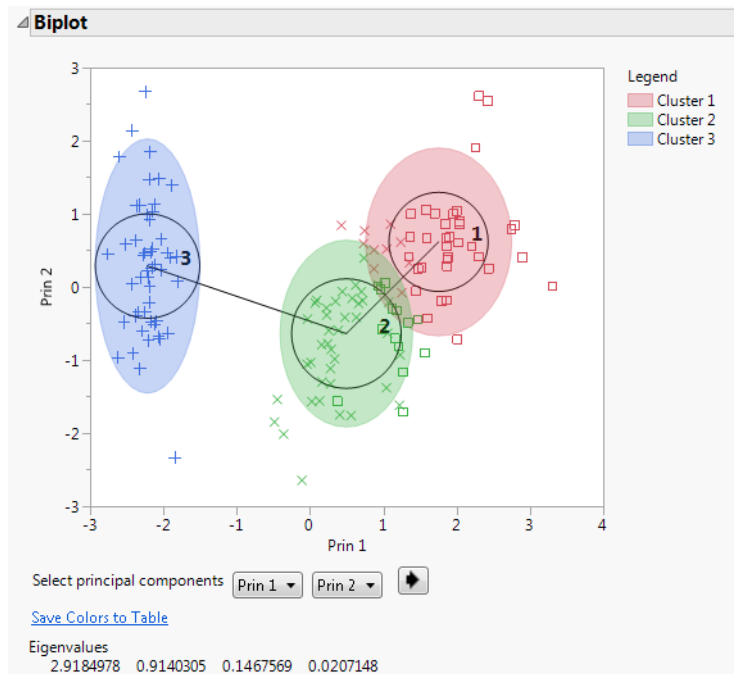
Cluster	Count	Step	Criterion
1	45	10	0
2	55		
3	50		

Cluster	Sepal length	Sepal width	Petal length	Petal width
1	6.81555029	3.09552593	5.54004585	1.98001626
2	5.81996034	2.7503061	4.29253797	1.39367974
3	5.00599574	3.427976	1.46203714	0.24601317

15. In the data table, select the Species column and select **Rows > Color or Mark by Column**.
16. Select the **Classic** option under Markers.
17. Click **OK**.
18. Click the red triangle menu next to SOM Grid 1 by 3 and select **Biplot**.

Figure 12.12 SOM Biplot



We can see that all rows from Cluster 3 are correctly identified as the setosa species. The other two species, virginica and versicolor, overlap slightly and can be mistaken for each other.

# Chapter 13

## Normal Mixtures

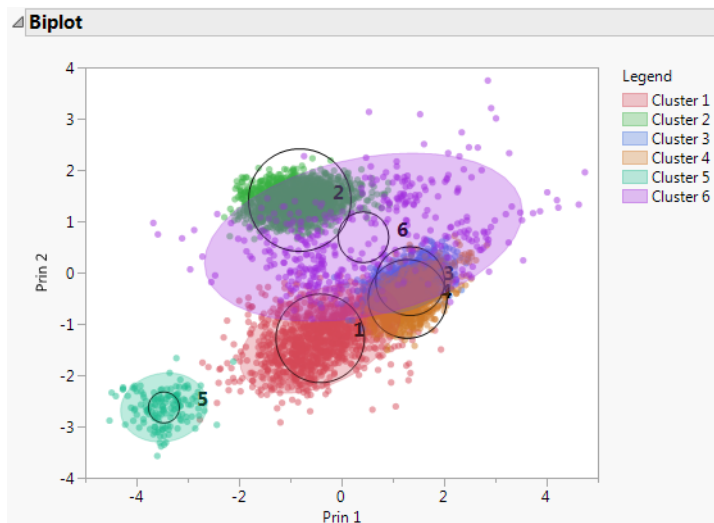
### Group Observations Using Probabilities

Use Normal Mixtures for clustering when your data come from overlapping normal distributions. You need to specify the number of clusters in advance.

Normal mixtures is an iterative technique based on the assumption that the joint probability distribution of the observations is approximated using a mixture of multivariate normal distributions. These mixtures represent different clusters. The individual clusters have multivariate normal distributions.

When clusters are well separated, hierarchical and  $k$ -means clustering work well. But when clusters overlap, normal mixtures provides a better alternative, because it is based on cluster membership probabilities, rather than arbitrary cluster assignments based on borders.

**Figure 13.1** Normal Mixtures Biplot



**Contents**

Overview of the Normal Mixtures Clustering Platform .....	265
Overview of Platforms for Clustering Observations .....	265
Example of Normal Mixtures Clustering .....	267
Launch the Normal Mixtures Clustering Platform .....	268
Model Based Clustering Report .....	269
Model Based Clustering Options .....	269
Model Based Clustering Control Panel .....	270
Normal Mixtures NCluster=<k> Report .....	271
Cluster Comparison Report.....	271
Normal Mixtures NCluster=<k> Report .....	272
Normal Mixtures NCluster=<k> Report Options .....	272
Statistical Details for the Normal Mixtures Clustering Platform.....	274



---

## Overview of the Normal Mixtures Clustering Platform

Normal Mixtures is one of four platforms that JMP provides for clustering observations. For a comparison of all four methods, see [“Overview of Platforms for Clustering Observations”](#) on page 265.

Normal mixtures is an iterative clustering technique for numerical variables. However, it also predicts the proportion of responses expected within each cluster. Normal mixtures assumes that the joint probability distribution of the measurement columns can be approximated using a mixture of multivariate normal distributions, which represent different clusters. Mean vectors and covariance matrices are estimated for each cluster. See McLachlan and Krishnan (1997) and Section 9.6 in Hand et al. (2001).

---

**Note:** The Normal Mixtures algorithm involves iterating through random guesses for the cluster centers. Because of this, results from different runs of the analysis might differ slightly.

---

If you suspect that you have multivariate outliers, you have two options. You can use an outlier cluster or the Explore Outliers Utility. The outlier cluster option assumes a uniform distribution and is less sensitive to outliers than the standard Normal Mixtures method. The Explore Outliers Utility enables you to explore and handle outliers prior to analysis. For details, see [“Outlier Cluster”](#) on page 271 and the Modeling Utilities chapter in the *Predictive and Specialized Modeling* book.

## 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 291.

---

JMP provides four platforms that you can use to cluster observations:

- Hierarchical Cluster is useful for smaller tables with up to several tens of thousands of rows and allows character data. Hierarchical clustering combines rows in a hierarchical sequence that is portrayed as a tree. You can choose the number of clusters that is most appropriate for your data after the tree is built.
- K Means Cluster is appropriate for larger tables with up to millions of rows and allows only numerical data. You need to specify the number of clusters,  $k$ , in advance. The

algorithm guesses at cluster seed points. It then conducts an iterative process of alternately assigning points to clusters and recalculating cluster centers.

- Normal Mixtures is appropriate when your data come from a mixture of multivariate normal distributions that might overlap and allows only numerical data. For situations where you have multivariate outliers, you can use an outlier cluster with an assumed uniform distribution.  
  
You need to specify the number of clusters in advance. Maximum likelihood is used to estimate the mixture proportions and the means, standard deviations, and correlations jointly. Each point is assigned a probability of being in each group. The EM algorithm is used to obtain estimates.
- Latent Class Analysis is appropriate when most of your variables are categorical. You need to specify the number of clusters in advance. The algorithm fits a model that assumes a multinomial mixture distribution. A maximum likelihood estimate of cluster membership is calculated for each observation. An observation is classified into the cluster for which its probability of membership is the largest.

**Table 13.1** Summary of Clustering Methods

Method	Data Type or Modeling Type	Data Table Size	Specify Number of Clusters
Hierarchical Cluster	Any	With Fast Ward, up to 200,000 rows  With other methods, up to 5,000 rows	No
K Means Cluster	Numeric	Up to millions of rows	Yes
Normal Mixtures	Numeric	Any size	Yes
Latent Class Analysis	Nominal or Ordinal	Any size	Yes

Some of the clustering platforms have options to handle outliers in the data. However, if your data has outliers, it is best to explore them first prior to analyzing. This can be done using the Explore Outliers Utility. For more information, see the Modeling Utilities chapter in the *Predictive and Specialized Modeling* book.

## 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 13.2 Normal Mixtures NCluster=6 Report

**Model Based Clustering**

**Cluster Comparison**

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

**Control Panel**

**Normal Mixtures NCluster=6**

**Cluster Summary**

Cluster	Count	Proportion
1	393	0.08550
2	944	0.18467
3	1194	0.24341
4	147	0.02932
5	720	0.14049
6	1602	0.31661

**Cluster Means**

Cluster	CD3	CD8	CD4	MCB
1	336.456533	193.385204	238.384846	206.503738
2	233.766979	140.184238	298.679748	256.78221
3	173.951704	109.760064	187.942955	195.539918
4	87.8647731	86.2326743	107.791654	10.1473172
5	338.61353	86.6669975	315.560339	208.345296
6	317.251206	306.00611	150.866465	189.617978

**Cluster Standard Deviations**

-LogLikelihood	BIC	AICc
103637.52	208033.06	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.

