バージョン13

多変量分析
第2版

「真の発見の旅とは、新しい風景を探すことではなく、新たな視点を持つことである。」
マルセル・プルースト

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

13.1

このマニュアルを引用する場合は、次の正式表記を使用してください: SAS Institute Inc. 2017.
『JMP® 13 多変量分析 第2版』Cary, NC: SAS Institute Inc.

JMP® 13 多変量分析 第2版

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

2017年2月

テクノロジーライセンスに関する通知

- Scintilla - Copyright © 1998-2014 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  何らかの目的でこのソフトウェアとそのマニュアルを手数料なしで使用、コピー、変更および配布することは、これをもって許可されます。ただし、すべてのコピーに上記の著作権に関する通知が記載されていること、および補助的なマニュアルに著作権に関する通知とこの許可に関する通知の両方が記載されていることを条件とします。

  NEIL HODGSONは、商業性および適合性の黙示的な保証を含め、このソフトウェアに関するすべての保証を放棄します。NEIL HODGSONは、いかなる場合においても、それが契約、過失、もしくは他の不法行為のどれであれ、このソフトウェアの使用もしくは性能から生じた、もしくはそれに関連して生じた使用、データ、もしくは利益の損失の結果として生じる特別損害、間接損害、もしくは付随的損害を始めとするいかなる損害に対しても責任を負いません。

- Telerik RadControls: Copyright © 2002-2012, Telerik. 含まれているTelerik RadControlsをJMP以外で使用することは許可されていません。
- ZLIB圧縮ライブラリ - Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Natural Earthを使用して作成。無料のベクトルおよびラスター地図データ @ naturalearthdata.com.
- パッケージ - Copyright © 2009-2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のためにWhiteBoxの名前やその貢献者の名前を使用することはできません。

  このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、著作権保有者または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- iODBC ソフトウェア - Copyright © 1995-2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  – 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために OpenLink Software Inc. の名前やその貢献者の名前を使用することはできません。

  このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、OPENLINK または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- bzip2、関連ライブラリの「libbzip2」、およびすべてのマニュアル: Copyright © 1996-2010, Julian R Seward. All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  このソフトウェアの供給源は正しく表記しなければならず、使用者が元のソフトウェアを記述したと主張することはできません。ある製品の中でこのソフトウェアを使用する場合は、その製品のマニュアルに謝辞を記載してもらえるとありがたいですが、必須ではありません。

  ソースに変更を加えたバージョンには、その旨を明記しなければならず、元のソフトウェアとは違うものであることを明確にしてください。

  事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために作成者の名前を使用することはできません。

  このソフトウェアは、作成者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、作成者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可

能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- R ソフトウェア : Copyright © 1999-2012, R Foundation for Statistical Computing.
- MATLAB ソフトウェア : Copyright © 1984-2012, The MathWorks, Inc. 米国特許法および国際特許法によって保護されています。www.mathworks.com/patents を参照してください。MATLAB および Simulink は、The MathWorks, Inc. の登録商標です。他の商標は、www.mathworks.com/trademarks に一覧されています。他の製品名やブランド名は、それぞれの所有者の商標または登録商標である可能性があります。
- libopc: Copyright © 2011, Florian Reuter. All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  – 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  – 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために Florian Reuter の名前やその貢献者の名前を使用することはできません。

  このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、著作権保有者または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

- libxml2 - ソースコードに特に記載がある場合を除く（たとえば、使用しているライセンスは類似しているが、著作権の通知が異なる hash.c、list.c ファイルや trio ファイル）、すべてのファイル :

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

  これをもって、このソフトウェアのコピーと関連する文書ファイル（「本ソフトウェア」）を入手した人すべてに対し、無料で本ソフトウェアを使用、コピー、変更、マージ、パブリッシュ、配布、サブライセンスする、もしくはコピーを販売する権利を含むがそれに限定せず、本ソフトウェアを制限なく取り扱う権利、および本ソフトウェアの供給相手に対してそうすることを許可する権利が付与されます。ただし、以下の条件を満たさなければなりません。

上記の著作権に関する通知とこの許可に関する通知が、本ソフトウェアのコピーのすべてまたは大部分に記載されていること。

このソフトウェアは、 現状のままで」提供され、商業性および特定の目的に対する適合性、および非侵害の保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。DANIEL VEILLARDは、いかなる場合においても、それが契約、過失、もしくは他の不法行為のどれであれ、本ソフトウェアから、もしくは本ソフトウェアに関連して、または本ソフトウェアの使用もしくは他の取り扱いに関連して生じた申し立て、損害賠償もしくは他の義務に対し、責任を負いません。

この通知に含まれているものを除き、Daniel Veillardから事前に書面による許可を得ることなく、本ソフトウェアの広告、またはその他の手段による本ソフトウェアの販売、使用もしくは他の取り扱いの宣伝にDaniel Veillardの名前を使用することはできません。

- UNIXファイルに使用された解凍アルゴリズムについて:

  Copyright © 1985, 1986, 1992, 1993

  カリフォルニア大学評議員。All rights reserved.

  このソフトウェアは、評議員および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、評議員または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

  1. 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  2. バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  3. 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために大学の名前や貢献者の名前を使用することはできません。

- Snowball - Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton.

  All rights reserved.

  ソースおよびバイナリの形で、そのまま、もしくは変更を加えて再配布および使用することは、次のような条件を満たす限り、許可されます。

  1. 再配布するソースコードには、上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

  2. バイナリ形式で再配布する場合は、共に提供されるマニュアルなどの資料に上記の著作権に関する通知、この条件リスト、これに続く放棄声明が記載されていなければなりません。

3. 事前に書面による許可を得ることなく、このソフトウェアから派生した製品の推奨または宣伝のために著作権保有者の名前や貢献者の名前を使用することはできません。

このソフトウェアは、著作権保有者および貢献者によって「現状のままで」提供され、商業性および特定の目的に対する適合性に関する黙示的な保証を含むがそれに限らない、いかなる明示的もしくは黙示的な保証も行われません。いかなる場合においても、著作権保有者または貢献者は、損害の原因が何であれ、そして法的責任の根拠が何であれ、つまり、契約、厳格責任、不法行為（過失その他を含む）のどれであれ、かかる損害の発生する可能性を事前に知らされていたとしても、このソフトウェアをどのように使用して生じた損害であれ、いかなる直接損害、間接損害、付随的損害、特別損害、懲罰的損害、もしくは結果損害（代替品または代替サービスの調達、使用機会、データもしくは利益の損失、業務の中断を含むがそれに限らない）に対しても責任を負いません。

目次
多変量分析

1 JMP の概要
マニュアルとその他のリソース 15
表記規則 16
JMP のマニュアル 17
JMP ドキュメンテーションライブラリ 17
JMP ヘルプ 23
JMP を習得するためのその他のリソース 23
チュートリアル 23
サンプルデータテーブル 24
統計用語と JSL 用語の習得 24
JMP を使用するためのヒント 24
ツールヒント 25
JMP User Community 25
JMPer Cable 25
JMP 関連書籍 25
「JMP スターター」ウィンドウ 26
テクニカルサポート 26

2 多変量分析について
多変量分析の概要 27

3 多変量の相関
多次元データの分布を検討する 29
「多変量の相関」プラットフォームの起動 30
「多変量」レポート 31
「多変量の相関」プラットフォームのオプション 33
ノンパラメトリック相関係数 37
散布図行列 38
外れ値分析 40
項目の信頼性 41
欠測データの補完 42
項目の信頼性の例 42
計算方法と統計的詳細 43
推定法について 43
Pearson の積率相関 44

ノンパラメトリックな相関 44
相関の逆行列 46
距離の計算 46
Cronbach の α 係数 48

4 主成分分析
多変量データの次元削減 49
主成分分析の概要 50
主成分分析の例 50
「主成分分析」プラットフォームの起動 51
欠測値のあるデータ 54
「主成分分析」レポート 54
「主成分分析」レポートのオプション 55
統計的詳細 63
推定法について 63
DModX の計算方法 64

5 線形判別分析
連続変数から分類変数を予測する 65
判別分析の概要 66
判別分析の例 66
「判別分析」起動ウィンドウ 68
ステップワイズ変数選択 69
判別法 72
共分散行列の縮小 75
「判別分析」レポート 75
主成分分析 76
正準プロットと正準構造 77
判別スコア 80
スコアの要約 81
判別分析のオプション 84
スコアオプション 85
正準オプション 87
三次元正準プロットの例 90
事前確率の指定 91
グループの追加 91
判別行列の保存 92
散布図行列 92
JMP と JMP Pro の違い 93
技術的詳細 94
「線形 横長データ」のアルゴリズムについて 94
保存される計算式 94
多変量検定 101

近似F検定 102
グループ間の共分散行列 103

6 PLS回帰
多重共線性がある場合の予測モデル 105
「PLS回帰」プラットフォームの概要 106
PLS回帰の例 107
「PLS回帰」プラットフォームの起動 110
中心化と尺度化 113
Xの標準化 113
モデルの設定パネル 113
「PLS回帰」レポート 115
モデル比較の要約 115
<検証法の名前>による検証手法 = <PLS法の名前> 116
あてはめレポート 120
PLS回帰のオプション 120
あてはめレポートのオプション 121
変数重要度のプロット 123
変数重要度 vs 係数プロット 123
列の保存 124
統計的詳細 126
PLS 126
van der Voet T^2 127
T^2プロット 128
Xスコア散布図行列の信頼楕円 128
予測値の標準誤差と信頼区間 128
標準化したスコアと負荷量 130
PLS判別分析（PLS-DA） 130

7 階層型クラスター分析
データ行をツリー構造にクラスタリング 131
階層型クラスター分析の概要 132
クラスター分析用プラットフォームの概要 132
クラスター分析の例 133
「階層型クラスター分析」プラットフォームの起動 136
クラスター分析の手法 137
距離の計算方法 137
データの構造 138
「Y, 列」変数の変換 139
「階層型クラスター分析」レポート 141
「樹形図」レポート 141
樹形図と距離グラフ 142
「クラスター分析の履歴」レポート 143

階層型クラスター分析のオプション ... 143
「階層型クラスター分析」プラットフォームのその他の例 ... 147
距離行列の例 ... 147
空間的な指標でウエハーをクラスタリングする例 ... 149
統計的詳細 ... 152
空間的な指標 ... 152
距離の手法の計算式 ... 154

8 K Means クラスター分析
データ行をクラスタリング ... 155
「K Means クラスター分析」プラットフォームの概要 ... 156
クラスター分析用プラットフォームの概要 ... 156
K-Means クラスターの例 ... 157
「K Means クラスター分析」プラットフォームの起動 ... 161
「反復クラスター分析」レポート ... 162
反復クラスター分析のオプション ... 163
「反復クラスター分析」設定パネル ... 163
「K Means 法クラスター数=<k>」レポート ... 165
「クラスターの比較」レポート ... 166
「K Means 法クラスター数=<k>」レポート ... 166
「K Means 法クラスター数=<k>」レポートのオプション ... 166
自己組織化マップ ... 167
自己組織化マップの設定パネル ... 168
「自己組織化マップ」レポート ... 169
自己組織化マップのアルゴリズムについて ... 169

9 正規混合分布法
多変量正規分布によりデータ行をクラスタリング ... 171
「正規混合」クラスター分析プラットフォームの概要 ... 172
クラスター分析用プラットフォームの概要 ... 172
正規混合クラスター分析の例 ... 173
「正規混合」クラスター分析プラットフォームの起動 ... 175
オプション ... 176
「反復クラスター分析」レポート ... 177
反復クラスター分析のオプション ... 177
「反復クラスター分析」設定パネル ... 177
「正規混合 クラスター数=<k>」レポート ... 179
「クラスターの比較」レポート ... 180
「正規混合 クラスター数=<k>」レポート ... 180
「正規混合 クラスター数=<k>」レポートのオプション ... 180
ロバスト正規混合 ... 182
ロバスト正規混合分布の設定パネル ... 182
ロバスト正規混合のレポート ... 183

正規混合法の統計的詳細 183
ロバスト正規混合の詳細 183

10 潜在クラス分析
カテゴリカルな変数のデータ行をクラスタリング 185
「潜在クラス分析」プラットフォームの概要 186
潜在クラス分析の例 186
「潜在クラス分析」プラットフォームの起動 189
「潜在クラス分析」レポート 190
「潜在クラスモデル（クラスター数: <k> 個）」レポート 190
「潜在クラス分析」プラットフォームのオプション 192
「潜在クラス分析」のオプション 192
潜在クラスモデルのオプション 192
「潜在クラス分析」プラットフォームの別例 193
クラスターメンバーの確率をプロットする 193
「潜在クラス分析」プラットフォームの統計的詳細 194

11 変数のクラスタリング
似通った変数をクラスターに分類 197
「変数のクラスタリング」プラットフォームの概要 198
「変数のクラスタリング」プラットフォームの例 198
「変数のクラスタリング」プラットフォームの起動 200
「変数のクラスタリング」レポート 200
相関のカラーマップ 201
クラスター要約 201
クラスターメンバー 202
標準化変数に対する係数 202
「変数のクラスタリング」プラットフォームのオプション 202
「変数のクラスタリング」プラットフォームの別例 203
相関のカラーマップの例 203
「変数のクラスタリング」プラットフォームの次元削減の例 204
「変数のクラスタリング」プラットフォームの統計的詳細 207

A 統計的詳細
多変量分析 209
「線形 横長データ」の手法と特異値分解 210
特異値分解 210
特異値分解と共分散行列 211
特異値分解および共分散行列の逆行列 211
特異値分解のアルゴリズム 212

B 参考文献

索引

多变量分析 217

第 1 章

JMP の概要
マニュアルとその他のリソース

この章には以下の情報が記載されています。

- 本書の表記法
- JMP のマニュアル
- JMP ヘルプ
- その他のリソース
  - その他の JMP のドキュメンテーション
  - チュートリアル
  - 索引
  - Web リソース
  - テクニカルサポートのオプション

表記規則

マニュアルの内容と画面に表示される情報を対応付けるために、次のような表記規則を使っています。

- サンプルデータ名、列名、パス名、ファイル名、ファイル拡張子、およびフォルダ名は「」で囲んで表記しています。
- スクリプトのコードは`Lucida Sans Typewriter`フォントで表記しています。
- スクリプトコードの結果（ログに表示されるもの）は*`Lucida Sans Typewriter`*（斜体）フォントで表記し、先に示すコードよりインデントされています。
- クリックまたは選択する項目は [] で囲んで太字で表記しています。これには以下の項目があります。
  - ボタン
  - チェックボックス
  - コマンド
  - 選択可能なリスト項目
  - メニュー
  - オプション
  - タブ名
  - テキストボックス
- 次の項目の表記規則は下記のとおりです。
  - 重要な単語や句、JMPに固有の定義を持つ単語や句は太字または「」で囲んで表記
  - マニュアルのタイトルは『』で囲んで表記
  - 変数名は斜体で表記
  - スクリプトの出力は斜体で表記
- JMP Proのみの機能にはJMP Proアイコン JMP PRO がついています。JMP Proの機能の概要については https://www.jmp.com/ja_jp/software/predictive-analytics-software.html をご覧ください。

メモ: 特別な情報および制限事項には、この文のように「メモ」という見出しがついています。

ヒント: 役に立つ情報には「ヒント」という見出しがついています。

JMP のマニュアル

JMP では、PDF 形式のマニュアルが用意されています。

- PDF 版は［**ヘルプ**］>［**ドキュメンテーション**］メニューまたは JMP オンラインヘルプのフッタから開くことができます。
- 検索しやすいようにすべてのドキュメンテーションが 1 つの PDF ファイルにまとめられた『JMP ドキュメンテーションライブラリ』と呼ばれるファイルがあります。『JMP ドキュメンテーションライブラリ』の PDF ファイルは［**ヘルプ**］>［**ドキュメンテーション**］メニューから開くことができます。

JMP ドキュメンテーションライブラリ

以下の表は、JMP ライブラリに含まれている各ドキュメンテーションの目的および内容をまとめたものです。

マニュアル	目的	内容
『はじめての JMP』	JMP をあまりご存知ない方を対象とした入門ガイド	JMP の紹介と、データを作成および分析し始めるための情報
『JMP の使用法』	JMP のデータテーブルと、基本操作を理解する	一般的な JMP の概念と、データの読み込み、列プロパティの変更、データの並べ替え、SAS への接続など、JMP 全体にわたる機能の説明
『基本的な統計分析』	このマニュアルを見ながら、基本的な分析を行う	［分析］メニューからアクセスできる以下のプラットフォームの説明： • 一変量の分布 • 二変量の関係 • 表の作成 • テキストエクスプローラ ［分析］>［二変量の関係］で二変量、一元配置分散分析、分割表に対する分析を実行する方法の説明。ブートストラップを使用した標本分布の近似方法やシミュレーションの機能を使用したパラメトリックな標本再抽出の実行方法の説明も含まれています。

マニュアル	目的	内容
『グラフ機能』	データに合った理想的なグラフを見つける	［グラフ］メニューからアクセスできる以下のプラットフォームの説明： • グラフビルダー • 重ね合わせプロット • 三次元散布図 • 等高線図 • バブルプロット • パラレルプロット • セルプロット • ツリーマップ • 散布図行列 • 三角図 • チャート このマニュアルには背景マップやカスタムマップの作成方法も記載されています。
『プロファイル機能』	対話式のプロファイルツールの使い方を学ぶ。任意の応答曲面の断面を表示できるようになります。	［グラフ］メニューに表示されるすべてのプロファイルについて。誤差因子の分析が、ランダム入力を使用したシミュレーションの実行とともに含まれています。
『実験計画(DOE)』	実験の計画方法と適切な標本サイズの決定方法を学ぶ	［実験計画（DOE）］メニューと［分析］>［発展的なモデル］メニューの「発展的な実験計画モデル」に関するすべてのトピックについて。

マニュアル	目的	内容
『基本的な回帰モデル』	「モデルのあてはめ」プラットフォームとその多くの手法について学ぶ	［分析］メニューの「モデルのあてはめ」プラットフォームで使用できる、以下の手法の説明： • 標準最小2乗 • ステップワイズ • 一般化回帰 • 混合モデル • MANOVA • 対数線形-分散 • 名義ロジスティック • 順序ロジスティック • 一般化線形モデル

マニュアル	目的	内容
『予測モデルおよび発展的なモデル』	さらなるモデリング手法について学ぶ	［分析］>［予測モデル］メニューで使用できる以下のプラットフォームの説明： • モデル化ユーティリティ • ニューラル • パーティション • ブートストラップ森 • ブースティングツリー • K近傍法 • 単純Bayes • モデルの比較 • 計算式デポ ［分析］>［発展的なモデル］メニューで使用できる以下のプラットフォームの説明： • 曲線のあてはめ • 非線形回帰 • Gauss過程 • 時系列分析 • 対応のあるペア ［分析］>［スクリーニング］メニューで使用できる以下のプラットフォームの説明： • 応答のスクリーニング • 工程のスクリーニング • 説明変数のスクリーニング • アソシエーション分析 ［分析］>［発展的なモデル］>［発展的な実験計画モデル］で使用できるプラットフォームについては、『実験計画(DOE)』に説明があります。

マニュアル	目的	内容
『多変量分析』	複数の変数を同時に分析するための手法について理解を深める	［分析］>［多変量］メニューで使用できる以下のプラットフォームの説明： • 多変量の相関 • 主成分分析 • 判別分析 • PLS ［分析］>［クラスター分析］メニューで使用できる以下のプラットフォームの説明： • 階層型クラスター分析 • K Means クラスター分析 • 正規混合 • 潜在クラス分析 • 変数のクラスタリング
『品質と工程』	工程を評価し、向上させるためのツールについて理解を深める	［分析］>［品質と工程］メニューで使用できる以下のプラットフォームの説明： • 管理図ビルダーと個々の管理図 • 測定システム分析 • 計量値/計数値ゲージチャート • 工程能力 • パレート図 • 特性要因図

マニュアル	目的	内容
『信頼性/生存時間分析』	製品やシステムにおける信頼性を評価し、向上させる方法、および人や製品の生存時間データを分析する方法について学ぶ	［分析］>［信頼性/生存時間分析］メニューで使用できる以下のプラットフォームの説明： • 寿命の一変量 • 寿命の二変量 • 累積損傷 • 再生モデルによる分析 • 劣化分析と破壊劣化 • 信頼性予測 • 信頼性成長 • 信頼性ブロック図 • 修理可能システムのシミュレーション • 生存時間分析 • 生存時間(パラメトリック)のあてはめ • 比例ハザードのあてはめ
『消費者調査』	消費者選好を調査し、その洞察を使用してより良い製品やサービスを作成するための方法を学ぶ	［分析］>［消費者調査］メニューで使用できる以下のプラットフォームの説明： • カテゴリカル • 多重対応分析 • 多次元尺度構成 • 因子分析 • 選択モデル • MaxDiff • アップリフト • 項目分析
『スクリプトガイド』	パワフルなJMPスクリプト言語（JSL）の活用方法について学ぶ	スクリプトの作成やデバッグ、データテーブルの操作、ディスプレイボックスの構築、JMPアプリケーションの作成など。
『スクリプト構文リファレンス』	JSL関数、その引数、およびオブジェクトやディスプレイボックスに送信するメッセージについて理解を深める	JSLコマンドの構文、例、および注意書き。

メモ：［**ドキュメンテーション**］メニューでは、印刷可能な 2 つのリファレンスカードも用意されています。『**メニューカード**』は JMP のメニューをまとめた表で、『**クイックリファレンス**』は JMP のショートカットキーをまとめた表です。

JMP ヘルプ

JMP ヘルプは、一連のマニュアルの簡易版です。JMP のヘルプは、次のいくつかの方法で開くことができます。

- Windows では、F1 キーを押すとヘルプシステムウィンドウが開きます。
- データテーブルまたはレポートウィンドウの特定の部分のヘルプを表示します。［**ツール**］メニューからヘルプツール を選択した後、データテーブルやレポートウィンドウの任意の位置でクリックすると、その部分に関するヘルプが表示されます。
- JMP ウィンドウ内で［**ヘルプ**］ボタンをクリックします。
- Windows の場合、［ヘルプ］メニューの［**ヘルプの目次**］、［**ヘルプの検索**］、［**ヘルプの索引**］の各オプションを使用して、JMP ヘルプ内を検索し、目的の内容を表示します。Mac の場合、［**ヘルプ**］>［**JMP ヘルプ**］を選択します。

JMP を習得するためのその他のリソース

JMP のマニュアルと JMP ヘルプの他、次のリソースも JMP の学習に役立ちます。

- チュートリアル（「チュートリアル」（23 ページ）を参照）
- サンプルデータ（「サンプルデータテーブル」（24 ページ）を参照）
- 索引（「統計用語と JSL 用語の習得」（24 ページ）を参照）
- 使い方ヒント（「JMP を使用するためのヒント」（24 ページ）を参照）
- Web リソース（「JMP User Community」（25 ページ）を参照）
- 専門誌『JMPer Cable』（「JMPer Cable」（25 ページ）を参照）
- JMP に関する書籍（「JMP 関連書籍」（25 ページ）を参照）
- JMP スターター（「「JMP スターター」ウィンドウ」（26 ページ）を参照）
- 教育用リソース（「サンプルデータテーブル」（24 ページ）を参照）

チュートリアル

［**ヘルプ**］>［**チュートリアル**］を選択して、JMP のチュートリアルを表示できます。［**チュートリアル**］メニューの最初の項目は［**チュートリアルディレクトリ**］です。この項目を選択すると、すべてのチュートリアルをカテゴリ別に整理した新しいウィンドウが開きます。

JMPに慣れていない方は、まず［**初心者用チュートリアル**］を試してみてください。JMPのインターフェースおよび基本的な使用方法を学ぶことができます。

他のチュートリアルでは、実験の計画、標本平均と定数の比較など、JMPの具体的な活用法を学習できます。

サンプルデータテーブル

JMPのマニュアルで取り上げる例は、すべてサンプルデータを使用しています。サンプルデータディレクトリを開くには、［**ヘルプ**］>［**サンプルデータライブラリ**］を選択します。

サンプルデータテーブルを文字コード順に並べた一覧を表示する、またはカテゴリごとにサンプルデータを表示するには、［**ヘルプ**］>［**サンプルデータ**］を選択します。

サンプルデータテーブルは次のディレクトリにインストールされています。

Windowsの場合: C:¥Program Files¥SAS¥JMP¥13¥Samples¥Data

Macintoshの場合: \Library\Application Support\JMP\13\Samples\Data

JMP Proでは、サンプルデータが（JMPではなく）JMPPROディレクトリにインストールされています。シングルユーザーライセンス版のJMP（JMP シュリンクラップ）では、サンプルデータがJMPSWディレクトリにインストールされています。

サンプルデータの使用例を参照するには、［**ヘルプ**］>［**サンプルデータ**］を選択し、教育用セクションから検索してください。教育用リソースについては、http://jmp.com/tools にも情報があります。

統計用語とJSL用語の習得

［**ヘルプ**］メニューには、次の索引が用意されています。

統計の索引　統計用語が説明されています。

スクリプトの索引　JSL関数、オブジェクト、ディスプレイボックスに関する情報を検索できます。スクリプトの索引からサンプルスクリプトを編集して実行することもできます。

JMPを使用するためのヒント

JMPを最初に起動すると、「使い方ヒント」ウィンドウが表示されます。このウィンドウには、JMPを使う上でのヒントが表示されます。

「使い方ヒント」ウィンドウを表示しないようにするには、［**起動時にヒントを表示する**］のチェックを外します。再表示するには、［**ヘルプ**］>［**使い方ヒント**］を選択します。または、「環境設定」ウィンドウで非表示に設定することもできます。詳細については、『JMPの使用法』を参照してください。

ツールヒント

次のような項目の上にカーソルを置くと、その項目を説明するツールヒントが表示されます。

- メニューまたはツールバーのオプション
- グラフ内のラベル
- レポートウィンドウ内の結果（テキスト）（カーソルで円を描くと表示される）
- 「ホームウィンドウ」内のファイル名またはウィンドウ名
- スクリプトエディタ内のコード

ヒント : Windows では、JMP 環境設定でツールヒントを表示しないよう設定できます。［**ファイル**］>［**環境設定**］>［**一般**］を選択し、［**メニューのヒントを表示**］の選択を解除します。このオプションは、Macintosh では使用できません。

JMP User Community

JMP User Community では、さまざまな方法で JMP をさらに習得したり、他の SAS ユーザとのコミュニケーションを図ったりできます。ラーニングライブラリには 1 ページガイド、チュートリアル、デモなどが用意されており、JMP を使い始める上でとても便利です。また、JMP のさまざまなトレーニングコースに登録して、自己教育を進めることも可能です。

その他のリソースとして、ディスカッションフォーラム、サンプルデータやスクリプトファイルの交換、Webcast セミナー、ソーシャルネットワークグループなども利用できます。

Web サイトの JMP リソースにアクセスするには、［**ヘルプ**］>［**JMP User Community**］を選択するか、https://community.jmp.com/ をご覧ください。

JMPer Cable

JMPer Cable は、JMP ユーザを対象とした年刊の専門誌です。JMPer Cable は次の JMP Web サイトで閲覧可能です。

http://www.jmp.com/about/newsletters/jmpercable/（英語）

JMP 関連書籍

JMP 関連書籍は、次の JMP Web ページで紹介されています。

https://www.jmp.com/ja_jp/academic/books-for-jmp-users.html

「JMP スターター」ウィンドウ

JMP またはデータ分析にあまり慣れていないユーザは、「JMP スターター」ウィンドウから開始するとよいでしょう。カテゴリ分けされた項目には説明がついており、ボタンをクリックするだけで該当の機能を起動できます。「JMP スターター」ウィンドウには、［分析］、［グラフ］、［テーブル］、および［ファイル］メニュー内の多くのオプションがあります。また、JMP Pro の機能やプラットフォームのリストも含まれています。

- 「JMP スターター」ウィンドウを開くには、［**表示**］（Macintosh では［**ウィンドウ**］）**>**［**JMP スターター**］を選択します。
- Windows で JMP の起動時に自動的に「JMP スターター」を表示するには、［**ファイル**］**>**［**環境設定**］**>**［**一般**］を選び、「開始時の JMP ウィンドウ」リストから［**JMP スターター**］を選択します。Macintosh では、［**JMP**］**>**［**環境設定**］**>**［**起動時に JMP スターターウィンドウを表示する**］を選択します。

テクニカルサポート

JMP のテクニカルサポートは、JMP のエンジニアが担当し、その多くは、統計学などの技術的な分野の知識を有しています。

http://www.jmp.com/japan/support には、テクニカルサポートへの連絡方法などが記載されています。

第 2 章

多変量分析について

多変量分析の概要

本書では、複数の変数をまとめて分析する以下の手法について取り上げます。

- 「多変量の相関」プラットフォームでは、複数の変数間における相関関係を調べます。第3章「多変量の相関」を参照してください。
- 「主成分分析」プラットフォームでは、複数の変数における変動をできるだけ説明する、少数の独立した線形結合（主成分）を求めます。主成分分析は探索的な手法であり、また、予測モデルを構築するときの手助けにもなります。第4章「主成分分析」を参照してください。
- 「判別分析」プラットフォームは、連続量の応答（Y）からカテゴリカルな分類変数（X）を予測する手法で、多変量分散分析（MANOVA）からの逆推定とみなすことができます。第5章「線形判別分析」を参照してください。
- 「PLS回帰」プラットフォームは、説明変数（X）の線形結合からなる因子に基づいて、応答変数（Y）を予測する線形モデルを構築します。PLS回帰は、XとYの関係を調べ、潜在的な因子を抽出します。第6章「PLS回帰」を参照してください。
- 「階層型クラスター分析」プラットフォームは、多変量データをもとに、値が近い行をグループにまとめます。クラスター分析は、データにおける塊を見つけ出すための探索的な統計手法です。第7章「階層型クラスター分析」を参照してください。
- 「K Means クラスター分析」プラットフォームも、多変量データをもとに、値が近い行をグループにまとめます。第8章「K Means クラスター分析」を参照してください。
- 「正規混合」プラットフォームは、重なりのある多変量正規分布のデータに対するクラスタリングです。第9章「正規混合分布法」を参照してください。
- 「潜在クラス分析」プラットフォームは、カテゴリカルなデータに対するクラスタリングです。このモデルでは、データが多項分布の混合分布に従っていると仮定します。第10章「潜在クラス分析」を参照してください。
- 「変数のクラスタリング」プラットフォームは、似通った変数をクラスターに分類します。この手法は、データの次元を減らすために使えます。この手法により、多変量データ全体の変動をうまく説明する、少数のクラスター成分や代表的な変数を探し出すことができます。すべての変数ではなく、それらのクラスター成分や代表的な変数を別のモデルで用いることが考えられます。第11章「変数のクラスタリング」を参照してください。

第 3 章

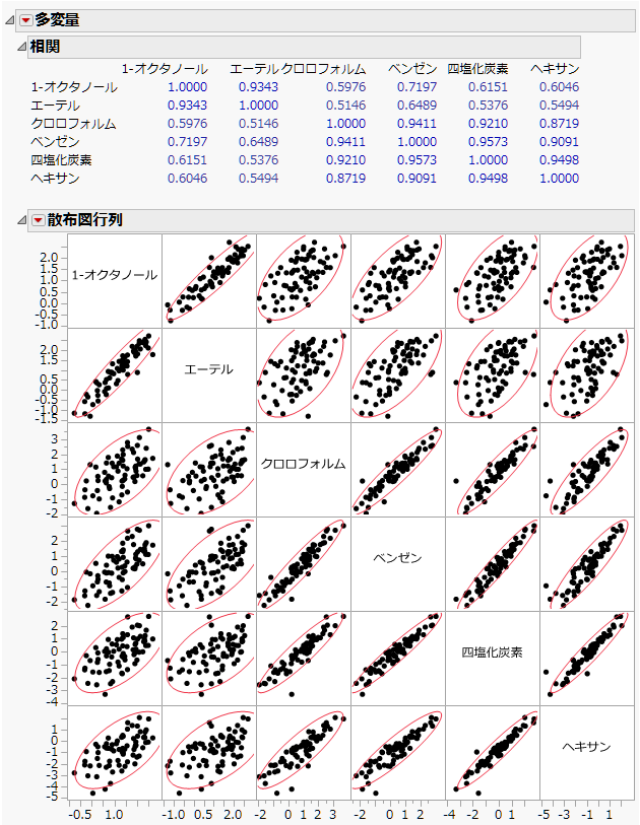
多変量の相関

多次元データの分布を検討する

「多変量の相関」プラットフォームでは、複数の変数間における相関関係を調べます。「多変量」とは、分析対象の変数が1つ（一変量）や2つ（二変量）ではなく、多数存在することを意味します。「多変量の相関」の結果から、次のようなことが分かります。

- 「相関」表では、応答変数（Y）のあいだに見られる線形関係の強さが分かります。
- 「散布図行列」表では、従属性、外れ値、クラスターを確認できます。
- これらのほかにも、偏相関、相関の逆行列、ペアごとの相関係数、共分散行列、主成分分析などによって、多変量の関係を調べることができます。

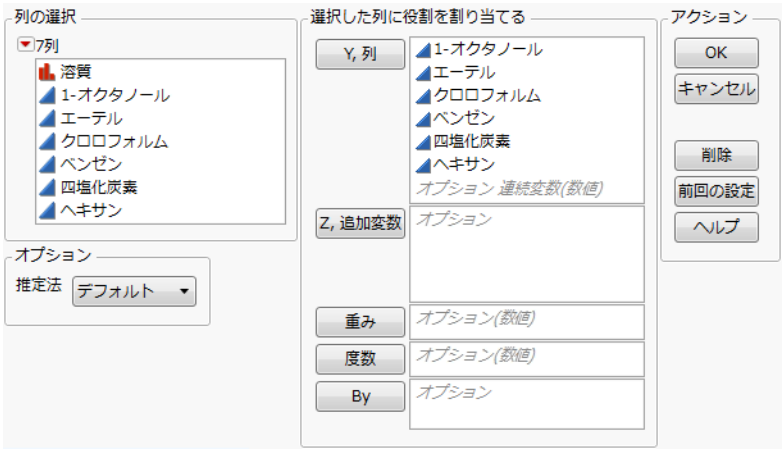
図3.1「多変量」レポートの例



「多変量の相関」プラットフォームの起動

「多変量の相関」プラットフォームを起動するには、［**分析**］>［**多変量**］>［**多変量の相関**］を選択します。

図 3.2「多変量の相関」起動ウィンドウ



Y, 列　応答変数（Y）の列を指定します。

重み　この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数　この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

By　By 変数の水準ごとに、個別に分析が行われます。

推定法　相関の計算方法を指定します。［**REML**］および［**ペアワイズ**］が、欠測があるデータから相関を計算するのに、よく使われている方法です。これらの一部は欠測値の処理に対応します。なお、［**欠測データの補完**］コマンドによって、推定した共分散行列に基づき欠測値を補完することもできます。「欠測データの補完」（42 ページ）を参照してください。

デフォルト　［**デフォルト**］オプションを選択した場合は、状況に応じて実際には［リストワイズ］、［ペアワイズ］、［REML］のいずれかが使用されます。

– 欠測値が 1 つもないデータテーブルの場合は、［**ローワイズ**］推定が使用されます。

– 欠測値があるデータテーブルで、列数が 10 列以上、行数が 5,000 行以上、または列数が行数より多い場合、［**ペアワイズ**］推定が使用されます。

– それ以外のデータテーブルには［**REML**］推定が使用されます。

REML　制限最尤法（REML）推定は、欠測値がある場合でも、すべてのデータを使用します。バイアス修正項の計算に時間がかかるため、データセットが大規模で欠測値が多い場合は、計算時間が長くなります。そのため、REML は小規模なデータセットに最も有効な手法です。データに欠測セルがない

場合、REML 法と ML 法の推定値は、通常の共分散行列の不偏推定値と同じです。欠測セルがある場合、REML 推定の分散と共分散推定値は、ML 推定の場合よりもバイアスが少なくなります。詳細は、「REML」（43 ページ）を参照してください。

最尤 最尤（ML）推定は、欠測値がある場合でも、すべてのデータを使用します。ML 推定の生成は速いので、欠測データのある大規模データテーブルでは、この方法が最も有効です。

ロバストな推定法 ロバスト推定法は、欠測値がある場合でも、すべてのデータを使用します。この方法は、極値をダウンウェイトするので、外れ値のあるデータテーブルでは有効です。統計量の詳細については、「ロバストな推定法」（43 ページ）を参照してください。

リストワイズ リストワイズ推定は欠測値のあるオブザベーションを使用しません。そのため、欠測セルのある行は、推定が行われる前に削除されます。この方法は欠測データのあるオブザベーションを除外するのに便利です。ローワイズ推定は JMP 8 以前では、使用できる唯一の推定法でした。従って、JMP 8 以前の JMP バージョンとの互換性を確認するのにも使用できます。

ペアワイズ ペアワイス推定は、欠測値がある場合でも、すべてのデータを使用します。この推定法は、列のペアごとに欠測値を考慮して、相関を計算します。データテーブルに欠測値があり、かつ、列が 10 列以上、行が 5000 行以上、または列数が行数より多い場合に最も効果的です。

メモ：［REML］、［ML］または［ロバスト］を選択した場合、データテーブルの列数が行数より多く、欠測値がある場合は、「推定法」が［ペアワイズ］に変わります。

配置の方法 散布図行列の配置のオプションを選択します。［正方形］オプションは、指定された列の順番で、すべての列のペアに対する散布図を表示します。［下三角］は、最初の n - 1 個の列を横軸に表示し、対角線の下側に散布図を表示します。［上三角］は、最初の n - 1 個の列を縦軸に表示し、対角線の上側に散布図を表示します。

「多変量」レポート

デフォルトの「多変量」レポートには、相関行列と散布図行列が表示されます。プラットフォームに用意されているコマンドによって、いくつかの分析を追加できます。「「多変量の相関」プラットフォームのオプション」（33 ページ）を参照してください。

図 3.3「多変量」レポートの例

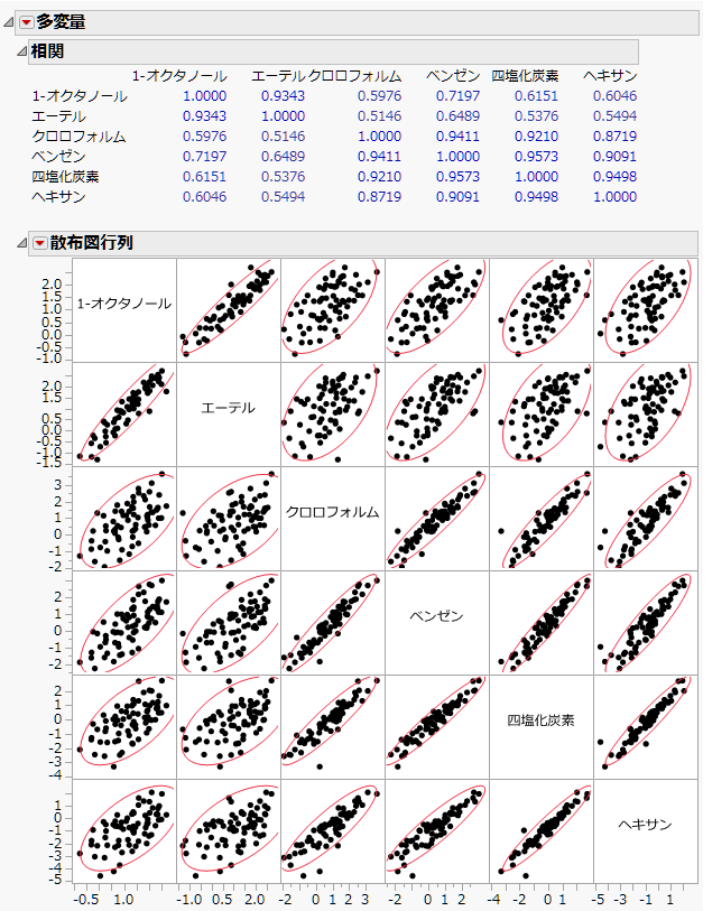


図 3.3 の作成方法

1. ［ヘルプ］ > ［サンプルデータライブラリ］を選択し、「Solubility.jmp」を開きます。
2. ［分析］ > ［多変量］ > ［多変量の相関］を選択します。
3. 「溶質」以外のすべての列を選択し、［Y, 列］をクリックします。
4. ［OK］をクリックします。

欠測値について

ほとんどのコマンドは、選択された推定法の結果に基づき、計算を行います。ただし、［ペアごとの相関係数］コマンドでは、変数のペアのうちのいずれかが欠測値である行が計算から除外されます。また、［基本統計量］ > ［単変量の基本統計量］では、平均などの単変量統計量が、1 列ごとに別々に計算され、他の列に含まれる欠測値は考慮されません。

「多変量の相関」プラットフォームのオプション

相関係数行列	「相関」表の表示／非表示を切り替えます。相関表は、相関係数を行列にしたもので、応答変数（Y）の各組み合わせに見られる線形関係の強さを示します。このオプションはデフォルトではオンになっています。「Pearsonの積率相関」（44ページ）を参照してください。 この相関行列は、起動ダイアログで指定した手法で計算されています。
相関のp値	*p*値の行列である「相関のp値」レポートを表示します。各*p*値は、「2つの変数間における母相関係数が0である」という帰無仮説の検定に該当します。この検定は、「2つの変数間には線形関係がない」という帰無仮説に対する検定にもなっています。この検定は、Pearsonの積率相関係数に対するものです。
相関の信頼区間	相関の両側信頼区間を示します。このオプションはデフォルトではオフになっています。 デフォルトの信頼係数は95%ですが、［**α水準の設定**］オプションを使えば値を変更できます。
相関係数の逆行列	相関行列の逆行列の表示／非表示を切り替えます。このオプションはデフォルトではオフになっています。 相関行列の逆行列の対角要素は、該当する変数が他の変数の線形関数によってどれぐらい説明されるかを表します。この対角要素は、他のすべての変数を説明変数とした回帰分析の結果から、$1/(1-R^2)$によって計算されます。重相関が0のときは、対角要素は1となります。重相関が1のときは、対角要素は無限大になります（このときは、レポートでは欠測値になります）。 相関行列の逆行列についての統計的詳細は、「相関の逆行列」（46ページ）を参照してください。
偏相関係数行列	「偏相関」表の表示／非表示を切り替えます。「偏相関」表には、2変数の関係を他のすべての変数の効果で調整した偏相関が表示されます。このオプションはデフォルトではオフになっています。 偏相関表は、相関の逆行列の対角要素が1になるように尺度化し、符号を逆にすることにより計算できます。
共分散行列	共分散行列の表示／非表示を切り替えます。共分散行列は、2変数が一緒に変化する程度を表します。このオプションはデフォルトではオフになっています。

ペアごとの相関係数

「ペアごとの相関係数」表の表示/非表示を切り替えます。この表には、ペアワイズ法によって計算されたPearsonの積率相関係数がリストされます。このオプションはデフォルトではオフになっています。

Y変数のペアごとに計算が行われており、ペアになっている変数のいずれかに欠測値がある場合は、それらは計算から除外されます。相関係数に対する有意確率も表示されます。また、相関の大きさが棒グラフで表されます。すべての結果はペアワイズ法に基づきます。

HotellingのT2乗検定

Y変数の多変量分布における平均に対し、1標本検定を実行します。各変数の仮説平均を入力するためのウィンドウが開きます。そのウィンドウで帰無仮説における平均ベクトルを指定します。この検定は、Y変数が多変量正規分布に従うという仮定のもとで導出されています。