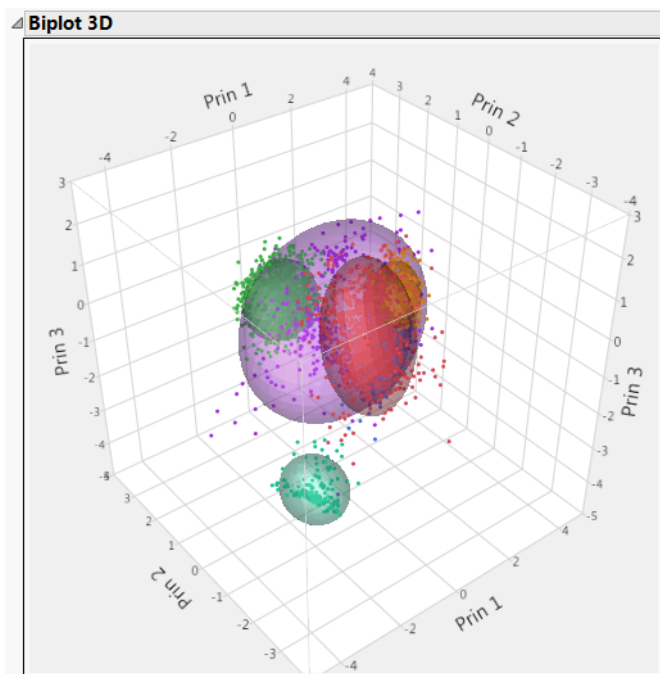
7. Click the red triangle next to Normal Mixtures NCluster=6 and select **Biplot 3D**.

---

**Note:** Your biplot 3D might appear differently because the algorithm has a random starting value.

---

**Figure 13.3** 3D Biplot of Cytometry Data



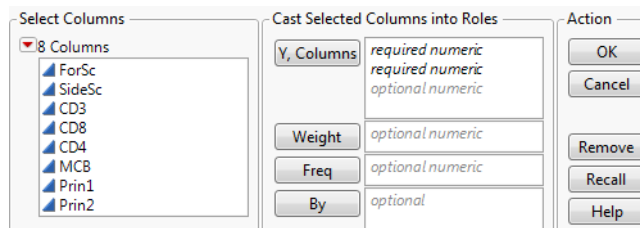
The plot shows contours for the normal densities that are fit to the clusters. Note that one cluster appears to be distinctly separated from the other clusters based on the first three principal components.

---

## Launch the Normal Mixtures Clustering Platform

Launch the Normal Mixtures Clustering platform by selecting Analyze > Clustering > Normal Mixtures. The Clustering launch window in Figure 13.4 uses the Cytometry.jmp sample data table.

**Figure 13.4** 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.

When you click OK, a Control Panel appears. See [“Model Based Clustering Control Panel”](#) on page 270.

---

## Model Based Clustering Report

When you click OK in the launch window, the Model Based Clustering report window opens, showing a Control Panel for fitting models. See [“Model Based Clustering Control Panel”](#) on page 270. As you fit models, additional reports are added to the window. See [“Normal Mixtures NCluster=<k> Report”](#) on page 271.

### Model Based Clustering Options

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.

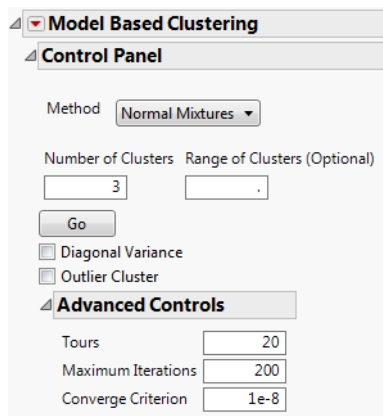
**Redo** Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

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

## Model Based Clustering Control Panel

The Control Panel for the Cytometry.jmp data table, with the variables CD3 through MCB as **Y, Columns**, is shown in Figure 13.5. 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 13.5** Control Panel for Normal Mixtures Method



The Model Based Clustering Control Panel has these options:

**Number of Clusters** Designates the number of clusters to form.

**Range of Clusters (Optional)** Provides an upper bound for the number of clusters to form. If a number is entered here, the platform creates separate analyses for every integer between **Number of clusters** and the value entered as **Range of Clusters (Optional)**.

**Go** Fits the clusters.

**Diagonal Variance** Constrains the off-diagonal elements of the covariance matrix to zero. The platform fits multivariate normal distributions that have no correlations between the variables.

---

**Note:** The Diagonal Variance option is sometimes necessary to avoid obtaining a singular covariance matrix when there are fewer observations than variables. It can also be used to avoid estimating very large covariance matrices for large numbers of variables.

---

**Outlier Cluster** Fits a cluster to catch outliers that do not fall into any of the normal clusters. If this cluster is created, it is designated Cluster 0, and the count of observations appears in the Cluster Summary report. The distribution of observations that fall in the outlier cluster is assumed to be uniform over the hypercube that encompasses the observations.

**Advanced Controls** The following advanced controls are available:

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

**Maximum Iterations** The maximum number of iterations of the convergence stage of the EM algorithm.

**Converge Criterion** The difference in the likelihood at which the EM iterations terminate.

---

## Normal Mixtures 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 271
- 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 in Figure 13.2 and [“Normal Mixtures NCluster=6 Report”](#) on page 267.

## Cluster Comparison Report

The Cluster Comparison report gives fit statistics to compare the various models. The fit statistics are BIC and AICc. Smaller values of each indicate better fit. The best fit is indicated in a column called Best.

## Normal Mixtures 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, along with a legend identifying the cluster colors. Circles are drawn around the cluster centers and the size of the circles is proportional to the count inside the cluster. The shaded area is the density contour around the mean. By default, this area indicates where 90% of the observations in that cluster would fall (Mardia et al. 1980). Use the list below the plot to change the plot axes to other principal components. Alternatively, use the arrow button to cycle through all possible axes combinations. An option to save the cluster colors to the data table is also located below the plot. For details, see [“Save Colors to Table”](#) on page 273. The eigenvalues are shown in decreasing order.

---

**Note:** The biplot always uses the correlation matrix to calculate the principal components.

---

**Biplot Options** Contains options for controlling the appearance of the Biplot.

**Show Biplot Rays** Shows the biplot rays. The labeled rays show the directions of the covariates in the subspace defined by the principal components. They represent the degree of association of each variable with each principal component.

**Biplot Ray Position** Enables you to specify the position and radius scaling of the biplot rays. By default, the rays emanate from the point (0,0). In the plot, you can drag the rays or use this option to specify coordinates. You can also adjust the scaling of the rays to make them more visible with the radius scaling option.

**Biplot Contour Density** Enables you to specify the confidence level for the density contours. The default confidence level is 90%.

**Mark Clusters** Assigns markers that identify the clusters to the rows of the data table.



**Biplot 3D** Shows a three-dimensional biplot of the data. Available only when there are three or more variables.

**Parallel Coord Plots** Creates a parallel coordinate plot for each cluster. The plot report has options for showing and hiding the data and means. For details, see the Parallel Plots chapter in the *Essential Graphing* book.

**Scatterplot Matrix** Creates a scatterplot matrix using all of the Y variables.

**Save Colors to Table** Assigns colors that identify the clusters to the rows of the data table. If there is a Biplot in the report window, the colors saved to the data table match the colors of the clusters in the Biplot. If the colors are changed in the Biplot and the Save Colors To Table option is selected again, the colors in the table will update to match those in the Biplot.

**Save Clusters** Adds a column called Cluster that contains the number of the cluster to which the given row is assigned to the data table. For normal mixtures, this is the cluster that is most likely.

**Save Cluster Formula** Adds a formula column called Cluster Formula to the data table. This formula identifies which cluster the row belongs to.

**Publish Cluster Formulas** Publishes to the Formula Depot the same scoring code used in the Save Cluster Formula option. If Publish Cluster Formulas is selected and Run Script is chosen from the model within the Formula Depot, the columns saved to the data table should match those that are saved when Save Cluster Formula is selected.

**Save Mixture Probabilities** Adds a column called Prob Cluster <k> for each cluster that contains the probability an observation belongs to that cluster.

**Save Mixture Formulas** Adds columns to the data table that contain the formulas used to calculate the mixture probabilities. Use these formula columns to score probabilities for excluded data, or data that you add to the table.

**Dist Formula <k>** The estimated multivariate normal density function for Cluster <k> evaluated at the observation.

**Dist Total** The sum of the distance formula columns. The formula in this column is equivalent to the formula in the Mixture Density column created by the Save Density Formula option.

**Prob Formula <k>** The probability that the observation belongs to Cluster <k>. These columns contain the formulas that give the values in the Prob Cluster <k> columns created by the Save Mixture Probabilities option. The column formula for calculating the mixture probabilities is:

$$\text{Prob Formula } \langle k \rangle = \frac{\text{Dist Formula } \langle k \rangle}{\text{Dist Total}}$$

**Save Density Formula** Adds a column called Mixture Density that contains the estimated density function for the normal mixture to the data table.

**Simulate Clusters** Uses the mixture density to simulate predictor values. Saves these and the clusters into which they are classified in a new data table.

**Remove** Removes the clustering report.

---

## Statistical Details for the Normal Mixtures Clustering Platform

Normal Mixtures uses the EM algorithm to do fitting because it is more stable than the Newton-Raphson algorithm. In addition, JMP uses a Bayesian regularized version of the EM algorithm, which allows smooth handling of cases where the covariance matrix is singular. Since the estimates are heavily dependent on initial guesses, the platform iterates through a number of tours. Each tour has randomly selected points for the initial center.

Doing multiple tours makes the estimation process somewhat expensive, so considerable patience is required for large problems. Controls enable you to specify the tour and iteration limits.

# Chapter 14

## 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 14.1** Example of Latent Class Analysis

Parameter Estimates													
Cluster	Overall	sex		marital status		country			size			type	
		Female	Male	Married	Single	American	European	Japanese	Large	Medium	Small	Family	Sporty
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
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
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
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
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

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					

**Contents**

Overview of the Latent Class Analysis Platform .....	277
Example of Latent Class Analysis.....	277
Launch the Latent Class Analysis Platform.....	280
The Latent Class Analysis Report.....	282
Latent Class Model for <k> Clusters Report .....	282
Latent Class Analysis Platform Options .....	284
Latent Class Analysis Options.....	285
Latent Class Model Options .....	285
Additional Example of the Latent Class Analysis Platform .....	286
Plot Probabilities of Cluster Membership .....	286
Statistical Details for the Latent Class Analysis Platform .....	287
Latent Class Model Fit .....	288
Maximum Number of Clusters.....	289

---

## Overview of the Latent Class Analysis Platform

The Latent Class Analysis platform fits a latent class model to categorical response variables and determines the most likely cluster or latent class for each observation. A *latent variable* is an unobservable grouping variable. Each level of the latent variable is called a *latent class*. For example, latent classes could be clusters of survey respondents that are grouped by their preference for risk.

The model takes the form of a multinomial mixture model. There are two sets of parameters in the model: the  $\gamma$  parameters and the  $\rho$  parameters. The  $\gamma$  parameters represent the overall probabilities of cluster membership. The  $\rho$  parameters represent the probabilities of observing a given response conditional on cluster membership. A latent class is characterized by a pattern of these conditional probabilities.

In order for the analysis results to be meaningful, a subject matter expert must interpret the clusters that the platform generates. This subject matter expert examines characteristics of the latent classes and constructs a definition for each class based on those characteristics.

---

**Note:** Rows with missing values in any of the response columns are excluded from the analysis.

---

For more information about latent class models, see Collins and Lanza (2010) and Goodman (1974).

---

## Example of Latent Class Analysis

This example uses the Latent Class Analysis platform to analyze responses to a 2005 survey of US high school students. The survey asked students a variety of multiple choice questions regarding health-risk behaviors.

In this example, you fit a latent class model to identify clusters of students based on their responses to 12 questions. The columns that you analyze were obtained from multiple choice survey questions by binning the responses into two classes (Yes/No).

1. Select **Help > Sample Data Library** and open Health Risk Survey.jmp.
2. In the Health Risk Survey data table, click the green triangle next to the Launch LCA Platform script.

The script selects the 12 columns of interest, opens the Latent Class Analysis launch window, and enters the 12 columns of interest as Y.

---

**Note:** To launch the LCA Platform on your own, select Analysis > Clustering > Latent Class Analysis.

---

3. Type 5 in the box next to **Up to**.  
This option fits latent class models for 3 and up to 5 clusters.
4. Click **OK**.

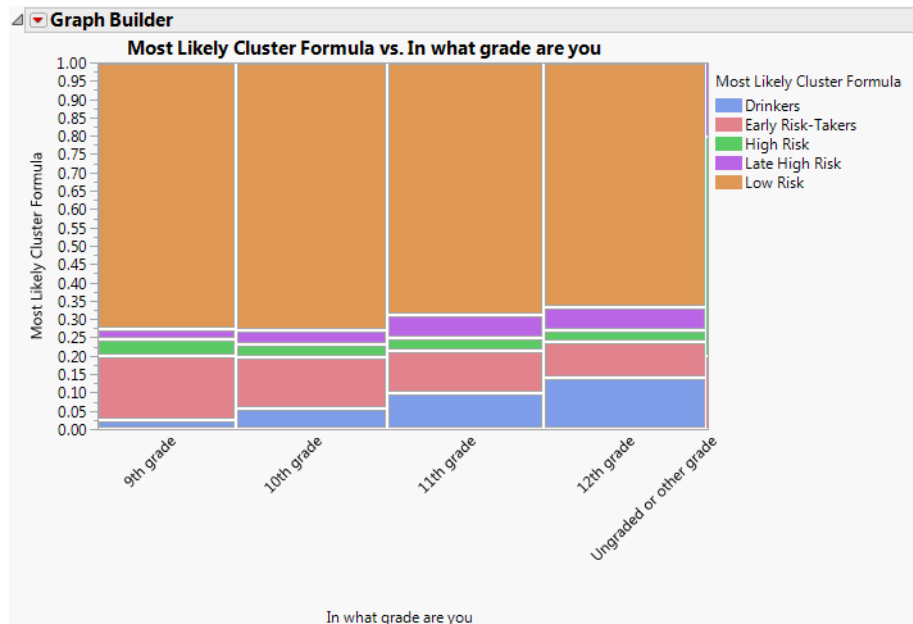
**Figure 14.2** Cluster Summary Report

Cluster Comparison				
NCluster	-LogLikelihood	BIC	AIC	Best
3	38713	77776.2	77501.9	
4	38207.1	76884.4	76516.3	
5	37964.8	76519.6	76057.6	Smallest BIC Smallest AIC

The Latent Class Analysis outline contains a Cluster Comparison report and three separate Latent Class Model reports. The Latent Class Model reports show the models for three, four, and five clusters. In the Cluster Comparison report, the model with five clusters has the smallest BIC and AIC, which indicates that it is the best fitting model out of the three. This is the model that you analyze.

5. In the Latent Class Model for 5 Clusters report, examine the bar charts under Parameter Estimates. Note the following:
  - Cluster 1 has mostly No answers to all of the risk behaviors.
  - Cluster 2 has high numbers of Yes answers for the four risk behaviors before the age of 13.
  - Cluster 3 has high numbers of Yes answers for driving when drinking and five or more drinks in the past 30 days.
  - Cluster 4 has high numbers of Yes answers for most of the risk behaviors except for the ones before the age of 13.
  - Cluster 5 has the highest number of Yes answers for most of the risk behaviors.
 Use this information to give the clusters meaningful names.
6. Click the red triangle next to Latent Class Model for 5 Clusters and select **Rename Clusters**:
  - Enter Low Risk for Cluster 1.
  - Enter Early Risk-Takers for Cluster 2.
  - Enter Drinkers for Cluster 3.
  - Enter Late High Risk for Cluster 4.
  - Enter High Risk for Cluster 5.
7. Click **OK**.



**Figure 14.4** Mosaic Plot of Cluster Membership versus Grade Level

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

---

## Launch the Latent Class Analysis Platform

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



**Figure 14.5** Latent Class Analysis Launch Window

Clusters rows based on categorical variables using multinomial mixtures. You must specify the number of latent classes (clusters) in advance.

Select Columns

▼ 204 Columns

- How old are you
- What is your sex
- In what grade are you
- Multiple Choice Questions (94/0)
- Dichotomous Response Questions (100/0)
- Body Mass Index Percentage
- Hidden Columns (6/0)

Cast Selected Columns into Roles

Y	required optional
Weight	optional numeric
Freq	optional numeric
ID	optional
By	optional

Action

OK

Cancel

Remove

Recall

Help

Number of Clusters

Up to

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

**Y** The column or columns that you want to analyze. You can analyze columns with nominal, ordinal, or multiple response modeling types. To analyze nominal or ordinal responses, two or more columns are required. Only one column is required if it contains multiple responses and has a multiple response modeling type.

**Weight** A column whose numeric values assign a weight to each row in the analysis.

**Freq** A column whose numeric values assign a frequency to each row in the analysis.

**ID** A column used to identify separate respondents. This identification is used in some output tables.

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

**Number of Clusters** The number of clusters to be computed in the analysis.

**Up to** Specifies a maximum number of clusters. If this number exceeds the value specified for Number of Clusters, a model report is produced with a number of clusters equal to each integer value in the range between Number of Clusters and Up to. These reports appear as part of the Latent Class Analysis report outline.