「Hotellingの*T*2乗検定」レポートには、次のものが表示されます。

変数 Yとして入力した変数のリスト。

平均 各変数の標本平均。

仮説平均 分析者が指定した、帰無仮説における平均。

検定統計量 Hotellingの*T*2乗の統計量の値。

F値 検定統計量の値。*n*個の行と*k*個の変数がある場合、*F*値は次のように求められます。

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

p値(Prob>F) 検定の*p*値。帰無仮説の下では、自由度が*n*および*n-k*のF分布に*F*値は従います。

基本統計量	このメニューには2つのオプションがあり、それぞれが各列の基本統計量（平均、標準偏差など）の表示／非表示を切り替えます。単変量および多変量の基本統計量は、欠測値がある場合やロバスト法が使用されているときに異なる場合があります。 **単変量の基本統計量**　他の列の値とは無関係に、各列で計算された統計量が表示されます。これらの値は、「一変量の分布」プラットフォームで算出される値と一致します。 **多変量の基本統計量**　起動ウィンドウで選択した推定法に従って計算された統計量が表示されます。［**REML**］、［**最尤**］、［**ロバスト**］のいずれかを選択した場合、それらの方法で推定された平均と分散が表示されます。［**リストワイズ**］を選択した場合は、1つでも欠測値のある行は除外した上で、平均と分散が計算されます。［**ペアワイズ**］を選択した場合は、平均と分散が各列から計算されます。 これらのオプションはデフォルトではオフになっています。
ノンパラメトリック相関係数	このメニューには、［Spearmanの順位相関係数(ρ)］［Kendallの順位相関係数(τ)］、［HoeffdingのD統計量］という3種類のノンパラメトリックな指標があります。これらのオプションはデフォルトではオフになっています。 詳細については、「ノンパラメトリック相関係数」（37ページ）を参照してください。
α水準の設定	相関の信頼区間のα水準（有意水準）を変更できます。 サブメニューには、［**0.01**］、［**0.05**］、［**0.10**］、［**0.50**］の4つの値が表示されています。それ以外の値を入力したい場合は［**その他**］を選択してください。
散布図行列	散布図行列の表示/非表示を切り替えます。散布図行列とは、応答変数の各ペアに対する散布図を行列形式で表したものです。このオプションはデフォルトではオンになっています。 詳細については、「散布図行列」（38ページ）を参照してください。

カラーマップ	［**カラーマップ**］メニューには3種類のカラーマップがあります。 **相関のカラーマップ**　-1～+1の相関係数を、青色～赤色で表したセルプロットが作成されます。 **p値のカラーマップ**　0～1の相関の有意度を、赤色～青色で表したセルプロットが作成されます。 **相関のクラスタリング**　似た相関を持つ変数をまとめたセルプロットが作成されます。相関は「相関のカラーマップ」と変わりませんが、変数の位置が異なる場合があります。 これらのオプションはデフォルトではオフになっています。
パラレルプロット	パラレルプロットの表示／非表示を切り替えます。このオプションはデフォルトではオフになっています。
三次元楕円プロット	95% 確率楕円体を表示した3次元散布図の表示／非表示を切り替えます。このオプションを選択すると、3つの変数を指定するためのダイアログが表示されます。このオプションはデフォルトではオフになっています。
主成分分析	このメニューには、主成分分析レポートの表示/非表示を切り替えるオプションが含まれています。相関行列、共分散行列、または原データの積和行列を選択できます。いずれかのレポートが表示されているときに別のオプションを選択すると、レポートが新しく選択したものに置き換わります。レポートを表示しない場合は［**なし**］を選択します。このオプションはデフォルトではオフになっています。 **主成分分析**は、元の変数の一次結合で合成変数を求める方法です。第1主成分は最大の分散を持つ一次結合、第2主成分は第1主成分に直交する一次結合のなかで最大の分散を持つもの、というように主成分が求められます。詳細は「主成分分析」（49ページ）章を参照してください。
外れ値分析	このメニューに含まれるオプションは、Mahalanobisの距離、ジャックナイフ法による距離、または T^2 のいずれかの手法を使って、変数間の相関を考慮して計算された中心からの距離をプロットしたグラフの表示/非表示を切り替えます。 詳細については、「外れ値分析」（40ページ）を参照してください。

項目の信頼性	このメニューのオプションは、「項目の信頼性」レポートの表示／非表示を切り替えます。「項目の信頼性」レポートは、複数の測定項目（変数の組）において、どれほど一貫した測定ができているかを示します。Cronbach の α 係数と、標準化した変数に対する Cronbach の α 係数の 2 種類が用意されています。これらのオプションはデフォルトではオフになっています。 詳細については、「項目の信頼性」（41 ページ）を参照してください。
欠測データの補完	元のデータテーブルにおける欠測値を推定値に置き換えて、データテーブルを新規に作成します。このオプションは、データテーブルに欠測値がある場合にのみ使用できます。 詳細については、「欠測データの補完」（42 ページ）を参照してください。
欠測データ補完の計算式を保存	欠測値を含む列に関して、欠測値の推定に使用した計算式を、データテーブル内の新しい列に保存します。新しい列には、「**補完された _<列名>**」という名前がつきます。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ　現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存　By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

ノンパラメトリック相関係数

［**ノンパラメトリック相関係数**］のサブメニューには 3 種類のノンパラメトリックな指標があります。各レポートには、関連の指標の有意確率が表示されます。また、相関係数が棒グラフで表示されます。

Spearman の順位相関係数（ρ）　データ値そのものではなく、データ値を順位に置き換えて計算した相関係数です。

Kendall の順位相関係数（τ） 大小関係が一致するペアと一致しないペアの個数から計算された相関係数です。データから 2 行を取り出して比較したときに、X が大きい方が、Y も大きくなっているペアを、**大小関係が一致するペア**といいます。逆に、X が大きい方が、Y が小さくなっているペアを、**大小関係が一致しないペア**といいます。なお同順位のペア（X または Y の値が等しいペア）がある場合、それらの影響も考慮して、修正された値が計算されます。

Hoeffding の D 統計量 –0.5 ～ 1 の値を取る指標です。値が正で大きいほど従属性が高いことを示します。この統計量は、2×2 の分割表から計算されたカイ 2 乗統計量の重み付き和の近似値です。2×2 の分割表は、各データ値を閾値として作成されます。この指標は、独立性からの全体的な逸脱を検出します。

メモ: ノンパラメトリックな相関係数は、起動ウィンドウで別の推定法を指定した場合でも、常にペアワイズ法によって計算されます。

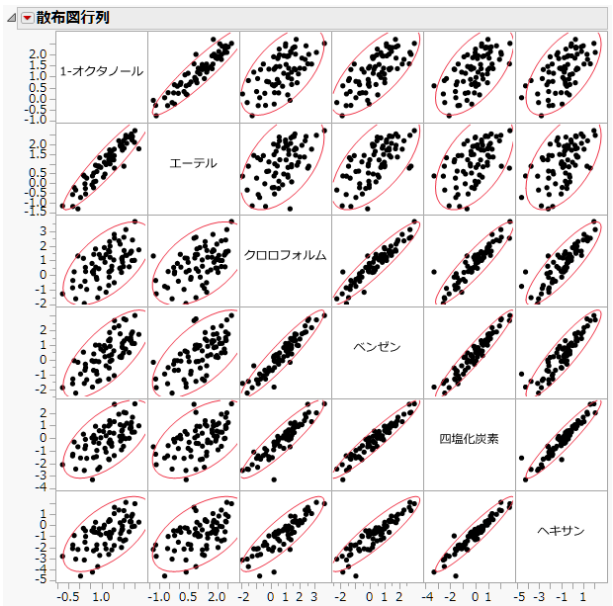
メモ: ［重み］変数を指定した場合、欠測値および重みが 0 である行は、ノンパラメトリックな相関係数の計算から除外されます。その他の重みの値はすべて 1 として扱われます。

これら 3 種類の相関係数に関する詳細については、「ノンパラメトリックな相関」(44 ページ)を参照してください。

散布図行列

散布図行列は、応答変数の各組み合わせに見られる相関を視覚的に確認するのに役立ちます。散布図行列はデフォルトで表示されますが、「多変量」の赤い三角ボタンのメニューにある［**散布図行列**］を使って表示/非表示を切り替えることができます。

図 3.4 散布図行列



デフォルトでは、各散布図の上に95%の二変量正規確率楕円が描かれます。変数の各ペアが二変量正規分布に従っていると仮定したとき、約95%の点がこの楕円形の中に含まれます。楕円形の幅は変数間の相関の度合に対応します。楕円形が円に近く、対角線方向に延びていない場合は、変数に相関関係がありません。楕円形が細く、対角線方向に延びている場合は、変数に相関関係があります。

散布図行列の操作

1つの散布図のサイズを変更すると、すべての散布図のサイズが変更されます。

対角線上の列名が書かれているセルを、別の列名のセルにドラッグすると、散布図行列の配列が変わります。

変数の相関がグループに分かれている場合には、その状況を散布図行列で見ることができます。たとえば、図3.4では、相関が2つのグループに分かれています。最初の2変数（左上）と、次の4変数（右下）に相関が分かれています。

散布図行列のオプション

「散布図行列」の赤い三角ボタンのメニューには、確率楕円を表示するオプション、確率楕円に色をつけるオプション、確率楕円の信頼水準を設定するオプションがあります。

表3.1 散布図行列のオプション

点の表示	散布図内の点の表示／非表示を切り替えます。
直線のあてはめ	回帰直線と、それに対する95%信頼区間の曲線の表示／非表示を切り替えます。
確率楕円	散布図内で95%の確率楕円の表示／非表示を切り替えます。信頼水準を変更するには、[**楕円のα**]メニューを使用します。
楕円内を塗る	すべての楕円内を塗ります。[**楕円の透明度**]と[**楕円の色**]を使って塗りの透明度と色を変更できます。
相関の表示	各散布図の左上隅に表示される相関係数の表示／非表示を切り替えます。
ヒストグラムの表示	列名が表示されているセルのX軸上またはY軸上に、ヒストグラムを表示します。ヒストグラムを表示した後、[**度数の表示**]を選択すると、ヒストグラムの各棒に度数が表示されます。[**X軸上**]または[**Y軸上**]を選択することで、ヒストグラムの向きを変えたり、ヒストグラムを削除したりできます。
楕円のα	楕円のα水準を設定します。標準的なα水準のどれかを選択するか、[**その他**]を選択して別の値を入力します。
楕円の透明度	楕円内が塗られている場合に、楕円の透明度を設定します。あらかじめ用意されている値のどれか1つを選択するか、[**その他**]を選択して別の値を入力します。デフォルトの値は0.2です。
楕円の色	楕円内が塗られている場合に、楕円の色を設定します。パレット内の色を選択するか、[**その他**]を選択して別の色を指定します。デフォルトの値は赤です。

表 3.1 散布図行列のオプション（続き）

ノンパラメトリック密度	密度等高線の表示／非表示を切り替えます。この密度等高線は、二変量におけるノンパラメトリックな密度を示す滑らかな曲面の等高線です。ノンパラメトリックな曲面の 10% と 50% の分位点の等高線が表示されます。

外れ値分析

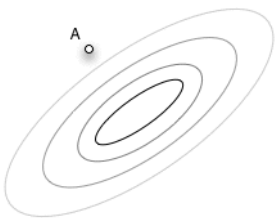
［**外れ値分析**］メニューのオプションは、次のいずれかの手法を使い、中心からの距離をプロットしたグラフの表示／非表示を切り替えます。

- Mahalanobis の距離
- ジャックナイフ法による距離
- T^2 統計量

これらのプロットは、中心からの距離を、変数間の相関を考慮して測定します。検定は、プロットの下部に表示される有意水準（α）で行われます。

図 3.5 において、A 点は外れ値ですが、その理由は、いずれかの座標において外れているからではなく、相関構造を考慮したときに外れているからです。

図 3.5 外れ値の例



Mahalanobis の距離

Mahalanobis の距離の外れ値プロットには、各点から平均ベクトル（重心）までの Mahalanobis の距離が表示されます。Mahalanobis の距離は、データの平均、標準偏差、相関によって変化します。オブザベーション番号に従って距離がプロットされるため、距離が最大の点を強調表示すれば、どのオブザベーションが極端な外れ値なのかがわかります。詳細は、「Mahalanobis の距離の計算」（46 ページ）を参照してください。

ジャックナイフ法による距離

「ジャックナイフ法による距離」の外れ値プロットには、ジャックナイフ法で計算されたMahalanobisの距離がプロットされます。この距離は、該当するオブザベーションを除いたときの平均、標準偏差、相関行列を使って計算されます。データに外れ値がある場合は、ジャックナイフ法が役立ちます。外れ値がある場合、通常のMahalanobisの距離は外れ値に影響され、外れ値を識別しにくくなったり、通常の点が実際よりも離れた位置にあるように見えてしまいます。詳細は、「ジャックナイフ法による距離の計算」（47ページ）を参照してください。

T^2

T^2プロットの距離は、Mahalanobisの距離の2乗で、多変量管理図に広く使用されています。プロットには、T^2統計量の計算値と上側管理限界が表示されます。この限界の外にある値は、外れ値の可能性があります。詳細は、「T2 乗の距離の計算」（47ページ）を参照してください。

距離と値の保存

プロットの赤い三角ボタンのメニューから［**保存**］オプションを選択すると、距離をデータテーブルに保存できます。

メモ：その際、計算式は保存されないため、データテーブルに変更を加えても距離の再計算は**行われません**。データテーブル内で、列の追加／削除や値の変更を行ったときは、［**分析**］>［**多変量**］>［**多変量の相関**］を選択して、もう一度計算をやり直してください。

各行の距離を保存するほか、［外れ値分析］で指定された上側管理限界（UCL）の値を保持する列プロパティを作成します。

項目の信頼性

項目の信頼性は、複数の測定項目（変数の組）において、どれほど一貫した測定ができているかを示します。信頼性の指標の1つに、Cronbachのα係数（Cronbach 1951）があります。Cronbachのα係数の主な用途として、工業分野における測定の信頼性分析と、質問票調査の分析が挙げられます。

Cronbachのα係数は、項目間にある相関の平均的な値を示します。Cronbachのα係数は、項目を2等分した時の相関（split half correlation）をすべての組み合わせで求めて、その平均を計算するのと等価です。各項目の分散が異なる場合のために、標準化したデータに対するα係数を求めることもできます。

メモ：Cronbachのα係数は、有意水準のαとは無関係です。また、「項目の信頼性」は、「生存時間や寿命の信頼性分析」と無関係です。

個別の項目の影響を調べるため、JMPでは、各項目を計算から除外したときのCronbachのα係数が表示されます。ある変数（項目）を除外したときにα係数が増加した場合、その変数と他の変数の間には強い相関はありません。α係数が減少した場合は、その変数が他の変数と相関していると結論できます。Nunnally（1979）は、Cronbachのα係数の経験則上受け入れられる下限として、0.7という値を提案しています。

計算方法の詳細については、「Cronbach のα係数」（48 ページ）を参照してください。

欠測データの補完

欠測データを補完するには、「多変量」の赤い三角ボタンのメニューから［**欠測データの補完**］を選択します。元のデータテーブルにおける欠測値を推定値に置き換えて、データテーブルを新規に作成します。

各行の非欠測値から計算された条件付き期待値によって、欠測値は補完されます。起動ウィンドウで指定した推定方法によって平均と共分散行列が推定され、それによって条件付き期待値が算出されます。欠測値が補完されたデータセットを用いれば、あらゆる種類の多変量検定や多変量分析を実行できます。

このオプションは、データテーブルに欠測値がある場合にのみ使用できます。

項目の信頼性の例

ここでは、サンプルデータのフォルダにある「Danger.jmp」データを使用します。このデータには、ある程度の危険性がある行為や物が、30 項目だけリストアップされています。異なった立場の 3 名（「学生」、「社会人（一般）」、「専門家」）に、危険性が高いと思う順に項目を並べ、順位をつけてもらいました。学生と社会人が、「非常に危険（1 位）」と思っている「原子力」に対して、専門家は 20 位と評価しています。一方、「オートバイ」はどのグループでも 5 位や 6 位となっています。

Cronbach のα係数を使うと、3 名の評価にどれぐらい一貫性があるかを知ることができます。この例のように、値がどのグループでも同じ（同じ 1 位～30 位のセット）場合は、データを標準化しても効果はありません。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Danger.jmp」を開きます。
2. ［**分析**］>［**多変量**］>［**多変量の相関**］を選択します。
3. **「行為／物」**以外のすべての列を選択し、［**Y, 列**］をクリックします。
4. ［**OK**］をクリックします。
5. 「多変量」の赤い三角ボタンをクリックし、開いたメニューから［**項目の信頼性**］>［**Cronbach のα係数**］を選択します。
6. （オプション）「多変量」の赤い三角ボタンをクリックし、開いたメニューから［**散布図行列**］を選択して散布図行列のプロットを非表示にします。

図 3.6「Cronbach の α 係数」レポート

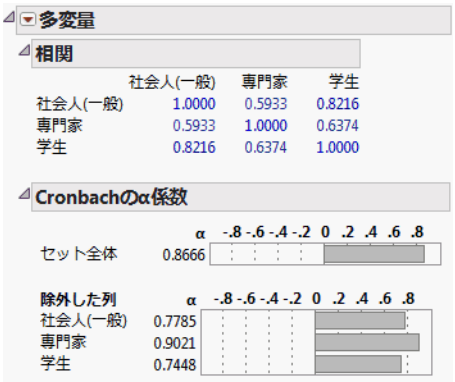


図 3.6 の結果から、全体の α 係数は 0.8666 で、3 名の評価には高い一貫性があることがわかります。さらに、「専門家」を除いた場合の、**「学生」**と**「社会人（一般）」**の Cronbach の α 係数は 0.9021 であり、この 2 名がほぼ同じ評価をしていることがわかります。

計算方法と統計的詳細

推定法について

REML

データに欠測値がある場合、REML 法（制限最尤法）は、最尤法に比べて、推定値のバイアスが小さいのが特徴です。REML 法は、誤差対比（error contrast）から導出された周辺尤度を最大化する推定方法です。REML 法は、分散および共分散を推定するのによく使われます。［多変量の相関］の［REML］は、反復測定データの相関構造に無構造（unstructured）を仮定した混合モデルの REML 推定と同じです。混合モデルについては、SAS システムの PROC MIXED に関するドキュメントを参照してください。

ロバストな推定法

この手法では、外れ値に対して小さな重みを与えることで、外れ値が実質的に無視されます。推定において、次式により重みを反復的に計算します。

$Q < K$ の場合は $w_i = 1.0$、そうでない場合は $w_i = K/Q$

ここで、K は、データの列数を自由度としたカイ 2 乗分布の 75% 点です。また、Q は次式により計算されます。

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

ここで、y_i は第 i オブザベーションの応答、μ は平均ベクトルの現在の推定値、S^2 は共分散行列の現在の

推定値です。また、Tは行列の転置を意味します。なお、各反復の最後において、分散共分散行列の推定値が不偏になるように調整が行われます。

ロバスト推定は、データに外れ値があまりない場合には、通常の推定方法に比べ、推定のばらつきが大きくなります。しかし、データに外れ値がある場合には、通常の方法に比べて、精度が高い推定値が得られます。

Pearsonの積率相関

Pearsonの積率相関係数は、二変量間の線形関係の強さを表します。応答変数をXとYとしたとき、この相関係数rは、次のように計算されます。

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

二変量間に完全な線形関係がある場合、相関係数は1（正の相関）または-1（負の相関）になり、線形関係がない場合は、0に近くなります。

ノンパラメトリックな相関

Spearman、Kendall、Hoeffdingの相関係数では、データにまず順位がつけられ、その順位に対して計算が行われます。値が同じデータには（同順位のデータには）、平均順位を使用します。

メモ: ［重み］変数を指定した場合、欠測値および重みが0である行は、ノンパラメトリックな相関係数の計算から除外されます。その他の重みの値はすべて1として扱われます。

Spearmanの順位相関係数（ρ）

Spearmanの順位相関係数（ρ）は、前述したPearsonの相関係数の計算式に、データの順位を代入して計算されます。

Kendallの順位相関係数（τ_b）

Kendallの順位相関係数（τ_b）は、大小関係の一致したペアと一致しないペアの数に基づいて計算されます。データから行のペアを取り出し、大小関係が両方の変数で一致しているとき、そのペアは**一致している**（concordant）といいます。大小関係が一致しているペア、一致しないペア、および、同順位のペアの個数から計算は行われます。

Kendallの順位相関係数（τ_b）は、次式により求められます。

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

この式における各記号は、次のとおりです。

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

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

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

この式で使われている記号の意味は、以下の通りです。

- sgn(z) は、z>0 のとき 1、z=0 のとき 0、z<0 のとき -1 です。
- t_i（u_i）は x（または y）が同順位である i 番目のグループにおけるオブザベーション数です。
- n は全部のオブザベーション数です。
- Kendall の順位相関係数（τ_b）は、-1 ～ 1 の間にあります。重み変数は、指定してあっても無視されます。

計算は次のように行われます。

- 第 1 変数の値に従って、オブザベーションの順位を求めます。
- 次に、第 2 変数の値に従って、オブザベーションの順位を求めます。
- 順位が変わった数をもとに、Kendall の順位相関係数（τ_b）が計算されます。

Hoeffding の D 統計量

Hoeffding の D 統計量（1948）は、次のような式で計算されます。

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

ここで

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

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

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

この式で使われている記号の意味は、以下の通りです。

- R_i と S_i は、x と y の順位です。
- Q_i（二変量順位とも呼ぶ）は、x と y の両方の値が i 番目の点より小さい点の個数を数え、それに 1 を足したものです。
- x 値または y 値のどちらか一方だけで同順位である点の場合、もう片方の値が i 番目の点より小さいときに、Q_i に 1/2 だけ寄与します。x と y の両方が同順位である点は、Q_i に 1/4 だけ寄与します。

オブザベーションの中に同順位がない場合、D 統計量の値は -0.5 ～ 1 で、1 の値は完全な従属を示します。重み変数は、指定してあっても無視されます。

相関の逆行列

相関行列の逆行列からは、多変量に関する有益な情報が得られます。相関行列の逆行列における対角要素は、VIF（Variance Inflation Factors）と呼ばれる統計量と等しく、該当する変数がその他の変数の線形式によってどれほど説明されるかを表します。相関行列を**R**、その逆行列を$\mathbf{R}^{-1}$とすると、その対角要素（r^{ii}）は次式で計算されます。

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

$\mathbf{R}_i^2$は、i番目の変数を応答変数とし、残りの変数を説明変数とした回帰モデルの寄与率（決定係数）です。そのため、r^{ii}が大きいときは、i番目の変数が残りの変数と強く相関していることを意味します。

距離の計算

外れ値のプロットには、各点からの距離が表示されます。

Mahalanobisの距離の計算

Mahalanobisの距離では、データの相関構造と個々の変数の分散が考慮されます。各変数のMahalanobisの距離M_iは、次のような式で計算されます。

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

ここで

Y_iは第i行のデータ

$\overline{\mathrm{Y}}$は平均ベクトル

Sは分散共分散行列

Mahalanobisの距離プロットに表示された上側管理限界の参照線（Mason and Young, 2002）は、次のように計算されます。

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

ここで

n = オブザベーションの数

p = 変数（列）の数

$\beta_{\left[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}\right]}$ = $\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$ をパラメータとするベータ分布の（1–α）分位点

ある変数がそれ以外の変数の一次結合になっている場合、相関行列は特異行列になります。その場合、相関行列におけるその変数の行と列は計算から除外されます。そのような変数は、Mahalanobis の距離には寄与しません。一般化逆行列を使えば、一次結合がある場合も Mahalanobis の距離は計算できます。

ジャックナイフ法による距離の計算

ジャックナイフ法による距離は、そのオブザベーションを除いたときの平均、標準偏差、相関行列の推定値を使って計算されます。各変数のジャックナイフ法による距離は、次のように計算されます。

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

ここで

n = オブザベーションの数

p = 変数（列）の数

M_i = i 番目のオブザベーションの Mahalanobis の距離

ジャックナイフ法による距離プロットに表示された上側管理限界の参照線（Penny, 1996）は、次のように計算されます。

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

T2 乗の距離の計算

T^2 の距離は、Mahalanobis の距離の 2 乗なので、$T_i^2 = M_i^2$ という式で計算されます。

T^2 の上側管理限界の計算式は次のとおりです。

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

ここで

n = オブザベーションの数

p = 変数（列）の数

$\beta_{\left[1-\alpha;\frac{p}{2},\frac{n-p-1}{2}\right]}$ = $\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$ をパラメータとするベータ分布の（1–α）分位点

多変量の距離を使うと、多次元で外れ値を見分けることができます。なお、変数の間に強い相関があるときは、部分空間では普通に見える点でも、多変量空間全体では外れ値とみなされることがあります。言い換えれば、値が相関関係にあるときは、1次元や2次元上で見ただけでは外れ値のように見えなくても、相関構造を考慮して多次元で見ると外れ値であることがあります。

Cronbach の α 係数

Cronbach の α 係数は、次の式で定義されています。

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

ここで

k = 測定に使われている項目の数

c = 項目間の共分散の平均値

v = 項目間の分散の平均値

全項目が標準化されていて、分散が1なら、次のようになります。

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

r は、項目間の相関の平均値です。

全体の α 係数が大きければ、その項目から作られる尺度やテストの信頼性が高いことを示唆しています。強い相関関係にある項目が多いと、α 係数は1.0に近くなります。

第 4 章

主成分分析
多変量データの次元削減

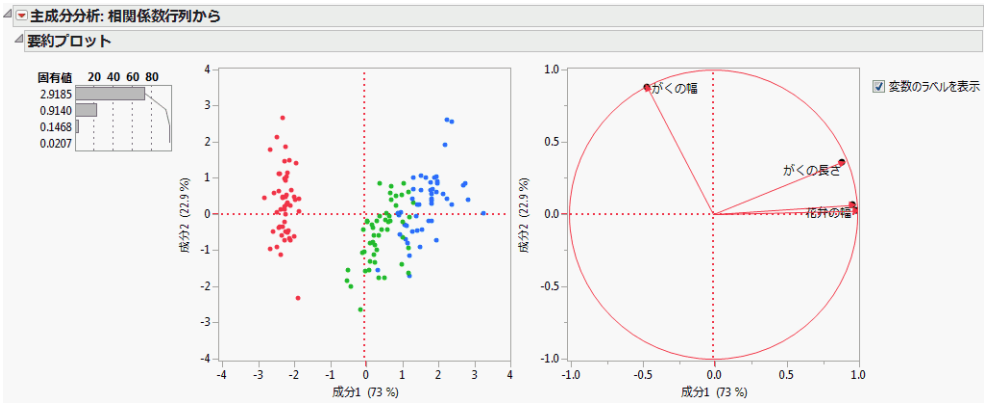
主成分分析（PCA; Principal Component Analysis）は、複数の変数における変動をできるだけ説明する、少数の独立した線形結合（**主成分**）を求めます。主成分分析は次元を削減する手法であり、探索的なデータ分析ツールの1つです。**主成分回帰**（PCA 回帰または PCR ともいう）によって予測モデルを作成するときにも使えます。

「主成分分析」プラットフォームには、変数の数が非常に多いデータ用に、「横長」（wide）と呼ばれる手法が用意されています。「横長」手法では、主成分を短時間で算出できます。これらの手法で求められた主成分は、主成分回帰に流用できます。

多くの0があるデータ（**疎データ**）用には、「疎」推定法が用意されています。「横長」手法と同様に、「疎」手法でも主成分を短時間で算出できます。ただし、「横長」手法とは異なり、「疎」手法では全体ではなくユーザが定義した数の主成分だけが算出されます。

また、「主成分分析」プラットフォームでは因子分析も行うことができます。JMP には、抽出した因子を解釈しやすくするために、いくつかの直交回転や斜交回転が用意されています。因子分析の詳細については、『消費者調査』の「因子分析」章を参照してください。

図 4.1 主成分分析の例



主成分分析の概要

主成分分析は、多変量データの変動を、なるべく少数の成分（**主成分**）で説明しようとする分析です。

主成分分析は、相関が高い多変量データの分布を調べるのに役立ちます。主成分分析は、多変量データにおいて最も変動が大きくなっている方向を抽出します。主成分分析を適用すれば、より少ない次元の成分によって、元の多変量データがもつ変動を把握することができます。主成分分析は、できるだけ少数の主成分によって、元データの変動をなるべく多く説明しようとする手法です。

変数が p 個あるとき、次のようにして p 個の主成分が形成されます。

- 第1主成分は、標準化された元の変数の線形結合のなかで、最大の分散を持つものです。
- それに続く主成分は、変数の線形結合のなかで、すでに定義されている主成分との相関がないもののうち、最大の分散を持つものです。

相関行列（もしくは、共分散行列、原データの積和行列）の固有ベクトルが、上記のような線形結合の係数になります。また、固有値が、各主成分の分散と等しくなります。

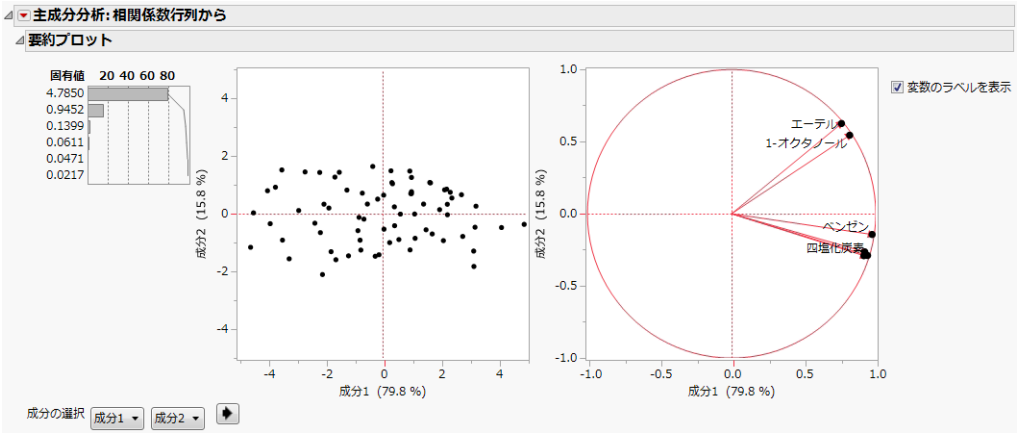
「主成分分析」プラットフォームの主成分分析では、相関行列・共分散行列・積和行列を分析対象とします。なお、このプラットフォームでは、因子分析も実行できます。詳細については、『消費者調査』の「因子分析」章を参照してください。

主成分分析の例

例として、2成分によってほぼ説明されるサンプルデータの「主成分分析」レポートを確認してみましょう。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Solubility.jmp」を開きます。
2. ［**分析**］>［**多変量**］>［**主成分分析**］を選択します。
3. 連続尺度の列をすべて選択し、［**Y, 列**］を選択します。
4. 「推定法」は［デフォルト］のままで、［**OK**］をクリックします。

図 4.2「主成分分析 : 相関係数行列から」レポート



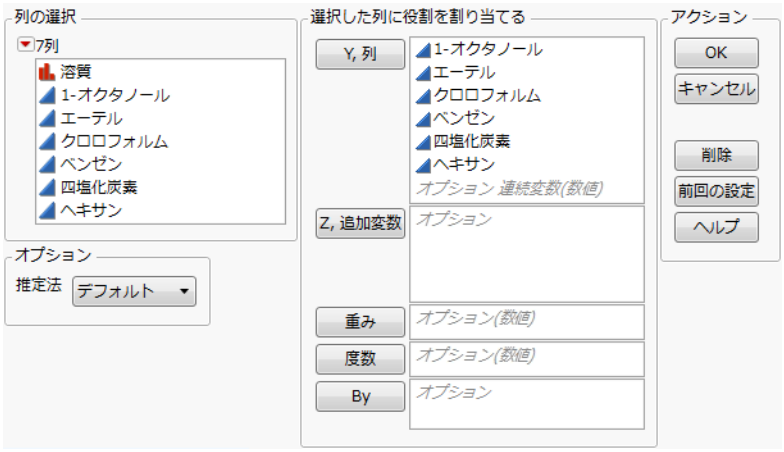
レポートには、固有値と、各主成分によって説明される変動の割合を示す棒グラフが表示されます。この例では、最初の主成分が、データ内の変動の約 80 %を説明しています。2 つめの主成分と合わせると、この 2 つの主成分でデータ内のほとんどすべての変動を説明しています (95.6 %)。また、スコアプロットと負荷量プロットも表示されます。詳細は、「「主成分分析」レポート」(54 ページ)の節を参照してください。

「主成分分析」プラットフォームの起動

「主成分分析」プラットフォームを起動するには、**[分析] > [多変量] > [主成分分析]** を選択します。主成分分析は、「多変量の相関」プラットフォームや「三次元散布図」プラットフォームでも実行できます。

「主成分分析の例」(50 ページ)で解説している例では、「Solubility.jmp」サンプルデータテーブルの連続変数をすべて使用しています。

図 4.3「主成分分析」起動ウィンドウ



Y, 列　主成分分析の対象となる変数を指定します。

Z, 追加変数　追加変数（supplementary variable）として使用する列を指定します。追加変数は主成分の計算には含まれません。追加変数を指定しても、元の主成分分析そのものには影響しません。連続尺度の追加変数に対しては、主成分負荷量が計算され、グラフにプロットされます。それらは主成分の解釈に役立てることができます。

重み　この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

メモ：「重み」の役割は、「横長」および「疎」の推定法では無視されます。

度数　この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

メモ：「度数」の役割は、「横長」および「疎」の推定法では無視されます。

By　［By］変数に指定した列の値ごとに「主成分分析」レポートが作成されるため、グループごとに分析を実行できます。

推定法　相関の計算方法を指定します。これらの一部は欠測値の処理に対応します。

デフォルト　［**デフォルト**］オプションを選択した場合は、状況に応じて実際には［リストワイズ］、［ペアワイズ］、［REML］のいずれかが使用されます。また、状況に応じて、「横長」手法への変更を促す警告が表示されます。

- 欠測値が1つもないデータテーブルの場合は、［**リストワイズ**］推定が使用されます。
- **[ペアワイズ]**推定は、欠測データとともにデータテーブルに使用します。10列以上、5,000行以上、または行数より多い列のいずれかです。
- それ以外のデータテーブルには［**REML**］推定が使用されます。

– 列数が 500 以上のデータテーブルの場合、［**横長**］推定を推奨する警告が表示されます。列数が非常に多い場合に他の方法を使用すると、計算に時間がかかるためです。［**横長**］をクリックして、横長の推定法に切り替えるか、または［**続ける**］をクリックして、最初に選択した方法を使用します。

REML 制限最尤法（REML）推定は、欠測値がある場合でも、すべてのデータを使用します。バイアス修正項の計算に時間がかかるため、データセットが大規模で欠測値が多い場合は、計算時間が長くなります。そのため、REML は小規模なデータセットに最も有効な手法です。データに欠測セルがない場合、REML 法と ML 法の推定値は、通常の共分散行列の不偏推定値と同じです。欠測セルがある場合、REML 推定の分散と共分散推定値は、ML 推定の場合よりもバイアスが少なくなります。詳細は、「REML」（63 ページ）を参照してください。

最尤 最尤（ML）推定は、欠測値がある場合でも、すべてのデータを使用します。ML 推定の生成は速いので、欠測データのある大規模データテーブルでは、この方法が最も有効です。

ロバストな推定法 ロバスト推定法は、欠測値がある場合でも、すべてのデータを使用します。この方法は、極値をダウンウェイトするので、外れ値のあるデータテーブルでは有効です。統計量の詳細については、「多変量の相関」章の「ロバストな推定法」（43 ページ）を参照してください。

リストワイズ リストワイズ推定は欠測値のあるオブザベーションを使用しません。そのため、欠測セルのある行は、推定が行われる前に削除されます。この方法は欠測データのあるオブザベーションを除外するのに便利です。ローワイズ推定は JMP 8 以前では、使用できる唯一の推定法でした。従って、JMP 8 以前の JMP バージョンとの互換性を確認するのにも使用できます。

ペアワイズ ペアワイス推定は、欠測値がある場合でも、すべてのデータを使用します。この推定法は、列のペアごとに欠測値を考慮して、相関を計算します。データテーブルに欠測値があり、かつ、列が 10 列以上、行が 5000 行以上、または列数が行数より多い場合に最も効果的です。

横長 横長推定は欠測値のあるオブザベーションを使用しません。そのため、欠測セルのある行は、推定が行われる前に削除されます。この推定法は完全な特異値分解を行いますが、内部計算において共分散行列を求めずに、効率的に主成分分析を行います。そのため、データに非常に多くの列があるときに役立ちます。詳細については、「横長」（63 ページ）を参照してください。

JMP PRO **疎** 疎推定は、欠測値がある場合でも、すべてのデータを使用します。この推定法は、特異値分解において、最初に指定された数の特異値および特異ベクトルだけを計算します。内部計算において共分散行列や不必要な主成分を求めずに、効率的に主成分分析を行います。データが疎の場合、つまりデータに多くの 0 を含む場合や、データに多数の列が存在する場合に便利です。詳細は、「疎」（64 ページ）を参照してください。

メモ : REML、ML またはロバストを選択した場合、データテーブルの列数が行数より多く、欠測値があり、JMP は推定法をペアワイズに切り替えます。

成分の数 （「推定法」に［疎］を指定した場合にのみ使用できます。）推定する成分の数を指定します。通常、「成分の数」はデータの次元よりもかなり小さい値です。

欠測値のあるデータ

さまざまな推定法で、欠測値のあるデータを扱う方法がいくつか用意されています。欠測値があるデータは、次のような方法で予め補完することもできます。

- **［多変量］>［多変量の相関］**で、［欠測データの補完］オプションを使用します。「多変量の相関」章の「欠測データの補完」（42 ページ）を参照してください。
- **［分析］>［スクリーニング］>［欠測値を調べる］**で、［多変量正規分布による補完］または［多変量の特異値分解補完］を使用します。詳細については、『予測および発展的なモデル』の「モデル化ユーティリティ」章を参照してください。

「主成分分析」レポート

［横長］または［疎］以外の手法を選んだ場合は、「主成分: 相関係数行列から」レポートがまず表示されます（「主成分分析」の赤い三角ボタンメニューから［共分散行列から］または［原点周りの積和行列から］を選択した場合はこのレポートのタイトルが異なります）。

［横長］手法を選択した場合は、「横長の主成分分析」レポートが表示されます。［疎］手法を選択した場合は、「疎の主成分分析」レポートが表示されます。

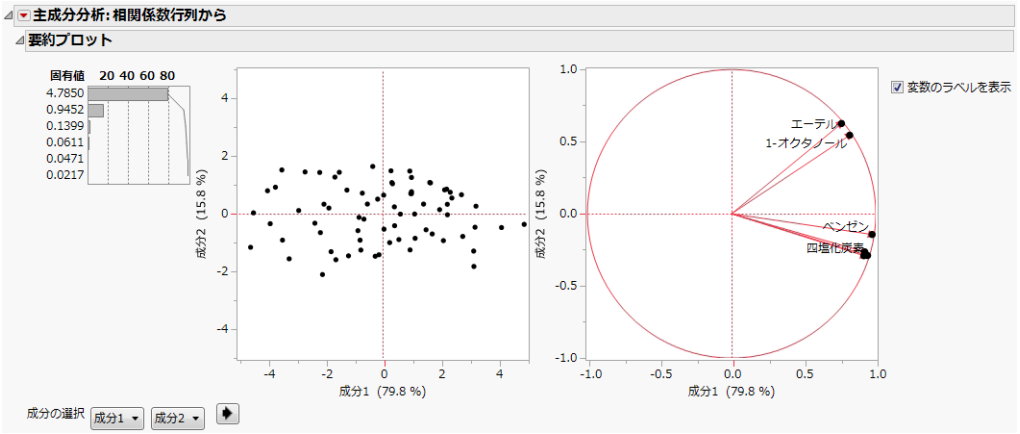
「主成分分析」レポートの最初に出力されている結果は、相関行列に対する主成分分析です。このレポートでは、指定した Y 変数のデータを各主成分がどれぐらい説明しているかが示されています。図 4.4 を参照してください。赤い三角ボタンのメニューから［主成分］オプションを選択することで、共分散行列または原データに基づいた分析に変更できます。

次のいずれに対する主成分分析にするかを、選ぶことができます。

- 相関行列
- 共分散行列
- 原データの積和行列

レポートを見ると、各主成分によってデータの変動がどの程度、説明されるかがわかります。なお、主成分スコアは、固有ベクトルを重みとした、変数の線形結合によって求められます。

図 4.4「主成分分析 : 相関係数行列から」レポート



レポートには、固有値と、各主成分によって説明される変動の割合を示す棒グラフが表示されます。また、スコアプロットと負荷量プロットも表示されます。固有値は、各主成分によって説明される変動の量を表しており、抽出すべき主成分の数を決めるときの目安となります。

スコアプロットは、初めの 2 成分の主成分スコアをプロットしたものです。相関行列に対する主成分分析では、主成分スコアの平均は 0 で、分散は固有値となっています。

負荷量プロットは、回転前の主成分負荷量をプロットしたものです。主成分負荷量は、主成分スコアと、各変数との相関です。値が 1 に近づくほど、主成分スコアと変数との間に強い関係があることを示します。

デフォルトでは、レポートには最初の 2 つの主成分のスコアプロットと負荷量プロットが表示されます。「成分の選択」の横のリストで、スコアプロットと負荷量プロットにグラフ表示する主成分を指定します。

「主成分分析」レポートのオプション

主成分分析のタイトルバーにある赤い三角ボタンをクリックすると、次のようなオプションが表示されます。

メモ : 一部のオプションは、「横長」または「疎」の推定法では表示されません。

主成分分析 （「横長」または「疎」の推定法では表示されません。）**[相関係数行列から]**、**[共分散行列から]**、または **[原点周りの積和行列から]** 主成分を作成できます。

相関 （「横長」または「疎」の推定法では表示されません。）相関係数行列の表示／非表示を切り替えます。

メモ : 対角要素の値は 1.0 になります。

共分散行列 （「横長」または「疎」の推定法では表示されません。）共分散行列の表示／非表示を切り替えます。

固有値　各主成分に対応する固有値が、大きい方から順に表示されます。固有値は、多変量データにおける分散の合計がどのように主成分によって分割されているかを表します。

固有値の尺度は、主成分の抽出にどの行列を選択するかによって異なります。

- ［相関係数行列から］オプションを選んだ場合、固有値の合計は変数の個数に一致します。
- ［共分散行列から］オプションを選んだ場合、固有値は尺度化されません。
- ［原点周りの積和行列から］オプションを選んだ場合、積和行列の固有値を標本サイズで割った値が表示されます。

赤い三角ボタンのメニューから［**Bartlett の検定**］オプションを選択した場合、固定値に対する仮説検定（図 4.6）が行われます（Jackson, 2003）。

図 4.5 固有値

固有値

番号	固有値	寄与率	20 40 60 80	累積寄与率
1	4.7850	79.750		79.750
2	0.9452	15.754		95.504
3	0.1399	2.331		97.835
4	0.0611	1.018		98.853
5	0.0471	0.785		99.638
6	0.0217	0.362		100.000

固有ベクトル　各主成分（左から順に第 1 主成分、第 2 主成分、と並ぶ）の固有ベクトルの値の表示／非表示を切り替えます。これらの係数を使って元の変数の線形結合を計算したものが、主成分スコアです。各固有ベクトルはノルムが 1 になるように標準化されています。

メモ : 表示される固有ベクトルの数は、相関行列のランクと同じか、または「疎」の推定法が選択されている場合は、起動ウィンドウで指定された成分の数です。

Bartlett の検定　（「横長」または「疎」の推定法では表示されません。）固有値の等質性に対する検定結果の表示／非表示を切り替えます（「固有値」表に追加されます）。この検定では、複数の固有値が等しいかどうかに関する検定に対して、カイ 2 乗値、自由度（DF）、*p* 値（prob > ChiSq）が計算されます。Bartlett（1937, 1954）を参照してください。

図 4.6 Bartlett の検定

固有値

番号	固有値	寄与率	20 40 60 80	累積寄与率	カイ2乗	自由度	p値(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	0.000	.	.