---

**Caution:** There is a maximum number of clusters that an LCA model can adequately fit. If you request more clusters than the maximum, a warning message appears in the report window. The LCA Platform fits up to the maximum number of clusters supported by the columns. For details about determining the maximum number of clusters, see [“Maximum Number of Clusters”](#) on page 289.

---

---

## The Latent Class Analysis Report

The initial Latent Class Analysis report contains a Cluster Comparison and Latent Class Model reports for each specified number of clusters.

### Cluster Comparison Report

The Cluster Comparison report gives fit statistics to compare the various models. The fit statistics are the negative log-likelihood (-LogLikelihood), BIC, and AIC. Smaller values of each indicate better fit. The best fit is indicated in a column called **Best**.

## Latent Class Model for <k> Clusters Report

The Latent Class Model Report for <k> Clusters contains the following results and outlines:

- [“Model Summary”](#) on page 282
- [“Parameter Estimates”](#) on page 283
- [“Transposed Parameter Estimates”](#) on page 283
- [“Effect Sizes”](#) on page 283
- [“MDS Plot”](#) on page 284
- [“Mixture Probabilities”](#) on page 284

### Model Summary

By default, a summary of the model for the specified number of clusters appears at the top of each Latent Class Model report. The model summary contains the -LogLikelihood, Number of Parameters, BIC, and AIC. These summary values can be used to determine how well the model fits the data. Lower values of -LogLikelihood, BIC, and AIC indicate better fits. For more information, see the Statistical Details Appendix in the *Fitting Linear Models* book. The Number of Parameters value gives the number of unique parameters in the latent class model. For more information, see [“Statistical Details for the Latent Class Analysis Platform”](#) on page 287.

## Parameter Estimates

The Parameter Estimates report contains tabular and graphical summaries of the parameter estimates and is displayed by default. Each summary contains rows corresponding to the model clusters.

The Overall column shows the probability of an observation belonging to each cluster. (These are the  $\gamma$  parameters. See [“Statistical Details for the Latent Class Analysis Platform”](#) on page 287.)

The remaining columns in the displays are grouped with vertical dividers according to the Y columns specified in the Latent Class Analysis launch window:

- Each group of categorical response columns has a column for each level within the respective response. In each group, the value in a given row and column is the conditional probability of the response indicated by the column, given that the observation belongs to the cluster identified by the row. (These are the  $p$  parameters.)
- Each group of multiple response columns has a column for each category within the multiple response. In each group, the value in a given row and column is the conditional probability of a response at the lower level of the indicated category, given that the observation belongs to the cluster identified by the row. (These are the  $p$  parameters.)

The graphical display shows the conditional probability values as *share charts*. For each cluster and each Y, the conditional probabilities given cluster membership are plotted as a horizontal stacked bar chart. For a binary or nominal response column, the percentages in these charts sum to one for each response. For a multiple response column, the percentages are of the lower level of each of the categories and do not sum to one. The stacking of bars follows the order of appearance of the variables in the table of values. You can also place your cursor over the bars to view the levels or categories of the variable.

---

**Tip:** You can select one or more rows in either table in the Parameter Estimates report to select the observations assigned to the corresponding clusters.

---

## Transposed Parameter Estimates

The Transposed Parameter Estimates report contains a table that is the transpose of the Parameter Estimates report table. Here the clusters are shown as columns. The conditional probabilities for each cluster are shown for each response category of each Y column in the analysis.

---

**Note:** The estimates from the Overall column are not included in the transposed table.

---

## Effect Sizes

The Effect Sizes table compares the Y columns across clusters and is displayed by default. The statistics in each row of this table are obtained from a contingency table analysis of expected

counts for cluster membership by levels or categories of a Y column. The expected counts are obtained by multiplying the number of observations in each cluster by the conditional probabilities for each level or category of the Y column.

For each response, the Pearson chi-square statistic,  $\chi^2$ , is calculated for the contingency table of expected counts for levels by clusters. Let  $n$  represent the number of observations. The value in the Effect Size column is defined as follows:

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

Each value in the LR Logworth column shows  $-\log_{10}(p_{LR})$  where  $p_{LR}$  is the likelihood ratio test  $p$ -value for the contingency table of expected counts. A Logworth value above 2 corresponds to significance at the 0.01 significance level.

---

**Tip:** You can select one or more rows in the Effect Sizes table to select the corresponding columns in the data table.

---

### MDS Plot

The MDS Plot contains one point for each cluster and is displayed by default. It is a two-dimensional representation of cluster proximity. Clusters that are closer together are more similar. The plot is created from a dissimilarity matrix of the  $p$  parameters. For more information about MDS plots, see the “[Multidimensional Scaling](#)” chapter on page 187.

### Mixture Probabilities

The Mixture Probabilities table displays probabilities of cluster membership for each row. The Most Likely Cluster column indicates the cluster with the highest probability of membership for each row.

---

**Note:** Rows that contain a missing value for one or more of the Y columns are excluded from the analysis and do not appear in the Mixture Probabilities table.

---



---

## Latent Class Analysis Platform Options

- “[Latent Class Analysis Options](#)”
- “[Latent Class Model Options](#)”

## Latent Class Analysis Options

The Latent Class Analysis red triangle menu contains the following options:

**New Number of Clusters** Enables you to run another analysis using a different number of clusters. The new analysis report is appended to the current report.

See 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.

## Latent Class Model Options

The Latent Class Model for <k> Clusters red triangle menu contains the following options:

**Model Reports** Enables you to show or hide the available model reports. For more information about the model reports, see [“The Latent Class Analysis Report”](#) on page 282.

**Color by Cluster** Colors each row in the data table according to its most likely cluster. For an example, see [“Additional Example of the Latent Class Analysis Platform”](#) on page 286.

**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.

---

**Remove Fit** Removes the specified clustering report from the report window.

---

## Additional Example of the Latent Class Analysis Platform

### Plot Probabilities of Cluster Membership

This example uses the Car Poll.jmp sample data table, which contains survey data for car owners and car makes. You are interested in classifying the car owners into three clusters and producing a plot to visualize the probabilities of cluster membership. A ternary plot provides a good visualization when you have three clusters.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Clustering > Latent Class Analysis**.
3. Select all of the columns except age and click **Y**.
4. Click **OK**.
5. Click the red triangle next to Latent Class Model for 3 Clusters and select **Color by Cluster**.
6. Click the red triangle next to Latent Class Model for 3 Clusters and select **Save Mixture Probabilities**.
7. In the Car Poll data table window, select the LCA Cluster Probabilities column group from the column list.
8. Select **Graph > Ternary Plot**.
9. Click **X, Plotting**.
10. Click **OK**.

**Figure 14.6** Ternary Plot of Cluster Membership Probabilities

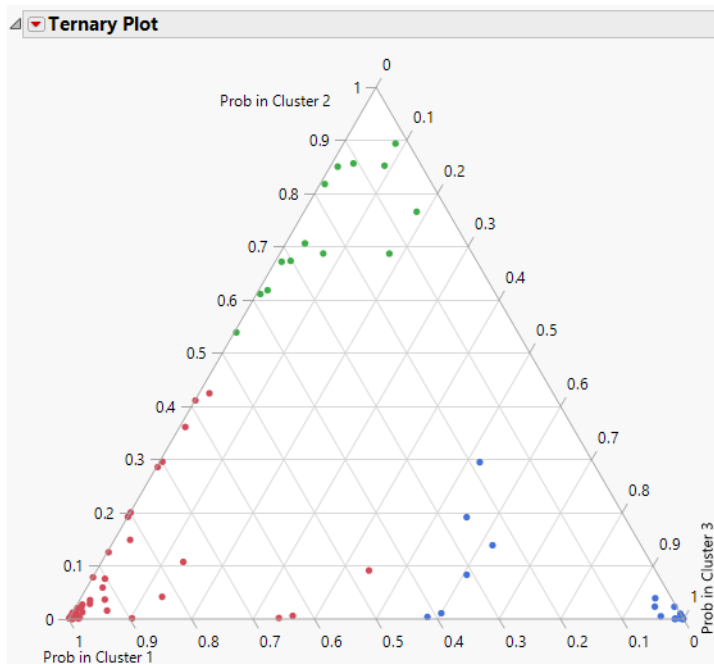


Figure 14.6 shows the ternary plot of cluster probabilities for each observation. Most of the cluster membership probabilities fall near the vertices, which indicates that they have high values for one cluster and lower values for the other two. However, there are some points in the middle of the plot, 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 a random seed was not specified.

## Statistical Details for the Latent Class Analysis Platform

- “Latent Class Model Fit”
- “Maximum Number of Clusters”

## Latent Class Model Fit

This section describes the latent class model that is fit in the Latent Class Analysis platform. For more information about latent class models, see Collins and Lanza (2010) and Agresti (2013).

---

**Note:** The LCA algorithm that is used in the Text Explorer platform takes advantage of the 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  $j = 1, \dots, J$  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  $W = R_1 * \dots * R_J$  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  $\mathbf{y} = y_1, \dots, y_J$ . Define  $\mathbf{Y}$  to be the  $W$  by  $J$  array of all the response patterns considered as row vectors. Each element,  $\mathbf{y}_w$ , in  $\mathbf{Y}$  has a probability  $\Pr(\mathbf{y}_w)$ . These probabilities sum to 1:

$$\sum_{w=1}^W \Pr(\mathbf{y}_w) = 1$$

Consider the following notation:

- $C$  is the number of clusters in the latent class model.
- $\gamma_c$  is the probability of membership in cluster  $c$ . (The  $\gamma_c$  are the latent class prevalences.) These parameters sum to 1.
- $r_{j,k}$  is the  $k^{\text{th}}$  level of the  $j^{\text{th}}$  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  $k^{\text{th}}$  level of the  $j^{\text{th}}$  response, and 0 otherwise.

The probability of observing a specific vector of responses  $\mathbf{y}_w = y_1, \dots, y_J$  is the sum of the conditional probabilities of observing that vector of responses for each of the  $C$  latent classes:

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



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  $\Pr(\text{Cluster} = c \mid \mathbf{y}_w)$ , which equals  $\Pr(\mathbf{y}_w \mid \text{Cluster} = c) / \Pr(\mathbf{y}_w)$ .

The  $\gamma$  and  $\rho$  parameters for latent class models are estimated using the iterative Expectation-Maximization (EM) algorithm. The number of unique parameters in a latent class model is defined as follows:

$$(C - 1) + C \sum_{j=1}^J (R_j - 1)$$

## Maximum Number of Clusters

The maximum number of clusters that can be fit in an LCA model depends on the model degrees of freedom. The degrees of freedom in an LCA model are based on the size of the contingency table created by the columns. The size of the contingency table is the number of cells in the table that contain at least one observation and is denoted as  $K$ . If all cells contain at least one observation,  $K$  is the product of the number of levels of the response columns. The formula for degrees of freedom is as follows:

$$\text{DF} = K - \{\text{nCluster} - 1 + \text{nCluster}(\text{nTotalLevels} - \text{nCols})\} - 1$$

where

nCluster = the number of clusters

nTotalLevels = the sum of the levels of the response columns

nCols = the number of response columns

In order for the LCA model to be adequately fit, the degrees of freedom must be positive. Therefore, to ensure  $\text{DF} > 0$ , the maximum number of clusters is defined as follows:

$$\max(\text{nCluster}) < \text{floor} \left[ \frac{K}{1 + \text{nTotal Levels} - \text{nCols}} \right]$$



# Chapter 15

## Cluster Variables

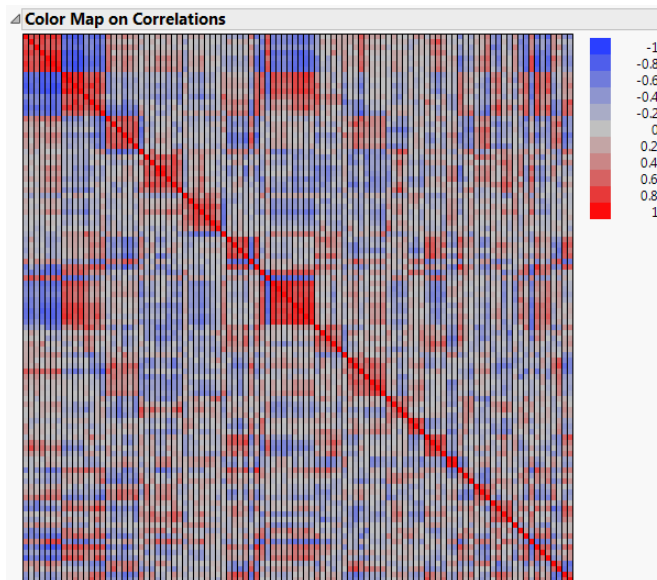
### Group Similar Variables into Representative Groups

---

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 15.1** Example of Correlation Map for Variables



**Contents**

Overview of the Cluster Variables Platform .....	293
Example of the Cluster Variables Platform .....	293
Launch the Cluster Variables Platform .....	295
The Cluster Variables Report .....	295
Color Map on Correlations .....	296
Cluster Summary .....	296
Cluster Members .....	297
Standardized Components .....	297
Cluster Variables Platform Options .....	297
Additional Examples of the Cluster Variables Platform .....	298
Example of Color Map on Correlations .....	298
Example of Cluster Variables Platform for Dimension Reduction .....	300
Statistical Details for the Cluster Variables Platform .....	303
Variable Clustering Algorithm .....	303

---

## Overview of the Cluster Variables Platform

Principal components analysis constructs components that are linear combinations of all the variables in the analysis. In contrast, the Cluster Variables option constructs components that are linear combinations of variables in a cluster of similar variables. The entire set of variables is partitioned into clusters. For each cluster, a *cluster component* is constructed using the first principal component of the variables in that cluster. This cluster component is the linear combination that explains as much of the variation as possible among the variables in that cluster.

You can use the Cluster Variables option as a dimension-reduction method. A substantial part of the variation in a large set of variables can often be represented by cluster components or by the most representative variable in the cluster. These new variables can then be used in predictive or other modeling techniques. The new cluster-based variables are usually more interpretable than principal components based on all the variables.

Principal components constructed from a common set of variables are orthogonal. However, cluster components are not orthogonal because they are constructed from distinct sets of variables.

When you have a large set of variables, the Cluster Variables platform uses an algorithm based on the singular value decomposition to shorten computation time. For additional background, see [“Wide Linear Methods and the Singular Value Decomposition”](#) on page 307 in the “Statistical Details” appendix.

---

## Example of the Cluster Variables Platform

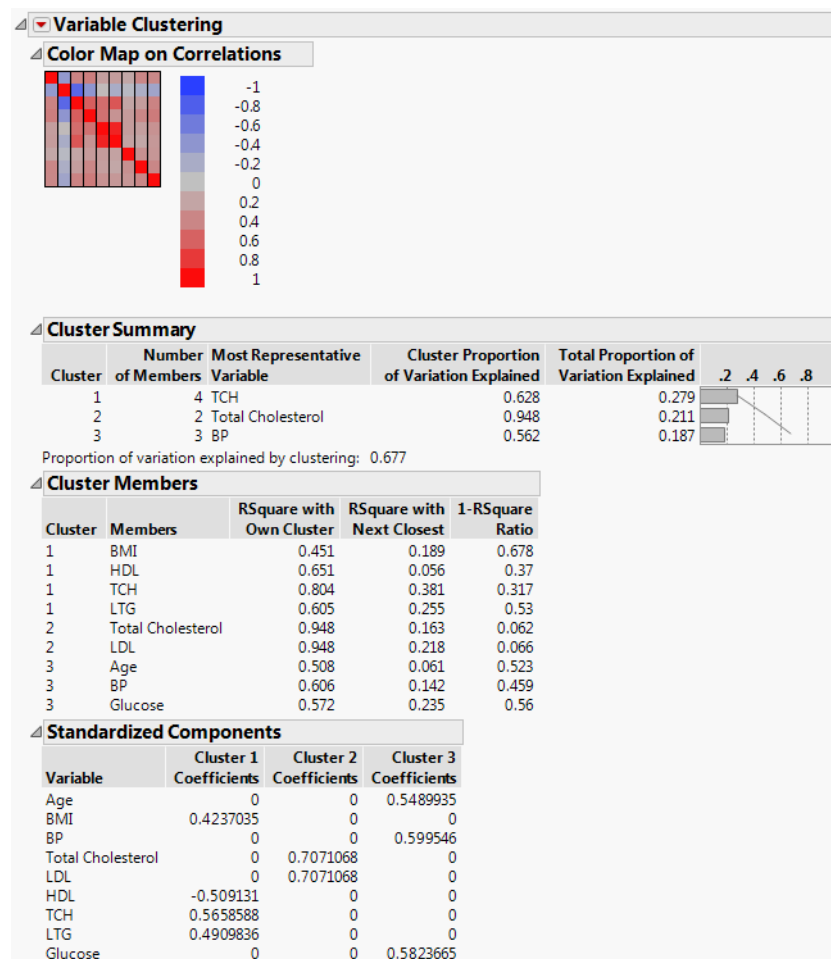
The Diabetes.jmp sample data table contains ten baseline variables used in modeling disease progression. In this example, you cluster the continuous baseline variables.