負荷量行列　各成分の負荷量の表の表示／非表示を切り替えます。これらの負荷量は、負荷量プロットに描かれています。表中の値の透明度は、負荷量の絶対値が、どれぐらいゼロに近いかを示しています。負荷量の絶対値がゼロに近いほど透明になります。

負荷量がどのように尺度化されるかは、主成分分析の対象となった行列で異なります。

- ［相関係数行列から］オプションを選んだ場合、負荷量の *i* 番目の列は、*i* 番目の固有ベクトルに *i* 番目の固有値の平方根を掛けたものとなります。*i,j* 番目の負荷量は、*i* 番目の変数と *j* 番目の主成分との相関です。
- ［共分散行列から］オプションを選んだ場合、第 *i* 列、第 *j* 行の負荷量は、*i* 番目の固有ベクトルに *i* 番目の固有値の平方根を掛けて、*j* 番目の変数の標準偏差で割ったものとなります。*i,j* 番目の負荷量は、*i* 番目の変数と *j* 番目の主成分との相関です。
- ［原点周りの積和行列から］オプションを選んだ場合、第 *i* 列、第 *j* 行の負荷量は、*i* 番目の固有ベクトルに *i* 番目の固有値の平方根を掛けて、*j* 番目の変数の標準誤差で割ったものとなります。ここで言う「*j* 番目の変数の標準誤差」とは、平方和と交差積行列の *j* 番目の対角要素を行数で割った値です（**X′X**/n）。

メモ： 原点周りの積和行列からの分析の場合、*i,j* 番目の負荷量は *i* 番目の変数と *j* 番目の主成分の間の相関ではありません。

濃淡表示の負荷量行列　各成分の負荷量の表の表示／非表示を切り替えます。この表の変数は、第 1 主成分の負荷量によって降順に並べ替えられています。

図 4.7 濃淡表示の負荷量行列

淡色表示の負荷量行列

	主成分1	主成分2	主成分3	主成分4	主成分5	主成分6
ベンゼン	0.974761	-0.143460	-0.081914	-0.023369	-0.108186	-0.101488
四塩化炭素	0.942849	-0.289099	0.069134	-0.059654	-0.099757	0.095746
ヘキサン	0.923483	-0.263641	0.256572	0.026770	0.099672	-0.034523
クロロフォルム	0.917422	-0.290351	-0.242516	0.075629	0.093454	0.027695
1-オクタノール	0.819019	0.544318	-0.041397	-0.162739	0.068711	0.002762
エーテル	0.761978	0.625283	0.044774	0.155130	-0.045337	0.016883

値が小さい負荷量を淡色表示: 閾値= 0.3

テキストの濃さ 0.4

値が小さい負荷量を淡色表示：閾値＝　「濃淡表示の負荷量行列」レポートに表示しない負荷量を決める値です。テキストボックスかスライダーを使って、選択した値より小さい絶対値の負荷量を淡色表示にします。

テキストの濃さ　「濃淡表示の負荷量行列」レポートで淡色表示する値の透明度。テキストボックスかスライダーを使って、淡色表示する負荷量の透明度を設定します。透明度は 0 ～ 1 です。低い値ほど透明度は高くなります。たとえば、透明度を 0 に設定すると、使用できない負荷量が行列から完全に削除されます。透明度を 1 に設定すると、負荷はなおも使用可能です。

変数の余弦２乗　各変数の余弦２乗を示した表の表示／非表示を切り替えます。各変数の余弦２乗を主成分全体で合計したものは、１になります（100% になります）。余弦２乗を見れば、各変数が該当の主成分によっていかに良く表されるかがわかります。また、ある変数を表すのにいくつの主成分が必要かを知る目安になります。このオプションを選ぶと、最初の３つの主成分の余弦２乗を描いたプロットも表示されます。

メモ：「疎」推定法を使用し、選択した成分が２つ以下の場合、指定した数の成分のみがプロットに表示されます。

変数の偏寄与率　各変数の偏寄与率を示した表の表示／非表示を切り替えます。偏寄与率を見ると、各変数が各主成分に寄与する割合がわかります。このオプションを選ぶと、最初の３つの主成分の偏寄与率を描いたプロットも表示されます。

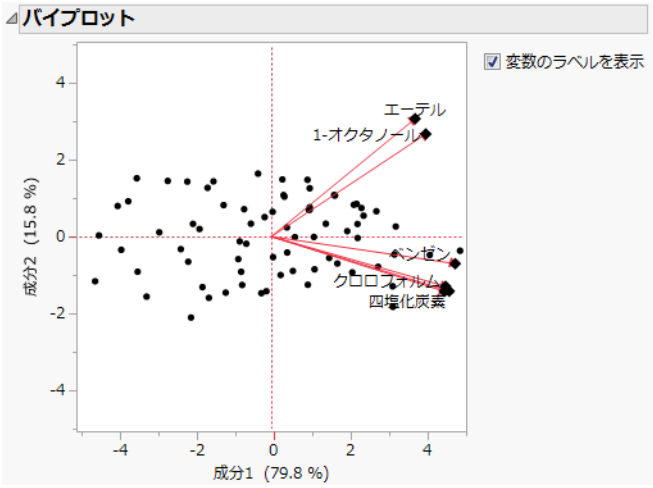
メモ：「疎」推定法を使用し、選択した成分が２つ以下の場合、指定した数の成分のみがプロットに表示されます。

要約プロット　デフォルトのレポートで作成された要約情報の表示／非表示を切り替えます。この要約情報には、固有値のプロット、スコアプロット、負荷量プロットが含まれます。デフォルトでは、レポートには最初の２つの主成分のスコアプロットと負荷量プロットが表示されます。プロットする主成分を指定するオプションがあります。「「主成分分析」レポート」（54 ページ）を参照してください。

ヒント：負荷量プロットの矢印の先を選択すると、データテーブル内の対応する列が選択されます。Ctrl キーを押したまま矢印の先をクリックすると、列の選択を解除できます。

バイプロット　スコアと負荷量を重ねて描いたプロットの表示／非表示を切り替えます。指定した成分数に対するバイプロットが描かれます。

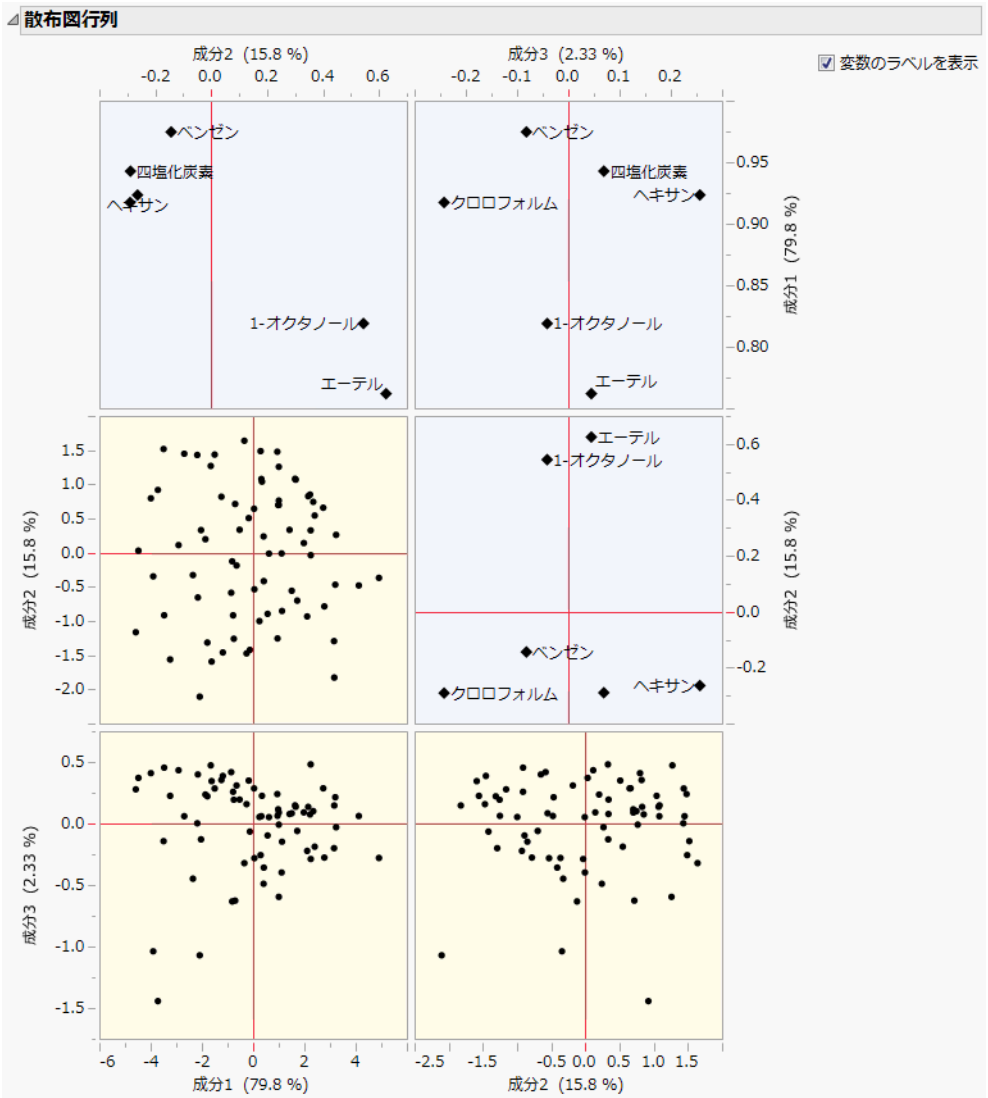
図 4.8 バイプロット



メモ: スコアプロットのマーカーは点で、負荷量プロットのマーカーはひし形です。

散布図行列　スコアと負荷量をプロットした散布図行列の表示／非表示を切り替えます。この散布図行列には、スコアプロットが左下に、負荷量プロットが右上に描かれています。スコアプロットの背景色は黄色、負荷量プロットの背景色は青色になっています。

図 4.9 散布図行列



メモ: 散布図行列に表示される負荷量プロット行列は、[負荷量プロット] オプションを選択したときに得られる負荷量プロット行列を転置したものです。

スクリープロット 各成分の固有値のグラフの表示／非表示を切り替えます。成分数に対して固有値をプロットしたものです。このプロットは、データ空間の次元数を決めるのに役立ちます。

スコアプロット 主成分スコアをプロットした散布図行列の表示／非表示を切り替えます。指定した成分数に対する散布図が描かれます。図 4.4 は主成分スコアをプロットした例です（一番左側のプロット）。

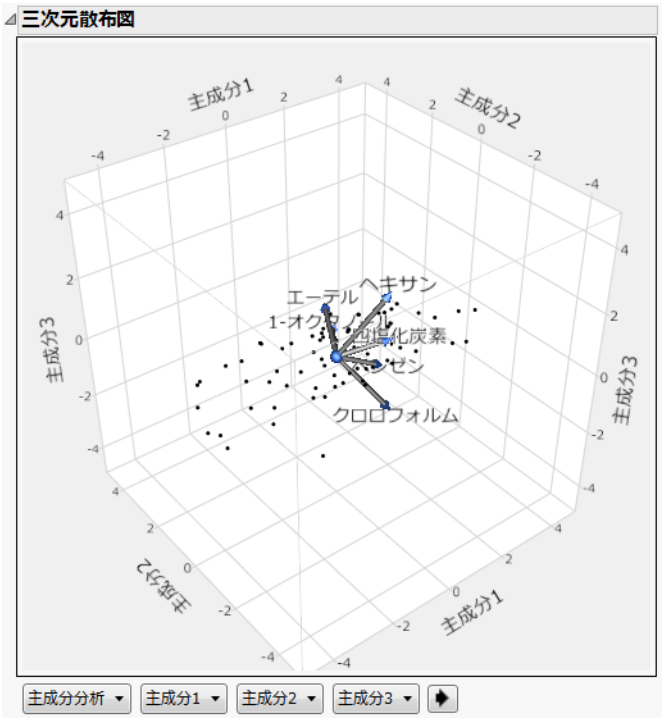
負荷量プロット 主成分負荷量のプロットの表示／非表示を切り替えます。指定した成分数に対するプロットが描かれます。負荷量プロットでは、変数の数が 30 個以下の場合は変数のラベルが表示されます。変数が 30 個より多い場合は、デフォルトでラベルが非表示になります。この情報は、図 4.4 に表示されています（一番右側のプロット）。

ヒント: 負荷量プロットの矢印の先を選択すると、データテーブル内の対応する列が選択されます。Ctrl キーを押したまま矢印の先をクリックすると、列の選択を解除できます。

補完したスコアプロット （「横長」または「疎」の推定法では表示されません。）欠測値をすべて補完したスコアプロットが作成されます。このオプションは、欠測値がある場合にのみ使用できます。

三次元スコアプロット （「横長」または「疎」の推定法では表示されません。）主成分スコアの三次元散布図が表示されます。1 回目の実行時には、最初の 3 つの主成分がプロットの対象となります。

図4.10 三次元スコアプロット



プロットのソース　プロット内のデータ点のソースです。使用できるオプションは、［主成分分析］、［回転後の主成分］、［データ列］です。

軸の調整　各軸の内容です。［主成分分析］オプションまたは［回転後の主成分］オプションを選択すると、［軸の調整］オプションが主成分になります。［データ列］オプションを選択すると、オプションは分析の変数となります。

サイクルボタン　使用できる軸の内容を順番に表示します。

このプロットでは、変数が中心からの線として表示されます。これは、**バイプロット線**といい、変数の向きを主成分の空間上で近似しています。変数が2つまたは3つしかない場合、バイプロット線は、近似ではなく、変数の正確な向きを表します。バイプロットの向きは、主成分の固有ベクトルに該当します。

表示オプション　矢印を表示できるプロットにおいて、矢印の表示／非表示を切り替えることができます。変数の数が1000個以下の場合、矢印が表示されます。変数が1000個より多い場合は、デフォルトでラベルが非表示になります。

矢印線　矢印線を表示できるすべてのプロットにおいて、変数の矢印線の表示／非表示を切り替えます。

追加変数の表示　（追加変数を指定したときだけ使用できます）バイプロット、スコアプロット、負荷量プロットにおいて、連続尺度の追加変数の矢印の表示／非表示を切り替えたり、カテゴリカルな追加変数にマーカーを表示したりできます。

因子分析 （「横長」または「疎」の推定法では表示されません。）主成分の回転もしくは因子分析が実行されます。詳細については、『消費者調査』の「因子分析」章を参照してください。

JMP PRO **変数のクラスタリング** （「横長」または「疎」の推定法では表示されません。）変数に対するクラスター分析を実行し、変数を重なりのないクラスターに分割します。似ているどうしの変数から構成されたクラスターに分類します。各クラスターが、1つの成分または変数で表せるようになります。成分は、クラスター内のすべての変数の線形結合です。また、クラスターは、クラスター内で最も代表的とみなされる変数で表すこともできます。詳細については、「変数のクラスタリング」（197ページ）章を参照してください。

メモ：［変数のクラスタリング］は、［共分散行列から］または［原点周りの積和行列から］オプションを選択した場合でも、すべての計算に相関行列が使用されます。

主成分の保存 指定した数の主成分を、各成分を計算するための計算式とともにデータテーブルに保存します。この計算式で計算される成分は、欠測値のあるデータに対しては欠測値になります。

主成分の計算は、主成分の抽出にどの行列を選択するかによって異なります。

- ［相関係数行列から］オプションを選んだ場合、*i*番目の主成分は、*i*番目の固有ベクトルを係数にして求められた、中心化かつ尺度化されたデータの線形結合です。
- ［共分散行列から］オプションを選んだ場合、*i*番目の主成分は、*i*番目の固有ベクトルを係数にして求められた、中心化されたデータの線形結合です。
- ［原点周りの積和行列から］オプションを選んだ場合、*i*番目の主成分は、*i*番目の固有ベクトルを係数にして求められた、生データの線形結合です。

メモ：指定した成分の数が相関行列のランクを超えている場合、保存される成分の数は相関行列のランクに設定されます。

予測値の保存 指定した数の主成分から計算される予測値を、データテーブル内の新しい列に保存します。

Xモデルまでの距離を保存 各データ行から主成分モデルまでの距離（DModX）を、データテーブルの新しい列に保存します。DModXの値が大きいデータ行は、指定された成分数の主成分モデルでは十分に説明されず、その主成分モデルからは外れ値となっている可能性があります。詳細は、「DModXの計算方法」（64ページ）を参照してください。

データ行の余弦2乗を保存 データ行の余弦2乗を、データテーブルの新しい列に保存します。

データ行の偏寄与率を保存 データ行の偏寄与率を、データテーブルの新しい列に保存します。

回転後の成分を保存 （「横長」または「疎」の推定法では表示されません。）回転後の成分の計算式がデータテーブルに保存されます。このオプションは、［因子分析］オプションを使用した場合にのみ表示されます。この計算式で計算される成分は、欠測値のあるデータに対しては欠測値になります。

補完して主成分を保存 （「横長」または「疎」の推定法では表示されません。）欠測値を補完した後に算出した主成分がデータテーブルに保存されます。列には、欠測値を補完し、主成分を算出する計算式が保存されます。このオプションは、欠測値がある場合にのみ使用できます。

補完して回転後の成分を保存 （「横長」または「疎」の推定法では表示されません。）欠測値を補完した後に算出した回転成分がデータテーブルに保存されます。列には、欠測値を補完し、回転後の成分を算出する計算式が保存されます。このオプションは、［因子分析］オプションを使用し、欠測値がある場合にのみ表示されます。

JMP PRO **成分計算式の発行** 指定した数の主成分計算式を作成し、それを「計算式デポ」レポート内の計算式列スクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションを選択した時点でレポートが作成されます。『予測および発展的なモデル』の「計算式デポ」章を参照してください。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

統計的詳細

推定法について

REML

データに欠測値がある場合、REML 法（制限最尤法）は、最尤法に比べて、推定値のバイアスが小さいのが特徴です。REML 法は、誤差対比（error contrast）から導出された周辺尤度を最大化する推定方法です。REML 法は、分散および共分散を推定するのによく使われます。［多変量の相関］の［REML］は、反復測定データの相関構造に無構造（unstructured）を仮定した混合モデルの REML 推定と同じです。混合モデルについては、SAS システムの PROC MIXED に関するドキュメントを参照してください。

横長

内部計算において共分散行列を求めずに、効率的に主成分分析を行います。このアルゴリズムは、特異値分解に基づきます。次の表記を使用します。

- n = 行の数
- p = 変数の数

- **X** = データ値の n x p 行列

0以外の固有値の数、およびその結果の主成分の数は、**X**の相関行列のランクと同じです。0以外の固有値の数は、n と p の小さい方を超えることはできません。

推定法として［横長］を選択した場合、データは常に標準化されます。データの標準化とは、データから平均を引き、それを標準偏差で割る変換を指します。標準化したデータの共分散行列は、**X**の相関行列となります。標準化したデータ値のn x p行列を**X**sとすると、次のように相関行列は求められます。

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

$\mathbf{X}_s$は、特異値分解した行列により $\mathbf{U}Diag(\mathbf{\Lambda})\mathbf{V}'$ と表されます。この特異値分解により、固有ベクトルと $\mathbf{X}_s'\mathbf{X}_s$ の固有値が求められます。なお、主成分（スコア）は $\mathbf{X}_s\mathbf{V}$ によって求められます。詳細については、「統計的詳細」の付録の「「線形 横長データ」の手法と特異値分解」（210ページ）を参照してください。

JMP PRO 疎

「横長」手法と同様に、「疎」手法は特異値分解に基づきます。そのため、「疎」手法のアルゴリズムでは共分散行列の計算が省略され、効率的に計算が行われます。

「横長」（63ページ）で説明した**X**の同じ表記と標準化を使用すると、**X**の相関行列は $\mathbf{X}_s$ の共分散行列によって次のように表されます。

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

「疎」手法は、特異値分解の計算方法において「横長」手法とは異なります。「横長」手法は完全な特異値分解を行いますが、「疎」手法は、特異値分解において、最初に指定された数の特異値および特異ベクトルだけを計算します。そのため、最初に指定した数の固有値と主成分が戻されます。アルゴリズムの詳細については、Baglama and Reichel（2005）を参照してください。

DModXの計算方法

DModXは、主成分モデルまでの距離で、次のように定義されます。

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

ここで

e_{ik} = モデルの残差

K = 変数の数

A = 主成分の数

大きなDModXの値は、データ内の外れ値がそれほど極端ではないことを示します。

第 5 章

線形判別分析

連続変数から分類変数を予測する

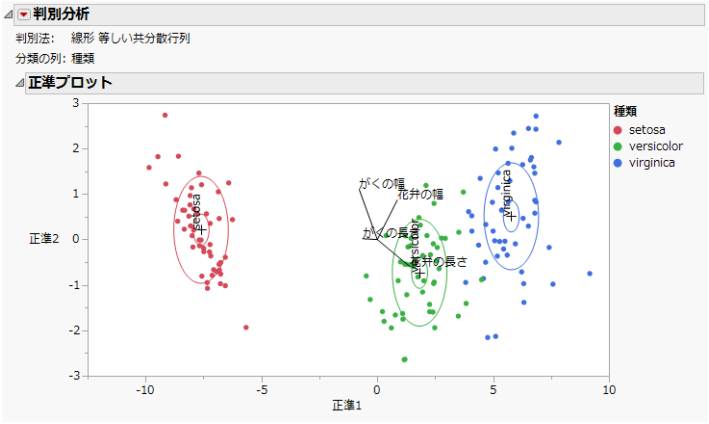
判別分析は、複数の連続変数から、どのグループ（カテゴリ）に属するかを予測します。判別分析は、連続尺度の応答変数（Y変数）から、カテゴリカルな分類変数（X変数）を予測します。判別分析では、所属するグループが既知で、そのグループを予測するために連続変数のデータを用います。

たとえば、ローンの申込者を、「リスク低」、「リスク中」、「リスク高」の3つのカテゴリ（X）に分類することを考えて見ましょう。収入、勤続年数、年齢、債務といった連続変数（Y）を使って、それらのカテゴリを予測するとします。この場合、連続変数によって各個人を各カテゴリに分類するための判別分析モデルを作成できます。

「判別分析」プラットフォームの特徴は次のとおりです。

- 判別に適した変数を選択するためのステップワイズ変数選択も行える
- 判別分析の手法として、線形、2次、正則化、横長データの手法から選ぶことができる
- 正準プロットと誤分類の要約
- 各点がどのグループに近いかを示す判別スコア
- データテーブルに予測距離と予測確率を保存するオプション

図5.1 正準プロット



判別分析の概要

判別分析は、連続変数によって各オブザベーションをグループに分類します。言い換えると、カテゴリカルな *X* 変数で示される所属先を、連続変数によって予測します。予測に使われる連続変数は、JMP の判別分析では、**「共変量」**と呼ばれており、「*Y*」と記されています。

判別分析はロジスティック回帰とは異なります。ロジスティック回帰では、連続変数を所与として、カテゴリカルな変数を確率変数として扱います。一方、判別分析では、カテゴリカルな変数を所与として、連続変数の共変量（*Y*）を確率変数として扱います。ただし、カテゴリカルな変数を連続変数で予測するという点では、これらの手法は似ています。

「判別分析」プラットフォームには、4 つの手法が用意されています。どの手法も、各オブザベーションから各グループの多変量平均（**重心**ともいう）までの距離を、Mahalanobis の距離で求めます。グループへ属する事前確率を指定することができ、それらの事前確率が距離の計算で考慮されます。各オブザベーションは、最も距離が近いグループに判別されます。

用意されている手法には次のものがあります。

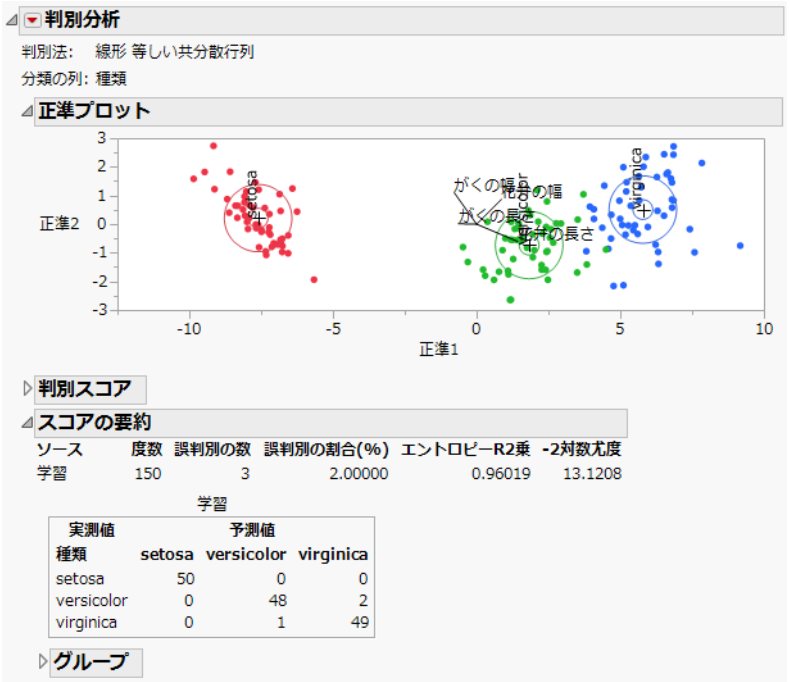
- 線形 — この手法では、群内共分散行列（グループ内共分散行列）がすべて等しいと仮定されます。そして、共変量 *X* の平均ベクトルだけが各群で異なると仮定されます。
- 2 次 — 群内共分散行列がすべて異なると仮定されます。この手法では、共分散行列を推定するのに、線形判別の場合よりも多くのパラメータを推定しなければいけません。ある群の標本サイズが小さい場合、推定値が不安定になってしまう危険があります。
- 正則化 — 群内共分散行列がすべて異なると仮定されますが、より安定した推定値を導き出すために 2 つの調整方法が用意されています。この手法は、群の標本サイズが小さいときに役立ちます。
- 横長データ — 共変量の数が多くて、他の手法では計算が難しい場合に役立ちます。この手法では、郡内共分散行列がすべて等しいと仮定されます。

判別分析の例

Fisher のあやめのデータでは、3 つの異なる品種のあやめについて、4 つの特性が測定されています。4 種類の測定値から精確に品種を予測するのが目標です。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Iris.jmp」を開きます。
2. ［**分析**］>［**多変量**］>［**判別分析**］を選択します。
3. **「がくの長さ」**、**「がくの幅」**、**「花弁の長さ」**、および **「花弁の幅」**を［**Y, 共変量**］に指定します。
4. **「種類」**を選択し、［**X, カテゴリ**］をクリックします。
5. ［**OK**］をクリックします。

図5.2「判別分析」レポートウィンドウ

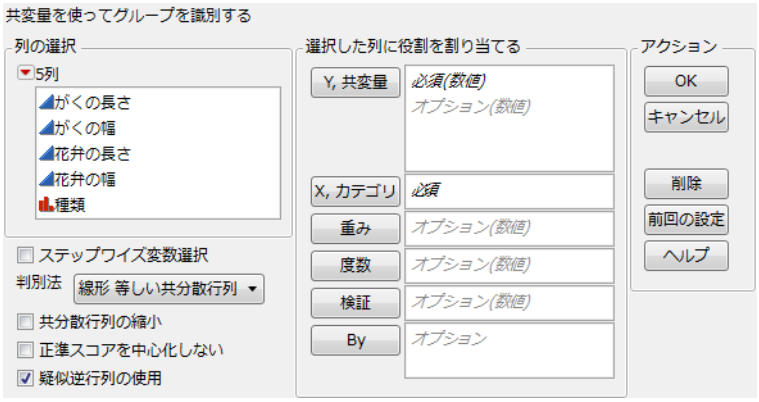


「種類」は3群なので、正準変数は全部で2つしかありません。「正準プロット」は、この2次元の正準座標に各オブザベーションをプロットしたものです。プロットを見ると、2次元の座標によって3つの品種が判別されていることがわかります。今回の例では検証セットがなかったため、「スコアの要約」レポートには学習セットのみの誤分類表が表示されています。検証セットがない場合、データセット全体が学習セットとみなされます。150個のオブザベーションのうち、3つだけが誤分類されています。

「判別分析」起動ウィンドウ

「判別分析」プラットフォームを起動するには、［分析］ > ［多変量］ > ［判別分析］を選択します。

図 5.3「Iris.jmp」の「判別分析」起動ウィンドウ



メモ: ［検証］ボタンは、JMP Pro のみに表示されます。JMP では、除外した行を使用して、検証セットを定義できます。「JMP と JMP Pro の違い」（93 ページ）を参照してください。

Y, 共変量 オブザベーションをカテゴリに分類するために使用する連続変数の列。

X, カテゴリ オブザベーションが属するカテゴリ（グループ）を含む列。

重み この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数 この役割を割り当てた列の数値は、分析において各行の度数として使用されます。度数列は、データテーブルの行を増やすような効果があります。つまり、ある行に度数として整数 k を指定すると、その行は k 個の行として扱われます。小数を含む度数も指定できます。

JMP PRO **検証** 2つまたは3つの異なる値を含む数値の列。

- 検証列の値が2つしかない場合は、小さい方の値が学習セット、大きい方の値が検証セットとして扱われます。
- 値が3つある場合は、値の小さい方から順に、学習セット、検証セット、テストセットとして扱われます。
- 値が3つ以上ある場合は、最も小さい3つの値以外は無視されます。

「列の選択」リストで列を選択せずに［検証］ボタンをクリックすることにより、データテーブルに検証列を追加できます。「検証列の作成」ユーティリティについての詳細は、『予測および発展的なモデル』の「モデル化ユーティリティ」章を参照してください。

By By 変数の水準ごとに、個別に分析が行われます。

ステップワイズ変数選択　共分散分析とｐ値を使って、ステップワイズ変数選択を実行します。詳細については、「ステップワイズ変数選択」（69ページ）を参照してください。

検証セットを指定した場合、検証セットから計算された統計量も表示されます。

メモ：このオプションは、「横長データ」判別手法では使用できません。

判別分析の手法　判別分析に関する４つの手法があります。「判別法」（72ページ）を参照してください。

共分散行列の縮小　プールして計算された群内共分散行列、もしくは、群ごとの群内共分散行列の非対角要素を縮小します。「共分散行列の縮小」（75ページ）を参照してください。

正準スコアを中心化しない　旧バージョンのJMPとの互換性のため、正準スコアを中心化しません。

疑似逆行列の使用　共分散行列が特異値である場合は、Moore-Penroseの疑似逆行列を使用します。このオプションをオンにすると、常に、スコアの計算式にすべての共変量が含まれます。このオプションをオフにすると、一次従属関係にある変数のうち、［**Y, 共変量**］リストで後に指定されたものが計算式に含まれない場合があります。

ステップワイズ変数選択

メモ：［ステップワイズ変数選択］は、「横長データ」手法では使用できません。

起動ウィンドウで［ステップワイズ変数選択］オプションを選択した場合、「判別分析」レポートに「列選択」パネルが表示されます。変数を選択するためのボタン、または、「ロック」および「追加」チェックボックスによって手動でステップワイズ分析を実行してください。選択に基づいて、*F*値と*p*値は更新されます。これらの値が更新される方法については、「F 値とｐ値の更新」（70ページ）を参照してください。

図5.4 検証セットを指定した場合の「Iris.jmp」の「列選択」パネル

メモ：［実行］ボタンは、JMPでいくつかの行を除外することにより検証セットを指定した場合、および、JMP Proで検証列を指定した場合にのみ表示されます。

F値とp値の更新

モデルに変数を追加または削除した場合、次の共分散分析モデルに基づいて、F値とp値が更新されます。

- 検討対象の共変量は応答変数とします
- モデルにすでに追加されている共変量は説明変数とします
- 分類変数（グループ変数）も説明変数とします

ステップワイズレポートの「F値」と「p値(Prob>F)」の値は、上記の共分散分析モデルにおけるグループ変数に対する *F* 値と *p* 値です。この分類変数に対する検定は、検討中の共変量がどれぐらい判別するのに寄与するかを示す指標となっています。

統計量

追加されている列 すでに判別モデルに含まれている列の数。現在選択されている列の数。

除外されている列 まだ判別モデルに含まれていない列の数。

次に追加する最小p値 まだ判別モデルに含まれていない共変量の *p* 値において、最小の *p* 値。

次に除外する最大p値 すでに判別モデルに含まれている共変量の *p* 値において、最大の *p* 値。

検証 エントロピー R2 乗 検証セットのエントロピー R2 乗。値が大きいほど、あてはまりが良いことを示します。エントロピー R2 乗が 1 の場合、すべての分類が正しく行われたことを意味します。一般に、判別分析モデルによる分類は 100% 的中することはなく、エントロピー R2 乗の値は小さくなる傾向にあります。

「エントロピー R2 乗」（83 ページ）を参照してください。検証セットが使用されている場合にのみ表示されます。

メモ:「検証 エントロピー R2 乗」は負の値である場合もあります。

検証 誤分類率 検証セットの誤分類率。値が小さいほど、分類が適切であることを示します。検証セットが使用されている場合にのみ表示されます。

ボタン

変数増加 モデルにまだ含まれていない共変量のうち、最も有意性の高いものを追加します。検証セットが使用されている場合でも、学習セットから計算されたp値に基づきます。

変数減少 すでにモデルに含まれている共変量でロックされていないもののうち、最も有意性の低いものを除外します。検証セットが使用されている場合でも、学習セットから計算されたp値に基づきます。

すべて追加 ロックされていないすべての共変量の「追加」列にチェックマークをつけ、モデルに追加します。

すべて削除 ロックされていないすべての共変量の「追加」列のチェックマークを外し、モデルから削除します。

このモデルを適用 「追加」列にチェックマークをつけた共変量に基づき、判別分析レポートを作成します。「列選択」アウトラインが閉じ、「判別分析」ウィンドウに、選択した判別法に基づいた分析結果が表示されます。

ヒント：［**このモデルを適用**］をクリックすると、選択した列が「スコアの要約」レポートの最上部に表示されます。

実行 「検証　エントロピー R2 乗」が減少し始めるまで、変数増加法により共変量を追加していきます。変数を追加しても、「検証　エントロピー R2 乗」が増加しないことが 2 回続いたら、処理を終了します。JMP では除外されている行がある場合にのみ、JMP Pro では検証列が使用されている場合にのみ表示されます。

列

ロック ボタンを使ってステップ処理が行われた場合でも、共変量を現在の状態のままにします。

次の点に注意してください。

- 共変量を追加し、それを**「ロック」**した場合、コントロールボタンを使った選択の内容に関わらず、共変量はモデル内に残ります。ロックした共変量の**「追加」**ボックスは薄く表示され、その共変量がモデル内にあることが示されます。
- 追加されていない共変量の**「ロック」**を選択すると、コントロールボタンを使った選択の内容に関わらず、モデルに追加されません。

追加 チェックマークが付いている列が、現在モデルに含まれている列です。チェックマークを付けたり外したりすることで、列を手動で追加または除外することができます。チェックマークが薄く表示されている場合は、共変量がロックされ、モデルに追加されていることを示します。

列 対象とする共変量。

F 値 共分散分析に基づいて計算された、グループ変数に対する検定の *F* 値です。詳細については、「F 値と p 値の更新」（70 ページ）を参照してください。

p 値 (Prob>F) 共分散分析に基づいて計算された、グループ変数に対する検定の *p* 値です。詳細については、「F 値と p 値の更新」（70 ページ）を参照してください。

ステップワイズ法の例

ステップワイズ法の使用法を説明するため、ここでは「Iris.jmp」サンプルデータを使用します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Iris.jmp」を開きます。
2. ［**分析**］>［**多変量**］>［**判別分析**］を選択します。
3. **「がくの長さ」**、**「がくの幅」**、**「花弁の長さ」**、および**「花弁の幅」**を［**Y, 共変量**］に指定します。
4. **「種類」**を選択し、［**X, カテゴリ**］をクリックします。
5. ［**ステップワイズ変数選択**］を選択します。
6. ［**OK**］をクリックします。

7. ［**変数増加**］を3回クリックします。

   3つの共変量がモデルに追加されます。［次に追加する最小p値］がパネルの一番上に表示されます。値が0.0103288であることから、残りの共変量である**「がくの長さ」**も**「種類」**の判別分析モデルにおいて重要である可能性があります。

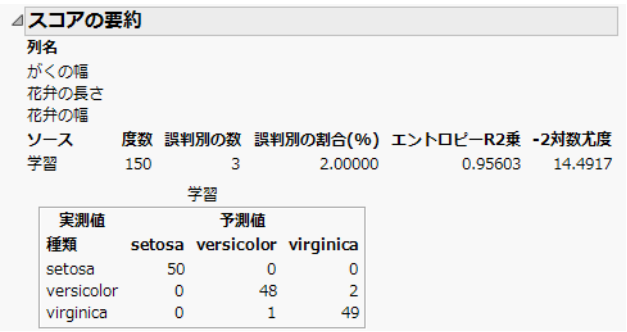
図5.5「Iris.jmp」のステップワイズ変数選択

8. ［**このモデルを適用**］をクリックします。

   「列選択」アウトラインが閉じ、ウィンドウが更新されます。追加した共変量と選択した判別法に応じたあてはめのレポートが表示されます。

   モデルに選択した共変量が、「スコアの要約」レポートにリストされていることを確認してください。

図5.6 選択した共変量を示す「スコアの要約」レポート



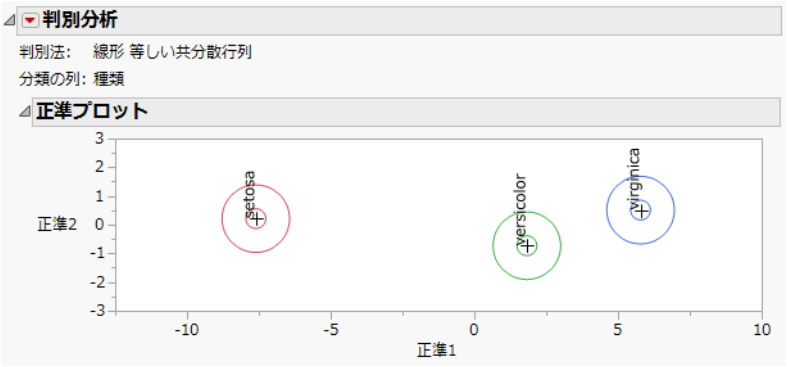
判別法

JMPには、［線形］、［2次］、［正則化］、および［横長データ］といった手法が用意されています。最初の3つの手法は、仮定されているモデルが異なります。［横長データ］手法も線形判別モデルをあてはめるのですが、共変量の数が多い状況において効率的に計算が行われます。

メモ： 500 個以上の共変量を追加すると、横長データ手法への変更を促す警告が表示されます。列数が非常に多い場合に他の方法を使用すると、計算に時間がかかるためです。横長データの手法に変更する場合は、**[線形 横長データ]** をクリックしてください。すでに選択している手法を使用する場合は、**[続行]** をクリックします。

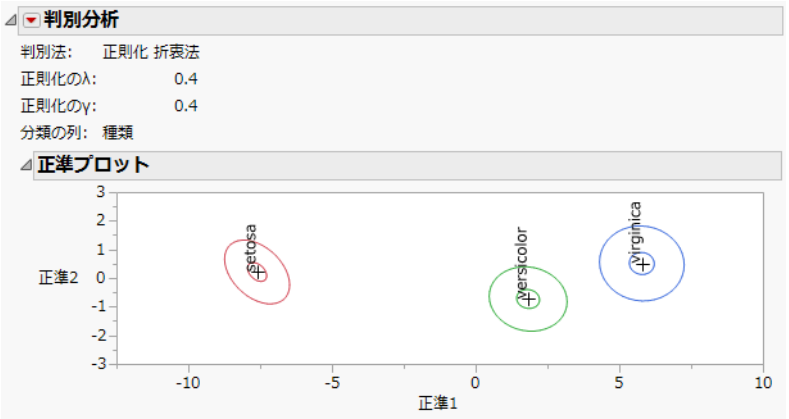
図 5.7 線形、2 次、正則化の判別法

線形



2 次

正則化
（λ=0.4、γ=0.4）



線形、2次、正則化の判別法を図5.7に示します。以下に、これらの手法について概説します。詳細については、「保存される計算式」（94ページ）を参照してください。

線形 等しい共分散行列　線形判別分析を実行します。この手法では、群内共分散行列（グループ内共分散行列）がすべて等しいと仮定されます。「線形判別法」（95ページ）を参照してください。

2次 異なる共分散行列　2次判別分析を実行します。この手法では、群内共分散行列がすべて異なると仮定されます。この手法では、共分散行列を推定するのに、線形判別の場合よりも多くのパラメータを推定しなければいけません。ある群の標本サイズが小さい場合、推定値が不安定になってしまう危険があります。「2次判別法」（96ページ）を参照してください。

あるグループ内で値がすべて同じ共変量が存在していると、その共変量との共分散がすべて0になります。このような状況でも群内共分散行列の逆行列を求めるため、0となっている共分散が、プールして計算された共分散に置き換えられます。この処理が行われたときは、レポートウィンドウに該当する共変量と群を示すメモが表示されます。

ヒント: 2次判別分析は、小規模なデータセットには適していません。逆行列が計算できなかったり、安定した共分散行列が得られなかったりします。このような問題を改良し、データが足りない場合でもグループ間で異なる共分散行列を仮定できるのが、正則化の手法です。

正則化 折衷法　この手法でも群内共分散行列がすべて異なると仮定されますが、より安定した推定値を導き出すために2つの調整方法が用意されています。この手法は、グループの標本サイズが小さいときに役立ちます。「正則化 折衷法」（74ページ）および「正則化判別法」（97ページ）を参照してください。

線形 横長データ　共変量の数が多くて、他の手法では計算が難しい場合に役立ちます。この手法では、郡内共分散行列がすべて等しいと仮定されます。この手法は、プールした群内共分散行列の逆行列を計算するのに、特異値分解を用います。「「線形 横長データ」のアルゴリズムについて」（94ページ）を参照してください。

メモ: ［線形 横長データ］オプションを使用した場合、他の判別法では用意されている機能の一部が使用できません。その理由は、横長データの場合にはプールした群内共分散行列が巨大になりますが、このアルゴリズムではその巨大な群内共分散を明示的に計算しないためです。

正則化 折衷法

正則化判別分析は、負でないパラメータを2つ使います。

- 最初のパラメータ（「**λ: 共通の共分散行列に近づける度合い**」）は、個別の共分散行列と共通の共分散行列をどのように混合するかを示します。値1は線形の判別法、0は2次の判別法に該当します。
- 第2のパラメータ（「**γ: 対角行列に近づける度合い**」）は、非対角要素、つまり変数間での共分散を収縮させる度合いを示します。値を1にすると、共分散行列が対角行列になります。

そのため、ラムダ（λ）とガンマ（γ）を0に指定した場合は、2次判別分析と同じ結果になります。同様に、ラムダ（λ）を1、ガンマ（γ）を0にすれば、線形判別分析が実行されます。正則化を指定する際、表5.1を参考にしてください。線形、2次、および正則化の判別法の例については、図5.7を参照してください。

表 5.1 正則化判別分析

ラムダが小さい	ラムダが大きい	ガンマが小さい	ガンマが小さい
共分散行列が異なる	共分散行列が同じ	変数間に相関がある	変数間に相関がない
行が多い	行が少ない		
変数が少ない	変数が多い		

共分散行列の縮小

「判別分析」起動ウィンドウには、［共分散行列の縮小］というオプションがあります。このオプションは、標本サイズが小さいグループがある場合に役立ちます。判別分析では、共分散行列の逆行列を計算する必要があります。非対角要素を縮小することで、それらの安定性を改善し、予測におけるばらつきを軽減します。［共分散行列の縮小］オプションは、Schafer and Strimmer（2005）によって説明されている手法で求められた係数によって、非対角要素を縮小します。

起動ウィンドウで、［共分散行列の縮小］オプションを選択し、線形判別を選択した場合は、適切なラムダとガンマの値を設定した正則化判別法が行われます。［共分散行列の縮小］オプションを選択して分析を実行した場合、「縮小率」レポートに「縮小率」と「λ」が表示されます。正則化法を選択して、「正則化パラメータ」ウィンドウでラムダを1に指定し、この「縮小率」レポートの「λ」の値をガンマに指定しすると、同じ結果が得られます。

「判別分析」レポート

「判別分析」レポートには、選択した判別法に基づいた分析結果が表示されます。「判別法」と「分類の列」が、レポートの最上部に表示されます。判別法に［正則化］を選択した場合は、それに関連するパラメータも表示されます。

「判別法」は、赤い三角ボタンのメニューからオプションを選択して変更することもできます。レポート内の結果は、選択した判別法に応じて更新されます。

図 5.8「判別分析」レポートの一例

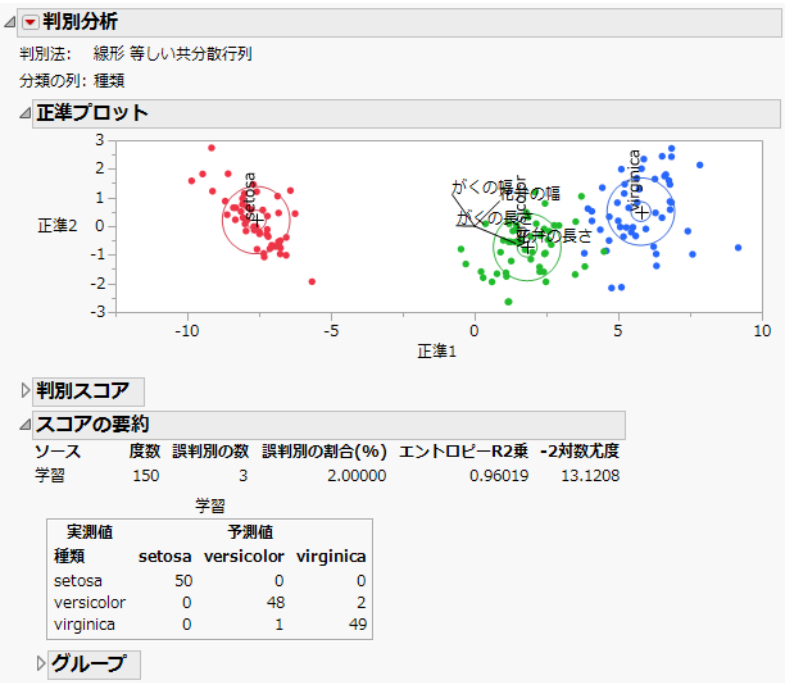


図 5.8 はデフォルトの「判別分析」レポートです。レポートには、次のような結果が表示されます。

- 判別法に［線形 横長データ］を選択した場合、「主成分分析」レポートが表示されます。「主成分分析」（76 ページ）を参照してください。
- 「正準プロット」は、グループを最もよく判別する 2 次元上に点と多変量平均をプロットしたものです。「正準プロットと正準構造」（77 ページ）を参照してください。
- 「判別スコア」レポートには、各オブザベーションがどのように分類されているかの詳細が表示されます。「判別スコア」（80 ページ）を参照してください。
- 「スコアの要約」レポートには、どれほど適切にオブザベーションが分類されているかが示されます。「判別スコア」（80 ページ）を参照してください。