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 15.2 Cluster Variables Report for Diabetes Data



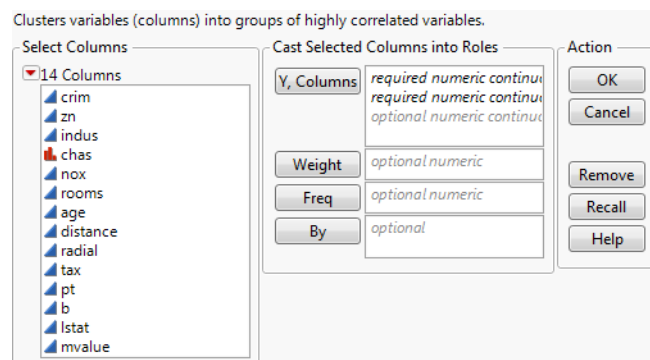
The Variable Clustering report is shown in Figure 15.2. The Cluster Summary report shows that the variables were grouped into three clusters:

- Cluster 1 consists of BMI, HDL, TCH, and LTG, as shown in the Cluster Members report. The Cluster Summary report shows that TCH is the most representative variable for Cluster 1 and that it explains 62.8% of the Cluster 1 variation.
- Cluster 2 consists of Total Cholesterol and LDL. The Cluster summary report shows that Total Cholesterol is the most representative variable for Cluster 2 and that it explains 94.8% of the Cluster 2 variation.
- Cluster 3 consists of Age, BP, and Glucose. The Cluster Summary report shows that the most representative variable is BP and it explains 56.2% 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 15.3.

**Figure 15.3** 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 296
- [“Cluster Summary”](#) on page 296
- [“Cluster Members”](#) on page 297
- [“Standardized Components”](#) on page 297

---

**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 298.

---

**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 [“Estimation Methods”](#) on page 50 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:

$$(1 - \text{RSquare with Own Cluster}) / (1 - \text{RSquare with Next Closest})$$

## 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 296.

**Cluster Summary** Shows or hides the Cluster Summary report. See [“Cluster Summary”](#) on page 296.

**Cluster Members** Shows or hides the Cluster Members report. See [“Cluster Members”](#) on page 297.

**Cluster Components** Shows or hides the Standardized Components report. See [“Standardized Components”](#) on page 297.

**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”](#)
- [“Example of Cluster Variables Platform for Dimension Reduction”](#)

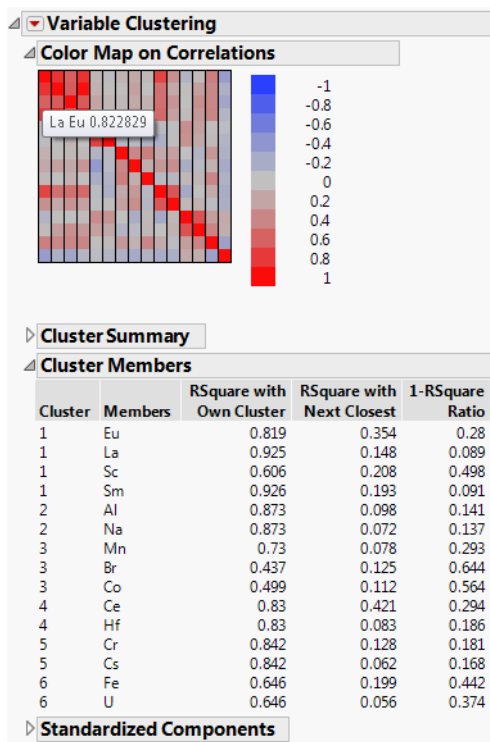
### Example of Color Map on Correlations

In this example, you construct and examine a Color Map on Correlations.

1. Select **Help > Sample Data Library** and open Cherts.jmp.
2. Select **Analyze > Clustering > Cluster Variables**.
3. Select all continuous variables and click **Y, Columns**.
4. Click **OK**.
5. Close the Cluster Summary and the Standardized Components reports.
6. Place your cursor over the cell in the second row and first column of the Color Map.

A tooltip appears, showing that the variables corresponding to this cell are La and Eu, and that their correlation is 0.822829.

**Figure 15.4** Color Map on Correlations for Cherts.jmp



The Cluster Members report shows that there are four variables in Cluster 1. In the Color Map on Correlations, the four-by-four square of cells in the upper left corner that corresponds to these five variables shows a distinct pattern of positive correlations. The color map also shows patterns of positive correlations for the variables in Clusters 2, 4, and 5. The two-by-two square of cells in the lower right corner of the color map that corresponds to the two Cluster 6 variables shows that they are negatively correlated. For more details, see [“Color Map on Correlations”](#) on page 296.

## Example of Cluster Variables Platform for Dimension Reduction

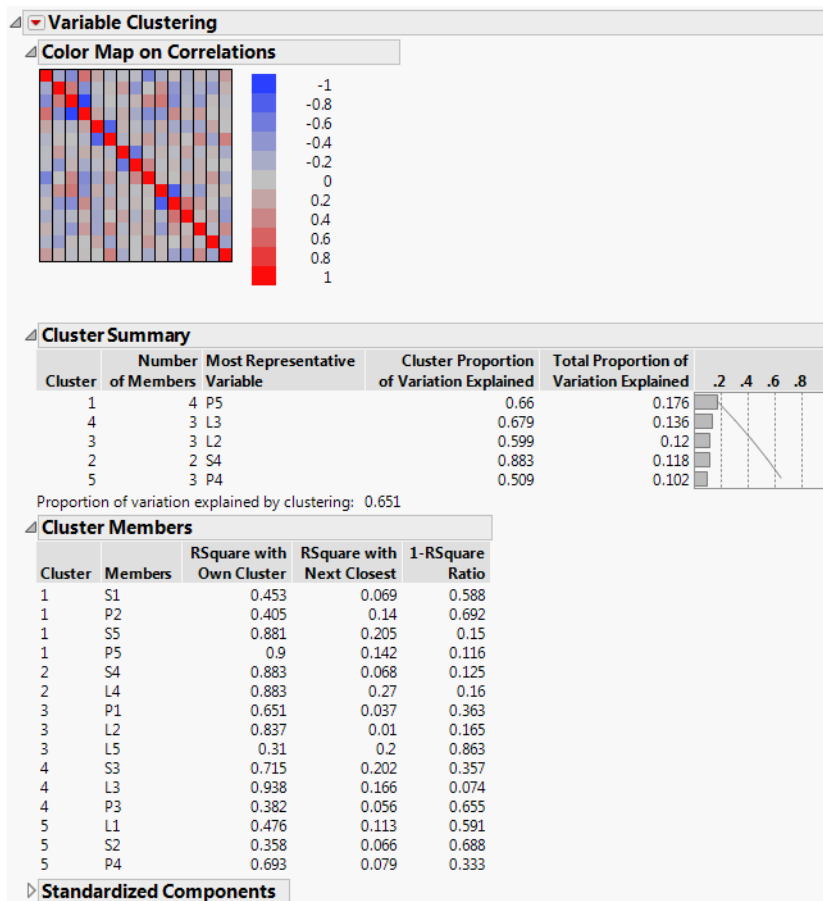
In this example, you use the Cluster Variables platform as a dimension-reduction tool for modeling. The Penta.jmp sample data table contains 15 variables used to predict the response variable, log RAI. Use Cluster Variables to reduce this number.

### Cluster Variables

1. Select **Help > Sample Data Library** and open Penta.jmp.
2. Select **Analyze > Clustering > Cluster Variables**.
3. Select all of the continuous variables, except logRAI and click **Y, Columns**.
4. Click **OK**.
5. Click the Variable Clustering red triangle and select **Save Cluster Components**.

Five formula columns are added to the data table.

Figure 15.5 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 15.6** 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 15.7** Fit Least Squares Report for Model with Cluster Components as Predictors

Response log RAI				
Effect Summary				
Summary of Fit				
RSquare		0.8214		
RSquare Adj		0.784191		
Root Mean Square Error		0.402125		
Mean of Response		0.734333		
Observations (or Sum Wgts)		30		
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	17.848635	3.56973	22.0757
Error	24	3.880902	0.16170	Prob > F
C. Total	29	21.729537		<.0001*
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.6651552	0.074349	8.95	<.0001*
Cluster 1 Components	-0.018483	0.056013	-0.33	0.7443
Cluster 2 Components	0.0035891	0.069032	0.05	0.9590
Cluster 3 Components	0.2043072	0.066714	3.06	0.0053*
Cluster 4 Components	0.5754553	0.065423	8.80	<.0001*
Cluster 5 Components	-0.046594	0.069786	-0.67	0.5107
Effect Tests				
Effect Details				

The model that includes the five Cluster Components as the only predictors explains a substantial amount of the variation in the response, with an adjusted Rsquare of 0.784. The model that uses all fifteen predictors has only a slightly higher adjusted Rsquare of 0.853 (Figure 15.6).