主成分分析

このレポートは、［線形 横長データ］を選択した場合にのみ表示されます。次の表記を使用します。

- 共変量の $n \times p$ 行列を $\mathbf{X}$ とします。ここで、n はオブザベーション数、p は共変量の個数です。
- $\mathbf{X}$ 内の各オブザベーションから共変量平均を引いて、その差を、共変量のプールした標準偏差で割ります。結果の行列を $\mathbf{X_s}$ とします。

レポートには次のものが表示されます。

個数 抽出された固有値の番号。固有値は「累積寄与率」が少なくとも 99.99% になる（つまり、変動の 99.99% が説明される）まで抽出されます。

固有値 $\mathbf{X_s}$ の共分散行列 $(\mathbf{X_s}'\mathbf{X_s})/(n - p)$ に対する固有値。これらの固有値は降順に並べられています。

累積寄与率 すべての固有値の合計に対する、固有値の累積合計の割合を、パーセントで表したもの。なお、すべての固有値の合計は、$\mathbf{X_s}'\mathbf{X_s}$ のランクと等しいです。

特異値 $\mathbf{X_s}$ の特異値。特異値も降順に並べられています。

正準プロットと正準構造

正準プロットは変数の正準相関構造を示すバイプロットです。

正準構造

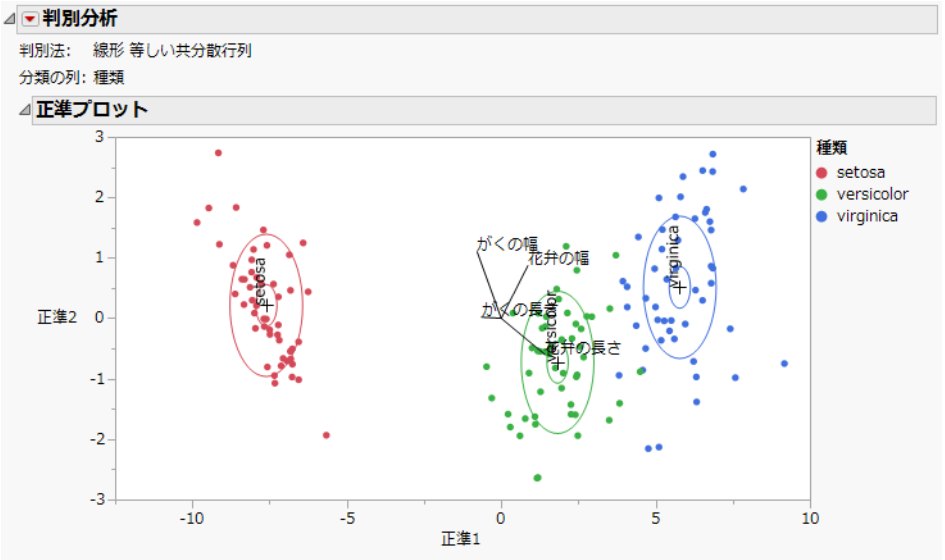
正準プロットでは、［X, カテゴリ］列から作成されるダミー変数（指示変数）の線形結合と、共変量の線形結合との相関が最大になるようなスコアが求められます。そうやって求められた共変量の線形結合（**正準スコア**）がプロットされます。このようにして求められた正準スコアは、グループ間の距離が最大となっています。

第1次元の正準スコアは、グループ（カテゴリ）から作成されたダミー変数の線形結合と、共変量の線形結合との相関が最大化となるときの、共変量の線形結合です。第2次元では、第1次元の正準スコアと直交するという条件のもとで、やはり相関が最大になるような線形結合が求められます。［X, カテゴリ］の列における水準が k 個の場合は、$k - 1$ 次元の正準スコアが求められます。

正準プロット

図5.9は、「Iris.jmp」の線形判別分析の「正準プロット」を示しています。点は、**「種類」**によって色分けされています。

図 5.9「Iris.jmp」の正準プロット



バイプロットの横軸と縦軸は、最初の 2 つの正準変数です。これらの正準変数は、グループ間の距離が最大となる 2 次元です。各正準変数はそれぞれ共変量の線形結合です（「正準構造」（77 ページ）を参照）。このバイプロットを見ると、各オブザベーションがどのように分布しているかや、各共変量が正準変数にどのように寄与しているかがわかります。

- バイプロットでは、各グループの多変量平均と各オブザベーションが点で示されます。これらが、最初の 2 つの正準変数に表現されています。
  - 各多変量平均に対応する点は、プラス記号（+）のマーカーで示されます。
  - 各平均の 95% 信頼楕円がプロットされます。2 つのグループが有意に異なる場合、信頼楕円は交わらない傾向にあります。
  - 各グループの 50% 確率楕円も描かれます。この楕円は、正規分布に従うと仮定したときの（かつ、線形判別分析では、等しい共分散行列に従うと仮定したときの）、データ点のおよそ 50% を含む領域を、2 次元の正準空間に射影したものを表しています。
- バイプロット上に描かれているバイプロット線は、共変量を表しています。
  - 共変量の線形結合（正準変数）の係数は、正準空間を構成するための**「重み」**と解釈できます。
  - この重みを解釈しやすくするために、各共変量は平均が 0、標準偏差が 1 に標準化されます。標準化された共変量に対する係数は、**正準重み**（canonical weight）と呼ばれています。正準重みが大きいほど、その共変量と正準変数との関係が大きいことを示します。
  - バイプロット線の長さと方向は、最初の 2 次元までの正準変数に対する重みを表しています。バイプロット線の長さは、正準重みのノルムに比例しています。
  - また、バイプロット線は、原点 (0,0) を出発点としています。この原点は、全体平均を示しています。

– 重みの数値を知るには、赤い三角ボタンのメニューから［**正準オプション**］>［**正準の詳細を表示**］を選択してください。そして、「正準の詳細」レポートの最下部で「標準化スコア係数」を開いてください。詳細は、「標準化スコア係数」（89ページ）の節を参照してください。

正準プロットの編集

その他のオプションを使ってバイプロットを編集できます。

- 95% 信頼楕円の表示／非表示は、［**正準オプション**］>［**平均の信頼限界楕円の表示**］で切り替えます。
- バイプロット線の表示／非表示は、［**正準オプション**］>［**バイプロット線の表示**］で切り替えます。
- バイプロット線の中心は、ドラッグして別の位置へ移動することができます。バイプロット線の位置およびスケールを指定するには、赤い三角ボタンのメニューから［**正準オプション**］>［**バイプロット線の位置**］を選択します。正準プロットに表示されるデフォルトの半径のスケールは、調整しないとバイプロット線が見えない場合を除き、1.5 です。
- 50% 等高線の表示／非表示は、［**正準オプション**］>［**正規 50% 等高線の表示**］で切り替えます。
- 楕円と一致するように点を色分けするには、赤い三角ボタンのメニューから［**正準オプション**］>［**プロット点の色分け**］を選択します。

3つ以上のカテゴリへの分類

「Iris.jmp」データの場合は3つの**「種類」**があるので、正準変数は2つだけです。図5.9のプロットを見ると、2つの正準変数で3つのグループがいい具合に分かれていることがわかります。

プロット内のバイプロット線は、次のことを示しています。

- **「花弁の長さ」**は、「正準 1」と正の関連性があり、「正準 2」と負の関連性があります。「正準 2」よりも「正準1」の定義における重みのほうが大きいです。
- **「花弁の幅」**は、「正準1」と「正準2」の両方に正の関連性があります。両正準変数の定義における重みは同じくらいです。
- **「がくの幅」**は、「正準1」と負の関連性があり、「正準2」と正の関連性があります。「正準1」よりも「正準2」の定義における重みのほうが大きいです。
- **「がくの長さ」**は、「正準1」と負の関連性があり、「正準2」との関連性はほとんどありません。

2つのカテゴリへの分類

分類変数の水準が2つだけの場合、第1次元の正準変数（「正準1」）だけに点がプロットされます。各共変量の正準重みは「正準1」とだけに関連があります。バイプロット線の縦軸（「正準2」）における座標は、特に意味がありません。横軸（「正準1」）にバイプロット線を射影して、第1次元の正準変数における座標を見てください。

図5.10は、「Fitness.jmp」サンプルデータの正準プロットです。被験者をM（男性）とF（女性）のカテゴリに分類するために、7つの連続変数が使用されます。分類変数には2つのカテゴリしかないので、正準変数は1つだけです。

図 5.10「Fitness.jmp」の正準プロット

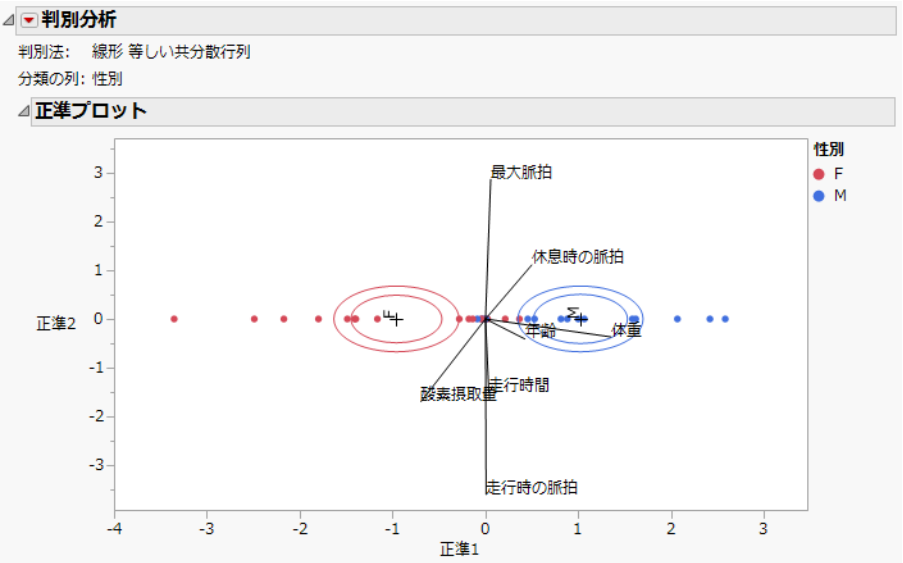


図 5.10 の点は、**「性別」**で色分けされています。2 つのグループが「正準 1」の値ではっきりと分かれていることに注目してください。

7 つの共変量に対応するバイプロット線には縦軸方向にも成分がありますが、この場合は、「正準 1」軸へ射影した座標だけを解釈しなければなりません。次のことを確認してください。

- **「最大脈拍」**、**「走行時間」**、**「走行時の脈拍」**は「正準 1」とあまり関連がありません。
- **「体重」**、**「休息時の脈拍」**、**「年齢」**は「正準 1」と正の関連があり、このうち**「体重」**が最も関連が強いです。共変量**「休息時の脈拍」と「年齢」**は共に関連性が低いです。
- **「酸素摂取量」**は「正準 1」と負の関連性があります。

判別スコア

「判別スコア」レポートでは、各オブザベーションの判別結果と、それを求めるために用いた情報がわかります。

行 データテーブルにおける、そのオブザベーションの行番号

実測値 そのオブザベーションがもつ、分類変数の実測値

平方距離 (実測値) 分類変数の実測値に対する SqDist**[< 水準 >]**の値。詳細については、「スコアオプション」（85 ページ）を参照してください。

確率 (実測値) 分類変数の実測値に所属する確率（の推定値）。

-Log(確率) 「確率(実測値)」の対数の符号を逆にしたもの。この値が大きいものは、「実際に観測されたカテゴリに判別されにくい」という意味でうまく予測されていません。

「-Log(確率)」の数値の横には、その棒グラフが描かれています。棒グラフの棒が長いものは、点の予測がよくないことを表します。また、誤判別されたオブザベーションにはアスタリスク (*) がついています。

検証セットまたはテストセットを使用した場合、検証セット内のオブザベーションには「v」、テストセット内のオブザベーションには「t」の印がつきます。

予測値 分類変数に対する予測値。この予測値は、推定された所属確率が最も大きいカテゴリです。

確率(予測値) 予測されたカテゴリに所属する確率（の推定値）。

その他 予測確率が0.1を超えるその他のカテゴリがあれば、それらのカテゴリをリストします。

図5.11は、「Iris.jmp」サンプルデータの線形判別法による「判別スコア」レポートです。[**スコアオプション**] > [**興味のある行だけを表示**] を選択して、誤判別の行、および、予測確率が0.05～0.95の範囲内の行だけを表示しています。

図5.11 興味のある行だけを表示

判別スコア

行	実測値	平方距離(実測値)	確率(実測値)	-Log(確率)		予測値	確率(予測値)	その他
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

'*'は誤判別されたものを表す

スコアの要約

「スコアの要約」レポートは、判別スコアを要約したものです。図5.12の表は、実測値と予測値との2元表になっています。すべてのオブザベーションが適切に判別されたとき、この表の非対角要素における度数が0になります。

図 5.12「Iris.jmp」のスコアの要約

スコアの要約

ソース	度数	誤判別の数	誤判別の割合 (%)	エントロピーR2乗	-2対数尤度
学習	65	1	1.53846	0.94487	7.81888
検証	49	3	6.12245	0.86644	
テスト	36	0	0.00000	0.98410	

学習

実測値	予測値 度数		
種類	setosa	versicolor	virginica
setosa	18	0	0
versicolor	0	22	1
virginica	0	0	24

検証

実測値	予測値 度数		
種類	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	15	2
virginica	0	1	15

テスト

実測値	予測値 度数		
種類	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	10	0
virginica	0	0	10

「スコアの要約」レポートには、次のような情報が表示されます。

列 ［ステップワイズ変数選択］を使用してモデルを作成した場合、モデルに追加された列がリストされます。図 5.6 を参照してください。

ソース 検証セットを使用していない場合、すべてのオブザベーションが学習セットに使われます。検証セットを使用している場合は、学習セットと検証セットの結果が表示されます。また、テストセットも用いた場合は、学習セット、検証セット、テストセットの結果が表示されます。

誤判別の数 指定のセットで誤判別されたオブザベーションの数。

誤判別の割合 (%) 指定のセットで誤判別されたオブザベーションの割合。

エントロピー R2 乗 適合度の指標。値が大きいほど、あてはまりが良いことを示します。エントロピー R2 乗が 1 の場合、すべての分類が正しく行われたことを意味します。一般に、判別分析モデルによる分類は 100% 的中することはなく、エントロピー R2 乗の値は小さくなる傾向にあります。

詳細については、「エントロピー R2 乗」（83 ページ）を参照してください。

メモ：「エントロピー R2 乗」は負の値である場合もあります。

-2 対数尤度 負の対数尤度を 2 倍したもの。モデルに基づいて、学習セットから算出されます。値が大きいほど、あてはまりが良いことを示します。学習セットに対してのみ計算されます。詳細は、『基本的な回帰モデル』を参照してください。

混同行列 この行列は、分類変数 X の各水準について、実測値と予測値との 2 元表となっています。JMP Pro で検証セットやテストセットを使用した場合、それらのセットに対しても混同行列が表示されます。JMP では、データテーブルで除外した行を使用している場合に、除外した行が検証セットとみなされ、学習セットと検証セットに対する混同行列が表示されます。詳細は、「JMP と JMP Pro の違い」（93 ページ）を参照してください。

エントロピー R2 乗

エントロピー R2 乗は適合度の指標です。検証セットやテストセットを使用している場合には、学習セットだけではなく、検証セットやテストセットに対してもエントロピー R2 乗が計算されます。

学習セットのエントロピー R2 乗

学習セットに対するエントロピー R2 乗は次のように求められます。

- 学習セットだけに対して、判別分析モデルがあてはめられます。
- モデルに基づいた予測確率が求められます。
- これらの予測確率を使って、学習セットの尤度が求められます。これを、$Likelihood\_Full_{Training}$ とします。
- 学習セットだけに対して、減少モデル（共変量を 1 つもたない判別分析モデル）があてはめられます。
- 減少モデルから求められる、X 水準に対する予測確率を使って、学習セットの尤度が計算されます。これを、$Likelihood\_Reduced_{Training}$ とします。
- 学習セットのエントロピー R2 乗は次のように求められます。

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

検証およびテストセットのエントロピー R2 乗

検証セットに対するエントロピー R2 乗は次のように求められます。

- 学習セットだけに対して、判別分析モデルがあてはめられます。
- 学習セットだけから推定されたモデルに基づき、検証セットで予測確率が求められます。
- これらの予測確率を使って、検証セットの尤度が求められます。これを、$Likelihood\_Full_{Validation}$ とします。
- 学習セットだけに対して、減少モデル（共変量を 1 つもたない判別分析モデル）があてはめられます。
- 減少モデルから求められる、X 水準に対する予測確率を使って、検証セットの尤度が計算されます。これを、$Likelihood\_Reduced_{Validation}$ とします。
- 検証のエントロピー R2 乗は次のように求められます。

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

テストセットのエントロピー R2 乗は、検証セットのエントロピー R2 乗と同様の方法で求められます。

判別分析のオプション

「判別分析」の赤い三角ボタンのメニューには、次のようなコマンドが表示されます。

ステップワイズ変数選択　ステップワイズ変数選択コントロールパネルの表示／非表示を切り替えます。「ステップワイズ変数選択」（69ページ）を参照してください。

判別分析の手法　判別法を選択します。「判別法」（72ページ）を参照してください。

判別スコア　レポートの「判別スコア」の表示／非表示を切り替えます。

スコアオプション　オブザベーションのスコア計算に関するオプションがあります。スコア計算の式を保存することができます。「スコアオプション」（85ページ）を参照してください。

正準プロット　「正準プロット」の表示／非表示を切り替えます。「正準プロットと正準構造」（77ページ）を参照してください。

正準オプション　正準プロットに関するオプションがあります。「正準オプション」（87ページ）を参照してください。

三次元正準プロット　三次元正準プロットを表示します。このオプションは、カテゴリカル変数Xの水準が4つ以上ある場合のみ使用できます。「三次元正準プロットの例」（90ページ）を参照してください。

事前確率の指定　X変数の各水準の事前確率を指定できます。「事前確率の指定」（91ページ）を参照してください。

グループの追加　一部の点が、データにある既知のグループではなく、新しいグループに属する可能性があるときに使用します。詳細については、「グループの追加」（91ページ）を参照してください。

グループ内共分散行列の表示　次のレポートの表示／非表示を切り替えます。

- プールした群内共分散行列と相関行列を表示する「共分散行列」レポート
- ［2次］と［正則化］の判別法では、群内相関行列を表示する「各グループの相関」レポートが表示されます。

  各グループに対して、群内共分散行列の行列式の対数も表示されます。
- ［2次］の判別法では、「共分散行列」レポートにおける「グループ共分散」アウトラインに、群内共分散行列も表示されます。

［グループ内共分散行列の表示］は、「横長データ」の判別法では使用できません。

グループ平均の表示　各共変量の平均を表示する「グループ平均の表示」レポートの表示／非表示を切り替えます。X変数の各水準の平均と全体平均が表示されます。

判別行列の保存　「判別分析の結果」という名前のスクリプトをデータテーブルに保存します。このスクリプトは、JSLで使用できる次のようなオブジェクトを一覧します。

- 共変量（Y）のリスト
- カテゴリカル変数X

– Xの水準のリスト
– （Xの各水準における）共変量の平均の行列
– プールした共分散行列

［判別行列の保存］は、「横長データ」の判別法では使用できません。「判別行列の保存」（92ページ）を参照してください。

散布図行列 共変量の各ペアに対する散布図を含んだ「散布図行列」レポートを、別のウィンドウに表示します。これは、「散布図行列」プラットフォームで、グループごとに陰影つき確率楕円を描くのと同じです。検証セットが使用されている場合でも、散布図にはすべてのデータがプロットされます。「散布図行列」（92ページ）を参照してください。

ただし、このメッセージも、［線形 横長データ］の判別法では使用できません。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

Byグループのスクリプトを保存 By変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウでBy変数を指定した場合のみ使用可能です。

スコアオプション

［スコアオプション］には、スコアを対象とした次のようなオプションがあります。

興味のある行だけを表示 「判別スコア」レポートに、誤判別された行と、予測確率が0.05～0.95の行だけを表示します。

度数の表示 「スコアの要約」レポートにおいて、混同行列の表示／非表示を切り替えます。混同行列は、カテゴリカル変数Xの各水準において、実測値と予測値の度数を示したものです。デフォルトの「スコアの要約」レポートには、混同行列が表示されます。JMP Proで検証セットやテストセットを使用した場合、それらのセットに対しても混同行列が表示されます。また、JMPで除外した行を使用している場合、除外した行が検証セットとみなされ、検証セットに対する混同行列が表示されます。詳細は、「JMPとJMP Proの違い」（93ページ）を参照してください。

各グループ平均への距離を表示 各オブザベーションから各グループ平均までのMahalanobisの距離を2乗したものを示す「各グループ平均への距離の2乗」レポートの表示／非表示を切り替えます。

各グループに属する確率を表示 カテゴリカル変数Xの各グループに、各オブザベーションが属する確率を示す「各グループに属する確率」レポートの表示／非表示を切り替えます。

ROC 曲線 「スコアの要約」レポートにROC曲線を追加します。ROC曲線の詳細については、『予測および発展的なモデル』の「パーティション」章を参照してください。

誤判別された行を選択 誤判別された行を、データテーブル内と、判別スコア内で選択します。

不確実な行を選択 判別が不確実な行を、データテーブル内と、判別スコア内で選択します。「判別が不確実な行」とは、**いずれの**グループに属する確率も0や1に近くない行です。

このオプションを選択すると、ウィンドウが表示され、そこで不確実さを示す予測確率の範囲を指定できます。デフォルトでは、予測確率が0または1から0.1以上離れている行が不確実な行とされます。したがって、デフォルトでは、0.1～0.9の間の確率の行が選択されます。

計算式の保存 距離、確率、および、どのグループに判別されるかの予測値を求める計算式を、データテーブルに保存します。詳細については、「保存される計算式」（94ページ）を参照してください。

- **SqDist[0]** および **SqDist[< 水準 >]** は、距離の計算式です。ここで、「< 水準 >」はXの水準です。距離の計算式は、Mahalanobisの距離に関連した計算式です。
- **Prob[< 水準 >]** は、確率の計算式です。ここで、「< 水準 >」はXの水準です。この確率は、Xの各水準にオブザベーションが属する事後確率です。確率の各列には「応答確率」列プロパティが保存されます。「応答確率」列プロパティについての詳細は、『JMPの使用法』を参照してください。
- **Pred <X>** は、予測値の計算式です。「最も属する確率が高い水準」を求めるための計算式になっています。
- ［線形 横長データ］の判別法では、共変量のベクトルと、**判別のための主成分**の計算式を含む**「判別データ行列」**列も保存します。「横長データに対する線形判別法」（98ページ）を参照してください。

メモ：［線形 横長データ］以外のすべての判別法では、計算式を保存すると、RowEdit Probスクリプトがデータテーブルに保存されます。このスクリプトはデータテーブル内の不確実な行を選択します。不確実な行と定義されるのは、予測確率が0または1から0.1以上離れている行です。このスクリプトが開く「行の編集」ウィンドウでは、不確実な行を調べることができます。（［線形 横長データ］以外の手法で）新しい判別分析を実行し、［計算式の保存］を選択した場合、既存のRowEdit Probスクリプトは、新しい判別分析の結果によって上書きされます。

計算スクリプトの作成 計算式列を作成するスクリプトを作成します。これらの計算式列は、［計算式の保存］オプションによって保存されるものです。このスクリプトを保存しておけば、他のデータテーブルにおいて、各グループに属する確率を計算する計算式列や、どのグループに属するかを予測する計算式列を作成できます。（標準版のJMPのみ）

JMP PRO **確率の計算式を発行** 確率の計算式を作成し、それを「計算式デポ」レポート内の計算式列スクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションを選択した時点でレポートが作成されます。『予測および発展的なモデル』の「計算式デポ」章を参照してください。

正準オプション

以下に述べるオプションのうち、最初のほうのものは、正準プロットや三次元正準プロットの外観に関するオプションです。その他のオプションは、プロットに関連する計算の詳細を表示するものです。

メモ: ［三次元正準プロット］は、共変量が3つ以上あり、グループ変数に4つ以上のカテゴリがある場合にのみ使用できます。

プロットの外観に関するオプション

点の表示 正準プロット内および三次元正準プロット内の点の表示／非表示を切り替えます。

平均の信頼限界楕円の表示 正準変数の平均に対する95%信頼楕円の表示／非表示を切り替えます。この信頼楕円は、正規分布に従うと仮定して求められています。また、三次元正準プロットでも、95%信頼楕円の表示／非表示を切り替えます。

正規50%等高線の表示 各グループにおける50%等高線を示す楕円の表示／非表示を切り替えます。各楕円は、正規分布に従うと仮定したときに、最初の2つの正準変数においてオブザベーションの約50%を含む領域を示しています。同様に、三次元正準プロットでの各楕円は、正規分布に従うと仮定したときに、最初の3つの正準変数においてオブザベーションの約50%を含む領域を示しています。

バイプロット線の表示 正準プロット内および三次元正準プロット内のバイプロット線の表示／非表示を切り替えます。バイプロット線は、正準空間における共変量の方向を示します。バイプロット線は、共変量の各正準変数に対する関連の度合を示します。

バイプロット線の位置 正準プロット内および三次元正準プロット内のバイプロット線の位置と半径のスケールを指定します。

- デフォルトでは、バイプロット線は全体平均を示す点(0,0)を出発点とした直線で表されています。正準プロットでバイプロット線をドラッグして移動するか、またはこのオプションによって出発点の座標を指定できます。
- 正準プロットに表示されるデフォルトの半径のスケールは、バイプロット線を表示するために調整の必要が生じない限り、1.5です。半径のスケールは、標準化スコア係数に対するものです。

プロット点の色分け 正準プロットおよび三次元正準プロットにおいて、*X*変数の水準ごとに色を付けます。また、このとき、データテーブルの行に対しても、色のマーカーが設定されます。これは、**［行］>［列の値による色/マーカー分け］**で*X*変数の列を選択するのと同じです。また、グラフを右クリックして**［行の凡例］**を選択し、X変数の水準ごとに色を設定するのとも同じです。

計算に関するオプション

正準の詳細を表示 「正準の詳細」レポートの表示／非表示を切り替えます。「正準の詳細を表示」(88ページ)を参照してください。

正準構造の表示 「正準構造の表示」レポートの表示／非表示を切り替えます。「正準構造の表示」(89ページ)を参照してください。ただし、このメッセージも、［線形 横長データ］の判別法では使用できません。

正準スコアの保存 データテーブルに新しい列を作成し、各オブザベーションの正準スコア計算式を保存します。*k* 番目の正準スコアの列は、**Canon[<k>]** という名前になります。

ヒント : スクリプトの場合は、Discriminant オブジェクトに **Save to New Data Table** コマンドを送ると、正準変数のグループ平均、標準化スコア係数の半径スケール 1.5 のバイプロット線、および正準スコアが新しいデータテーブルに保存されます。ただし、このメッセージも、[線形 横長データ] の判別法では使用できません。

正準の詳細を表示

「正準の詳細」レポートには、共変量とグループ変数 *X* の関係を示す検定が表示されます。レポートの下部に、関連する行列が表示されます。

図 5.13「Iris.jmp」の正準の詳細

正準の詳細

グループ内共分散行列から計算された正準の詳細

固有値	寄与率	累積寄与率	正準相関	尤度比	近似のF検定	分子自由度	分母自由度	p値(Prob>F)
32.1919292	99.1213	99.1213	0.98482089	0.02343863	199.1453	8	288	<.0001*
0.28539104	0.8787	100.0000	0.47119702	0.77797337	13.7939	3	145	<.0001*

検定	値	近似のF検定	分子自由度	分母自由度	p値(Prob>F)
Wilksのλ	0.0234386	199.1453	8	288	<.0001*
Pillaiのトレース	1.1918988	53.4665	8	290	<.0001*
Hotelling-Lawley	32.47732	582.1970	8	203.4	<.0001*
Royの最大根	32.191929	1166.9574	4	145	<.0001*

▷ グループ内共分散行列
▷ グループ間共分散行列
▷ スコア係数
▷ 標準化スコア係数

メモ : この正準分析で使用される共分散行列は、常にプールした群内共分散行列（グループ内共分散行列）です。どの判別法を選択したとしても、「正準の詳細」レポートでは、プールした群内共分散行列が常に使われます。そのため、「正準の詳細」レポートの統計量や検定は、すべての判別法で共通です。

統計量と検定

「正準の詳細」レポートには、固有値がリストされ、「固有値がゼロである」という帰無仮説に対する尤度比検定が表示されます。また、「正準相関がゼロである」という帰無仮説に対して、4 つの検定が表示されます。

固有値 グループ間共分散行列（群間共分散行列）と、グループ内共分散行列（群内共分散行列）の逆行列を掛け合わせた行列の固有値。固有値は、大きい順に表示されます。固有値の大きさは、その次元における、判別で説明される分散の量を反映しています。

寄与率 固有値の合計に対する、該当の固有値が占める割合。

累積寄与率 寄与率の累積合計。

正準相関 カテゴリカル変数 X のグループと、共変量との間の正準相関。まず、X のグループから、指示変数（ダミー変数）を作成します。そして、一方の変数の組を指示変数とし、もう一方の変数の組を共変量として、正準相関分析を行います。「正準相関」に表示されている値は、この正準相関分析における正準相関の値です。

尤度比 現在の次元以降の母正準相関がすべてゼロかどうかを調べる検定の尤度比統計量。この尤度比統計量は、現在の次元以降に関して、(1 - 正準相関^2) を掛け合わせたものです。

検定 「共変量の平均はグループ間で等しい」という帰無仮説に対する検定で、Wilks の Λ、Pillai のトレース、Hotelling-Lawley のトレース、および Roy の最大根の 4 つが計算されます。「多変量検定」（101 ページ）および「線形判別分析」の付録の「近似 F 検定」（102 ページ）を参照してください。

近似の F 検定 対応する検定の F 値。一部の検定では、F 値は近似値または上限値です。「線形判別分析」の付録の「近似 F 検定」（102 ページ）を参照してください。

分子自由度 対応する検定の分子自由度。

分母自由度 対応する検定の分母自由度。

p 値 (Prob>F) 対応する検定の p 値。

行列

レポートの下部に、正準構造に関連する 4 つの行列が表示されます。行列を表示するには、それぞれの名前の横にある開閉アイコンをクリックしてください。また、非表示にするには、行列の名前をクリックしてください。

グループ内共分散行列 プールしたグループ内共分散行列（群内共分散行列）。

グループ間共分散行列 グループ間共分散行列（群間共分散行列）、$\mathbf{S}_B$。「グループ間の共分散行列」（103 ページ）を参照してください。

スコア係数 生データから正準スコアを計算する際に使用する係数。この係数が、**[正準オプション] > [正準スコアの保存]** オプションに使用されます。これらの計算方法については、SAS Institute Inc. の「The CANDISC Procedure」（2011）を参照してください。

標準化スコア係数 標準データから正準スコアを計算する際に使用する係数。この係数は、一般に、**正準重み**（canonical weight）と呼ばれています。これらの計算方法については、SAS Institute Inc. の「The CANDISC Procedure」（2011）を参照してください。

正準構造の表示

「正準構造」レポートには、正準変数と共変量との間の相関を示す 3 つの行列が表示されます。また、グループ変数の各水準における平均も表示されます。行列を表示するには、それぞれの名前の横にある開閉アイコンをクリックしてください。また、非表示にするには、行列の名前をクリックしてください。

図 5.14「Iris.jmp」の相関構造

全体の正準構造　正準変数と共変量との間の相関。**負荷量**ともいいます。

グループ間正準構造　正準変数のグループ平均と共変量のグループ平均との間の相関。

プールしたグループ内正準構造　グループ変数によって調整された、正準変数と共変量との間の偏相関。

正準変数のクラス平均　グループ変数の各水準における、各正準変数の平均。

三次元正準プロットの例

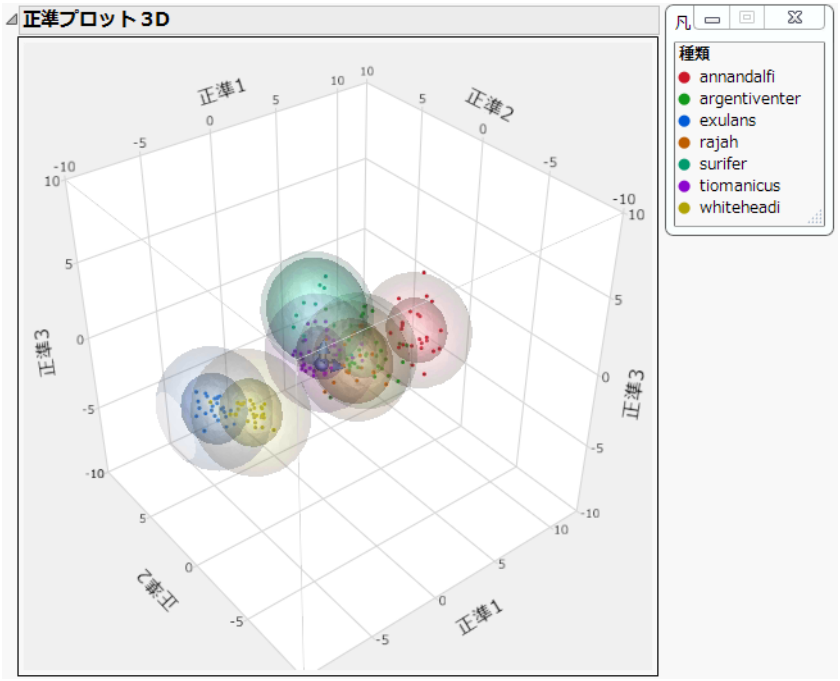
1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Owl Diet.jmp」を開きます。
2. 180 行目から 294 行目までを選択します。

   これらの行は「**種類**」の値が欠測値となっています。これらの選択した行に、非表示かつ除外の行属性を設定します。
3. ［**行**］>［**非表示かつ除外**］を選択します。
4. ［**行**］>［**列の値による色/マーカー分け**］を選択します。
5. 「**種類**」を選択します。
6. 「色」メニューから、［**JMP ダーク**］を選択します。
7. ［**凡例のウィンドウを表示**］にチェックマークをつけます。
8. ［**OK**］をクリックします。

   小さい凡例ウィンドウが表示されます。データテーブル内の行には、「**種類**」別に色が割り当てられています。
9. ［**分析**］>［**多変量**］>［**判別分析**］を選択します。
10. 「**頭蓋長**」、「**歯列長**」、「**口蓋孔**」、「**顎長**」を［**Y, 共変量**］に指定します。
11. 「**種類**」を［**X, カテゴリ**］に指定します。
12. ［**OK**］をクリックします。
13. 「判別分析」の赤い三角ボタンのメニューから、［**三次元正準プロット**］を選択します。

ヒント:「凡例」内のカテゴリをクリックすると、三次元正準プロット内の対応する点が強調表示されます。三次元正準プロット内をクリックしてドラッグすると、プロットが回転します。

図 5.15 三次元正準プロットと凡例ウィンドウ



事前確率の指定

事前確率を指定するのに、次のようなオプションが用意されています。

等しい確率　すべてのグループに等しい事前確率を割り当てます。このオプションはデフォルトです。

発生頻度に比例　観測されたデータ内の発生頻度に比例する事前確率をグループに割り当てます。

その他　任意の事前確率を指定できます。

グループの追加

判別結果の一部が、データに存在しているグループには属さずに、別のグループに属すると考えられる場合には、このオプションを使用してください。このオプションを選択すると、追加する水準の事前確率を指定するためのウィンドウが表示されます。

追加された新しいグループに属する確率のほうが高いと考えられるオブザベーションが、その新しいグループに割り当てられます。この新しいグループは、「その他」と呼ばれます。「その他」グループに属する確率は、**「データはグループごとに分かれていない」**と仮定したときの分布から求められます。この分布は、特定の共分散行列をもつ正規分布であり、より広い領域に広がった等高線になります。なお、指定した事前確率によって、距離の計算は調整されます。

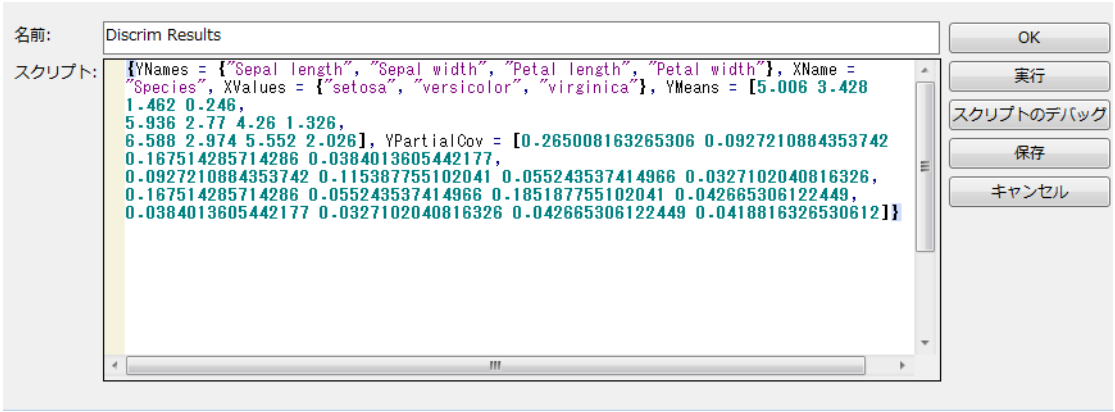
判別行列の保存

［**判別行列の保存**］を選択すると、JMP スクリプト言語で使用するグローバルのリスト（DiscrimResults）が作成されます。このリストには、学習セットから計算された次のものが含まれます。

- YNames。共変量（Y）のリスト
- XName。カテゴリカル変数
- XValues。X の水準のリスト
- YMeans。（X の水準ごとの）共変量の平均の行列
- YPartialCov。グループ内共分散行列（群内共分散行列）

「Iris.jmp」サンプルデータ内の「**判別分析**」スクリプトで得た分析結果を検討してみましょう。赤い三角ボタンのメニューから［**判別行列の保存**］を選択すると、データテーブルに「**判別分析の結果**」という名前のスクリプトが保存されます。図 5.16 は、このスクリプトを示しています。

図 5.16「Iris.jmp」の「判別分析の結果」スクリプト



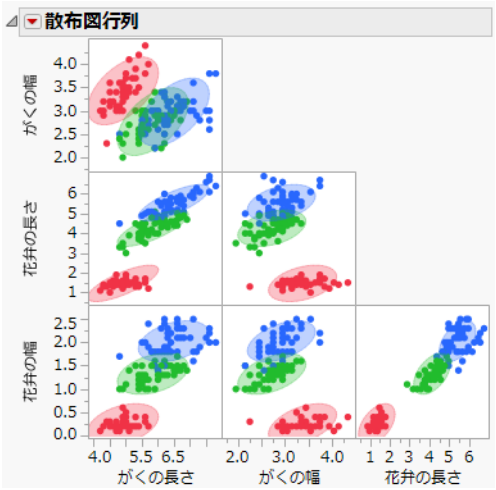
メモ: スクリプト内で、Discriminant プラットフォームオブジェクトに Get Discrim Matrices コマンドを送ることができます。この場合、［判別行列の保存］を使用するのと同様の結果が得られますが、データテーブルには保存されません。

散布図行列

「散布図行列」コマンドを選択すると、共変量の各ペアに対する下三角の散布図行列を含むウィンドウが別に表示されます。この散布図では、データテーブルのすべてのオブザベーションが点で表示されます。

カテゴリカル変数 X の各グループに対して、90% の領域を表す楕円が描かれます。線形判別分析の場合、楕円は、プールした群内共分散行列から計算されています。図 5.17 は、「Iris.jmp」サンプルデータの「散布図行列」ウィンドウです。

図 5.17「Iris.jmp」の散布図行列



レポートの赤い三角ボタンメニューのオプションについては、『グラフ機能』を参照してください。

JMP と JMP Pro の違い

JMP で検証セットを用いたい場合には、検証セットとしたい行を除外します。それには、検証セットとしたい行を選択し、[行] > [除外する/除外しない] を選択します。このとき、除外していない行が学習セットとして使われます。

メモ: JMP Pro では、「判別分析」起動ウィンドウで検証列を指定します。検証列のデータタイプは数値でなければならず、また、異なる値が少なくとも 2 つ含まれている必要があります。

JMP Pro で検証列を用いた場合は、次のように検証セットが設定されます。

- 検証列の値が 2 つしかない場合は、小さい方の値が学習セット、大きい方の値が検証セットとして扱われます。
- 値が 3 つある場合は、値の小さい方から順に、学習セット、検証セット、テストセットとして扱われます。
- 値が 3 つ以上ある場合は、最も小さい 3 つの値以外は無視され、小さい値から順に学習セット、検証セット、テストセットになります。

検証セットを指定した場合、「判別分析」プラットフォームは次のことを実行します。

- 学習データを使用してモデルをあてはめます。
- [ステップワイズ変数選択] オプションは、モデルの「検証 エントロピー R2 乗」と「検証 誤分類率」の統計量を表示します。詳細については、「統計量」(70 ページ) および 「検証およびテストセットのエントロピー R2 乗」(83 ページ) を参照してください。

- ［判別スコア］レポートには、検証セットとテストセットの行を識別する印が表示されます。
- 「スコアの要約」レポートには、学習セット、検証セット、テストセットに関して、実測値と予測値の混同行列が表示されます。

技術的詳細

「線形 横長データ」のアルゴリズムについて

「線形 横長データ」オプションを選択したとき、次のように分析が実行されます。

- グループ平均を引き、プールした標準偏差で割ることによってデータを標準化します。
- 判別分析用の主成分（スコア）を求めるために、特異値分解によって、特異ベクトルを算出します。
- 特異値の平方和が99.99%となる次元までを用います。
- グループ平均によってシフトされていないデータに変換します。その変換されたデータに対して、線形判別分析を実行します。こうして変換されたデータでは、プールされた群内分散行列が対角行列となるため、計算が早く済みます。

保存される計算式

ここでは、［**スコアオプション**］>［**計算式の保存**］で保存される計算式について説明します。計算式は判別法によって異なります。

カテゴリカル変数Xによって定義される各グループについて、共変量のオブザベーションは、p次（pは共変量の数）の多変量正規分布に従うと仮定されます。計算式で使用される記号は、表5.2のとおりです。

表5.2［計算式の保存］で保存される計算式の記号

p	共変量の数
T	グループの総数（Xの水準数）
$t = 1, ..., T$	Xによって定義されるグループを示す添え字
n_t	グループt内のオブザベーション数
$n = n_1 + n_2 + ... + n_T$	オブザベーションの総数
$\mathbf{y}$	オブザベーションの共変量のp x 1ベクトル
$\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \cdots, y_{ipt})$	p個の共変量の値で構成される、グループtにおけるi番目のオブザベーションのベクトル
$\bar{\mathbf{y}}_t$	グループtにおける共変量$\mathbf{y}$の平均を示すp x 1ベクトル

表 5.2［計算式の保存］で保存される計算式の記号（**続き**）

$\mathbf{y}_{bar}$	データ全体における共変量の平均を示す p x 1 ベクトル
$\mathbf{S}_t = \frac{1}{n_t - 1} \sum_{i=1}^{n_t} (\mathbf{y}_{it} - \bar{\mathbf{y}}_t)(\mathbf{y}_{it} - \bar{\mathbf{y}}_t)'$	グループ t における群内共分散行列。p x p の行列。
$\mathbf{S}_p = \frac{1}{n - T} \sum_{t=1}^{T} (n_t - 1)\mathbf{S}_t$	プールした群内共分散行列。p x p の行列。
q_t	グループ t に属する事前確率
$p(t\|\mathbf{y})$	$\mathbf{y}$ がグループ t に属する事後確率
$\|\mathbf{A}\|$	行列 A の行列式