---

## Statistical Details for the Cluster Variables Platform

### Variable Clustering Algorithm

The clustering algorithm iteratively splits clusters of variables and reassigns variables to clusters until no more splits are possible. The initial cluster consists of all variables. The algorithm was developed by SAS and is implemented in PROC VARCLUS (SAS Institute Inc. 2017e).

**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:

- For all clusters, do the following:
  - Compute the principal components for the variables in each cluster.

2. 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:
  1. Rotate the principal components for the variables in the current cluster using an orthoblique rotation.
  2. 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.
  3. 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.
  4. 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:
  1. Compute the variable's squared correlation with the first principal component for each cluster.
  2. Place the variable in the cluster for which its squared correlation is the largest.

---

**Note:** An orthoblique rotation is also known as a raw quartimax rotation. See Harris and Kaiser ([1964](#)).

---



# Appendix **A**

## **Statistical Details** **Multivariate Methods**

---

This appendix discusses Wide Linear methods and the use of the singular value decomposition.

Contents

Wide Linear Methods and the Singular Value Decomposition ..... 307

    The Singular Value Decomposition ..... 307

    The SVD and the Covariance Matrix ..... 308

    The SVD and the Inverse Covariance Matrix ..... 308

    Calculating the SVD ..... 309

# 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 309.

## 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:

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

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

The matrices  $\mathbf{U}$ ,  $\text{Diag}(\Lambda)$ , 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  $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$

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

The  $\lambda_i$  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  $\lambda_i$  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 “Principal Components Report” on page 63 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:

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

If  $\mathbf{S}$  is of full rank, then  $\mathbf{V}$  is a  $p$  by  $p$  orthonormal matrix, and you can write  $\mathbf{S}^{-1}$  as follows:

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

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

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

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 117 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  $\mathbf{M}$  to a bidiagonal matrix  $\mathbf{J}$ . The second step consists of computing the singular values of  $\mathbf{J}$ , which are the same as the singular values of the original matrix  $\mathbf{M}$ . The columns of the matrix  $\mathbf{M}$  are usually standardized in order to equalize the effect of the variables on the calculation. The Golub and Kahan method is computationally efficient.



# Appendix **B**

## References

---

- Agresti, A. (2013). *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.
- Baglama, J., and Reichel, L. (2005). "Augmented implicitly restarted Lanczos bidiagonalization methods." *SIAM Journal on Scientific Computing* 27:19–42.
- Ballard, D. H. (1981). "Generalizing the Hough Transform to Detect Arbitrary Shapes." *Pattern Recognition* 13:111–122.
- Bartlett, M. S. (1937). "Properties of sufficiency and statistical tests." *Proceedings of the Royal Society of London, Series A* 160:268–282.
- Bartlett, M. S. (1954). "A Note on the Multiplying Factors for Various Chi Square Approximations." *Journal of the Royal Statistical Society, Series B* 16:296–298.
- Benzécri, J. P. (1979). "Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.]." *Cahiers de l'Analyse des Données* 4:377–378.
- Borg, I., and Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications*. 2nd ed. New York: Springer.
- Boulesteix, A.-L., and Strimmer, K. (2007). "Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data." *Briefings in Bioinformatics* 8:32–44.
- Browne, M. (2001). "An Overview of Analytic Rotation in Exploratory Factor Analysis." *Multivariate Behavioral Research* 36:111–150.
- Collins, L., and Lanza, S. (2010). *Latent Class and Latent Transition Analysis*. Hoboken NJ: John Wiley & Sons.
- Cox, I., and Gaudard, M. (2013). *Discovering Partial Least Squares with JMP*. Cary, NC: SAS Institute Inc.
- Cronbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16:297–334.
- Cudeck, R., and MacCallum, R. C., eds. (2007). *Factor Analysis at 100, Historical Developments and Future Directions*. Mahwah, NJ: Lawrence Erlbaum Associates.
- de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.
- De Jong, S. (1993). "SIMPLS: An Alternative Approach to Partial Least Squares Regression." *Chemometrics and Intelligent Laboratory Systems* 18:251–263.
- Denham, M. C. (1997). "Prediction Intervals in Partial Least Squares." *Journal of Chemometrics* 11:39–52.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C., and Wold, S. (2006). *Multi- and Megavariate Data Analysis Basic Principles and Applications (Part I)*. Chapter 4. Umetrics.

- Fisher, L., and Van Ness, J. W. (1971). "Admissible Clustering Procedures." *Biometrika* 58:91–104.
- Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951a). "Sur la liaison et la division des points d'un ensemble fini." *Colloquium Mathematicae* 2:282–285.
- Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951b). "Taksonomia Wroclawska." *Przegląd Antropologiczny* 17:193–211.
- Frank, I. E., and Todeschini, T. (1994). *The Data Analysis Handbook*. New York: Elsevier.
- Friedman, J. H. (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association* 84:165–175.
- Garthwaite, P. (1994). "An Interpretation of Partial Least Squares." *Journal of the American Statistical Association* 89:122–127.
- Golub, G. H., and Kahan, W. (1965). "Calculating the singular values and pseudo-inverse of a matrix." *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 2:205–224.
- Goodman, L. A. (1974). "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215–231.
- Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.
- Harris, C. W., and Kaiser, H. F. (1964). "Oblique Factor Analytic Solutions by Orthogonal Transformation." *Psychometrika* 32:363–379.
- Hartigan, J. A. (1981). "Consistency of Single Linkage for High-Density Clusters." *Journal of the American Statistical Association* 76:388–394.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Verlag.
- Hoskuldsson, A. (1988). "PLS Regression Methods." *Journal of Chemometrics* 2:211–228.
- Hoeffding, W. (1948). "A Non-Parametric Test of Independence." *Annals of Mathematical Statistics* 19:546–557.
- Huber, P. J. (1964). "Robust Estimation of a Location Parameter." *Annals of Mathematical Statistics* 35:73–101.
- Huber, P. J. (1973). "Robust Regression: Asymptotics, Conjecture, and Monte Carlo." *Annals of Statistics* 1:799–821.
- Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Jackson, J. E. (2003). *A User's Guide to Principal Components*. Hoboken, NJ: John Wiley & Sons.
- Jardine, N., and Sibson, R. (1971). *Mathematical Taxonomy*. New York: John Wiley & Sons.
- Jöreskog, K. G. (1977). "Factor Analysis by Least-Squares and Maximum Likelihood Methods." In *Statistical Methods for Digital Computers*, edited by K. Enslein, A. Ralston, and H. Wilf, 125 - 165. New York: John Wiley & Sons.



- Kohonen, T. (1989). *Self-Organization and Associative Memory*. 3rd ed. Vol. 8 of Springer Series in Information. Berlin: Springer-Verlag.
- Kohonen, T. (1990). "The Self-Organizing Map." *Proceedings of the IEEE* 78:1464–1480.
- Lindberg, W., Persson, J.-A., and Wold, S. (1983). "Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate." *Analytical Chemistry* 55:643–648.
- Mardia, K., Kent, J., and Bibby, J. (1980). *Multivariate Analysis*. New York: Academic Press.
- Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia: SIAM.
- McLachlan, G. J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons.
- McQuitty, L. L. (1957). "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies." *Educational and Psychological Measurement* 17:207–229.
- Milligan, G. W. (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika* 45:325–342.
- Nelson, P. R. C., Taylor, P. A., and MacGregor, J. F. (1996). "Missing Data Methods in PCA and PLS: Score calculations with incomplete observations." *Chemometrics and Intelligent Laboratory Systems* 35:45–65.
- Nunnally, J. C. (1978). *Psychometric theory*. 2nd ed. New York: McGraw-Hill.
- Penny, K. I. (1996). "Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance." *Journal of the Royal Statistical Society, Series C* 45:73–81.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1998). *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. Cambridge, England: Cambridge University Press.
- Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.
- SAS Institute Inc. (1983). *SAS Technical Report A-108: Cubic Clustering Criterion*. Cary, NC: SAS Institute Inc. Retrieved December 16, 2015 from [https://support.sas.com/documentation/onlinedoc/v82/techreport\\_a108.pdf](https://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf)
- SAS Institute Inc. (2017a). "The CANDISC Procedure." *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. <http://support.sas.com/documentation/onlinedoc/stat/143/candisc.pdf>
- SAS Institute Inc. (2017b). "The FACTOR Procedure." *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. <http://support.sas.com/documentation/onlinedoc/stat/143/factor.pdf>
- SAS Institute Inc. (2017c). "The FASTCLUS Procedure." *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. <http://support.sas.com/documentation/onlinedoc/stat/143/fastclus.pdf>
- SAS Institute Inc. (2017d). "The PLS Procedure." *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. <http://support.sas.com/documentation/onlinedoc/stat/143/pls.pdf>
- SAS Institute Inc. (2017e). "The VARCLUS Procedure." *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. <http://support.sas.com/documentation/onlinedoc/stat/143/varclus.pdf>