線形判別法

線形判別法では、「群内共分散行列はすべてのグループで等しい」と仮定されます。この共通した共分散行列は、$\mathbf{S}_p$ と推定されます。以下の式で用いている記号については、表5.2を参照してください。

オブザベーション $\mathbf{y}$ からグループ t への Mahalanobis の距離は、次のように定義されます。

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

グループ t 内のオブザベーション $\mathbf{y}$ の尤度の推定値は、次のように求められます。

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

推定されるパラメータの個数は、プールした共分散行列における $p(p+1)/2$ 個と、平均ベクトルにおける Tp 個です。推定されるパラメータの総数は、$p(p+1)/2 + Tp$ 個です。

グループ t に属する事後確率は、次のように求められます。

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

オブザベーション $\mathbf{y}$ は、事後確率の値が最も大きいグループに割り当てられます。

線形判別法で保存される計算式は、次のように定義されます。

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*>	$t = 1, \ldots, T$ に関して、$p(t\|\mathbf{y})$ が最大となるような t

2 次判別法

2 次判別法では、「グループごとに群内共分散行列が異なる」と仮定されます。グループ t における群内共分散行列は、$\mathbf{S}_t$ と推定されます。つまり、推定されるパラメータの個数は、群内共分散行列における $Tp(p+1)/2$ 個と、平均ベクトルにおける Tp 個です。推定されるパラメータの総数は、$Tp(p+3)/2$ 個です。

グループの標本サイズが p と比べて小さい場合、群内共分散行列の推定値はかなり不安定になります。そして、判別スコアは、群内共分散行列の逆行列における最小固有値から大きな影響を受けます。Friedman（1989）を参照してください。そのため、グループの標本サイズが p に比べて小さい場合は、「正則化判別法」（97 ページ）で説明されている正則化判別法を用いることを検討してください。

以下の式で用いている記号については、表 5.2 を参照してください。オブザベーション $\mathbf{y}$ からグループ t への Mahalanobis の距離は、次のように定義されます。

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

グループ t 内のオブザベーション $\mathbf{y}$ の尤度の推定値は、次のように求められます。

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

グループ t に属する事後確率は、次のように求められます。

$$p(t|\mathbf{y}) = (q_t l_t(\mathbf{y}))/\left(\sum_{u=1}^{T} q_u l_u(\mathbf{x})\right)$$

$$= \frac{1}{1 + \sum_{u \neq t} \exp(-[(d_u^2 + \log|\mathbf{S}_u| - 2\log(q_u)) - (d_t^2 + \log|\mathbf{S}_t| - 2\log(q_t))]/2)}$$

オブザベーション $\mathbf{y}$ は、事後確率の値が最も大きいグループに割り当てられます。

2 次判別法で保存される計算式は、次のように定義されます。

SqDist[<group t>]	$d_t^2 + \log\|\mathbf{S}_t\| - 2\log(q_t)$
Prob[<group t>]	$p(t\|\mathbf{y})$
Pred <X>	$t = 1, \ldots, T$ に関して、$p(t\|\mathbf{y})$ が最大となるような t

メモ: SqDist[<group t>] は負になる場合もあります。

正則化判別法

正則化判別法では、λ と γ の 2 つのパラメータを使用します。

- パラメータ λ は、プールして計算された群内共分散行列と、（グループごとに異なると仮定されて）各グループごとに計算された群内共分散行列との重みのバランスを取ります。
- パラメータ γ は、対角行列への縮小の度合いを決定します。

正則化判別法では、上記した 2 つの正則化によって、2 次判別分析の推定結果を安定させます。Friedman（1989）を参照してください。以下の式で用いている記号については、表 5.2 を参照してください。

正則化判別法の場合、グループ t の共分散行列は次のように求められます。

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

オブザベーション $\mathbf{y}$ からグループ t への Mahalanobis の距離は、次のように定義されます。

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

グループ t 内のオブザベーション $\mathbf{y}$ の尤度の推定値は、次のように求められます。

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

グループ t に属する事後確率は、次のように求められます。

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

オブザベーション $\mathbf{y}$ は、事後確率の値が最も大きいグループに割り当てられます。

正則化判別法で保存される計算式は、次のように定義されます。

SqDist[<group *t*>]	$d_t^2+\log\left\|\mathbf{\Sigma}_t\right\|-2\log(q_t)$
Prob[<group *t*>]	$p(t\|\mathbf{y})$
Pred <*X*>	$t = 1, \ldots, T$ に関して、$p(t\|\mathbf{y})$ が最大となるような t

メモ: SqDist[<group *t*>] は負になる場合もあります。

横長データに対する線形判別法

［線形 横長データ］オプションによって実行される判別法は、共変量の個数が多い場合、特に、共変量の個数がオブザベーション数より多い場合（$p > n$）に役立ちます。この手法では、プールした群内共分散行列 $\mathbf{S}_p$ の逆行列やその転置行列を、$p > n$ の場合に計算負荷がない方式で計算します。特異値分解によって、大規模な共分散行列の逆行列を計算することを回避します。

［線形 横長データ］の判別法では、「すべてのグループにおける群内共分散行列は等しい」と仮定します。オブザベーション数が共変量の個数と等しいかそれ以上の場合、この手法は線形判別法とまったく同じです。

横長データに対する線形判別法の計算式

以下の式で用いている記号については、表5.2を参照してください。［線形 横長データ］の判別法は、以下の手順で算出されています。

1. 各グループの標本平均を含んだ、$T \times p$行列$\mathbf{M}$を計算します。$\mathbf{M}$の(t,j)番目の要素m_{tj}は、グループtにおける、j番目の共変量の標本平均です。
2. 各共変量jについて、グループ全体のプールした標準偏差を計算します。これを、s_{jj}とします。
3. 対角要素s_{jj}を持つ対角行列を$\mathbf{S}_{diag}$とします。
4. 各共変量の値を、次のようにして中心化および尺度化します。
   - オブザベーションが属するグループの平均を引きます。
   - 差を、プールした標準偏差で割ります。

   これを式で表すと、グループtに属するオブザベーションiの、j番目の共変量を標準化した値は、次式のようになります。

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

   この式で、$t(i)$は、オブザベーションiが属するグループtを示します。
5. y^{*}_{ij}の値の行列を$\mathbf{Y}_s$とします。
6. グループで標準化した共変量から計算された、プールした群内共分散行列を$\mathbf{R}$とします。この行列$\mathbf{R}$は、次のように表せます。

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

7. $\mathrm{Y}_{\mathbf{s}}$を特異値分解します。

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

   この式で、$\mathbf{U}$と$\mathbf{V}$の各ベクトルは、正規直交しています。また、$\mathbf{D}$は、対角要素が正の特異値となっている対角行列です。「統計的詳細」の付録の「特異値分解」（210ページ）を参照してください。

   $\mathbf{R}$は次のように表せます。

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

8. $\mathbf{R}$がフルランクの場合には、$\mathbf{R}^{-1/2}$は次のように表せます。

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

   この式で、$\mathbf{D}^{-1}$は、$\mathbf{D}$の対角要素の逆数を対角要素にもつ対角行列です。

R がフルランクではない場合、**R** の疑似逆行列は次のように定義されます。

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

これにより、**R** の平方根の逆数に相当する行列を、次のように定義します。

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

9. **R** がフルランクの場合には、$\mathbf{R}^{-} = \mathbf{R}^{-1}$ です。そこで、どのような場合でも式が使えるように、常に疑似逆行列を使用します。

   ここで p x p の行列 $\mathbf{T}_s$ を次のように定義します。

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

   このとき、次のような式が成立します。

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

   この式で、$\mathbf{S}_p^{-}$ は、元データのプールされた群内共分散行列の一般化逆行列です。これは上式により、特異値分解で計算できます。

Mahalanobis の距離

Mahalanobis の距離、尤度、および事後確率の計算式は、「線形判別法」（95 ページ）と同じです。ただし、$\mathbf{S}_p$ の逆行列には、特異値分解によって算出された一般化逆行列が使われます。

計算式を保存すると、Mahalanobis の距離は分解によって求められます。オブザベーション **y** のグループ t までの距離は、次のようにして求められます。最後の等式における *SqDist*[0] と *「判別主成分」* は、「保存される計算式」（101 ページ）で定義されているものです。

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

保存される計算式

［線形 横長データ］オプションの判別分析で保存される計算式は、次のとおりです。

判別データ行列	共変量のベクトル
判別主成分	主成分スコアを求める行列によって変換されたデータ。このデータは、グループ内で相関していないデータになります。$\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}})$ によって求められます。この式で、$\bar{\mathbf{y}}$ は、全体平均を表す p x 1 ベクトルです。
SqDist[0]	$(\mathbf{y}-\bar{\mathbf{y}})'\mathbf{T}'_s\mathbf{T}_s'(\mathbf{y}-\bar{\mathbf{y}})$
SqDist[<group *t*>]	オブザベーションからグループの重心までの Mahalanobis の距離。「Mahalanobis の距離」（100 ページ）を参照してください。
Prob[<group *t*>]	$p(t\|\mathbf{y})$。「線形判別法」（95 ページ）を参照してください。
Pred <*X*>	$t = 1, \ldots, T$ に関して、$p(t\|\mathbf{y})$ が最大となるような t

多変量検定

以下のセクションにおいて、**E** は残差交差積行列、**H** はモデル交差積行列です。**E** の対角要素は、各変数の残差平方和です。**H** の対角要素は、各変数のモデルの平方和です。判別分析に関する文献では、**E** は、「within」の頭文字を取って **W** と呼ばれることもあります。

多変量検定の結果表に表示される統計量は、$\mathbf{E}^{-1}\mathbf{H}$ の固有値 λ の関数です。次のリストは、各検定統計量の計算方法をまとめたものです。

メモ : 応答に対する計画行列を指定すると、**E** 行列と **H** 行列の前から **M**' が、後ろから **M** が掛けられます。

- Wilks の λ

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

- Pillai のトレース

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

- Hotelling-Lawley のトレース

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

- Roy の最大根

  $\Theta = \lambda_1$、$\mathbf{E}^{-1}\mathbf{H}$ の最大固有値

E と **H** は次のように定義されます。

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

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

ここで、**b** はモデル係数の推定されたベクトル、$\mathbf{A}^{-}$は、行列 **A** の一般化逆行列です。

モデル全体の **L** 行列は、（切片に対する）ゼロの列に、モデルパラメータと同数の行と列を持つ単位行列を連結したものです。各効果の **L** 行列は、モデル全体の **L** 行列から、該当する行を抜き出したものです。

近似 F 検定

次表に示す F 値と自由度を計算において、p を $\mathbf{H}+\mathbf{E}$ のランクとします。また、q を $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$ のランクとします。ここで、**L** 行列は、$\mathbf{X}'\mathbf{X}$ の要素のうち、どこを検定するかを特定するための行列です。v を誤差の自由度とし、s を p と q のうちの小さい方とします。さらに、$m = 0.5(|p-q|-1)$ と $n = 0.5(v-p-1)$ とします。

表5.3（102ページ）で示される F 値は、近似的に F 分布に従います。

表 5.3 近似 F 統計量

検定	近似 F	分子自由度	分母自由度
Wilks のラムダ	$F = \left(\frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\right)\left(\frac{rt-2u}{pq}\right)$	pq	$rt-2u$
Pillai のトレース	$F = \left(\frac{V}{s-V}\right)\left(\frac{2n+s+1}{2m+s+1}\right)$	$s(2m+s+1)$	$s(2n+s+1)$
Hotelling-Lawley のトレース	$F = \frac{2(sn+1)U}{s^2(2m+s+1)}$	$s(2m+s+1)$	$2(sn+1)$
Roy の最大根	$F = \frac{\Theta(v-\text{最大}(p,q)+q)}{\text{最大}(p,q)}$	最大(p,q)	$v-\text{最大}(p,q)+q$

グループ間の共分散行列

グループ間の共分散行列（群間共分散行列）は、表5.2の記号によって、次のように表されます。

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

第 6 章

PLS 回帰

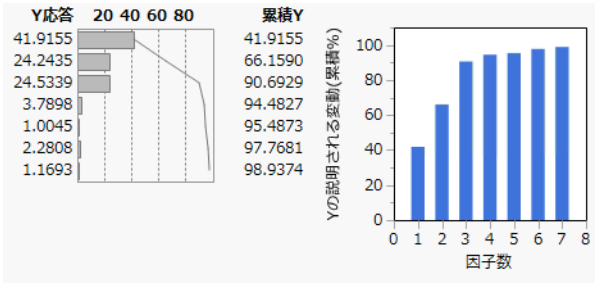
多重共線性がある場合の予測モデル

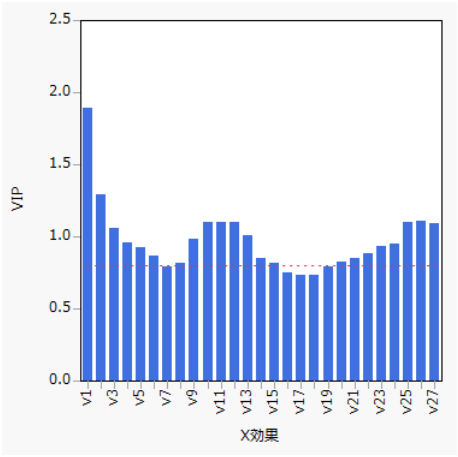
「PLS 回帰」プラットフォームは、説明変数（X）の線形結合からなる因子に基づいて応答変数（Y）を予測する線形モデルを構築します。PLS 回帰における因子は、X の線形結合と Y の線形結合との共分散が最大になるようなものです。PLS 回帰は、X と Y の関係を調べ、潜在的な因子を抽出します。

JMP PRO JMP Pro は、さらに多くの機能を備えています。JMP Pro では、PLS 判別分析（PLS-DA）を行ったり、さまざまなモデル効果を含めたり、複数の検証法を使用したり、欠測データを補完したり、各種統計量に対してブートストラップ推定を行ったりできます。

PLS 回帰は、通常の最小 2 乗法が失敗するような次のような場面で利用することが考えられます。X 変数の個数がデータの行数よりも多い場合、X 変数の間に高い相関がある場合、X 変数の個数が多い場合、および、Y 変数の個数が多い場合などです。

図 6.1「PLS 回帰」レポートの一部





「PLS回帰」プラットフォームの概要

PLS回帰は、通常の最小2乗とは異なり、説明変数の個数がデータの行数よりも多い場合でも使用できま す。PLS回帰は、分光測定、計量化学、ゲノミクス、心理学、教育学、経済学、政治学、環境科学といった分野で、多変量データをモデル化するのに幅広く使用されています。

PLS回帰は、説明変数の個数がデータの行数よりも多い場合や、説明変数の間に高い相関がある場合に特に役立ちます。また、PLS回帰は、複数の応答変数を1つのモデルでモデル化できます。Garthwaite（1994）、Wold（1995）、Wold et al.（2001）、Eriksson et al.（2006）、およびCox and Gaudard（2013）を参照してください。

PLS回帰モデルの手法には、NIPALS（Nonlinear Iterative Partial Least Squares）と、SIMPLS（Statistically Inspired Modification of PLS）の2つがあります。（NIPALSについてはWold, H., 1980を、SIMPLSについてはDe Jong, 1993を参照してください。両手法の説明については、Boulesteix and Strimmer, 2007を参照してください。）SIMPLS法は、目的関数を明確に示して、それを最適化するという考えに基づき、導出された方法です。応答が1つの場合は、どちらの手法も同じ結果となります。応答が複数の場合は、結果は少し異なります。

JMPで「PLS回帰」プラットフォームを開くには、［分析］>［多変量］>［PLS回帰］を選択します。JMP Proでは、その他に、［分析］>［モデルのあてはめ］の［PLS回帰］手法を選択することでも、「PLS回帰」プラットフォームを起動できます。

JMP PRO JMP Proでは、次のことが行えます。

- ［モデルのあてはめ］の［PLS 回帰］手法で応答変数に名義尺度を指定することにより、PLS-DA（PLS判別分析）を行えます。
- ［モデルのあてはめ］の［PLS 回帰］手法で多項式項、交互作用項、およびカテゴリカル項といった効果を指定できます。
- 複数の検証法や交差検証法が用意されています。
- 欠測データを補完できます。
- 各種統計量に対してブートストラップ推定を行えます。ブートストラップ推定を行うには、関心のあるレポート内で右クリックしてください。詳細は、『基本的な統計分析』を参照してください。

PLSプラットフォームでは、van der Voet T^2検定と交差検証法によって、抽出する因子数を決めることができます。

- JMPの標準版では、交差検証として、1つ取って置き法（LOOCV; Leave-One-Out Cross Validation）を行えます。検証をしないことも選択できます。
- JMP PRO JMP Proでは、K分割交差検証法、一つ取って置き法、無作為抽出による検証法を選択できます。また、JMP Proでは、検証列を指定することもできます。検証をしないことも選択できます。

PLS回帰の例

分光測定における較正の例を取り上げます。バルト海の水質汚染について調査するため、海水標本のスペクトルを測定しました。

次の3成分の量を調べています。

- 製紙工場の廃棄物に含まれるリグニン・スルホネート（「**ls**」）
- 自然林から出るフミン酸（「**ha**」）
- 洗剤に含まれる蛍光増白剤（「**dt**」）

各標本に含まれるこれらの成分量を応答変数とし、各波長（**v1** ～ **v27**）におけるスペクトルの強度を説明変数とします。

ここでは測定の較正が目的なので、成分量がわかっている標本を使用します。データには16標本があり、「ls」、「ha」、「dt」の濃度と、27の波長におけるスペクトルの強度が記録されています。「PLS回帰」プラットフォームによって、分光計で測定されたスペクトルの強度から、成分量を予測するモデルを作成します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Baltic.jmp」を開きます。

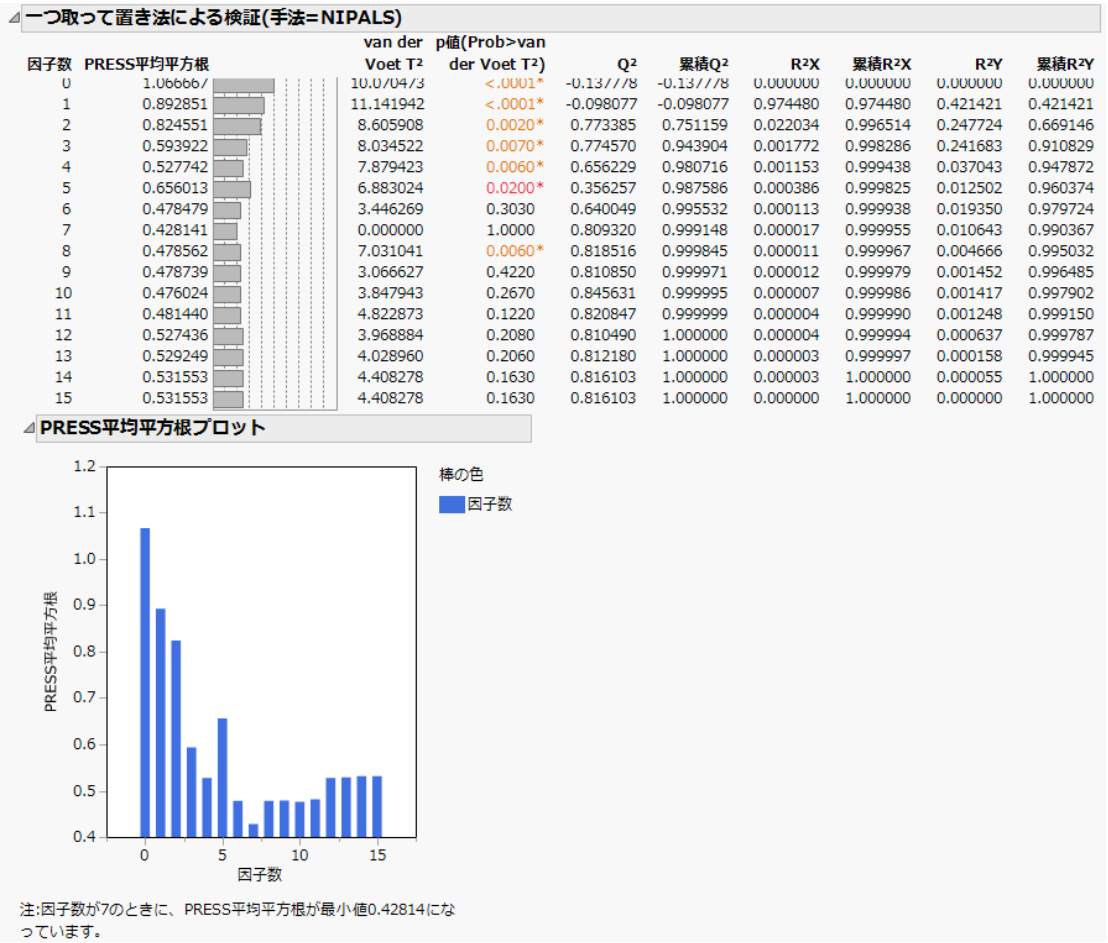
   メモ:「Baltic.jmp」のデータは、Umetrics（1995）で報告されています。原典は、Lindberg, Persson, and Wold（1983）です。

2. ［**分析**］>［**多変量**］>［**PLS回帰**］を選択します。
3. 「**ls**」、「**ha**」、「**dt**」を［**Y, 目的変数**］に指定します。
4. 「**Intensities**」グループにある「**v1**」～「**v27**」を［**X, 説明変数**］に指定します。
5. ［**OK**］をクリックします。

   PLS回帰の「モデルの設定」パネルが表示されます。
6. 「検証法」として［**一つ取って置き法**］を選択します。
7. ［**実行**］をクリックします。

   レポートの一部を図6.2に示します。van der Voet検定は無作為化検定（ランダム化検定）であり、また、その計算に乱数を用いています。そのため、「p値(Prob > van der Voet T^2)」に実際に表示される値は、図6.2の値と若干異なります。

図6.2「PLS回帰」レポート



一つ取って置き法による検証(手法=NIPALS)

因子数	PRESS平均平方根	van der Voet T²	p値(Prob>van der Voet T²)	Q²	累積Q²	R²X	累積R²X	R²Y	累積R²Y
0	1.06666	10.07473	<.0001*	-0.13778	-0.13778	0.000000	0.000000	0.000000	0.000000
1	0.892851	11.141942	<.0001*	-0.098077	-0.098077	0.974480	0.974480	0.421421	0.421421
2	0.824551	8.605908	0.0020*	0.773385	0.751159	0.022034	0.996514	0.247724	0.669146
3	0.593922	8.034522	0.0070*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0060*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0200*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3030	0.640049	0.995532	0.000113	0.999938	0.019350	0.979724
7	0.428141	0.000000	1.0000	0.809320	0.999148	0.000017	0.999955	0.010643	0.990367
8	0.478562	7.031041	0.0060*	0.818516	0.999845	0.000011	0.999967	0.004666	0.995032
9	0.478739	3.066627	0.4220	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2670	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822873	0.1220	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.2080	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2060	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1630	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1630	0.816103	1.000000	0.000000	1.000000	0.000000	1.000000

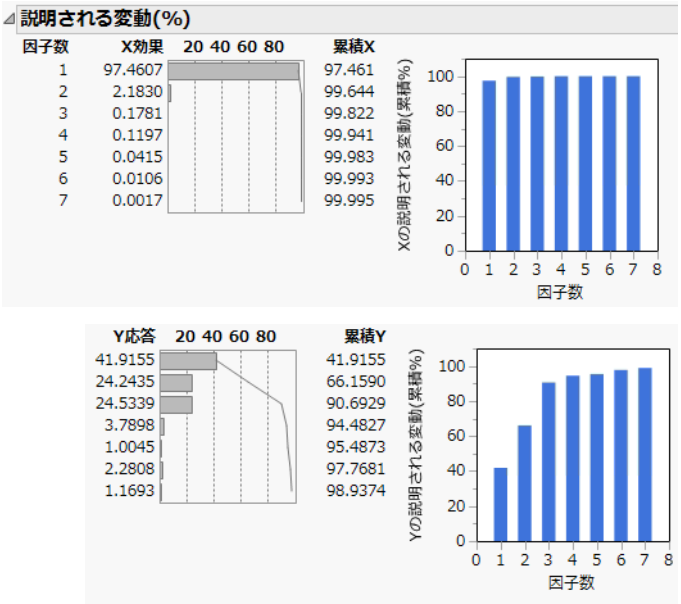
PRESS平均平方根プロット

注:因子数が7のときに、PRESS平均平方根が最小値0.42814になっています。

PRESS（Predicted REsidual Sum of Squares）平均平方根プロットを見ると、因子数が7のときにPRESS平均平方根が最小になっていることがわかります。このことは、PRESS平均平方根プロットの下に注として記載されています。レポートの下部には、**「NIPALSによるあてはめ(7因子)」**という名前のレポートが作成されます。そのレポートの一部を図6.3に示します。

van der Voet T^2統計量は、それぞれの因子数のモデルが、PRESS平均平方根が最小値になるモデルと、有意に異なるかどうかを検定します。van der Voet検定の有意水準が0.10を超える前の最小因子数を抽出するのがよいと提案する人もいます（SAS Institute Inc, 2011 and Tobias, 1995）。この例でこの提案に従うとすると、6因子を採用することになるので、**「モデルの設定」**パネルで**「因子数」**に「6」と入力します。

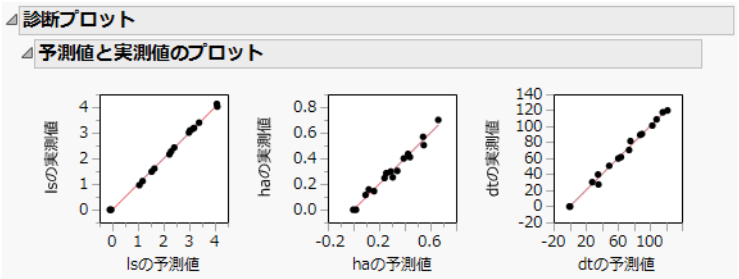
図6.3 抽出された7つの因子



8. 「NIPALSによるあてはめ(7因子)」の赤い三角ボタンのメニューから［**診断プロット**］を選択します。

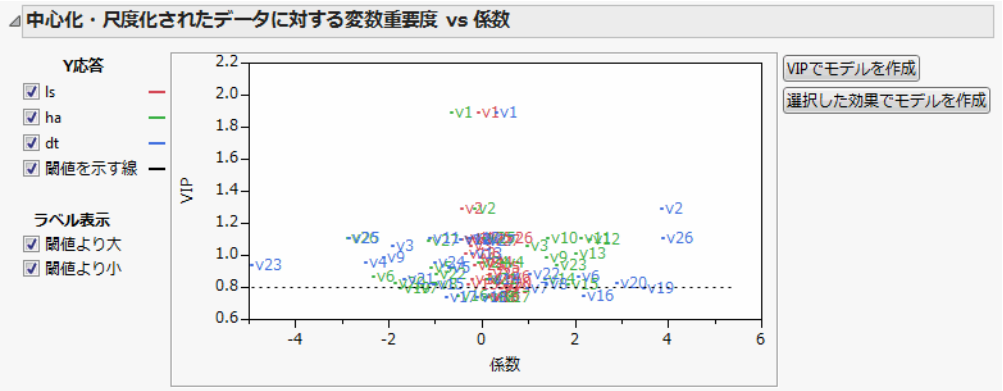
   これにより、予測値と実測値のプロット、および、3種の残差プロットが表示されます。予測値と実測値のプロット（図6.4）を見ると、予測値と実測値がどれぐらい近いかが分かります。

図6.4 診断プロット



9. 「NIPALSによるあてはめ(7因子)」の赤い三角ボタンのメニューから［**変数重要度 vs 係数プロット**］を選択します。

図6.5 変数重要度 vs 係数プロット



「変数重要度 vs 係数」プロットは、応答に影響のある変数を特定するのに役立ちます。たとえば、**v23**、**v2**、**v26**は、変数重要度（VIP）が0.8を超えており、かつ、係数も比較的大きくなっています。

「PLS回帰」プラットフォームの起動

「PLS回帰」プラットフォームを起動するには、次の2つの方法があります。

- ［**分析**］>［**多変量**］>［**PLS回帰**］を選択します。
- JMP PRO ［**分析**］>［**モデルのあてはめ**］を選択し、「手法」で［**PLS回帰**］を選択します。この方法では、次のことができます。
  - カテゴリカル変数をYやXに指定する。Y変数がカテゴリカルな場合の分析は、「PLS判別分析」と呼ばれています。
  - モデルに交互作用項や多項式項を追加する。
  - ［Xの標準化］オプションを選択して、中心化および尺度化した列によって高次の効果を構成する。
  - モデルを指定するためのスクリプトを保存する。

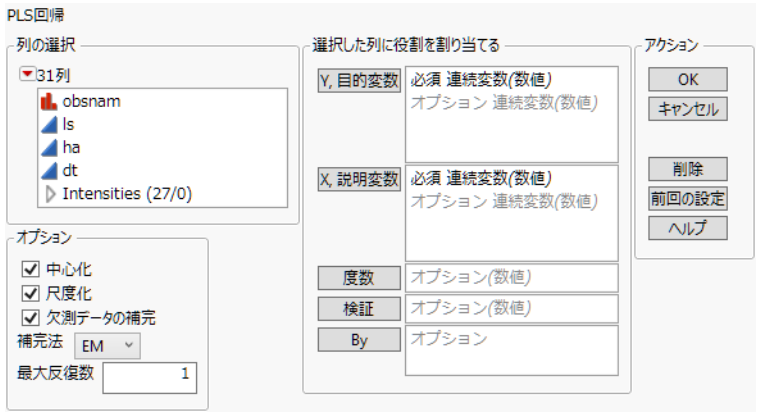
以下の「モデルのあてはめ」起動ウィンドウの機能は、［PLS回帰］手法では使用できません。

- 重み、枝分かれ、属性、変換、切片なし

ヒント: ただし、「列の選択」ボックス内を右クリックし、表示されるメニューから変換オプションを選択する操作によって、変数を変換できます。

- 配合応答曲面、Scheffeの3次多項式、動径の各マクロ

図6.6 JMP Proの「PLS回帰」起動ウィンドウ（補完法としてEMを選択）



「PLS回帰」起動ウィンドウには次のオプションがあります。

Y, 目的変数　数値の応答列を指定します。複数の列を指定した場合は、複数の応答変数に対するPLSモデルが構築されます。

JMP PRO JMP Proでは、「モデルのあてはめ」起動ダイアログで名義尺度の列を指定し、PLS判別分析を実行することもできます。詳細については、「PLS判別分析（PLS-DA）」（130ページ）を参照してください。

X, 説明変数　予測変数の列を指定します。なお、「PLS回帰」起動ダイアログでは、予測変数はすべて連続尺度でなければいけません。

JMP PRO JMP Proでは、「モデルのあてはめ」起動ダイアログで名義尺度の列も指定できます（なお、順序尺度の列も、名義尺度として取り扱われます）。

度数　データが要約されている場合は、各行の度数を含む列を指定します。

JMP PRO **検証**　（オプション）検証列を指定します。検証列には、連続する整数値が含まれている必要があります。次のような検証が行われます。

- 検証列の値が2つしかない場合は、小さい方の値が学習セット、大きい方の値が検証セットとして扱われます。
- 値が3つある場合は、値の小さい方から順に、学習セット、検証セット、テストセットとして扱われます。
- 値が 4 つ以上ある場合は、K 分割検証法が行われます。その他の検証法については、「検証法」（114ページ）を参照してください。

メモ：「列の選択」リストで列を選択せずに［検証］ボタンをクリックすることにより、データテーブルに検証列を追加できます。「検証列の作成」ユーティリティの詳細については、『基本的な統計分析』を参照してください。

By　指定した列の水準ごとに、個別に分析が行われます。1つ1つの分析に対して、個別にオプションを適用できます。

中心化 各列から平均を減算し、Y変数とモデル効果をすべて中心化します。「中心化と尺度化」（113ページ）を参照してください。

尺度化 各列を標準偏差で除算し、Y変数とモデル効果をすべて尺度化します。「中心化と尺度化」（113ページ）を参照してください。

JMP PRO **Xの標準化** （「モデルのあてはめ」ウィンドウのみ）このオプションを選択すると、モデル効果の構成に使用されるすべての列が中心化および尺度化されます。このオプションを選択しない場合は、元のままのデータテーブル列を用いて高次の効果が構成されます。その後、それぞれの高次の効果が、［中心化］および［尺度化］のオプションの選択状況に応じて中心化または尺度化されます。［Xの標準化］オプションは、Y変数に対しては中心化や尺度化を行いません。「Xの標準化」（113ページ）を参照してください。

JMP PRO **欠測データの補完** Y変数およびX変数の欠測値を非欠測値で置き換えます。**「補完法」**のリストから補完法を選択してください。

［**欠測データの補完**］が選択されていない場合、説明変数が欠測値となっている行は、分析から除外され、予測値が計算されません。説明変数には欠測値がなく、応答変数だけに欠測値がある行も分析から除外されますが、予測値の計算は行われます。

JMP PRO **補完法** （［**欠測データの補完**］が選択されている場合にのみ表示）次の補完法の中から選択します。

- ［**平均**］：各モデル効果と各応答列について、欠測値を非欠測値の平均値で置き換えます。
- ［**EM**］：反復法の一種であるEM（Expectation-Maximization）法を用いて欠測値を補完します。反復計算の1回目では、効果と応答における欠測値を平均値で置き換えた後に、指定したモデルをデータにあてはめます。そして、推定されたモデルの予測値を使って欠測値を補完します。反復計算の2回目以降では、現在のモデル推定値を使って得られる条件付き期待値で、欠測値を補完していきます。

欠測値の補完においては、多項式の各項は、別々の1つの説明変数として扱われます。まず、多項式の各項は、元のデータから計算されるか、または［Xの標準化］チェックボックスがオンの場合は、標準化した列の値から計算されます。この時、多項式の項に関係する変数が欠測値であった場合、その多項式の項も欠測値となります。そして、このように定義された多項式の項に対して、欠測値の補完が行われます。

EM法の詳細については、Nelson, Taylor, and MacGregor（1996）を参照してください。

JMP PRO **最大反復数** （「補完法」として［EM］が選択されている場合にのみ表示）EM法による反復計算の最大回数を設定できます。欠測値の現在推定値と前回推定値の相対的差における最大値が10^-8を下回ると、EM法は終了します。

起動ウィンドウでの設定が終わったら［**OK**］をクリックします。「モデルの設定」パネルが表示されます。「モデルの設定パネル」（113ページ）を参照してください。

中心化と尺度化

［中心化］および［尺度化］オプションはデフォルトで選択されています。つまり、説明変数と応答変数は、平均が0、標準偏差が1になるように、中心化および尺度化されます。中心化は、説明変数と応答変数の平均を原点に移動します。中心化を行わなかった場合、平均周りの変動ではなく、原点周りの変動を説明する因子が抽出されていきます。尺度化を行うと、異なるばらつきをもつ変数の変動を統一できます。たとえば、説明変数の中に**「時間」**と**「温度」**があったとします。中心化と尺度化を行うと、**「時間」**における1標準偏差の変化と、**「温度」**における1標準偏差の変化が、どちらも1に変換されます。

Xの標準化

JMP PRO「モデルのあてはめ」ウィンドウで［PLS回帰］手法を選択すると、デフォルトで［Xの標準化］が選択されます。この場合、モデル効果として指定されたすべての列と、交互作用項または多項式項に関与するすべての列が標準化されます。

X1とX2の2列があり、「モデルのあてはめ」ウィンドウでモデル効果として交互作用項「X1*X2」を入力したとします。ここで［Xの標準化］を選択した場合、X1とX2が両方とも、交互作用項の作成前に中心化および尺度化されます。作成される交互作用項は、次式で計算されます。

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

この後、すべてのモデル項が、**［中心化］**オプションと**［尺度化］**オプションの選択状況に従って、モデルに組み込まれる前に、再度、中心化と尺度化されます。

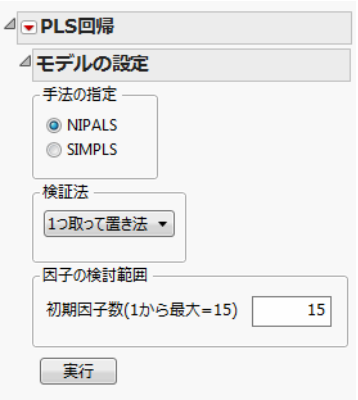
［Xの標準化］が選択されておらず、**［中心化］**と**［尺度化］**が両方とも選択されている場合、モデルに組み込まれる項は次式で計算されます。

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

モデルの設定パネル

プラットフォームの起動ウィンドウで**［OK］**（［モデルのあてはめ］の場合は**［実行］**）をクリックすると、「モデルの設定」パネルが表示されます。

図6.7 PLS回帰の「モデルの設定」パネル



メモ: JMP Pro では、「モデルの設定」パネルの「検証法」部分の表示形態が異なります。

「モデルの設定」パネルでは以下を指定できます。

手法の指定 モデルをあてはめる方法を選択します。[**NIPALS**] と [**SIMPLS**] の2つがあります。応答変数が1つしかない場合は、2つの手法のいずれで推定しても、推定されたモデルは同じになります。2つのアルゴリズムの違いについては、「統計的詳細」（126ページ）を参照してください。

検証法 検証法を選択します。選択された検証法によって、最適な因子数が決められます。JMP Pro で、プラットフォームの起動ウィンドウで検証列を指定した場合、これらのオプションは表示されません。

JMP PRO **保留** 指定された割合のデータを検証セットに使用し、残りのデータをモデルのあてはめに使用します。

JMP PRO **K分割法** まず、元のデータをk個に**分割**します。そして、順番に、(K-1) 個分のデータにモデルがあてはめられ、残っているデータでモデルが検証されます。全部でK回モデルがあてはめられます。この方法は、少ないデータを効果的に利用するので、小規模なデータセットに適しています。

1つ取って置き法 1つ取って置きの交差検証法（LOOCV; Leave-One-Out Cross Validation）を実行します。

なし 最適な因子数を決めるのに、検証法を使用しません。因子数は［因子の検討範囲］で指定します。

因子の検討範囲 どの検証法も使用しない場合は、因子をいくつにするかを指定します。いずれかの検証法を使用する場合は、ここで指定された値が、検証するためにあてはめられるモデルの因子数の上限となります。

因子の指定 ［**実行**］をクリックして最初のモデルをあてはめた後に表示されます。新しいモデルのあてはめに使用する因子数を指定します。

「PLS回帰」レポート

「モデルの設定」パネル（図6.7）で最初に［**実行**］をクリックすると、「モデルの起動」ウィンドウから［検証法］が削除されます。検証列を指定するか、「検証法」で［保留］を選択した場合、レポートに含まれるモデルのあてはめは、すべて学習データに基づきます。そうでない場合、モデルのあてはめは、すべてデータセット全体に基づきます。

検証を使用した場合、3つのレポートが表示されます。

- モデル比較の要約
- <検証法の名前>による検証 手法 = <PLS法の名前>
- NIPALS（またはSIMPLS）によるあてはめ(<N>因子)

検証法として［**なし**］を選択した場合、2つのレポートが表示されます。

- モデル比較の要約
- NIPALS（またはSIMPLS）によるあてはめ(<N>因子)

別の因子数でもモデルをあてはめる場合には、「モデルの設定」パネルで希望の因子数を指定してください。

モデル比較の要約

「モデル比較の要約」には、あてはめたモデルごとに結果の要約情報が表示されます。

図6.8 モデル比較の要約

モデル比較の要約

手法	行数	因子数	Xの説明される変動(累積%)	Yの説明される変動(累積%)	VIPの数>0.8
NIPALS	16	6	99.993471	97.768092	22
NIPALS	16	7	99.995152	98.937438	22

図6.8のレポートは、7因子のモデルをあてはめた後に、6因子のモデルをあてはめたときのものです。このレポートには、次のような結果が含まれます。

手法 「モデルの設定」パネルで指定した分析手法が表示されます。

行数 学習セットで使用されたオブザベーションの数が表示されます。

因子数 モデルで使われている因子の数が表示されます。

Xの説明される変動(累積%) モデルによって説明されるXの変動が、パーセント単位で表示されます。

Yの説明される変動(累積%) モデルによって説明されるY変動が、パーセント単位で表示されます。

VIPの数>0.8 VIP（Variable Importance for Projection; 射影における変数重要度）の値が、0.8よりも大きい説明変数の個数が表示されます。VIPは、XおよびYをモデル化する上で、X変数の重要度を表す指標です（Wold, 1995およびEriksson et al., 2006）。

<検証法の名前>による検証手法 = <PLS法の名前>

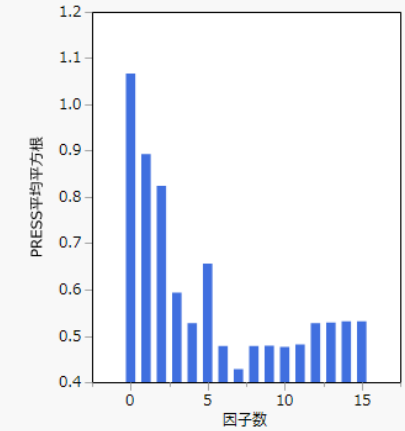
「モデルの設定」パネルで［検証法］として、何らかの検証法を選択した場合には、検証結果のレポートも表示されます。「モデルの設定」パネルでの指定に基づき、0から最大数までの因子の各モデルに関して、要約統計量が表示されます。このレポートには、PRESS平均平方根の棒グラフも表示されます。「PRESS平均平方根プロット」（118ページ）を参照してください。PRESS平均平方根が最小になっているモデルが、最適なモデルであると判断できます。

図6.9 交差検証のレポート

一つ取って置き法による検証(手法=NIPALS)

因子数	PRESS平均平方根	van der Voet T²	p値(Prob>van der Voet T²)	Q²	累積Q²	R²X	累積R²X	R²Y	累積R²Y
0	1.06667	10.07073	<.0001*	-0.13778	-0.13778	0.000000	0.000000	0.000000	0.000000
1	0.892851	11.141942	<.0001*	-0.098077	-0.098077	0.974480	0.974480	0.421421	0.421421
2	0.824551	8.605908	0.0020*	0.773385	0.751159	0.022034	0.996514	0.247724	0.669146
3	0.593922	8.034522	0.0070*	0.774570	0.943904	0.001772	0.998286	0.241683	0.910829
4	0.527742	7.879423	0.0060*	0.656229	0.980716	0.001153	0.999438	0.037043	0.947872
5	0.656013	6.883024	0.0200*	0.356257	0.987586	0.000386	0.999825	0.012502	0.960374
6	0.478479	3.446269	0.3030	0.640049	0.995532	0.000113	0.999938	0.019350	0.979724
7	0.428141	0.000000	1.0000	0.809320	0.999148	0.000017	0.999955	0.010643	0.990367
8	0.478562	7.031041	0.0060*	0.818516	0.999845	0.000011	0.999967	0.004666	0.995032
9	0.478739	3.066627	0.4220	0.810850	0.999971	0.000012	0.999979	0.001452	0.996485
10	0.476024	3.847943	0.2670	0.845631	0.999995	0.000007	0.999986	0.001417	0.997902
11	0.481440	4.822873	0.1220	0.820847	0.999999	0.000004	0.999990	0.001248	0.999150
12	0.527436	3.968884	0.2080	0.810490	1.000000	0.000004	0.999994	0.000637	0.999787
13	0.529249	4.028960	0.2060	0.812180	1.000000	0.000003	0.999997	0.000158	0.999945
14	0.531553	4.408278	0.1630	0.816103	1.000000	0.000003	1.000000	0.000055	1.000000
15	0.531553	4.408278	0.1630	0.816103	1.000000	0.000000	1.000000	0.000000	1.000000

PRESS平均平方根プロット



注:因子数が7のときに、PRESS平均平方根が最小値0.42814になっています。

JMP PRO ［Xの標準化］オプションが選択されている場合、標準化はデータテーブル全体に対して一度だけ適用され、個々の学習セットには再適用されません。ただし、［中心化］や［尺度化］のオプションが選択されている場合は、交差検証における中心化や尺度化は各学習セットに適用されます。これらのオプションが選択されている場合は、それぞれ個別に中心化や尺度化された学習セットによって交差検証が進められます。

レポートには、次の統計量が表示されます。検証や交差検証のいずれかが使用された場合にレポートに表示されている要約統計量は、学習セットに対するものです。

因子数 モデルのあてはめに使用された因子の数。

PRESS平均平方根 すべての応答値のPRESSを平均し、その平方根を求めたものです。詳細については、「PRESS平均平方根」（118ページ）を参照してください。

van der Voet T^2 van der Voet検定は、各モデルが最適なモデルと有意に異なるかどうかを検定します。それぞれのvan der Voet T^2検定の帰無仮説は、「この因子数に基づくモデルは、最適なモデルと異ならない」です。対立仮説は「モデルは最適なモデルと異なる」です。詳細については、「van der Voet T^2」（127ページ）を参照してください。

p値 (Prob > van der Voet T^2) van der Voet T^2検定のp値。詳細については、「van der Voet T^2」（127ページ）を参照してください。

Q^2 モデルがもつ予測能力を測定する無次元の指標。PRESSをYの平方和で割ったものを、1から引いた値。

$$1 - PRESS/SSY$$

詳細については、「Q^2の計算」（119ページ）を参照してください。

累積Q^2 当該の因子数以下であるモデルがもつ予測能力の指標。因子数fに対し、累積Q^2は次のように計算されます。

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

この式で、$PRESS_i$とSSY_iは、因子数がiであるモデルの統計量です。

R^2X 当該の因子によって説明されるXの変動の割合。R^2Xの値が大きい成分は、X変数の変動の大半を説明します。「検証が使用された場合のR^2XとR^2Yの計算」（119ページ）を参照してください。

累積R^2X 当該の因子数のモデルによって説明されるXの変動の割合。$i = 1$から当該の因子数までのR^2Xの合計。

R^2Y 当該の因子によって説明されるYの変動の割合。R^2Yの値が大きい成分は、Y変数の変動の大半を説明します。「検証が使用された場合のR^2XとR^2Yの計算」（119ページ）を参照してください。

累積R^2Y 当該の因子数のモデルによって説明されるYの変動の割合。$i = 1$から当該の因子数までのR^2Yの合計。

Q^2および累積R^2Yの解釈

Q^2と累積R^2Yは両方ともモデルの予測能力を測定する統計量ですが、その方法は異なります。

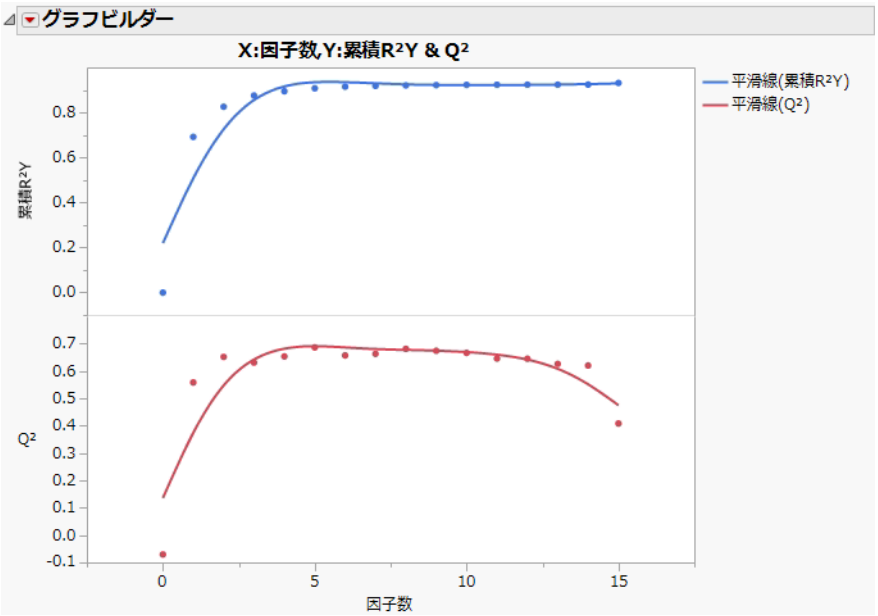
- 累積R^2Yは、因子数が増えるにつれて増加します。これは、より多くの因子がモデルに追加されるにつれて、より多くの変動が説明されるからです。

- Q^2 の場合は、因子数が増えるにつれて増加した後、減少するか少なくとも増加を止めます。これは、より多くの因子が追加されるにつれてモデルは学習セットに合わせられ、新しいデータに対して適切に一般化されないため、PRESS統計量が減少するからです。

Q^2 と累積R^2Yの分析は、モデルにいくつの因子を含めるかを決定するvan der Voet検定の代わりに使用できます。Q^2値が大きく、減少が始まっていない因子数を選択します。また、累積R^2Y値も大きいものを選択します。

図6.10は、「Penta.jmp」データテーブルの累積R^2YとQ^2を、因子数に対してプロットしたものです。検証法には「1つ取って置き法」を使用しています。累積R^2Yは、因子数4のあたりまで増加し、その後は平らになっています。統計量Q^2は、因子数2で最も大きく、その後は平らになっています。このプロットから、因子数2のモデルがYの変動の多くを説明し、データのオーバーフィットも回避することがわかります。

図6.10「Penta.jmp」の累積R^2YとQ^2



PRESS平均平方根プロット

PRESS平均平方根プロットは、横軸に因子数、縦軸にPRESS平均平方根を示した棒グラフです。これは、「交差検証」レポートの「PRESS平均平方根」の右側に表示される横向きの棒グラフと同じです。図6.9を参照してください。

PRESS平均平方根

因子数aに対するPRESS平均平方根は、次のように計算されます。

1. a個の因子のモデルが各学習セットにあてはめられます。

2. 得られた予測式を検証セットのデータに適用します。
3. 各Y変数に対し、次の計算が行われます。
   - 検証セットごとに、各観測値とその予測値の差の2乗（予測誤差の2乗）を求めます。
   - 各応答変数に関して「予測誤差の2乗」の平均を求め、さらにそれを次のように除算します。検証法が［K分割］と［1つ取って置き法］の場合は、応答変数全体の分散で除算します。検証法が［保留］の場合は、学習セット内の応答値の分散で除算します。
   - これらを合計します。複数の検証セットがある場合は、この合計を、検証セットの数から1を引いた数で割ります。これが、Y変数のPRESS統計量です。
4. 「PRESS平均平方根」は、すべての応答変数のPRESSを平均し、その平方根を求めたものです。
5. Y変数が複数ある場合には、ステップ3で得られたPRESS統計量を全応答変数で平均したものが使われます。

Q^2の計算

統計量Q^2は、$1 - PRESS/SSY$と定義されています。この式で、$PRESS$は、学習データから推定されたモデルを検証セットで評価したときの予測誤差平方和を、全応答変数で平均したものです。SSYは、検証セットにおけるYの平方和を、全応答変数で平均したものです。

「交差検証」レポートの統計量Q^2は、選択した「検証法」に応じて、次のように計算されます。

1つ取って置き法 Q^2は、オブザベーションを一度に1つずつ除外することで構築したモデルを各検証セットに適用して計算された$1 - PRESS/SSY$の平均です。

K分割法 Q^2は、K個の各分割を除外することで構築したk個のモデルを各検証セットに適用して計算された$1 - PRESS/SSY$の平均です。

保留法や検証列の使用 Q^2は、1つの学習セットから構築されたモデルを検証セットに適用して計算された$1 - PRESS/SSY$です。

検証が使用された場合のR^2XとR^2Yの計算

「交差検証」レポートの統計量R^2XとR^2Yは、選択した「検証法」に応じて、次のように計算されます。

メモ: R^2Yの計算も、以下と同様です。

1つ取って置き法 R^2Xは、オブザベーションを一度に1つずつ除外することで構築したモデルの「Xの説明される変動(%)」の平均です。

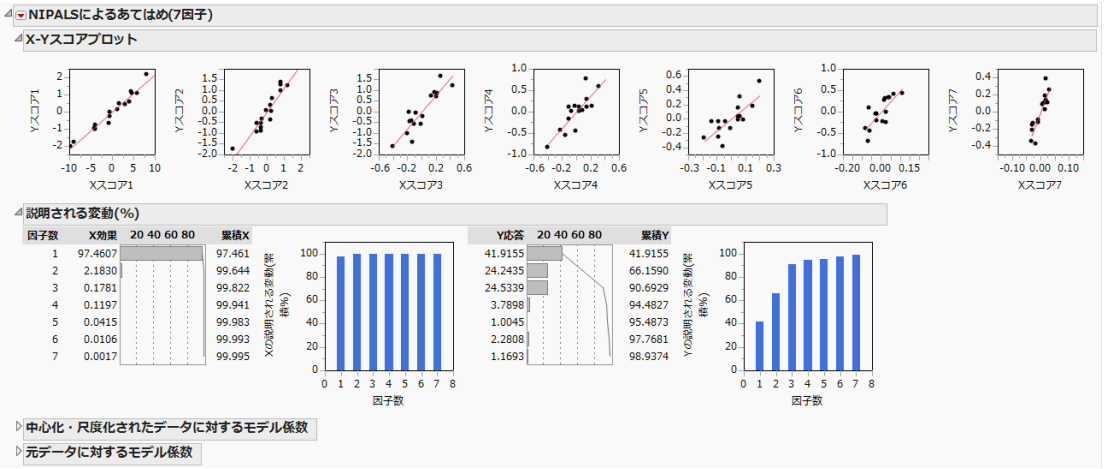
K分割法 R^2Xは、各分割を除外することで構築したモデルの「Xの説明される変動(%)」の平均です。

保留法や検証列の使用 R^2Xは、学習データを使用して構築したモデルの「Xの説明される変動(%)」です。

あてはめレポート

モデルのあてはめ結果に関するレポートには、あてはめたモデルごとに結果の詳細が表示されます。指定された検証法により最適と判断された因子数のモデルがあてはめられます。検証法を指定しなかった場合は、指定した因子数のモデルがあてはめられます。レポートのタイトルには、NIPALSまたはSIMPLSのどちらの手法を用いたかと、あてはめられたモデルの因子数が示されます。

図6.11 あてはめレポート



モデルのあてはめ結果に関するレポートには、次のような要約情報が表示されます。

X-Yスコアプロット　X因子スコアとY因子スコアの散布図です。

説明される変動(%)　XとYの説明される変動が、パーセントおよび累積パーセントで表示されます。これらの値は、抽出された因子ごとに算出されます。

中心化・尺度化されたデータに対するモデル係数　中心化・尺度化されたデータから計算された、各Yに対するXのモデル係数が表示されます。

PLS回帰のオプション

「PLS回帰」の赤い三角ボタンのメニューには、次のようなオプションがあります。

JMP PRO **乱数シード値の設定**　［K分割］と［保留］の検証法で使う乱数シード値を設定します。分析を再現する場合に便利です。シード値を正の値に設定してスクリプトを保存すると、指定したシード値がスクリプトに自動的に保存されます。このスクリプトを実行すると、毎回同じ検証セットに基づいて分析が実行され、同じ結果となります。このオプションは、「検証法」を［なし］に設定した場合や、検証列を使用している場合は表示されません。

以下のオプションについて詳しくは、『JMPの使用法』の「JMPのレポート」章を参照してください。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

あてはめレポートのオプション

あてはめレポートのタイトルバーにある赤い三角ボタンをクリックすると、次のようなオプションが表示されます。

変動プロット 「X の説明される変動 (%)」と「Y の説明される変動 (%)」の 2 つのプロットを表示します。これらのプロットは、各因子によって X および Y の変動がどれぐらい説明されているかを表す積み重ねた棒グラフを表示します。

変数重要度のプロット 各 X 変数に対する VIP のグラフ表示します。VIP の数値は、「変数重要度表」に表示されます。「変数重要度のプロット」（123 ページ）を参照してください。

変数重要度 vs 係数プロット モデル係数に対して、VIP をプロットしたグラフを表示します。選択した Y に対応する点のみを表示することもできます。ラベル表示のオプションも用意されています。元のデータと中心化・尺度化されたデータの両方のプロットが表示されます。「変数重要度 vs 係数プロット」（123 ページ）を参照してください。

VIP 閾値の設定 「変数重要度のプロット」、「変数重要度表」、「変数重要度 vs 係数プロット」の閾値レベルを設定します。

係数プロット 各 X 変数に対する各応答のモデル係数の重ね合わせプロットを表示します。選択した Y に対応する点のみを表示することもできます。元のデータと中心化・尺度化されたデータの両方のプロットが表示されます。

負荷量プロット 各因子に対する X 負荷量と Y 負荷量を表示します。X と Y、それぞれのプロットが表示されます。

負荷量散布図行列 X 負荷量と Y 負荷量の散布図行列を表示します。

負荷量の相関図 X 負荷量と Y 負荷量とを重ね合わせた散布図を、単一の散布図もしくは散布図行列で表示します。このオプションを選択した場合、プロットしたい因子の数を指定します。

– 因子を2つ指定すると、単一の散布図が表示されます。プロットの下で、軸を定義する2つの因子を選択します。右矢印ボタンをクリックすると、因子のさまざまな組み合わせを順番に表示させることができます。

– 因子を3つ以上指定した場合、指定した数までの因子の各ペアを描いた散布図行列が表示されます。

どちらの場合も、チェックボックスを使ってラベルを制御することができます。

X-Yスコアプロット 次の2つのオプションがあります。

［**直線のあてはめ**］は、X-Yスコアプロット上であてはめ線の表示／非表示を切り替えます。

［**信頼区間の表示**］は、X-Yスコアプロット上で95%信頼区間の表示／非表示を切り替えます。この信頼区間は、外れ値を検出する目的だけに使用してください。

スコア散布図行列 Xスコアの散布図行列と、Yスコアの散布図行列を別々に表示します。各Xスコアの散布図行列には、95%信頼楕円が表示されます。この信頼楕円は、外れ値の検出に役立ちます。信頼楕円の詳細については、「Xスコア散布図行列の信頼楕円」（128ページ）を参照してください。

距離プロット 次のプロットを表示します。

– 各オブザベーションからXモデルまでの距離

– 各オブザベーションからYモデルまでの距離

– XモデルとYモデルの両方までの距離の散布図

良いモデルでは、モデルからXおよびYまでの距離が短くなり、最後の散布図において、原点（0,0）周りに点が分布します。これらの散布図では、XやYの外れ値を見つけられます。また、いくつかの点が他のデータから離れて分布している場合、それらの点には別の共通点があると考えられるので、別に分析した方がよいかもしれません。検証データが使用されている場合や、検証データとテストデータが使用されている場合は、学習データに対する結果とともに、それらのデータに対する結果も出力されます。

T2乗プロット 各オブザベーションのT^2統計量をプロットしたグラフを表示します。グラフには、それらに対する管理限界も描かれます。各オブザベーションのT^2統計量は、そのオブザベーションの因子スコアから計算されます。T^2および管理限界の計算方法については、「T^2プロット」（128ページ）を参照してください。

診断プロット モデルのあてはめを評価する診断プロットを表示します。プロットの種類は、予測値と実測値、予測値と残差、行番号と残差、残差の正規分位点プロットの4つです。プロットは応答ごとに作成されます。検証データが使用されている場合や、検証データとテストデータが使用されている場合は、学習データに対する結果とともに、それらのデータに対する結果も出力されます。

プロファイル 各Y変数のプロファイルを表示します。

スペクトルプロファイル プロファイルを表示します。このプロファイルでは、最初のセルに、すべての応答変数の値が描かれています。これにより、X変数における変化がY変数にどのように影響を与えるかを視覚的に確認できます。

列の保存　各種計算式と結果を保存するオプションがあります。「列の保存」（124ページ）を参照してください。

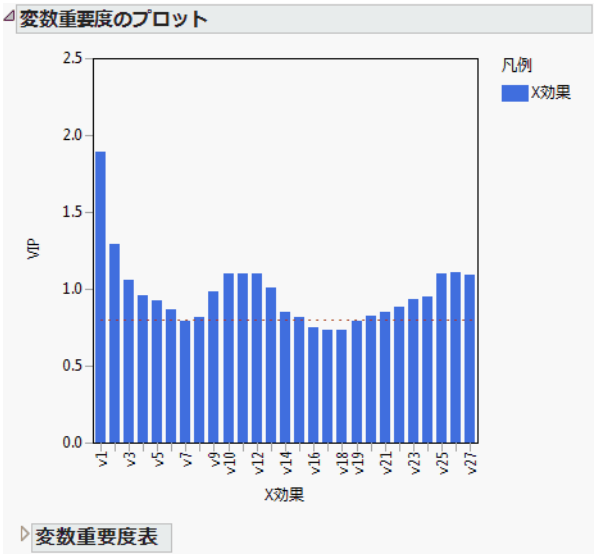
あてはめの削除　メインのプラットフォームレポートからモデルのレポートを削除します。

VIPでモデルを作成　起動ウィンドウが開き、現在のモデルで使われている応答がYに、VIPの値が指定の閾値を超える変数だけがXに入力された起動ダイアログが呼び出されます。「中心化・尺度化されたデータに対する変数重要度 vs 係数」レポート内のボタンと同様の役目を果たします。「変数重要度 vs 係数プロット」（123ページ）を参照してください。

変数重要度のプロット

「変数重要度のプロット」には、各X変数のVIPが表示されます。「変数重要度表」には、VIPスコアが表示されます。VIPは、XおよびYをモデル化する上で、X変数の重要度を表す指標です。係数とVIPの値が小さい変数は、モデルから削除する候補となります（Wold, 1995）。1つの目安として、VIPが0.8以下のX変数は重要でないと考えられている（Eriksson et al, 2006）ので、プロットの0.8の位置に赤い点線が引かれています。

図6.12 変数重要度のプロット



変数重要度 vs 係数プロット

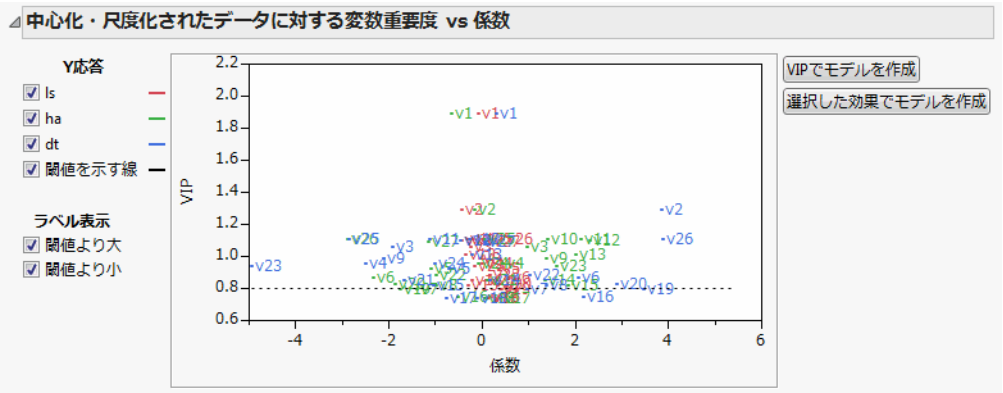
プロットの右にある2つのオプションを使えば、簡単に、変数を減らしたり、モデルを構築したりできます。

- ［**VIPでモデルを作成**］ボタンをクリックすると、起動ウィンドウが開き、現在のモデルで使われている応答がYに、VIPの値が指定の閾値を超える変数だけがXに入力された起動ダイアログが呼び出されます。

- ［選択した効果でモデルを作成］をクリックすると、プロット上で選択された X だけが入力された起動ダイアログが呼び出されます。

  現在選択している列を別のプラットフォームで使いたい場合は、目的のプラットフォームを開きます。選択していた列は、新しいプラットフォームの起動ウィンドウでもそのまま維持されます。役割のボタンをクリックすると、選択されている列にその役割が割り当てられます。

図6.13 中心化・尺度化されたデータに対するモデル係数プロット



列の保存

予測式の保存　各Y変数に対して、予測式を含む「**予測式 <応答>**」という列をデータテーブルに保存します。

予測値をXスコアの計算式として保存　各Y変数に対して、Xスコアで表した予測式を含む「**予測式 <応答>**」という列をデータテーブルに保存します。

予測値の標準誤差の計算式を保存　各Y変数に対して、応答平均の標準誤差を含む「**予測値の標準誤差 <応答>**」という列をデータテーブルに保存します。詳細については、「予測値の標準誤差と信頼区間」（128ページ）を参照してください。

平均の信頼限界の計算式を保存　各Y変数に対して、「**平均 <応答> の下側95%**」および「**平均 <応答> の上側95%**」という2つの列をデータテーブルに保存します。これらは、応答平均に対する95%信頼区間です。詳細については、「予測値の標準誤差と信頼区間」（128ページ）を参照してください。

個別の信頼限界の計算式を保存　各Y変数に対して、「**個別 <応答> の下側95%**」および「**個別 <応答> の上側95%**」という2つの列をデータテーブルに保存します。これらは、個別の応答値に対する95%信頼区間です。詳細については、「予測値の標準誤差と信頼区間」（128ページ）を参照してください。

スコアの計算式を保存　データテーブルに次の2つの列を保存します。

- 各Xスコアの計算式を含む「**Xスコア <N> の計算式**」という列。
- 各Yスコアの計算式を含む「**Yスコア <N> の計算式**」という列。

「PLS」（126ページ）を参照してください。

Y予測値の保存　各Y変数の予測値がデータテーブルに保存されます。

Y残差の保存　各Y変数の残差がデータテーブルに保存されます。

X予測値の保存　各X変数の予測値がデータテーブルに保存されます。

X残差の保存　各X変数の残差がデータテーブルに保存されます。

Xの説明される変動を保存　因子全体に対する各X変数の説明される変動が、新しいデータテーブルに保存されます。

Yの説明される変動を保存　因子全体に対する各Y変数の説明される変動が、新しいデータテーブルに保存されます。

スコアの保存　抽出された各因子のXスコアとYスコアがデータテーブルに保存されます。

負荷量の保存　XとYの負荷量が新しいデータテーブルに保存されます。

標準化したスコアを保存　「負荷量 相関図」の作成に使用したXとYの標準化したスコアがデータテーブルに保存されます。計算式の詳細については、「標準化したスコアと負荷量」(130ページ)を参照してください。

標準化した負荷量を保存　「負荷量 相関図」の作成に使用したXとYの標準化した負荷量がデータテーブルに保存されます。計算式の詳細については、「標準化したスコアと負荷量」(130ページ)を参照してください。

T2乗の保存　T^2の値をデータテーブルに保存します。T2乗プロットに使用される値です。

距離の保存　Xモデルまでの距離（**DModX**）とYモデルまでの距離（**DModY**）がデータテーブルに保存されます。距離プロットに使用される値です。

X重みの保存　各因子に対するX変数の重みが新しいデータテーブルに保存されます。

JMP PRO **検証の保存**　各オブザベーションが検証でどのように使用されたかを示す列がデータテーブルに保存されます。検証法として［保留］を使用した場合、この列は、各行が学習と検証のどちらに使用されたかを示します。［K分割］を使用した場合、この列は、行が割り当てられたサブグループの番号を示します。

JMP PRO **補完の保存**　［欠測データの補完］が選択されている場合に、XとYに指定された列の欠測値を補完値で置き換え、新しいデータテーブルを作成します。多項式項の列は含まれません。［検証］列が指定されている場合は、その［検証］列も含まれます。

JMP PRO **予測式を発行**　予測式を作成し、その予測式の計算式列を生成するスクリプトを「計算式デポ」レポートに保存します。「計算式デポ」レポートが開いていない場合は、このオプションを選択した時点でレポートが作成されます。『予測および発展的なモデル』の「計算式デポ」章を参照してください。

JMP PRO **スコアの計算式を発行**　XとYのスコアの計算式を作成し、そのスコアの計算式列を生成するスクリプトを「計算式デポ」レポートに保存します。「計算式デポ」レポートが開いていない場合は、このオプションを選択した時点でレポートが作成されます。『予測および発展的なモデル』の「計算式デポ」章を参照してください。

統計的詳細

ここでは、「PLS回帰」プラットフォームで使用されている手法の一部を詳しく説明します。詳細については、Hoskuldsson（1988）、Garthwaite（1994）、またはCox and Gaudard（2013）を参照してください。

PLS

PLS回帰では、説明変数（X）の線形結合からなる因子に基づいて応答変数（Y）を予測する線形モデルを構築します。PLS回帰における因子は、Xの線形結合とYの線形結合との共分散が最大になるようなものです。このようにして、PLS回帰では、XとYの関係を調べ、潜在的な因子を抽出します。PLS回帰では、「応答の変動」と「説明変数の変動」の両方を説明するような因子が抽出されます。PLS回帰は、データの行数より説明変数の方が多い場合や、説明変数に強い相関が見られる場合に役立ちます。

NIPALS

NIPALS法は、以下に述べる反復計算によって、1つずつ因子を抽出していきます。$\mathbf{X}_0$を中心化および尺度化した説明変数の行列、$\mathbf{Y}_0$を中心化および尺度化した応答変数の行列とします。PLS法は、初めに、説明変数の線形結合である$\mathbf{t} = \mathbf{X}_0\mathbf{w}$を求めます。ここで、**t**は**スコアベクトル**、**w**はその**重みベクトル**です。そして、$\mathbf{X}_0$および$\mathbf{Y}_0$を求められたスコア**t**に回帰することにより、それらの予測値を計算します。

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

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

ベクトル**p**と**c**は、それぞれXの**負荷量**およびYの**負荷量**と呼ばれます。

これらのスコア（線形結合）は、説明変数の線形結合$\mathbf{t} = \mathbf{X}_0\mathbf{w}$が、応答変数の線形結合$\mathbf{u} = \mathbf{Y}_0\mathbf{q}$と最大の共分散$\mathbf{t}'\mathbf{u}$を持つように、求められます。共分散が最大になるとき、XとYの重みである**w**と**q**は、共分散行列$\mathbf{X}_0'\mathbf{Y}_0$の第1左特異ベクトルおよび第1右特異ベクトル、つまりそれぞれ$\mathbf{X}_0'\mathbf{Y}_0\mathbf{Y}_0'\mathbf{X}_0$と$\mathbf{Y}_0'\mathbf{X}_0\mathbf{X}_0'\mathbf{Y}_0$の第1固有ベクトルに比例したベクトルになっています。

以上の方法により、第1因子が計算されます。続いて、第2因子は、第1因子の計算における$\mathbf{X}_0$と$\mathbf{Y}_0$を、それぞれXとYの残差に置き換えて、同様に計算されます。

$$\mathbf{X}_1 = \mathbf{X}_0 - \hat{\mathbf{X}}_0$$

$$\mathbf{Y}_1 = \mathbf{Y}_0 - \hat{\mathbf{Y}}_0$$

これらの残差は、**収縮(deflate)された**XおよびYともいいます。スコアベクトルを抽出し、データ行列を収縮する過程が、抽出する因子の数だけ繰り返されます。

SIMPLS

SIMPLS法は、統計基準を最適化するという考えに基づいて導出された手法であり、XとYの線形結合に対し、共分散が最大となるスコアベクトルを抽出します。ただし、Xスコアは直交するように決定されます。各反復において、NIPALS法では、行列$\mathbf{X}_0$と$\mathbf{Y}_0$を収縮（deflate）していきますが、SIMPLS法では、交差積行列$\mathbf{X}_0'\mathbf{Y}_0$を収縮してきます。

単変量のY変数に関しては、これら2つのアルゴリズムは同等ですが、多変量のYについては、モデルが異なります。SIMPLSはDe Jong（1993）によって提唱されました。

van der Voet T^2

van der Voet T^2検定は、ある因子数のモデルが、最適とみなされた因子数のモデルと有意に異なるかどうかを検定します。この検定は、「両モデルの残差の2乗は、まったく同じ分布に従っている」という帰無仮説を、無作為化検定（ランダム化検定）の枠組みで検定します。この帰無仮説は、言い換えれば、「両方のモデルの予測能力は同じである」ということになります。

「交差検証」レポートに表示されるvan der Voet T^2の統計量を求めるために、以下で述べる方法が計算が各検証セットに対して実行されます。検証セットが1つの場合、その1つの検証セットに対する結果がレポートされます。[1つ取って置き法]と[K分割]の場合は、各検証セットの結果を平均したものがレポートされます。

因子数がiであるモデルの第k応答変数の第j行における予測残差を$R_{i,jk}$とします。因子数がoptであるモデルの、第k応答変数の第j行における予測残差を$R_{opt,jk}$とします。optは、最適とみなされた因子数です。検定統計量は次の差に基づきます。

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

K個の応答があるとします。次の表記を使用します。

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

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

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

i個の因子のvan der Voet統計量は、次のように定義されます。

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

データから得られたC_iを、$R^2_{i,jk}$と$R^2_{opt,jk}$を無作為に交換して得られる分布に照らし合わせて、有意水準が近似的に算出されます。モンテカルロシミュレートによって無作為に交換された分布が求められ、それらの標本のうちで検定統計量がC_i以上である割合が、有意水準として表示されます。

T²プロット

i番目のオブザベーションのT^2値は次のような式で表されます。

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

ここで、t_{ij}は、第i行における第j因子のXスコア、pは因子数、nはモデルの学習に使用されたオブザベーションの個数です。検証データが使われていない場合、nはオブザベーションの総数です。

T^2プロットの管理限界は次のように計算されます。

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

ここでも、pは因子数、nはモデルの学習に使用されたオブザベーションの個数です。検証データが使われていない場合、nはオブザベーションの総数です。Tracy, Young, and Mason（1992）を参照してください。

Xスコア散布図行列の信頼楕円

［スコア散布図行列］オプションを選択すると、Xスコア散布図行列に95%信頼楕円が表示されます。NIPALSアルゴリズムでも、SIMPLEアルゴリズムでも、直交するようにXスコアは抽出されるため、Xスコアは無相関です。楕円は、Xスコアの各ペアが、相関ゼロの二変量正規分布に従うものと仮定して求められます。

縦軸がスコアi、横軸がスコアjの散布図について検討してみましょう。楕円の上下左右における極値の座標は、次のようにして求められます。

- 上下の極値は、+/-sqrt(スコアiの分散×z)
- 左右の極値は、+/-sqrt(スコアjの分散×z)

ここで、$z = ((n\text{-}1)*(n\text{-}1)/n)$*BetaQuantile(0.95, 1, $(n\text{-}3)/2$)です。このzの背景については、Tracy, Young, and Mason（1992）を参照してください。

予測値の標準誤差と信頼区間

Xを予測変数の行列、**Y**を応答変数の行列とします。これらは、起動ウィンドウでの選択内容に基づいて中心化・尺度化される場合があります。**Y**の成分は互いに独立で、共通の分散σ^2の正規分布に従うと仮定します。

Hoskuldsson（1988）では、スコアを通常の予測変数として見れば、PLSモデルは線形重回帰モデルと同じであると述べています。彼は、この類似性を使って、予測値の分散の近似式を紹介しています。Umetrics（1995）も参照してください。ただし、Denham（1997）は、PLSの予測値は、Yの非線形関数であると指摘しています。彼は、予測値の信頼区間を求めるのに、ブートストラップや交差検証の手法を推奨しています。「PLS回帰」プラットフォームでは、Umetrics（1995）で説明されている正規分布に基づく手法を用いています。

以下では、**X**スコアの行列を**T**とし、Xの新しい観測値$\mathbf{x}_0$を考えます。**Y**に対する予測値は、**T**に対する**Y**の回帰モデルで算出されます。$\mathbf{x}_0$に対応したスコアベクトルを$\mathbf{t}_0$と記します。

aを因子の数とします。s^2を、データが中心化されている場合は残差平方和を$df = n - a - 1$で除算したもの、データが中心化されていない場合は残差平方和を$df = n - a$で除算したものとします。このs^2は、σ^2の推定値です。

予測式の標準誤差

$\mathbf{x}_0$における平均に対する予測値の標準誤差は、次式で推定されます。

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

平均の信頼限界の計算式

$t_{0.975,df}$を、データが中心化されている場合は自由度$df = n - a - 1$のt分布の0.975分位点、データが中心化されていない場合は$df = n - a$のt分布の0.975分位点とします。

平均の95%信頼区間は、次式で求められます。

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

個別の信頼限界の計算式

$\mathbf{x}_0$における個別の応答値に対する予測値の標準誤差は、次式で推定されます。

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

$t_{0.975,df}$を、データが中心化されている場合は自由度$df = n - a - 1$のt分布の0.975分位点、データが中心化されていない場合は$df = n - a$のt分布の0.975分位点とします。

個別の応答値に対する95%信頼区間は、次式で求められます。

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

標準化したスコアと負荷量

次の表記を使用します。

- n_{tr}は、学習セットのオブザベーション数です。
- mは、効果Xの個数です。
- kは、応答Yの個数です。
- $VarX_i$は、第i因子によって説明されるXの変動(%)です。
- $VarY_i$は、第i因子によって説明されるYの変動(%)です。
- $\mathbf{XScore}_i$は、第i因子における、Xスコアのベクトルです。
- $\mathbf{YScore}_i$は、第i因子における、Yスコアのベクトルです。
- $\mathbf{XLoad}_i$は、第i因子における、X負荷量のベクトルです。
- $\mathbf{YLoad}_i$は、第i因子における、Y負荷量のベクトルです。

標準化したスコア

第i因子における、標準化したXスコアのベクトルは次のように定義されます。

$$\frac{\mathbf{XScore}_i}{(n_{tr}-1)\sqrt{mVarX_i/n_{tr}}}$$

第i因子における、標準化したYスコアのベクトルは次のように定義されます。

$$\frac{\mathbf{YScore}_i}{(n_{tr}-1)\sqrt{kVarY_i/n_{tr}}}$$

標準化した負荷量

第i因子における、標準化したX負荷量のベクトルは次のように定義されます。

$$\mathbf{XLoad}_i\sqrt{mVarX_i}$$

第i因子における、標準化したY負荷量のベクトルは次のように定義されます。

$$\mathbf{YLoad}_i\sqrt{kVarY_i}$$

PLS判別分析（PLS-DA）

起動ウィンドウでカテゴリカル変数をYに指定すると、その変数は指示変数（ダミー変数）に変換されます。その変数にk水準がある場合、該当する水準では1、それ以外は0の値をもつ、k個の指示変数に変換されます。そして、そのk個の指数変数が連続変数として扱われ、Yが連続変数のときと同様のPLS分析が行われます。

第 7 章

階層型クラスター分析
データ行をツリー構造にクラスタリング

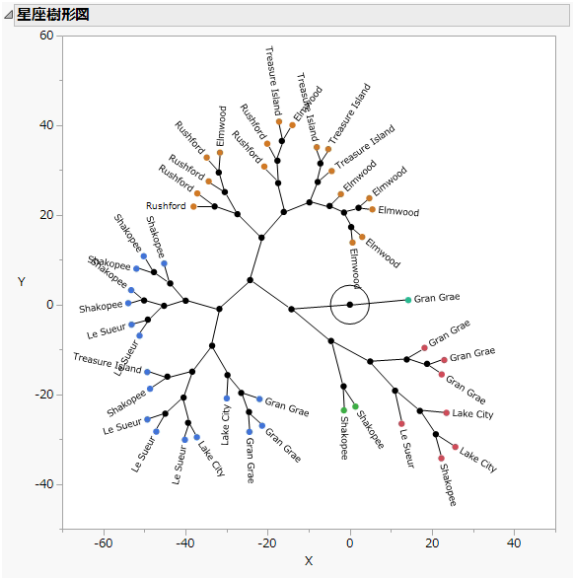
クラスタリングは、多変量データをもとに、値が近い行をグループにまとめていく手法です。データにおける塊を見つけ出すために使用します。

階層型クラスター分析では、逐次的にクラスターを結合していきます。階層型クラスター分析では、まず、データ行の1つ1つそれぞれが1つのク ラスターとみなされます。そして、1ステップごとに距離が最も近い2つのクラスターが1つに結合されていきます。この結合過程は最終的にツリー（**樹形図**）として描かれます。

階層型クラスター分析は、数千行までの小さなテーブルに適しています。計算には時間を要する場合があります。大きなデータテーブルの場合は、K Means クラスター分析または正規混合を使用してください。

メモ：「階層型クラスター分析」プラットフォームは、文字型の列にも対応しています。「K Means クラスター分析」と「正規混合」のプラットフォームは数値型の列しか対応していません。

図 7.1 正座樹形図の例



階層型クラスター分析の概要

JMP には、データ行をクラスタリングするためのプラットフォームが 4 つ用意されています。「階層型クラスター分析」は、そのなかの 1 つです。4 つの手法の比較については、「クラスター分析の例」（133 ページ）を参照してください。

階層型クラスター分析では、まず、データ行の 1 つ 1 つそれぞれが 1 つのクラスターとみなされます。そして、1 ステップごとに距離が最も近い 2 つのクラスターが 1 つに結合されていきます。この結合処理が繰り返され、最後にはすべてのデータが 1 つのクラスターにまとめられます。階層型クラスター分析は、結合していく処理を行うため、**凝集型クラスター分析**（agglomerative clustering）とも呼ばれています。

結合過程はツリー（樹形図）として描かれます。また、JMP には、クラスターの個数を決めるのに役立つように、距離グラフが用意されています。クラスター間の距離があまり大きくなっていない段階を特定することで、クラスター数を決めることができます。

階層型クラスター分析では文字列データの列にも対応し、その場合、距離は次のように定義されます。

- 文字型の列が順序尺度である場合は、低いほうのカテゴリから順番に通し番号が付けられ、その通し番号が連続尺度のデータのように扱われます。これらの値は、連続尺度のデータのように標準化されます。
- 文字型の列が名義尺度の場合は、カテゴリが一致した場合は距離を 0、一致しない場合は距離を 1 として計算が行われます。

階層型クラスター分析では、クラスター間の距離を定義する方法として、群平均法、重心法、Ward 法、最短距離法、最長距離法の 5 つから選ぶことができます。どの方法を選ぶかによって、クラスター分析の結果が変わってきます。

ヒント : 階層型クラスター分析の処理は、高速 Ward 法を除いて、n 個のオブザベーションに対する $n(n+1)/2$ の距離の計算から開始されます。そのため、n が大きいと、計算に時間がかかる場合があります。オブザベーションの数が多い場合は、K Means クラスター分析または正規混合を使用することを検討してください。

クラスター分析用プラットフォームの概要

クラスタリングは、多変量データをもとに、値が近い行をグループにまとめていく手法です。通常、データ点は n 次元空間全体に均等に散らばっておらず、いくつかの塊（クラスター）になっているでしょう。それらのクラスターを見つけ出すと、データをよりよく理解できるようになるでしょう。

メモ : JMP には、変数をクラスタリングするためのプラットフォームも用意されています。「変数のクラスタリング」（197 ページ）章を参照してください。

JMP には、データ行（オブザベーション）をクラスタリングするためのプラットフォームが 4 つ用意されています。

- 「階層型クラスター分析」は、数千行までの小さなテーブルに適しており、文字データにも対応します。行がツリー型の階層構造にまとめられます。クラスタリングの処理が終わった後でも、クラスターの個数を変更することができます。

- 「K Means クラスター分析」は、数十万行までの大きいデータに適しています。この分析は、数値データだけに対応しています。処理を開始する前に、クラスターの数 *k* を指定する必要があります。まず、適切と思われるシード点が推定されます。その後、各点をクラスターに割り当てる作業とクラスター中心を再計算する作業が交互に繰り返されます。
- 「正規混合」は、複数の多変量正規分布の混合分布から得られた、重なりがあるデータに適しています。この分析は、数値データだけに対応しています。外れ値があるような場面では、それらの外れ値を表すために、一様分布に従うと仮定したクラスターを使用できます。また、［ロバスト正規混合］オプションによりロバストな推定を行うこともできます。

  この手法では、処理を開始する前に、クラスターの個数を指定しておく必要があります。最尤法によって、混合割合、平均、標準偏差、相関係数といったパラメータが同時に推定されます。各点に、それぞれの各グループに属する事後確率が計算されます。推定値の反復計算には EM アルゴリズムが使用されています。
- カテゴリカルデータの場合は、「潜在クラス分析」が適しています。この手法では、処理を開始する前に、クラスターの個数を指定しておく必要があります。多項分布の混合分布がモデルとして仮定されます。各データ行に対して、各クラスターに属する事後確率が計算されます。そして、属する事後確率が最も高いクラスターに分類されます。

表 7.1 クラスター分析の手法のまとめ

手法	データタイプまたは尺度	データテーブルのサイズ	クラスター数の指定
階層型クラスター分析	すべて	高速 Ward 法の場合、200,000 行まで その他の手法の場合、5,000 行まで	なし
K Means クラスター分析	数値	数百万行まで	あり
正規混合分布法	数値	制限なし	あり
潜在クラス分析	名義尺度または順序尺度	制限なし	あり

クラスター分析の例

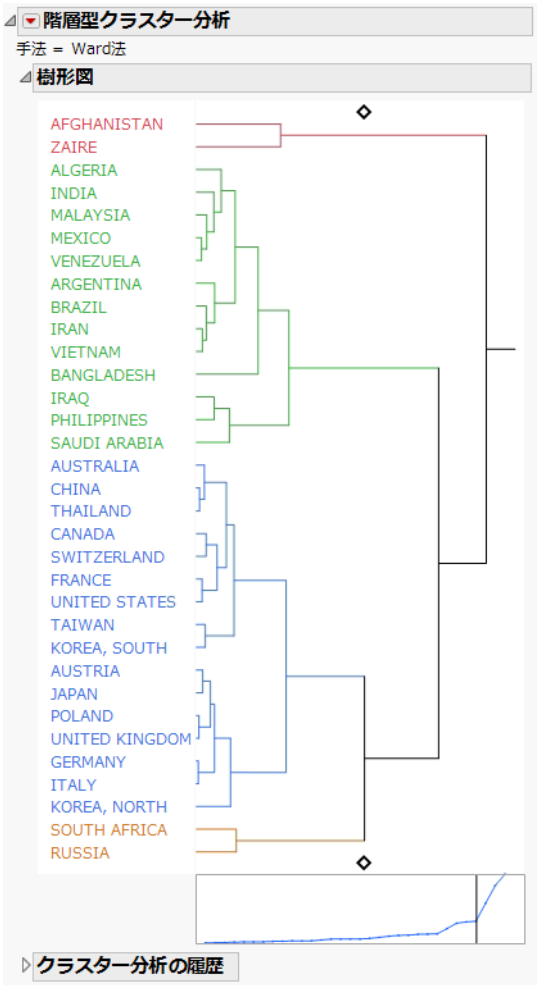
この例では、1976 年の 10 万人あたりの租出生率と租死亡率によって国をグループ分けします。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Birth Death Subset.jmp」を開きます。
2. ［**分析**］>［**クラスター分析**］>［**階層型クラスター分析**］を選択します。
3. 「**租出生率**」と「**租死亡率**」を選択し、［**Y, 列**］をクリックします。
4. 「**国**」を選択して［**ラベル**］をクリックします。

   これにより、［OK］をクリックして表示される樹形図には行番号ではなく「**国**」列の値がラベルとして表示されます。

5. ［OK］をクリックします。
6. 「階層型クラスター分析」の赤い三角ボタンをクリックし、［**クラスターの色分け**］を選択します。

図7.2「階層型クラスター分析」レポート



樹形図はクラスタリングの実行結果を示しています。クラスタリングの過程は、樹形図を左から右へとたどると確認できます。各ステップで、**最も距離が近い**2つのクラスターを、1つのクラスターに結合しています。

樹形図では、クラスター間の相対的な距離は、クラスターを結合している縦線間の横の距離によって判断できます。たとえば、Afghanistan と Zaire の違いは、Mexico と Venezuela で構成されるクラスターと Malaysia との違いより大きいことを示しています。

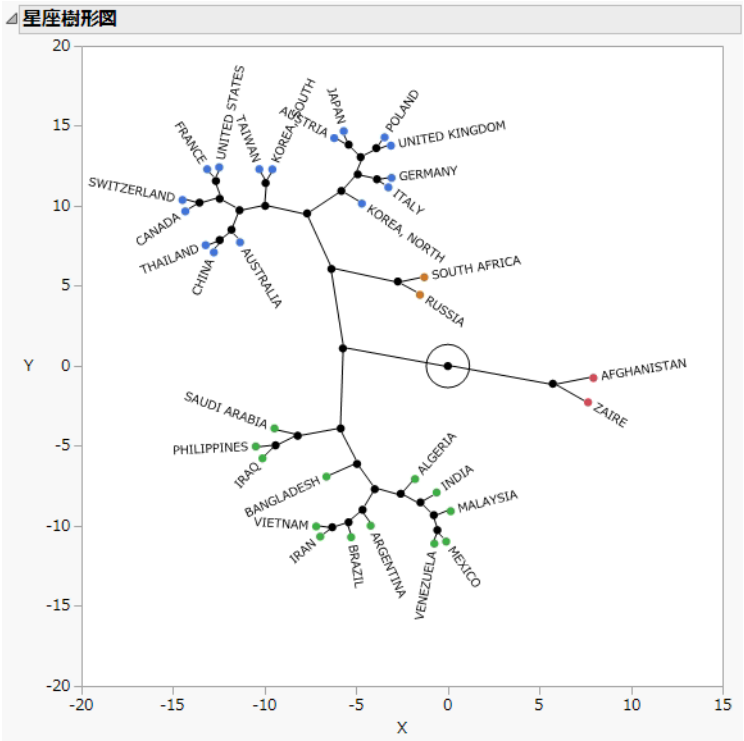
樹形図の下に表示されているプロットには、2つのクラスターが1つのクラスターに結合される各ステップの点があります。横軸はクラスターの数で、左から右にいくにつれ減少します。縦軸は、該当の数のク