- Schafer, J., and Strimmer, K. (2005). "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics." *Statistical Applications in Genetics and Molecular Biology* 4 Article 32.
- Sneath, P. H. A. (1957). "The Application of Computers to Taxonomy." *Journal of General Microbiology* 17:201–226.
- Sokal, R. R., and Michener, C. D. (1958). "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Science Bulletin* 38:1409–1438.
- Tobias, R. D. (1995). "An Introduction to Partial Least Squares Regression." In *Proceedings of the Twentieth Annual SAS Users Group International Conference*, 1250–1257. Cary, NC: SAS Institute Inc. <http://www.sascommunity.org/sugi/SUGI95/Sugi-95-210%20Tobias.pdf>
- Tracy, N. D., Young, J. C., and Mason, R. R. (1992). "Multivariate Control Charts for Individual Observations." *Journal of Quality Technology* 24:88–95.
- Umetrics. (1995). *Multivariate Analysis (3-day course)*. Winchester, MA.
- Waern, Y. (1972). "Structure in Similarity Matrices: A Graphic Approach." *Scandinavian Journal of Psychology* 13:5–16.
- White, K. P., Jr., Kundu, B., and Mastrangelo, C. M., (2008). "Classification of Defect Clusters on Semiconductor Wafers Via the Hough Transform." *IEEE Transactions on Semiconductor Manufacturing* 21:272–278.
- Wold, S. (1994). "PLS for Multivariate Linear Modeling." In *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry*, edited by H. van de Waterbeemd, pp. 195–218. Weinheim, Germany: Verlag-Chemie.
- Wold, S., Sjostrom, M., and Eriksson, L. (2001). "PLS-Regression: A Basic Tool of Chemometrics." *Chemometrics and Intelligent Laboratory Systems* 58:109–130.

# Index

## Multivariate Methods

---

### Numerics

95% bivariate normal density ellipse [45](#)

### A

agglomerative clustering [219](#)

algorithms [307](#)

approximate F test [121](#)

Average Linkage [225](#), [243](#)

### B

Baltic.jmp [126](#)

bar chart of correlations [39](#)

Biplot [256](#), [272](#)

Biplot 3D [256](#), [273](#)

Biplot Options [256](#), [273](#)

Biplot Ray Position [104](#)

biplot rays [71](#)

bivariate normal density ellipse [45](#)

Burt table, Multiple Correspondence  
Analysis [164](#)

By variable [61](#), [176](#)

### C

calculation details [307](#)

Canonical 3D Plot [101](#)

centroid [47](#)

Centroid Method [225](#), [243](#)

classical test theory [201](#)

Cluster Criterion [232](#)

Cluster platform [217](#)

compare methods [219](#), [247](#), [265](#)

hierarchical [219–243](#)

introduction [219](#), [247](#), [265](#)

k-means [247–274](#)

launch [223](#)

normal mixtures [263](#)

Cluster the Correlations [42](#)

Clustering History [232](#)

Color Clusters [232](#)

Color Map [233](#)

Color Map On Correlations [42](#)

Color Map On p-values [42](#)

Color Points [104](#)

Common Factor Analysis [179](#)

common factor analysis [171](#)

Complete Linkage [226](#), [243](#)

computational details [307](#)

Consider New Levels [101](#)

Constellation Plot [233](#)

contrast M matrix [120](#)

correlation matrix [37](#)

Correspondence Analysis Options [160](#)

Cronbach's alpha [48–49](#)

statistical details [55](#)

Cross Table options, Multiple Correspondence  
Analysis [159](#)

### D

dendrogram [217](#), [219](#), [229](#)

Dendrogram Scale command [233](#)

Density Ellipse [45–46](#)

dimensionality [59](#)

Discriminant Analysis, PLS [152](#)

Distance Graph [233](#)

Distance Scale [233](#)

### E

E matrix [120](#)

Effect Sizes, Latent Class Analysis [283](#)

eigenvalue decomposition [59](#), [63](#)

Eigenvalues [181](#)

Eigenvectors [65](#)

Ellipse alpha [46](#)

**Ellipse Color** 46

**Ellipses Transparency** 46

EM algorithm 247

**Even Spacing** 233

Expectation Maximization algorithm 247

exploratory factor analysis 171

## F

factor analysis 59

Factor Analysis platform

By variable 61, 176

Freq variable 176

Weight variable 176

factor analysis, overview 59

**Factor Rotation** 71

Fit Line 46

formulas used in JMP calculations 307

Freq variable 176

## G

**Geometric Spacing** 233

group similar rows *see* Cluster platform

## H

H matrix 120

**Hoeffding's D** 44, 52

## I

Impute Missing Data in PLS 131

Inverse Corr table 39

inverse correlation 39, 53

IRT 31, 201

Item Analysis Platform 206

options 210

report 207

item characteristic curve 207

item reliability 48–49

Item Response Theory 31, 201

## J

Jackknife Distances 47

## K

**Kendall's Tau** 44

Kendall's tau-b 51

**KMeans** 247

K-Means Clustering Platform  
SOMs 257

## L

L matrix 120

latent trait 201

**Legend** 233

linear combination 59

**Loading Plot** 70

## M

M matrix 120

Mahalanobis distance 47, 53

**Mark Clusters** 232

MDS Plot 284

missing data imputation, PLS 131

missing value 38

missing values 39

Mixture Probabilities 284

multinomial mixture model 277

Multiple Correspondence Analysis

Burt table 164

Cross Table options 159

launch window options 156

platform options 159

Supplementary ID variable 157

Tests for Independence 159

**Multivariate** 33, 35

multivariate mean 47

multivariate outliers 47

Multivariate platform 305

principal components 59

## N

Nonpar Density 46

**Nonparametric Correlations** 43

Nonparametric Measures of Association  
table 43

normal density ellipse 45

Number of factors 179

## O

oblique transformation [180](#)  
orthogonal transformation [180](#)  
**Other** [46](#)  
**Outlier Analysis** [46](#)  
Outlier Distance plot [53](#)

## P

**Pairwise Correlations** [38](#)  
Pairwise Correlations table [39](#)  
**Parallel Coord Plots** [256](#), [273](#)  
Partial Corr table [39](#)  
partial correlation [39](#)  
Partial Least Squares platform  
validation [130](#)  
PCA [59](#)  
Pearson correlation [39](#), [51](#)  
PLS [123–150](#)  
Statistical Details [147–151](#)  
Principal Components [179](#)  
principal components analysis [59](#)  
product-moment correlation [39](#), [51](#)

## Q

questionnaire analysis [48–49](#)

## R

reduce dimensions [59](#)  
reliability analysis [48–49](#)  
**ROC Curve** [102](#)  
Rotation method [180](#)

## S

**Save Canonical Scores** [105](#)  
**Save Cluster Hierarchy** [234](#)  
**Save Cluster Tree** [234](#)  
**Save Clusters** [234](#), [257](#), [273](#)  
**Save Density Formula** [274](#)  
**Save Display Order** [234](#)  
**Save Distance Matrix** [234](#)  
**Save Formula for Closest Cluster** [234](#)  
**Save Mixture Formulas** [273](#)  
**Save Mixture Probabilities** [273](#)

**Scatterplot Matrix** [44](#), [101](#)  
scatterplot matrix [37](#)  
**Score Plot** [69](#)  
Scree Plot [181](#)  
**Scree Plot** [68](#)  
**Shaded Ellipses** [46](#)  
**Show Biplot Rays** [104](#)  
**Show Canonical Details** [105](#)  
**Show Correlations** [46](#)  
**Show Dendrogram** [232](#)  
**Show Distances to each group** [102](#)  
**Show Group Means** [101](#)  
**Show Histogram** [46](#)  
**Show Means CL Ellipses** [104](#)  
**Show Normal 50% Contours** [104](#)  
**Show Points** [45](#), [104](#)  
**Show Probabilities to each group** [102](#)  
**Show Within Covariances** [101](#)  
significance probability [39](#)  
**Simulate Clusters** [274](#)  
**Single Linkage** [226](#), [243](#)  
Solubility.jmp [60](#)  
SOMs [257](#)  
Spearman's Rho [51](#)  
**Spearman's Rho** [43](#)  
**Standardize Data** [228](#)  
statistical details [307](#)  
Supplementary ID variable, Multiple  
Correspondence Analysis [157](#)

## T

$T^2$  Statistic [47](#)  
Tests for Independence, Multiple  
Correspondence Analysis [159](#)  
Transposed Parameter Estimates [283](#)  
**Two way clustering** [233](#)

## U

Univariate [38](#)

## W-Z

**Ward's** [226](#), [243](#)  
Weight variable [176](#)