ラスターを構成するために結合された2つのクラスター間の距離です。樹形図のひし形のどちらかをクリックしてドラッグし、データを最も良く表しているクラスターの数を選択できます。赤い三角ボタンメニューの［クラスターの数］オプションを使用して、クラスターの数を選択することもできます。

距離グラフでは、クラスターが4つのところで線の傾きが急になっています。この傾きの変化は、クラスターが4つになるまでに結合されたクラスター間の違いが比較的小さいことを示しています。つまり、クラスター数4が最良の選択であることを示しています。このクラスター数は、デフォルトで示された数であることに注意してください。

7. 「階層型クラスター分析」の赤い三角ボタンをクリックし、［**星座樹形図**］を選択します。

図7.3 星座樹形図

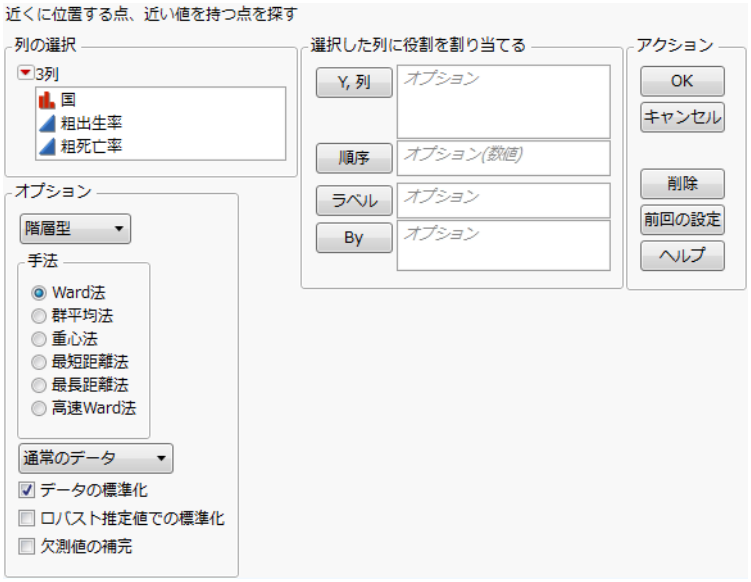


この星座樹形図は、国を端点、各クラスター結合を新しい点として表しています。線はクラスター内の所属関係を表します。クラスター結合間の線の長さは、結合されたクラスター間のおおよその距離を表します。星座樹形図から、Afghanistan と Zaire を含むクラスターが、プロットの上半分と下半分にある残りの国で構成された2つのクラスターから、等しく離れていることがわかります。

「階層型クラスター分析」プラットフォームの起動

「階層型クラスター分析」プラットフォームを起動するには、［分析］>［クラスター分析］>［階層型クラスター分析］を選択します。図7.4は、「Birth Death Subset.jmp」データテーブルの「クラスター分析」起動ウィンドウです。

図7.4「クラスター分析」起動ウィンドウ



Y, 列　クラスター分析の対象となる変数を指定します。

順序　ここで指定された列に基づいて、クラスターを平均値の順に並べることができます。

> **ヒント:** たとえば、主成分分析で得られた第1主成分スコアを、［順序］列に指定すると、主成分スコアでクラ スターが並べられます。

属性のID　（データ構造に［**積み重ねたデータ**］を選択した場合のみ表示されます。）積み重ねた変数を指定します。

対象のID　（データ構造に［**要約したデータ**］または［**積み重ねたデータ**］を選択した場合のみ表示されます。）測定値が積み重なった各ユニットの一意の識別子を含む列を指定します。

ラベル　レポートの樹形図のラベルに使用する列を指定します。

By　この列の水準に従ってデータがグループ化され、それぞれ個別に分析されます。指定した列の水準ごとに、対応する行が分析されます。分析結果は、個別のレポートにまとめられます。複数のBy変数を指定した場合は、By変数の水準のすべての組み合わせごとに分析が行われます。

起動ウィンドウには次のようなメニューとオプションがあります。

クラスター分析の手法

デフォルトは［階層型］で、他に［K-Means 法］または［正規混合］を選択できます。［K-Means 法］または［正規混合］を選択した場合は、［OK］をクリックするとコントロールパネルが表示されます。ここで、次のいずれかの手法を選択できます。

k-means クラスター分析 「K Means クラスター分析」（155 ページ）章を参照してください。

正規混合分布法 「正規混合分布法」（171 ページ）章を参照してください。

ロバスト正規混合 「正規混合分布法」章の「「正規混合 クラスター数=<k>」レポート」（179 ページ）を参照してください。

自己組織化マップ 「K Means クラスター分析」章の「自己組織化マップ」（167 ページ）を参照してください。

距離の計算方法

階層型クラスタリングのときに用いる距離としては、次のようなものが用意されています。詳細については、「距離の手法の計算式」（154 ページ）を参照してください。

Ward 法 2 つのクラスター間の距離は、分散分析のクラスター間平方和をすべての変数について合計したものとして計算されます。クラスター内平方和が最小化されるように、クラスターを結合していきます。クラスター間平方和は、全体平方和で割って分散比（半偏相関の 2 乗）を求めると解釈しやすくなります。

Ward 法は、多変量の正規混合分布、球面性の共分散行列、等しい抽出確率という仮定のもとでの尤度が最大になるようにクラスターを結合していきます。

Ward 法では、オブザベーション数が少ないクラスターが結合される傾向にあり、オブザベーション数がほぼ同じクラスターができてしまいます。また、外れ値に対して非常に敏感です。Milligan（1980）を参照してください。

群平均法 2 つのクラスター間の距離は、各クラスターに属する点のペアの距離を平均したものです。群平均法では、分散の小さいクラスターが結合され、クラスターの分散が等しくなってしまう傾向が多少あります。Sokal and Michener（1958）を参照してください。

重心法 2 つのクラスター間の距離は、その平均間のユークリッド距離として定義されます。重心法は、他の階層型クラスター分析方法より外れ値に対して頑健性がありますが、それ以外の点では Ward 法や群平均法に劣ることがあります。Milligan（1980）を参照してください。

最短距離法 2 つのクラスターから 1 点ずつを選択したときに距離が最短になる 2 点間の距離を、クラスター間の距離とします。最短距離法は、論理的に見て望ましい性質を持っています。しかし、モンテカルロ実験では良い結果が出ていません。それについては、Jardine and Sibson（1976）、Fisher and Van Ness（1971）、Hartigan（1981）、Milligan（1980）を参照してください。この方法は Florek et al.（1951a, 1951b）によって考案され、後に McQuitty（1957）と Sneath（1957）が再考案しました。

クラスターの形状が制約されないため、長く延びた不規則なクラスターができがちで、コンパクトなクラスターを形成することができません。最短距離法では、大きなクラスターに分離する前に、分布の裾が分離する傾向があります。Hartigan（1981）を参照してください。

最長距離法　2つのクラスターから1点ずつを選択したときに距離が最長になる2点間の距離を、クラスター間の距離とします。この方法ではクラスターの直径がほぼ同じになってしまう傾向が強く、それほど極端でない外れ値にも大きく影響されてしまうことがあります。Milligan（1980）を参照してください。

高速Ward法　行数が大量のデータ向けに、計算時間が速いアルゴリズムを用いたWard法です。このアルゴリズムでは距離行列の計算を必要としないため、計算時間が短縮されます。データが2,000行を超える場合に、自動的に使用されます。

データの構造

以下のオプションから、分析に使用されるデータがどのような形式となっているかを選択してください。

通常のデータ　分析に用いるデータが、オブザベーションごとに1行ずつあり、変数ごとに1列ずつある、通常の矩形データの場合には、このオプションを選択します。

要約したデータ　グループごとに平均を計算して、その要約された平均でクラスター分析したい場合には、このオプションを選択します。このオプションを選択すると、起動ウィンドウに［対象のID］ボックスが表示されます。グループ別にしたい列を［対象のID］に指定します。**［要約したデータ］**オプションは、それらの水準ごとに平均を計算し、それらを入力データとして扱います。

データは距離行列　データが距離を表している場合には、このオプションを選択します。対象がn個の場合、この距離データにはn個の行と$n+1$個の列が必要です。1つの列（通常は第1列目）には、n個の対象それぞれを一意に識別する値が含まれている必要があります。それ以外の列は、対象と対象との間の距離を表すn個の値が含まれている必要があります。次の点に留意してください。

- ある点とその点自身との距離は0であるため、距離データの対角要素はすべて0です。0以外の値または欠測値は0として扱われ、レポートにその旨が記載されます。
- 距離データは、対称な正方行列か、欠測値を下側に含む上三角行列か、または、欠測値を上側に含む下三角行列でなければいけません。正方行列を使用する場合は、行列が対称でないと警告が表示されます
- 別の形式のデータで分析をして、そこで距離行列を保存することもできます。「距離行列の保存」（145ページ）を参照してください。

  ［データは距離行列］オプションを選択した場合は、距離を含む列を［Y, 列］に指定し、識別する値を含む列を［ラベル］に指定します。［ラベル］に指定する列には文字型のデータが含まれている必要があります。例については、「距離行列の例」（147ページ）を参照してください。

積み重ねたデータ　関心のある応答が1つだけで、各対象に複数の行があるデータの場合は、このオプションを選択します。

［積み重ねたデータ］オプションを選択すると、起動ウィンドウに［属性のID］テキストボックスと［対象のID］テキストボックスが表示されます。

– **1 つ**の列を［Y, 列］に指定します。

– ［Y, 列］変数のグループ化を説明する列を［属性の ID］に指定します。2 列だけを入力し、［空間的な指標の計算］を選択した場合は、クラスターの分析に使用する空間的な指標を追加できます。「空間的な指標の計算」（140 ページ）を参照してください。

– 識別する値を含む列を［対象の ID］に指定します。

このオプションで実行される分析の結果は、「Y, 列」変数を「属性の ID」列で分割し、応答列を標準化せずに階層型クラスター分析を実行した場合と同じです。

ヒント : 2 次元座標での観測値を分析する場合には、このオプションとともに［空間的な指標の計算］オプションも役立ちます。たとえば、ウエハーのダイごとに、1 行ずつデータが記録されていたとします。この機能を用いると、ウエハーを空間的な指標を用いてクラスタリングできます。「空間的な指標でウエハーをクラスタリングする例」（149 ページ）を参照してください。

注意 : [積み重ねたデータ] で分析されるデータは、観測値が共通した 1 変数だけで測定されているので、多くの場合、［データの標準化］オプションは適切ではありません。

「欠測値でないデータが不足しています。」の警告

［**要約したデータ**］または［**積み重ねたデータ**］を使用している場合は、**「欠測値でないデータが不足しています。」**という JMP の警告の意味がわかりにくいかもしれません。この警告は、以下のような場合に表示されます。

- ［**通常のデータ**］では、0 ～ 1 行以外のすべての行において、「Y, 列」変数の少なくとも 1 つが欠測値となっている場合。
- ［**要約したデータ**］では、「対象の ID」列で要約したときに、0 ～ 1 行以外のすべての行において、要約された「Y, 列」変数の少なくとも 1 つが欠測値の場合。クラスター分析されるデータを確認するには、［テーブル］>［要約］を選択し、「対象の ID」列を［グループ化］に、「Y, 列」変数を［統計量］>［平均］に指定してください。
- ［**積み重ねたデータ**］では、「属性の ID」列で分割したときに、0 ～ 1 行以外のすべての行において、分割された「Y, 列」変数の少なくとも 1 つが欠測値の場合。クラスター分析されるデータの構を確認するには、［**テーブル**］>［**列の分割**］を選択し、「属性の ID」列を［分割する列］に、「Y, 列」変数を［分割する列］に、「対象の ID」列を［グループ化］に指定してください。

「Y, 列」変数の変換

以下のオプションによって、クラスター分析に使用する「Y, 列」変数を事前にどのように処理するかを変更できます。

データの標準化　連続尺度や順序尺度の列に対して、ばらつきを揃えます。［**積み重ねたデータ**］オプションが選択されている場合を除き、各列の値は、列の平均を引いて列の標準偏差で割ることにより標準化されます。［データの標準化］チェックボックスの選択を解除すると、標準化した値ではなく、生データによって距離が計算されます。

ロバスト推定値での標準化 連続尺度と順序尺度の列に対して、外れ値にあまり影響されない方法で平均と標準偏差を推定します。このオプションは、Huber のM 推定（Huber 1964, Huber 1973, Huber and Ronchetti 2009）によって平均や標準偏差を推定します。このオプションを用いた場合、外れ値となっている点を含む列が、通常の標準化を行ったときよりも、距離の計算に大きく寄与します。

メモ: ［データの標準化］と［ロバスト推定値での標準化］のチェックボックスの両方ともオンにした場合は、列ごとにロバストな推定値で標準化されます。すなわち、列ごとに、その列のロバストな平均を引いた後、その列のロバストな標準偏差で割られます。このような標準化は、各列がそれぞれ異なる尺度で測定されている場合、**または**、特定の列にのみ外れ値がある場合に有用です。

メモ: ［データの標準化］チェックボックスがオフで、［ロバスト推定値での標準化］チェックボックスがオンの場合には、すべての列の値から計算されたロバストな平均と標準偏差が各列の標準化に使用されます。これは、すべての列が同じ尺度が測定されていて、**かつ**、すべての次元で外れ値となるデータがあるような場合に有用です。

欠測値の補完 欠測値を補完します。変数の数が50以下か、行数の半分より少ない場合は、多変量正規分布による補完が行われます。その他の場合は、多変量の特異値分解による補完が行われます。

多変量正規分布による補完は、まず、ペアごとの共分散を計算して共分散行列を求めます。そして、各行において、欠測値を含まない列を説明変数として、欠測値部分の予測値を線形回帰モデルで求めます。ただし、各行での補完の計算で使われる共分散行列が正値定符号行列でない場合は、欠測値は列平均によって補完されます。

多変量の特異値分解による補完では、共分散行列を計算するのを避けるために、特異値分解を用います。詳細については、『予測および発展的なモデル』の「モデル化ユーティリティ」章を参照してください。

注意: このような欠測値補完では、データにはクラスターが存在せず1つの塊であること、データが単一の多変量正規分布に従っていること、および、欠測値が完全にランダムであることが仮定されています。これらの仮定は現実的ではないので、この機能には注意が必要です。しかし、欠測値を含むデータ行を破棄するよりは、有益な結果が出る可能性があります。

空間的な指標の計算 （データの構造に［**積み重ねたデータ**］を選択した場合にのみ表示されます。）データが積み重ねデータで、2つの属性が指定されており、それらが空間的な座標（たとえば、X座標とY座標）である場合には、［**空間的な指標の計算**］オプションが有用です。このオプションでは、不適合や不良のパターンをクラスタリングするのに、円・扇形・筋といった空間的指標のどれを用いるかを選択できます。これは特定の応用分野に対する手法で、半導体のウエハーなどの限られた分野でのみ役立ちます。「空間的な指標」（152ページ）および「空間的な指標でウエハーをクラスタリングする例」（149ページ）を参照してください。

「階層型クラスター分析」レポート

「階層型クラスター分析」レポートには、使用した手法、樹形図、および「クラスター分析の履歴」表が表示されます。起動ウィンドウでラベル列を指定したときは、そのラベル列の値が樹形図に表示され、各データ行をラベルによって識別できます。

「樹形図」レポート

樹形図は、各行を結合した集まりを表現したツリー状の図です。樹形図では、クラスター間の距離がどれぐらい離れているかも判断できます。

クラスタリングの過程は、樹形図を左から右へとたどると確認できます。各ステップで、**最も距離が近い**2つのクラスターを、1つのクラスターに結合しています。

- 2つのクラスターが1つに結合したことは、2本の横線が縦線で繋がれることによって樹形図で表されています。
- 縦線のX座標は、そこで結合された2つのクラスター間の距離を表しています。また、そこまでの結合により、樹形図上で繋がっているだけの個数だけクラスターが形成されています。

メモ: データの行数が256行より少ない場合、「距離グラフ」のY座標は距離に比例します。それ以外の場合は、「距離グラフ」のY座標には幾何級数が使用されます。「樹形図のスケール」（144ページ）を参照してください。

ここでは、次のような操作ができます。

- 樹形図の上下にあるひし形のハンドルのいずれかを左右にドラッグすると、その位置でのクラスターの個数が表示されます。
- クラスターの幹のどれかをクリックすると、そのクラスターに属するすべての点が、樹形図上とデータテーブル内の両方で強調表示されます。

距離グラフ

樹形図の下には、「距離グラフ」が描かれます。この「距離グラフ」には、2つのクラスターが1つのクラスターに結合される箇所で、点がプロットされています。X軸はクラスターの個数を表しており、左から右にいくにつれ減少します。Y軸は、その個数になるのに結合されたクラスター間の距離を表しています。

樹形図の上下にあるひし形のハンドルのいずれかを左右にドラッグすると、クラスターの個数を変更できます。ひし形をクリックしてドラッグすると、樹形図に縦線が表示され、クラスターの個数に対応して移動します。多くの場合、距離があまり急激に変化しなくなる箇所があります。そのような情報を参考にしてクラスターの個数を決めるとよいでしょう。

樹形図と距離グラフ

「クラスター分析の例」（133ページ）の「Birth Death Subset.jmp」について検討してみましょう。

図7.5「Birth Death Subset.jmp」の「樹形図」レポート

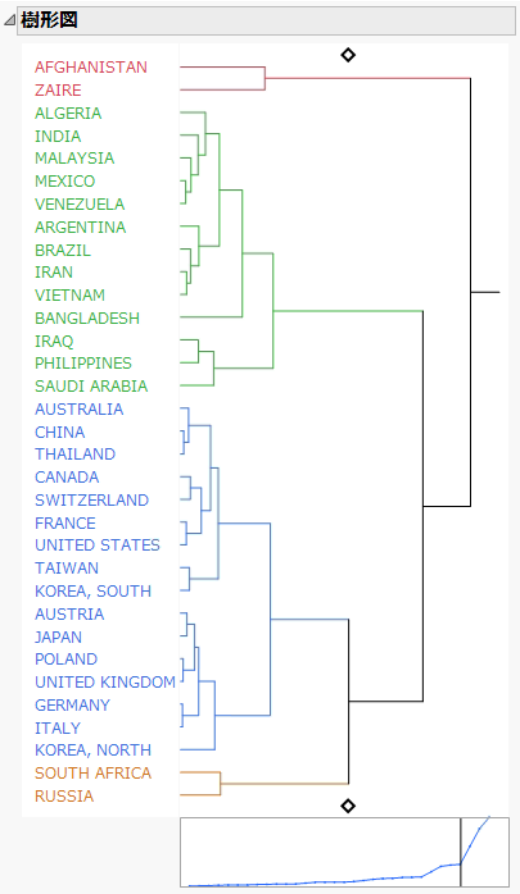


図7.5では、クラスターが4個である位置にひし形が設定されています。クラスターが4個になる直前で結合されたクラスターは、「Algeria」から「Bangladesh」までで構成されるクラスターと、「Iraq」から「Saudi Arabia」までで構成されるクラスターです。ひし形でクラスターを4個にしたときの距離グラフに、これら2個のクラスター間の距離が示されています。この距離の数値は、「クラスター分析の履歴」レポートにおいて、「クラスター数」が「4」である行に表示されています。その数値は1.618708760です。これは、4個のクラスターが形成されるときに結合された2つのクラスター（「Algeria」で始まるクラスターと、「Iraq」で始まるクラスター）の距離です。

クラスターが5個になる直前に結合された2つのクラスターは、「Australia」から「Korea, South」までで構成されるクラスターと、「Austria」から「Korea, North」までで構成されるクラスターです。これら2つのクラスター間の距離は、「Algeria」から「Bangladesh」までで構成されるクラスターと、「Iraq」から「Saudi Arabia」までで構成されるクラスターとの距離とあまり変わりません。縦線の位置を見ると、ほぼ同じ距離であることが分かります。つまり、クラスター数が4個の場合と、クラスター数が5個の場合では、その直前に結合されたクラスター間の距離にはさほど違いはありません。

距離グラフを見ると、4個のクラスターで折れ線が平らになっています。クラスターを4個から5個にしても、直前に結合されたクラスター間の距離はあまり変化しないことがわかります。しかし、クラスターを3個から4個にすると、距離は大きく減少します。

「クラスター分析の履歴」レポート

「クラスター分析の履歴」の表には、クラスター分析の履歴が含まれます。

クラスターの数 「結合先」と「結合者」が結合された**後の**、クラスターの個数を示します。クラスターの個数は、（最初の結合が行われた後の）*n - 1* 個から始まります。ここで、*n* はデータの行数です。そして、すべての行が1つのクラスターに含まれるまで降順にリストされます。このように、履歴の情報は樹形図でいうと左から右の順序で表示されています。

距離 起動ウィンドウで選択した距離の手法に従って計算されたクラスター間の距離。「距離の計算方法」（137ページ）を参照してください。

結合先 結合されるクラスターを示します。結合される2つのクラスターのうち、樹形図で上のほうに表示されているクラスターです。樹形図でのクラスターの表示順序や、「結合先」列の値は、データの並び順に左右されるもので、本質的な意味はありません。

結合者 結合されるクラスターを示します。結合される2つのクラスターのうち、樹形図で下のほうに表示されるクラスターです。樹形図でのクラスターの表示順序や、「結合者」列の値は、データの並び順に左右されるもので、本質的な意味はありません。

階層型クラスター分析のオプション

「階層型クラスター分析」の赤い三角ボタンのメニューには、次のようなオプションがあります。

クラスターの色分け 所属しているクラスターに従って、樹形図のラベルや結合線を色分けします。また、データテーブルの行も色分けされます。クラスターの数を変更すると、色が更新されます。このオプションの選択を解除すると、クラスターの数を変更しても色は更新されません。

クラスターのマーカー分け 所属しているクラスターに従って、データテーブルの各行にマーカーをつけます。クラスターの数を変更すると、マーカーが更新されます。このオプションの選択を解除すると、クラスターの数を変更してもマーカーは更新されません。

クラスターの数 表示されたウィンドウにクラスター数を入力すると、樹形図上のひし形ハンドルの位置がそれに合わせて調整されます。

クラスター数の選択規準 各クラスター数に対して、立方体クラスター規準（CCC; Cubic Clustering Criterion）を計算します。CCCはクラスター数の推定値を計算するために使用されます。これは、距離をベースとする任意のクラスタリングアルゴリズムとともに使用できます。CCCの値が大きいほど、クラスターの数に関してあてはまりが良いことを示します。SAS（1983）を参照してください。（[**データは距離行列**]を選択した場合は表示されません。）

樹形図の表示 樹形図の表示／非表示が切り替わります。

樹形図のスケール 樹形図のスケールを設定する以下のようなオプションがあります。

距離スケール 樹形図を、距離に比例した長さで描きます。使用される距離は、起動ウィンドウで指定された距離です。水平方向の結合点間の長さが、距離に比例したものになります。これは「距離グラフ」で使用されるスケールと同じです。このオプションは、樹形図のデフォルトです。

等間隔 樹形図において、2つの結合点間の水平方向の長さを、等間隔で表示します。

幾何級数 樹形図において、2つの結合点間の水平方向の長さを、クラスター数だけに応じて設定します。このオプションは、データの行数が多いときに、小さなクラスターの部分を詳しく見たい場合に便利です。

距離グラフ 樹形図の下にある距離グラフの表示／非表示が切り替わります。

クラスター数ハンドルの表示 樹形図上で、クラスター数の変更に使うハンドルの表示／非表示が切り替わります。

選択した行にズーム 樹形図でクラスターを選択し、このオプションを選択すると、そのクラスターを中心に樹形図がズームされます。クラスターをダブルクリックしても、同様にズームされます。元の表示に戻すには［ズームの解除］を選択します。

ズームの解除 樹形図のズームを元に戻します。

選択したクラスターでピボット 選択したクラスターの下位にある2つのクラスターの順序を入れ替えます。

カラーマップ 各「Y, 列」変数を値によって色分けするカラーマップ（ヒートマップともいう）を追加するためのオプションを表示します。サブメニューには、色のパターンの選択肢があります。

変数間クラスター 行と同様に「Y, 列」に指定した変数もクラスタリングします。列の樹形図に加え、カラーマップも表示されます。通常、列の値は同じスケールで測定されていなければならず、［データの標準化］を選択してはいけません。（[**積み重ねたデータ**]を選択した場合は表示されません。）

位置 樹形図のラベルなどの位置を変更するオプションを表示します。

凡例 カラーマップに使用されている色の凡例の表示／非表示を切り替えます。このオプションは、カラーマップが有効な場合のみ使用できます。

カラーマップに列の追加 指定された列のカラーマップを追加します。（[**要約したデータ**]、[**データは距離行列**]、または［**積み重ねたデータ**］を選択した場合は表示されません。）

星座樹形図 階層クラスター分析の星座樹形図の表示／非表示を切り替えます。個々のデータ行を端点、各クラスター結合を新しい点として表し、線で結んで所属関係を表します。線の長さはクラスター間の距離を表し、線が長いほど、クラスター間の距離が離れていることを示します。

星座樹形図の線にマウスを置けば、長さが表示されます。ただし、長さは相対値です。軸のスケール、点の方向、線の角度は恣意的です。ノードの末端が離れたり、プロットが雑然としたものにならないように決定されます。これは大規模なデータセットで重要になります。

星座樹形図内を右クリックし、[**ラベル表示**] を選択解除すると、端点のラベルが表示されなくなります。

星座樹形図の保存 星座樹形図の座標をデータテーブルに保存します。([**要約したデータ**]、[**データは距離行列**]、または [**積み重ねたデータ**] を選択した場合は表示されません。)

クラスターの保存 クラスター番号を含むデータテーブル列を作成します。起動ウィンドウで [空間的な指標の計算] を選択した場合、クラスターメンバーはハフデータテーブルにも保存されます。

最も近いクラスターの計算式を保存 新しいデータテーブル列を作成し、最も近いクラスターの計算式を保存します。このオプションは、各クラスターの重心までのユークリッド距離を計算し、最も近いクラスターを選択します。この計算式で求められるクラスターの割り当ては、「階層型クラスター分析」で求められるものとは必ずしも一致しません。クラスターの決定方法が異なるためです。ただし、かなり似通ったものとなります。([**要約したデータ**]、[**データは距離行列**]、または [**積み重ねたデータ**] を選択した場合は表示されません。)

表示順序の保存 樹形図内での行の順序を含むデータテーブル列を作成します。

クラスター階層の保存 スクリプトで樹形図を描くために必要な情報がデータテーブルに保存されます。結合ごとに、結合先・結合者・結果を表す3つの行が作成され、クラスター中心・サイズなどの情報が保存されます。

樹形図データの保存 JMP と SAS でクラスター分析の樹形図を比較する場合に必要な情報をデータテーブルに保存します。結合ごとに、新しいクラスターにつき1行ずつ、クラスターサイズなどの情報が保存されます。

距離行列の保存 行間の距離を含んだデータテーブルを新たに作成します。

クラスター平均の保存 クラスターごとの行数（オブザベーション数）と平均を含んだデータテーブルを新たに作成します。

クラスター要約 ([**データは距離行列**] を選択した場合は表示されません。) 以下の情報を表示します。

クラスター平均 クラスターごとの行数と平均を含んだ表です（ただし、積み重ねデータの場合は、行数ではなくて、「対象の ID」の個数です)。

クラスター標準偏差 クラスターごとの行数と標準偏差を含んだ表です（ただし、積み重ねデータの場合は、行数ではなくて、「対象の ID」の個数です)。

クラスター平均プロット 平均のパラレルプロットまたは2次元ヒートマップを表示します。

[**積み重ねデータ**] で「属性の ID」列を2つ指定したとき以外は、パラレルプロットが表示されます。パラレルプロットでは、各変数の軸のスケールが次のように設定されます。

– ［データの標準化］を選択した場合、軸は平均±2 標準偏差の範囲となります。この場合、標準偏差と平均は生データに対して計算されます。クラスター平均がこの範囲を超えた場合、軸は延長されます。

– ［データの標準化］を選択しなかった場合は、共通したスケールの縦軸が使われます。（このスケールはグラフビルダーの［スケールの統一］オプションと同じです。）

［**積み重ねデータ**］で「属性の ID」列を 2 つ指定したときは、Y 変数の平均を座標ごとに描いた 2 次元プロットが表示されます。このプロットは、［青 -> グレー -> 赤］のグラデーションによって、色付けされます。

列の要約　各変数に対して、その変数の変動のうちクラスターで説明される割合を表示します。この値は、クラスターを説明変数としたときの R2 乗です。このオプションでは、棒グラフも表示されます。

散布図行列　変数をすべて使用した散布図行列を作成します。（［**要約したデータ**］、［**データは距離行列**］、または［**積み重ねたデータ**］を選択した場合は表示されません。）

パラレルプロット　クラスターごとにパラレルプロットを作成します。（［**要約したデータ**］、［**データは距離行列**］、または［**積み重ねデータ**］を選択した場合は表示されません。）軸のスケールは、クラスター平均プロットの場合と同様です。「クラスター平均プロット」（145 ページ）を参照してください。

クラスター調整の処置比較　（Shift キーを押しながら赤い三角ボタンをクリックした場合にのみ表示されます。）1 つの応答列と 2 水準の処置列を選択します。「クラスターで調整した平均の差」レポートが表示されます。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ　現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し　分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存　レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存　By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

「階層型クラスター分析」プラットフォームのその他の例

この節では、2つの例を見ていきます。

- 「距離行列の例」（147ページ）
- 「空間的な指標でウエハーをクラスタリングする例」（149ページ）

距離行列の例

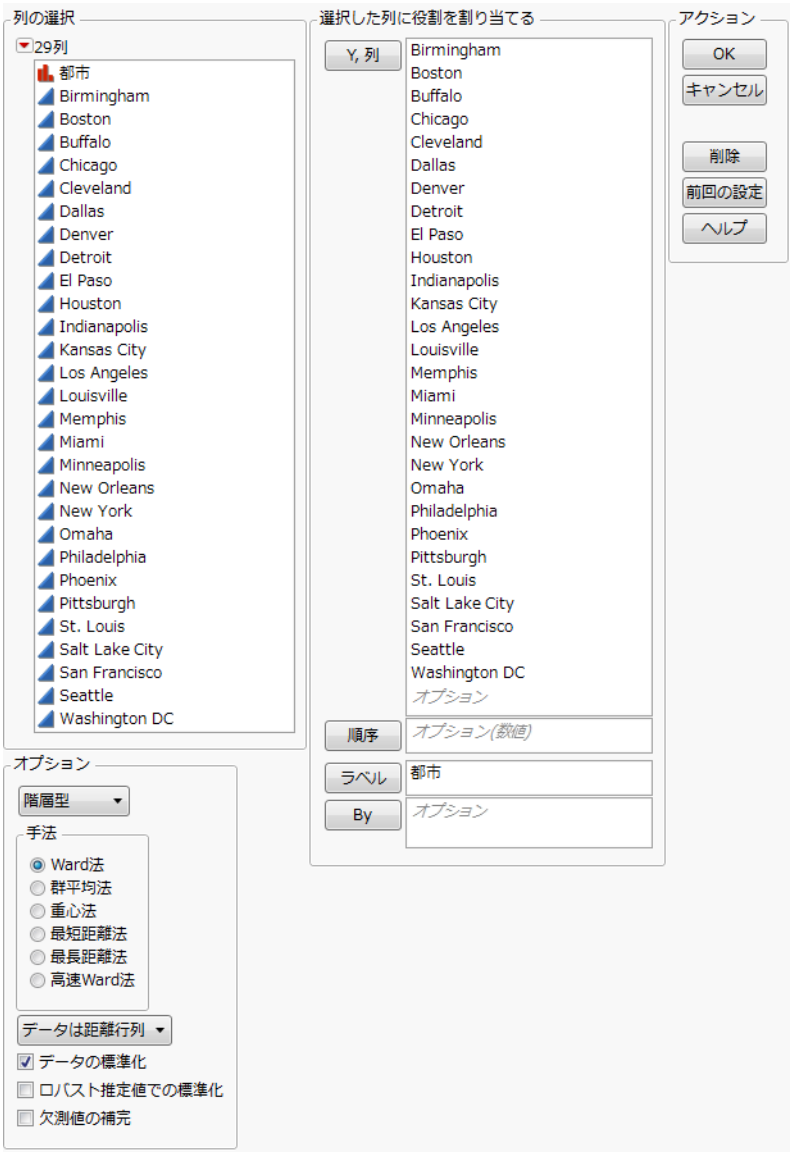
分析対象のデータが距離行列である場合、次のような列を含むデータになっていなければいけません。

- データタイプが文字のID列（通常は最初の列）。
- 距離を含む、*n* 列の数値列。*n* は行数と一致していなければいけません。これら *n* 列のデータは、0 または欠測値を対角に含む対称行列になっていなければいけません。

「Flight Distances.jmp」の距離行列で、この構造を確かめてみましょう。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Flight Distances.jmp」を開きます。
2. ［**分析**］>［**クラスター分析**］>［**階層型クラスター分析**］を選択します。
3. 起動ウィンドウの左下のリストで、［**通常のデータ**］を［**データは距離行列**］に変更します。
4. 「**都市**」を選択して［**ラベル**］をクリックします。
5. 残りのすべての列を選択し、［**Y, 列**］をクリックします。

図 7.6 設定後の距離行列の起動ウィンドウ



6. ［**OK**］をクリックします。
7. 「階層型クラスター分析」の赤い三角ボタンをクリックし、［**クラスターの色分け**］を選択します。

図 7.7「Flight Distances.jmp」の「樹形図」レポート

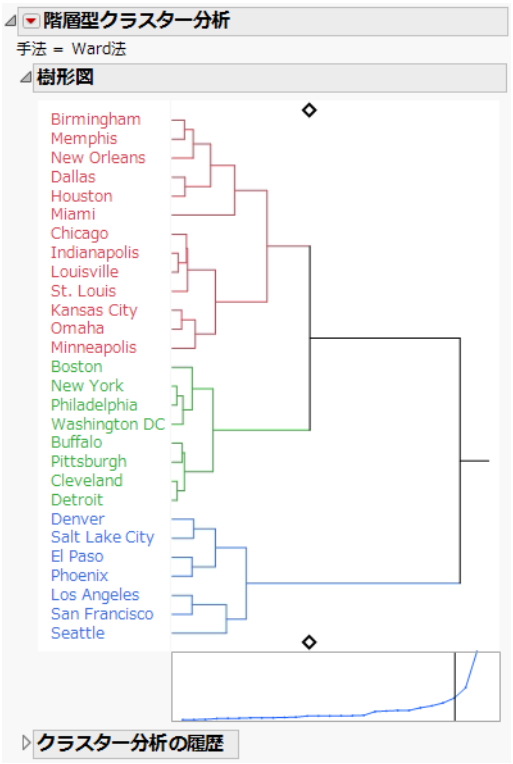


図 7.7 は、「Flight Distances.jmp」の「樹形図」レポートです。ひし形の位置は、このモデルでは都市を 3 つのクラスターにグループ化していることを示し、樹形図でこれら 3 つのクラスターが色分けされています。レポートの解釈方法の詳細については、「「樹形図」レポート」（141 ページ）を参照してください。

空間的な指標でウエハーをクラスタリングする例

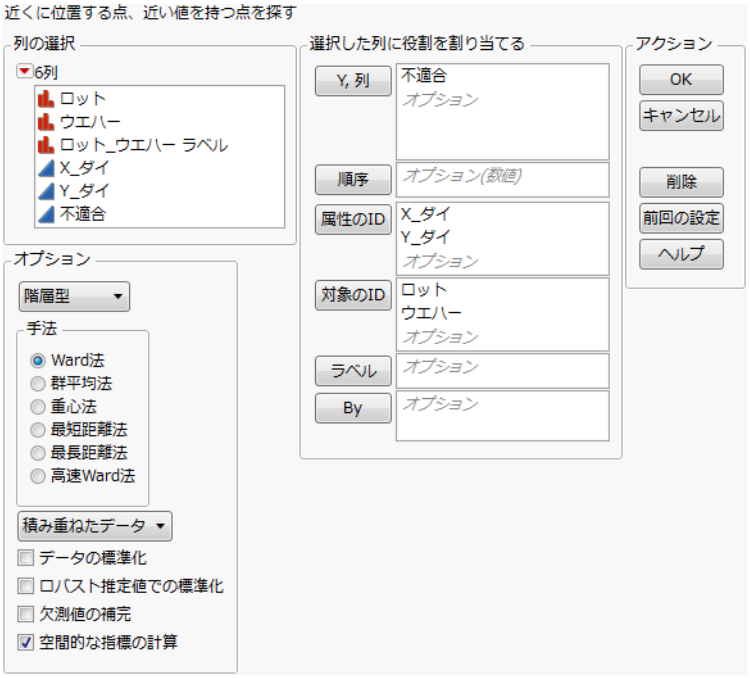
「階層型クラスター分析」プラットフォームでは、［空間的な指標の計算］というオプションを使用できます。この例では、このオプションを使ってウエハーのクラスタリングを行いましょう。このオプションの詳細については、「空間的な指標」（152 ページ）を参照してください。

1. ［**ヘルプ**］＞［**サンプルデータライブラリ**］を選択し、「Wafer Stacked.jmp」を開きます。
2. ［**分析**］＞［**クラスター分析**］＞［**階層型クラスター分析**］を選択します。
3. 起動ウィンドウの左下のリストで、［**通常のデータ**］を［**積み重ねたデータ**］に変更します。

   積み重ねたデータ用の追加のオプションが起動ウィンドウに表示されます。
4. 「**不適合**」を選択して［**Y, 列**］をクリックします。
5. 「**X_ ダイ**」と「**Y_ ダイ**」を選択して、［**属性の ID**］をクリックします。
6. 「**ロット**」と「**ウエハー**」を選択して、［**対象の ID**］をクリックします。

7. 左下隅にあるオプションリストから、[**空間的な指標の計算**] を選択します。

図 7.8 設定後の「クラスター分析」起動ウィンドウ



8. [**OK**] をクリックします。

図 7.9「空間的な指標」ウィンドウ

1423 箇所において、「**不適合**」か否かが測定されたため、属性は 1423 個あります。

9. ［**OK**］をクリックして、「空間的な指標」ウィンドウの選択を確定します。

   「階層型クラスター分析」レポートと「Wafer Stacked 不適合 空間的な指標」データテーブルの2つのウィンドウが表示されます。

10. 樹形図の最上部にあるひし形のハンドルをクリックしてドラッグし、さまざまなクラスター数について検討してみます。

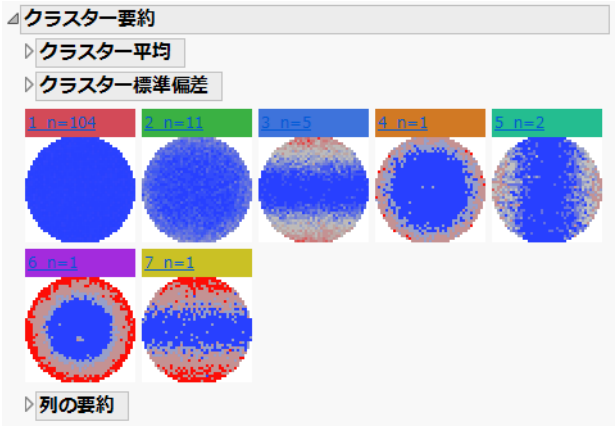
    ハンドルをドラッグすると、樹形図の下にある距離グラフの縦線が、対応するクラスター数の位置に移動します。縦軸は該当のステップで結合されたクラスター間の距離です。グラフでは、クラスター数が7のときに傾きが横ばいになっているようです。

11.「階層型クラスター分析」の赤い三角ボタンをクリックし、［**クラスターの数**］を選択します。

12.「7」と入力し、［**OK**］をクリックします。

13.「階層型クラスター分析」の赤い三角ボタンをクリックし、［**クラスターの要約**］を選択します。

図7.10「クラスター要約」レポート



ウエハーマップは、クラスターごとに不適合がどの位置で生じているかを描いています。クラスター1には、104個のウエハーが含まれており、どの位置でも不適合は少ないです。クラスター3には、5個のウエハーが含まれており、上部と下部に不適合が生じています。同時に作成されたデータテーブルで、各ウエハーに対して、ウエハーマップとホフ空間マップを確認できます。詳細については、「空間的な指標」（152ページ）を参照してください。

統計的詳細

ここでは、「階層的クラスター分析」プラットフォームの統計的詳細について説明します。

空間的な指標

［空間的な指標の計算］オプションを使用するには、データは積み重ねたデータでなければならず、また、座標（X 座標と Y 座標）が含まれている 2 つの列を [属性の ID] に指定する必要があります。いくつかの空間的な指標は、ハフ変換に基づいて計算されます。詳しくは、White et al.（2008）および Ballard（1981）を参照してください。また、「空間的な指標でウエハーをクラスタリングする例」（149 ページ）も参照してください。

「空間的な指標の選択」ウィンドウ

起動ウィンドウにおいて次の設定をした後に［OK］ボタンをクリックすると、「空間的な指標の選択」ウィンドウが呼び出されます。

- データ形式として［積み重ねたデータ］を選択する
- X 座標と Y 座標が含まれている 2 つの列を、［属性の ID］に指定する
- ［対象の ID］列を 1 つ指定する
- ［空間的な指標の計算］チェックボックスにチェックする

「空間的な指標の選択」ウィンドウでは、どの空間的な指標をクラスター分析に用いるかを選択します。また、それらの変数に対する重みを設定します。

変数　どの変数や指標を、クラスター分析で使用するかを選択します。空間的な指標は、応答 Y から算出されます。

属性　これは Y 変数の値そのものです。各位置（2 つの「属性の ID」変数で定義されている位置）での Y 変数の値です。

扇形の角度　扇形や半球形に分布していることを示す指標です。

円の半径　円状（ドーナッツ状）に分布していることを示す指標です。

筋の角度　同じ角度である筋（直線）を表す指標です

筋の位置　同じ距離である筋（直線）を表す指標です。

ショット 領域をいくつかの四角形に分けたときに、特定の四角形に分布していることを示す指標です。計算に用いる四角形の横と縦の長さを指定してください。**「ショット」**という用語は、半導体ウエハーで使用されているものです。「ショット」は、複数のダイを含む矩形領域を意味します。

「ショットの横サイズ」と「ショットの縦サイズ」に値を入力してください。横サイズを4、縦サイズを5にすると、1つのショットに20個のダイが含まれることを意味します。一方、ショットの総数は次のように計算されます。

floor[(hSize+hShotSize-1)/hShotSize]* floor[(vSize+vShotSize-1)/vShotSize]

ここで、hSizeはウェーハーの横サイズ、vSizeはvSizeは縦サイズです。また、hShotSizeは「ショットの横サイズ」、vShotSizeは「ショットの縦サイズ」です。

メモ: ショット指標の変数名は、Shot[vert, horiz]のように付けられます。ここで、vertは垂直方向におけるダイの位置、horizは水平方向におけるダイの位置です。

個数 指定された指標の個数です。指定された個数の指標がクラスター分析に使われます。

重み クラスター分析に使用される各指標にどれだけの重みを与えるかを決めるもの。（クラスター分析での各変数に対する重み）

空間的な指標のレポート

以上のような設定をした後に「空間的な指標の選択」ウィンドウで［OK］をクリックすると、レポートとデータテーブルが表示されます。

「階層型クラスター分析」レポート

積み重ねたデータを使い、2つの「属性のID」列を指定して分析を行った場合、［クラスターの要約］を選択すると、マップ（ウェーハーマップ）も描かれます。このマップは、「属性のID」列で定義された各座標でのクラスター平均を描いたものですです。これらのマップでは、「青-> グレー -> 赤」のカラーグラデーションで、分位点をもとに色を付けています。分位点を色付けに用いるのは、外れ値による影響を軽減するためです。

空間的な指標のデータテーブル

新たに作成される空間的な指標のデータテーブルには、「対象のID」の一意な組み合わせごとに1つの行で構成されています。データの各セルには、「青-> グレー -> 赤」のカラーグラデーションが使われています。このデータテーブルには以下の列があります。

対象 2つの「属性のID」変数によって定義されたX座標とY座標での応答値を示したヒートマップ。列のタイプは、「式」となっています。

ハフ ハフ空間のヒートマップ。列のタイプは、「式」となっています。ハフ空間に関しては、White et al.（2008）を参照してください。

空間的な指標 空間的な指標ごとに1列ずつ作成されます。それぞれに計算された空間的な指標が含まれています。セルは値によって色分けされます。

距離の手法の計算式

ここでは、「階層型クラスター分析」プラットフォームで使用される距離を説明します。どの距離を分析に使用するかは、起動ウィンドウで選択できます。それぞれの距離の特徴については、「距離の計算方法」（137 ページ）も参照してください。

計算式では、次のような記号を使います。小文字はオブザベーション（データ行）に関連するもの、大文字はクラスターに関連するものです。

n: オブザベーションの数

v: 変数の数

x_i: i 番目のオブザベーション

C_K: K 番目のクラスター。{1, 2, ..., n} の部分集合を含む。

N_K: C_K に含まれるオブザベーションの数

$\bar{x}$: 標本平均ベクトル

$\bar{x}_K$: クラスター C_K の平均ベクトル

$\|x\|$: $\mathbf{x}$ の要素の平方和の平方根（ベクトル $\mathbf{x}$ のユークリッド距離）

$d(x_i,x_j)$: $\|x_i - x_j\|^2$

群平均法　群平均法の距離（の 2 乗）は、次の式で計算されます。

$$D_{KL} = \sum_{i \in C_K} \sum_{j \in C_L} \frac{d(x_i, x_j)}{N_K N_L}$$

重心法　重心法の距離（の 2 乗）は、次の式で計算されます。

$$D_{KL} = \left\| \bar{x}_K - \bar{x}_L \right\|^2$$

Ward 法　Ward 法の距離（の 2 乗）は、次の式で計算されます。

$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

最短距離法　最短距離法の距離（の 2 乗）は、次の式で計算されます。

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

最長距離法　最長距離法の距離（の 2 乗）は、次の式で計算されます。

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

第 8 章

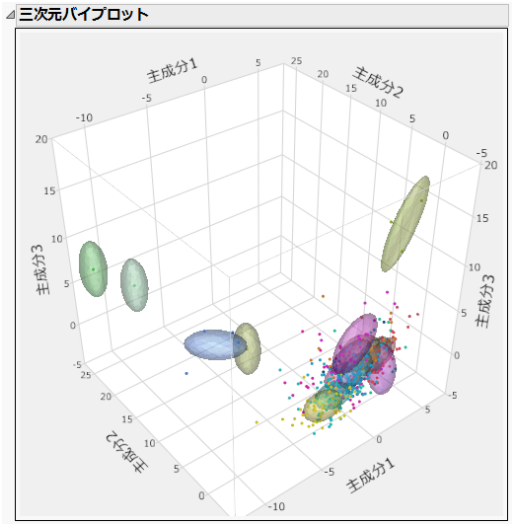
K Means クラスター分析

データ行をクラスタリング

k-means 法も、多変量データをもとに、値が近い行をグループにまとめる手法です。*k*-means 法は、200～100,000 行ほどある大きなデータテーブルに適しています。

k-means 法は、事前に指定されたクラスター数（クラスターの個数）に対して、反復アルゴリズムを用います。**k-means** は、クラスターの重心からの距離が最小になるように、各データ行をクラスターに分類します。反復計算を行う前に、クラスター数（*k*）を指定しておく必要があります。ただし、さまざまなクラスター数（*k*）を試してみて、その後に現在のデータに最もふさわしいだろうクラスター数を選ぶことはできます。

図 8.1 三次元バイプロット



「K Means クラスター分析」プラットフォームの概要

JMP には、データ行をクラスタリングするためのプラットフォームが 4 つ用意されています。「K Means クラスター分析」は、そのなかの 1 つです。4 つの手法の比較については、「クラスター分析用プラットフォームの概要」（156 ページ）を参照してください。

「K Means クラスター分析」プラットフォームは、事前に指定されたクラスター数（クラスターの個数）に対して、反復アルゴリズムを用います。まず、クラスター数と同数の k 個の点が選択されます。この点は、**「クラスターシード」**と呼ばれており、クラスターの平均を示す最初の推定値です。そして、最も近くにあるクラスターシードに各データ行が割り振られます。次に、クラスターごとに平均を計算し、既存のクラスターシードをそれらの新しく計算された平均に置き換えます。そして、そのように新しく計算されたクラスターシードに、データ行が再び割り振られます。この処理が反復されると、最後にはクラスターシードの平均や割り振りに変化が生じない状態になります。

k-means 法のこのようなアルゴリズムは、**EM アルゴリズム**の特殊形態です。EM アルゴリズムの *E* は期待値（Expectation）、*M* は最大化（Maximization）を意味します。*k*-means 法のアルゴリズムでは、クラスター平均の計算が「期待値」の E ステップで、最も近いクラスターへの点の割り当てが「最大化」の M ステップになっています。

「*K* Means クラスター分析」プラットフォームは、数値の列しか使用できません。列の尺度（名義尺度、順序尺度）は無視され、数値列がすべて連続尺度として処理されます。

「K Means クラスター分析」プラットフォームでは、事前にクラスター数 k（または k の範囲）を指定しておく必要があります。ただし、さまざまな k の結果を比べて、データに最もふさわしいだろうクラスター数を後から選択することはできます。

k-means 法の背景については、SAS Institute Inc.（2005）および Hastie et al.（2009）を参照してください。

クラスター分析用プラットフォームの概要

クラスタリングは、多変量データをもとに、値が近い行をグループにまとめていく手法です。通常、データ点は n 次元空間全体に均等に散らばっておらず、いくつかの塊（クラスター）になっているでしょう。それらのクラスターを見つけ出すと、データをよりよく理解できるようになるでしょう。

メモ： JMP には、変数をクラスタリングするためのプラットフォームも用意されています。「変数のクラスタリング」（197 ページ）章を参照してください。

JMP には、データ行（オブザベーション）をクラスタリングするためのプラットフォームが 4 つ用意されています。

- 「階層型クラスター分析」は、数千行までの小さなテーブルに適しており、文字データにも対応します。行がツリー型の階層構造にまとめられます。クラスタリングの処理が終わった後でも、クラスターの個数を変更することができます。
- 「K Means クラスター分析」は、数十万行までの大きいデータに適しています。この分析は、数値データだけに対応しています。処理を開始する前に、クラスターの数 k を指定する必要があります。まず、適切

と思われるシード点が推定されます。その後、各点をクラスターに割り当てる作業とクラスター中心を再計算する作業が交互に繰り返されます。

- 「正規混合」は、複数の多変量正規分布の混合分布から得られた、重なりがあるデータに適しています。この分析は、数値データだけに対応しています。外れ値があるような場面では、それらの外れ値を表すために、一様分布に従うと仮定したクラスターを使用できます。また、［ロバスト正規混合］オプションによりロバストな推定を行うこともできます。

  この手法では、処理を開始する前に、クラスターの個数を指定しておく必要があります。最尤法によって、混合割合、平均、標準偏差、相関係数といったパラメータが同時に推定されます。各点に、それぞれの各グループに属する事後確率が計算されます。推定値の反復計算にはEMアルゴリズムが使用されています。

- カテゴリカルデータの場合は、「潜在クラス分析」が適しています。この手法では、処理を開始する前に、クラスターの個数を指定しておく必要があります。多項分布の混合分布がモデルとして仮定されます。各データ行に対して、各クラスターに属する事後確率が計算されます。そして、属する事後確率が最も高いクラスターに分類されます。

表 8.1 クラスター分析の手法のまとめ

手法	データタイプまたは尺度	データテーブルのサイズ	クラスター数の指定
階層型クラスター分析	すべて	高速 Ward 法の場合、200,000 行まで その他の手法の場合、5,000 行まで	なし
K Means クラスター分析	数値	数百万行まで	あり
正規混合分布法	数値	制限なし	あり
潜在クラス分析	名義尺度または順序尺度	制限なし	あり

K-Means クラスターの例

この例では、「Cytometry.jmp」サンプルデータテーブルを使って、データ行のクラスタリングを行います。サイトメトリー（cytometry）は、細胞の表面のマーカーを検出するのに使用されています。これらのマーカーは、特定の疾病を診断するのに役立ちます。この例では、サイトメトリー測定で読み取った 4 つのマーカーに基づいて、データ行をグループに分けます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Cytometry.jmp」を開きます。
2. ［**分析**］>［**クラスター分析**］>［**K Means クラスター分析**］を選択します。
3. 「**CD3**」、「**CD8**」、「**CD4**」、「**MCB**」を選択して［**Y, 列**］をクリックします。
4. ［**OK**］をクリックします。
5. 「**クラスターの数**」の横に「3」と入力します。

6. 「**オプション クラスターの最大数**」の横に「15」と入力します。

   クラスターの最大数を 15 に設定したため、3 ～ 15 個のクラスターのあてはめが行われます。その後、最も適しているであろうクラスター数を選択できます。

7. ［**実行**］をクリックします。

図 8.2「クラスターの比較」レポート

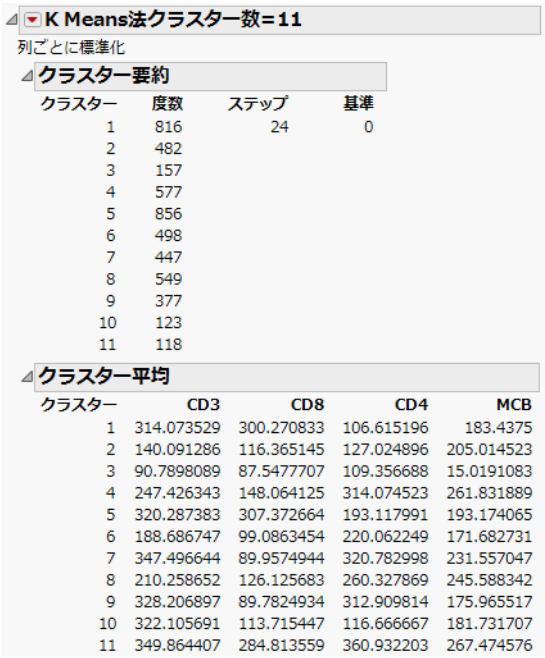
クラスターの比較

方法	クラスター数	CCC	最適
K-Meansクラスター分析	3	23.1784	
K-Meansクラスター分析	4	8.80709	
K-Meansクラスター分析	5	29.5123	
K-Meansクラスター分析	6	52.5517	
K-Meansクラスター分析	7	49.5876	
K-Meansクラスター分析	8	56.5308	
K-Meansクラスター分析	9	54.053	
K-Meansクラスター分析	10	69.8707	
K-Meansクラスター分析	11	70.5239	最適CCC
K-Meansクラスター分析	12	61.5326	
K-Meansクラスター分析	13	68.1277	
K-Meansクラスター分析	14	66.4044	
K-Meansクラスター分析	15	69.9928	

「クラスターの比較」レポートは、レポートウィンドウの最上部に表示されます。CCC（Cubic Clustering Criterion：立方クラスタリング規準）の最も大きいときのクラスタリングが、最も適していると考えられるものです。この例では、クラスター数が 11 のときに、CCC が最大となっています。

8. 「K Means 法クラスター数=11」のレポートまでスクロールします。

図 8.3「K Means 法クラスター数 =11」レポート



K Means法クラスター数=11

列ごとに標準化

クラスター要約

クラスター	度数	ステップ	基準
1	816	24	0
2	482		
3	157		
4	577		
5	856		
6	498		
7	447		
8	549		
9	377		
10	123		
11	118		

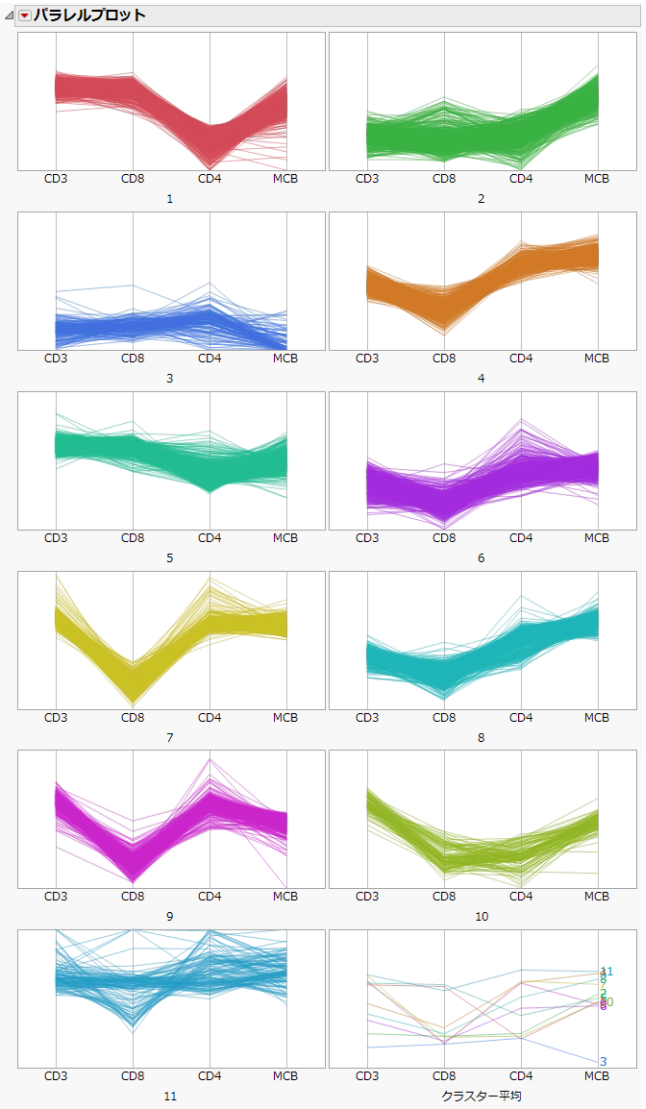
クラスター平均

クラスター	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

「クラスター要約」レポートには、11 個のクラスターそれぞれのオブザベーション数（データの行数）が表示されています。「クラスター平均」レポートには、4 つのマーカーの読み取り値の、クラスターごとの平均が表示されています。

9. 「K Means 法クラスター数=11」の赤い三角ボタンをクリックし、［**パラレルプロット**］を選択します。

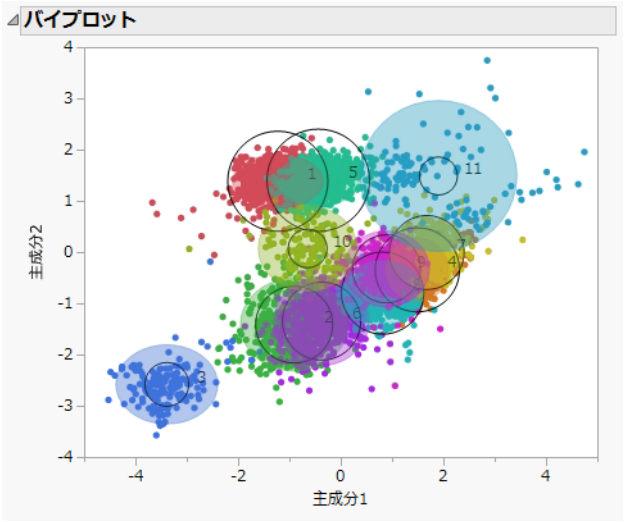
図 8.4 Cytometry データのパラレルプロット



パラレルプロットからは、どのようにデータがクラスターごとに分布しているかが分かります。これらのプロットを使って、クラスター間でどのような違いがあるかを確認しましょう。クラスター 4、6、7、8、9 では、**CD8** の値が比較的低く、**CD4** の値が高いことがわかります。一方、クラスター 1 では **CD8** の値が高く、**CD4** の値が低くなっています。

10.「K Means 法クラスター数=11」の赤い三角ボタンをクリックして［**バイプロット**］を選択します。

図 8.5 Cytometry データのバイプロット

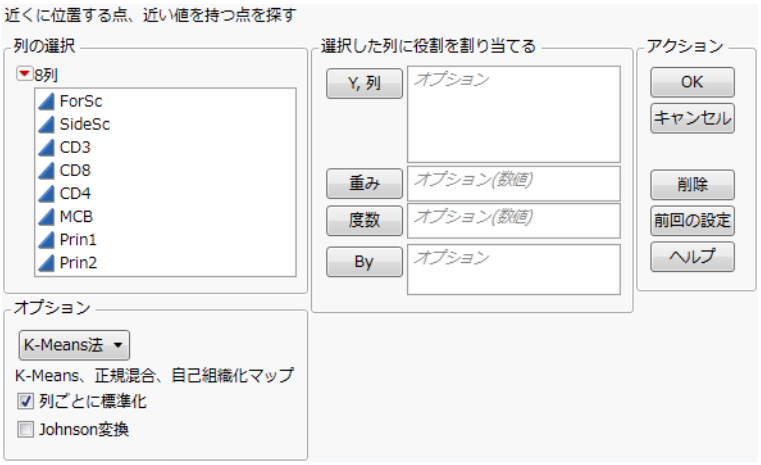


最初の 2 つの主成分に基づいて、他から最も離れているように見えるクラスターは、クラスター 3、10、11 です。これらは、図 8.4 のパラレルプロットを見ても、他のクラスターのプロットと異なっていることが確認できます。

「K Means クラスター分析」プラットフォームの起動

「K Means クラスター分析」プラットフォームを起動するには、[**分析**] > [**クラスター分析**] > [**K Means クラスター分析**] を選択します。図 8.6 は、「Cytometry.jmp」に対する起動ウィンドウです。

図 8.6 K Means クラスター分析の起動ウィンドウ



Y, 列 クラスター分析の対象となる変数を指定します。

メモ : *k*-means クラスター分析では、数値の列しか使用できません。

重み この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数 この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

By この列の水準に従ってデータがグループ化され、それぞれ個別に分析されます。指定した列の水準ごとに、対応する行が分析されます。分析結果は、個別のレポートにまとめられます。複数の By 変数を指定した場合は、By 変数の水準のすべての組み合わせごとに分析が行われます。

起動ウィンドウのオプション

デフォルトの手法として［K-Means 法］が選択されていますが、［階層型］または［正規混合］を選択することもできます。［K-Means 法］または［正規混合］を選択し、［OK］をクリックすると、設定パネルが表示されます。「「反復クラスター分析」設定パネル」（163 ページ）を参照してください。

列ごとに標準化 列ごとに個別に標準化します。変数の測定尺度が異なる場合に、1 つの変数が結果に強い影響をもつのを防ぐのに用います。たとえば、ある変数が 0 から 1000 までに分布しており、別の変数が 0 から 10 までに分布していたとします。標準化すれば、前者の変数によってクラスター分析の結果がほとんど決まってしまう現象を回避できます。

Johnson 変換 Y 変数に Johnson 分布をあてはめ、その結果をもとに正規分布に近づくようにデータを変換します。［列ごとに標準化］を選択した場合は、Y 変数ごとに Johnson 変換が行われます。［列ごとに標準化］を選択しなかった場合は、すべての Y 変数の値に単一の Johnson 変換が行われます。Johnson Sb 分布と Johnson Su 分布が検討され、最尤法によって推定されます。Johnson Sb 分布と Johnson Su 分布については、『基本的な統計分析』の「一変量の分布」章を参照してください。

Johnson 変換はデータが正規分布に近づき、歪度が小さくなるように、また、変換後に外れ値が分布の中央により近づくように、データの変換を試みます。

「反復クラスター分析」レポート

起動ウィンドウで［OK］をクリックすると、以下のものを表示した「反復クラスター分析」レポートウィンドウが開きます。

- 「変換」レポート （［Johnson 変換］オプションを選択した場合にのみ表示されます。）「変換」レポートには、［Y, 列］に指定した変数ごとに、変換の情報が表示されています。Y 変数ごとに、Johnson Sb 分布と Johnson Su 分布のいずれかのうち、あてはまりが良かったほうの $\gamma, \delta, \theta, \sigma$ に対するパラメータ推定値が表示されています。Johnson 分布の詳細については、『基本的な統計分析』の「一変量の分布」章を参照してください。
- モデルのあてはめに関する「設定パネル」。「設定パネル」については、「「反復クラスター分析」設定パネル」（163 ページ）を参照してください。

モデルをあてはめるごとに、ウィンドウにレポートが追加されていきます。「「K Means 法クラスター数=<k>」レポート」（165 ページ）を参照してください。

反復クラスター分析のオプション

変換の計算式を保存 （Johnson 変換オプションを選択した場合にのみ表示されます。）変換したデータをデータテーブルの新しい式列として保存します。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

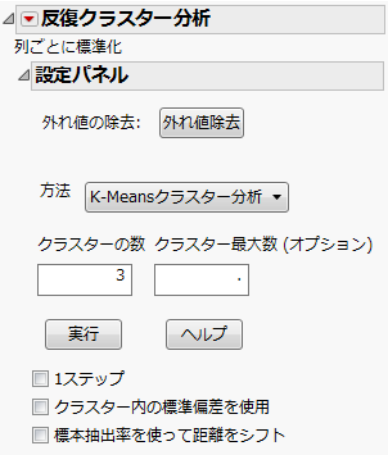
スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

「反復クラスター分析」設定パネル

図 8.7 は、「Cytometry.jmp」データテーブルの「設定パネル」です。クラスター数を 1 つずつ指定して 1 回、1 回、実行していくことも、［クラスター最大数］オプションを使って、あてはめるクラスター数の範囲を指定し、まとめて実行することもできます。

図 8.7「反復クラスター分析」設定パネル



この設定パネルには次のようなオプションがあります。

外れ値の除去　多変量の外れ値を検出します。調べたい近傍点の最大数を入力します。[OK] を入力すると、各点から第 1 近傍点への距離、各点から第 2 近傍点への距離、… とグラフが描かれていき、指定した最大数までのグラフが描かれます。

注意 : 近傍点における個数を表すには、「*k*」という記号が使われることが多いです。一方、*k*-means 法ではクラスター数が「*k*」と表されることが多いです。この両者は関係ありません。

「第 k 番目の近傍点への距離」プロットの下に、次のようなオプションが表示されます。

除外のリセット　同じレポート内で新しいクラスター分析を実行する場合に、現在の除外状態を反映させます。たとえば、近傍点プロットや散布図から、データテーブルのいくつかの行の属性を「除外」にしたとします。同じレポート内の設定パネルで新たなクラスター分析を実行するときに、それらの行を計算から除外したい場合には、このボタンをクリックしてください。

散布図行列　すべての Y 変数に対する散布図行列を描きます。この散布図行列は、現在のレポートとは別のウィンドウに表示されます。

近傍距離の保存　各点から *k* 番目の近傍点までの距離を、新しい列に保存します。

閉じる　外れ値除去の結果をレポートから削除します。

手法　次のクラスター分析手法を使用できます。

K-Means クラスター分析　この章で説明しています。

正規混合分布法　「正規混合分布法」章の「「反復クラスター分析」設定パネル」（177 ページ）で説明しています。

ロバスト正規混合　「正規混合分布法」章の「「正規混合 クラスター数=<k>」レポート」（179 ページ）で説明しています。

自己組織化マップ 「自己組織化マップの設定パネル」（168 ページ）で説明しています。

クラスターの数 形成されるクラスターの数。

オプション クラスター最大数 形成されるクラスター数の上限値。この値を指定した場合、[**クラスターの数**] の値とこの値の間にあるすべての整数に対し、分析が行われます。

実行 ［1 ステップ］が選択されている場合を除き、k-means 法の反復計算を最後まで実行します。

1 ステップ k-mean 法の反復計算を 1 ステップずつ実行したい場合には、このチェックボックスにチェックします。［1 ステップ］を選択した場合、［実行］ボタンをクリックしてもクラスター分析は実行されません。レポートに［実行］ボタンと［ステップ］ボタンという 2 つのボタンが表示されます。

- 一度に 1 ステップずつ反復を実行するには［ステップ］ボタンをクリックします。
- 反復計算を最後まで行うには［実行］ボタンをクリックします。

クラスター内の標準偏差を使用 このオプションを使うと、距離は各クラスター内で推定された標準偏差によって標準化して計算されます。このオプションを使わない場合、距離は変数ごとに全体の標準偏差の推定値によって標準化して計算されます。

標本抽出率を使って距離をシフト クラスターのサイズ（該当するクラスターに属するデータの行数）に基づいて距離を調整します。クラスターのサイズが異なる場合、小さなクラスターより大きなクラスターに属する事前確率の方が高いと仮定されます。そのため、サイズが大きなクラスターに割り当てられる確率のほうが高くなります。

「K Means 法クラスター数=<k>」レポート

設定パネルで［実行］をクリックすると、次のレポートが表示されます。

- 「クラスターの比較」レポート 「「クラスターの比較」レポート」（166 ページ）を参照してください。
- 1 つまたは複数の「K Means 法クラスター数=<k>」レポート。ここで、*k* はクラスターの個数です。「K Means 法クラスター数=<k>」レポートは、あてはめを実行する度に表示されます。

「Cytometry.jmp」データテーブルで変数「**CD3**」から「**MCB**」を［Y, 列］に指定して実行した結果の「クラスターの比較」レポートと「K Means 法クラスター数=<k>」レポートは、図 8.2 および「「K Means 法クラスター数=11」レポート」（159 ページ）のようになります。

「クラスターの比較」レポート

「クラスターの比較」レポートには、異なるクラスター数のモデルを比較するための統計量が表示されます。この統計量は「立方体クラスター規準」(CCC: Cubic Clustering Criterion) と呼ばれているものです。CCC が大きいほど、クラスターによって適切に分けられているであろうことを示しています。最良のクラスター数には、「最適」列に「最適 CCC」と記述されます。CCC についての詳細は、SAS Institute Inc.（1983）を参照してください。なお、分析対象のデータに同じ数値しかもたない列があった場合、その列は CCC の計算から除外されます。

「K Means 法クラスター数 =<k>」レポート

「K Means 法クラスター数 =<k>」レポートには、クラスターごとに次のような要約統計量が表示されます。

- 「クラスター要約」レポートには、クラスター番号、クラスターごとのオブザベーション数（データの行数）、および、反復計算が収束するまでにかかった反復回数が表示されます。
- 「クラスター平均」レポートには、クラスターごとに分けて算出された、各変数の平均が表示されます。
- 「クラスター標準偏差」レポートには、クラスターごとに分けて算出された、各変数の標準偏差が表示されます。

「K Means 法クラスター数 =<k>」レポートのオプション

「K Means 法クラスター数 =<k>」レポートには、次のようなオプションがあります。

バイプロット　データの主成分のうち最初の 2 つを軸にして、点とクラスターをプロットします。クラスター中心の周りに円が描かれます。円の大きさはクラスター内のデータ数に比例します。陰影つきの領域は、平均を中心とした 50% の等密度面で、そのクラスター内のオブザベーションのうち 50% がその領域内に収まることを示しています（Mardia et al., 1980）。プロットの下に、クラスターの色をデータテーブルに保存するオプションが表示されます。固有値は降順に表示されます。

バイプロットオプション　バイプロットの外観をコントロールするための次のようなオプションがあります。

バイプロット線の表示　バイプロット線を表示します。ラベルの付いたバイプロット線は、主成分を基底とした部分空間における共変量の方向を示します。これは、各変数の各主成分に対する関連の度合を示します。

バイプロット線の位置　バイプロット線の位置と半径のスケールを指定できます。デフォルトでは、点 (0,0) を出発点とします。プロットでバイプロット線をドラッグして移動するか、またはこのオプションによって出発点の座標を指定できます。半径のスケールオプションを使ってバイプロット線のスケールを調整し、バイプロット線を見やすくすることもできます。

クラスターのマーカー分け　データテーブルの各行に、クラスターに応じたマーカーをつけます。

三次元バイプロット　データの 3 次元バイプロットを表示します。3 つ以上の変数がある場合のみ使用できます。

パラレルプロット　クラスターごとにパラレルプロットを作成します。オプションで、プロットにおけるデータや平均の表示／非表示を切り替えることができます。詳細については、『グラフ機能』の「パラレルプロット」章を参照してください。

散布図行列　すべてのＹ変数に対する散布図行列を描きます。この散布図行列は、現在のレポートとは別のウィンドウに表示されます。

色をテーブルに保存　データテーブルの各行に、クラスターに応じたマーカーをつけます。

クラスターの保存　データテーブルに次の2つの列を保存します。

- 各行に割り振られたクラスターの番号を含む**「クラスター」**列。
- （[自己組織化マップ] を選択した場合を除きます。）各オブザベーション（各データ行）と、それが属するクラスターの平均との間の、標準化した距離を含む**「距離」**列。まず、変数ごとに、観測値とクラスター平均との差を、その変数全体の標準偏差で割ります。これらの標準化した差から変数全体の平方和を求めます。その平方和の平方根が**「距離」**として求められます。なお、[Johnson 変換] を選択した場合、**「距離」**は変換後の変数から計算されます。

クラスター計算式の保存　「クラスター計算式」という計算式列をデータテーブルに保存します。この計算式は、どのクラスターに属するかを求めるものになっています。

クラスターのシミュレーション　クラスターの平均と標準偏差の推定値を使ってシミュレーションしたデータを含む、新しいデータテーブルを作成します。

削除　クラスター分析のレポートを削除します。

自己組織化マップ

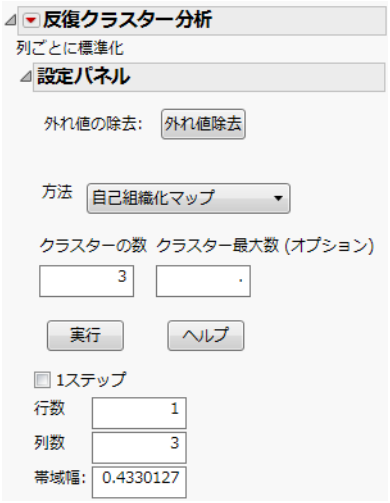
自己組織化マップ（SOM）は Teuvo Kohonen（1989、1990）によって開発され、その他のニューラルネット専門家や統計学者によって拡張されてきました。本来の自己組織化マップは、元々のニューラルネットアルゴリズムのような学習プロセスとして構築されましたが、JMP の自己組織化マップは、*k*-means クラスター分析の特殊形態に過ぎません。これは、自己組織化マップに関する文献でいう**「局所重み付き線形平滑化を行う一括学習型アルゴリズム」**に該当します。

自己組織化マップは、ただクラスターを形成するだけではありません。クラスターグリッドを作成し、結果的にグリッド上で近くに位置するクラスターの点が多変量空間においても近くに位置するよう、点を配置します。従来の *k*-means クラスター分析では、クラスターの構造が不定でしたが、自己組織化マップのクラスターはグリッド構造を持っています。このグリッド構造は、クラスターを2次元で解釈するときに大変役立ちます。クラスター間の距離が短いほど、それらのクラスターは類似するためです。「自己組織化マップのアルゴリズムについて」（169 ページ）を参照してください。

自己組織化マップの設定パネル

反復クラスター分析パネルの「方法」リストから［自己組織化マップ］オプションを選択します（図 8.8）。

図 8.8 自己組織化マップの設定パネル



パネルのオプションの一部は、「「反復クラスター分析」設定パネル」（163 ページ）で説明されています。以下、その他のオプションについて説明します。

クラスターの数　SOM では使用しません。

オプション クラスター最大数　SOM では使用しません。

行数　クラスターグリッドの行数。

列数　クラスターグリッドの列数。

帯域幅　帯域幅　(バンド幅)　は、近隣のクラスターが重心を予測するのに与える影響を指定します。帯域幅が小さいほど、より近いクラスターの重みが大きくなります。

「自己組織化マップ」レポート

図 8.9「自己組織化マップ」レポート

「自己組織化マップ」レポートには、要求されたグリッドのサイズに応じた名前が付けられます。指定された帯域幅（バンド幅）が、自己組織化マップのグリッドレポートの最上部に表示されます。レポートに示されるその他の統計的な結果は、k-means 法の結果と同じです。「「K Means 法クラスター数=<k>」レポート」（166 ページ）を参照してください。

自己組織化マップの赤い三角ボタンのメニューについては、「「K Means 法クラスター数=<k>」レポートのオプション」（166 ページ）を参照してください。

自己組織化マップのアルゴリズムについて

JMP での自己組織化マップ処理は、次のように行われます。

- 最初のクラスターシードは、多次元空間をほぼ網羅する方法で選択します。JMP では、第 2 次元までの主成分をもとに、シードを決めます。
- 次に、この主成分空間にグリッドを配置します。グリッドの一番外側の辺は、各次元で中心から 2.5 標準偏差だけ離れたところに位置します。このグリッドを元の変数空間に変換し、クラスターのシードを決定します。
- *k*-means 法と同じように、各点を最も近いクラスターに割り当てていきます。

- *k*-means 法と同じように、各クラスターの平均を計算します。そして、各変数におけるクラスター平均を応答変数、自己組織化マップのグリッド座標を説明変数とした重み付き回帰を行います。重み関数として「カーネル」関数を用いており、中心を推定するために用いたクラスターに対して大きな重みを与え、グリッド内で離れたクラスターほど重みは小さくなるようにしています。この回帰による予測値が新しいクラスター平均です。
- 以上の手順が、処理が収束するまで反復されます。

第 9 章

正規混合分布法

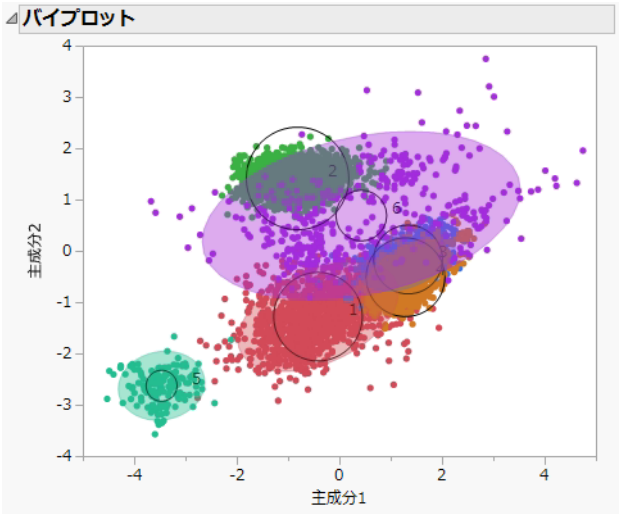
多変量正規分布によりデータ行をクラスタリング

データが重なりのある多変量正規分布から得られたものである場合は、「正規混合」プラットフォームを用いるとよいでしょう。また、このプラットフォームでは、外れ値がある場合のために、「ロバスト正規混合」というオプションも用意されています。「正規混合」プラットフォームでは、処理を開始する前に、クラスターの個数を指定する必要があります。

「正規混合」プラットフォームは、「データが多変量正規分布の混合分布に従っている」と仮定した方法です。計算には、反復アルゴリズムが使われています。複数の混合されている各多変量正規分布が、それぞれ、1つのクラスターを表します。

クラスターがはっきりと分かれている場合は、階層型クラスター分析や *k*-means 法が妥当です。しかし、クラスターが重なり合っている場合は正規混合法の方が妥当です。なぜなら、正規混合法は、境界によってグループを排他的に分類するのではなく、各クラスターに所属する確率を求めるからです。

図 9.1 正規混合バイプロット



「正規混合」クラスター分析プラットフォームの概要

JMP には、データ行をクラスタリングするためのプラットフォームが 4 つ用意されています。「正規混合」は、そのなかの 1 つです。4 つの手法の比較については、「クラスター分析用プラットフォームの概要」（172 ページ）を参照してください。

正規混合法は、数値変数に対するクラスター分析手法のひとつで、計算には反復アルゴリズムが使われています。正規混合法は、各クラスター内に分類される確率も予測します。正規混合法では、「分析対象の多変量データが、多変量正規分布の混合分布に従っている」と仮定しています。各クラスターの平均ベクトルと共分散行列の推定値が求められます。McLachlan and Krishnan（1997）および Hand et al.（2001）の第 9.6 節を参照してください。

メモ: 正規混合法のアルゴリズムでは、反復計算におけるクラスター中心の開始値に乱数が使われています。そのため、分析を実行する度に少し異なる結果になります。

多変量の外れ値があると疑われる場合は、「外れ値のクラスター」を仮定するか、または「ロバスト正規混合」を使用してください。「外れ値のクラスター」オプションは、ひとつのクラスターが一様分布に従うと仮定します。また、「ロバスト正規混合」オプションはロバストな方法でパラメータを推定します。これらの方法では、通常の正規混合法よりも外れ値による影響が少なくなります。詳細については、「外れ値のクラスター」（179 ページ）および「ロバスト正規混合の詳細」（183 ページ）を参照してください。

クラスター分析用プラットフォームの概要

クラスタリングは、多変量データをもとに、値が近い行をグループにまとめていく手法です。通常、データ点は n 次元空間全体に均等に散らばっておらず、いくつかの塊（クラスター）になっているでしょう。それらのクラスターを見つけ出すと、データをよりよく理解できるようになるでしょう。

メモ: JMP には、変数をクラスタリングするためのプラットフォームも用意されています。「変数のクラスタリング」（197 ページ）章を参照してください。

JMP には、データ行（オブザベーション）をクラスタリングするためのプラットフォームが 4 つ用意されています。

- 「階層型クラスター分析」は、数千行までの小さなテーブルに適しており、文字データにも対応します。行がツリー型の階層構造にまとめられます。クラスタリングの処理が終わった後でも、クラスターの個数を変更することができます。
- 「K Means クラスター分析」は、数十万行までの大きいデータに適しています。この分析は、数値データだけに対応しています。処理を開始する前に、クラスターの数 k を指定する必要があります。まず、適切と思われるシード点が推定されます。その後、各点をクラスターに割り当てる作業とクラスター中心を再計算する作業が交互に繰り返されます。
- 「正規混合」は、複数の多変量正規分布の混合分布から得られた、重なりがあるデータに適しています。この分析は、数値データだけに対応しています。外れ値があるような場面では、それらの外れ値を表すため

に、一様分布に従うと仮定したクラスターを使用できます。また、［ロバスト正規混合］オプションによりロバストな推定を行うこともできます。

この手法では、処理を開始する前に、クラスターの個数を指定しておく必要があります。最尤法によって、混合割合、平均、標準偏差、相関係数といったパラメータが同時に推定されます。各点に、それぞれの各グループに属する事後確率が計算されます。推定値の反復計算にはEMアルゴリズムが使用されています。

- カテゴリカルデータの場合は、「潜在クラス分析」が適しています。この手法では、処理を開始する前に、クラスターの個数を指定しておく必要があります。多項分布の混合分布がモデルとして仮定されます。各データ行に対して、各クラスターに属する事後確率が計算されます。そして、属する事後確率が最も高いクラスターに分類されます。

表 9.1 クラスター分析の手法のまとめ

手法	データタイプまたは尺度	データテーブルのサイズ	クラスター数の指定
階層型クラスター分析	すべて	高速 Ward 法の場合、200,000 行まで その他の手法の場合、5,000 行まで	なし
K Means クラスター分析	数値	数百万行まで	あり
正規混合分布法	数値	制限なし	あり
潜在クラス分析	名義尺度または順序尺度	制限なし	あり

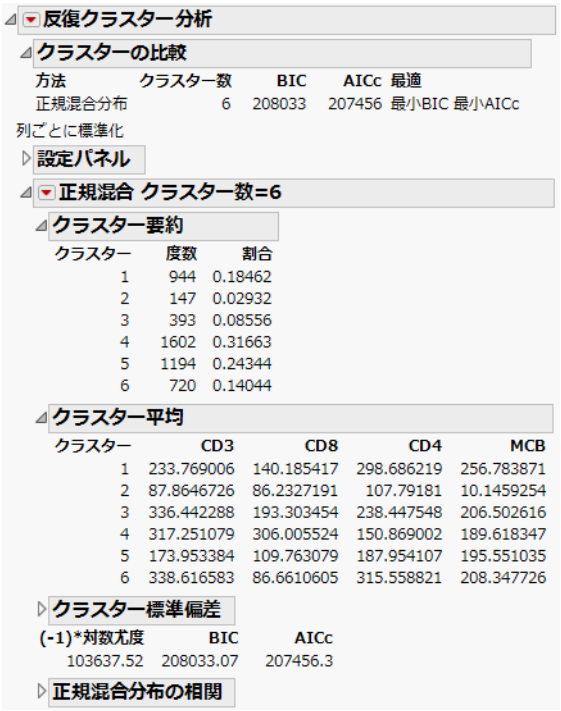
正規混合クラスター分析の例

サイトメトリー（Cytometry）は、細胞のさまざまな特徴を測定するのに使用されます。細胞マーカーを測定することは、特定の疾病を診断するのに役立ちます。この例では、サイトメトリーで測定した 4 つのマーカーに基づいてオブザベーション（データ行）をクラスターに分類します。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Cytometry.jmp」を開きます。
2. ［**分析**］>［**クラスター分析**］>［**正規混合**］を選択します。
3. 「**CD3**」、「**CD8**」、「**CD4**」、「**MCB**」を選択して［**Y, 列**］をクリックします。
4. ［**OK**］をクリックします。
5. 「**クラスターの数**」の横に「6」と入力します。
6. ［**実行**］をクリックします。

メモ : このアルゴリズムではランダムな初期値が使用されるため、結果は分析の度にわずかに異なります。

図 9.2 正規混合 クラスター数=6



反復クラスター分析

クラスターの比較

方法	クラスター数	BIC	AICc	最適
正規混合分布	6	208033	207456	最小BIC 最小AICc

列ごとに標準化

設定パネル

正規混合 クラスター数=6

クラスター要約

クラスター	度数	割合
1	944	0.18462
2	147	0.02932
3	393	0.08556
4	1602	0.31663
5	1194	0.24344
6	720	0.14044

クラスター平均

クラスター	CD3	CD8	CD4	MCB
1	233.769006	140.185417	298.686219	256.783871
2	87.8646726	86.2327191	107.79181	10.1459254
3	336.442288	193.303454	238.447548	206.502616
4	317.251079	306.005524	150.869002	189.618347
5	173.953384	109.763079	187.954107	195.551035
6	338.616583	86.6610605	315.558821	208.347726

クラスター標準偏差

(-1)*対数尤度	BIC	AICc
103637.52	208033.07	207456.3

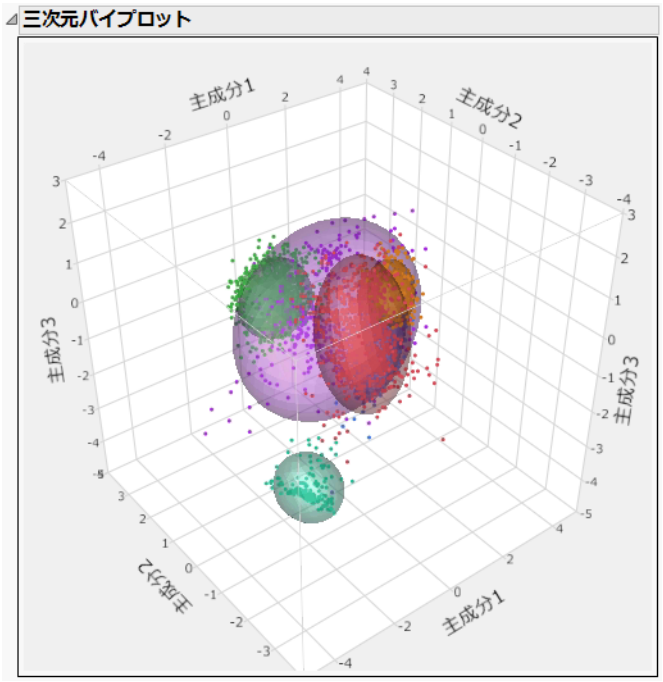
正規混合分布の相関

「クラスター要約」レポートには、6 個の各クラスターに含まれるオブザベーションの数（データの行数）が表示されます。「クラスター平均」レポートには、4 変数それぞれの、クラスターごとの平均が表示されます。

7. 「正規混合 クラスター数=6」の横の赤い三角ボタンをクリックし、[**三次元バイプロット**] を選択します。

メモ : このアルゴリズムではランダムな初期値が使用されるため、結果の三次元バイプロットは分析を行う度にわずかに異なります。

図 9.3 Cytometry データの三次元バイプロット

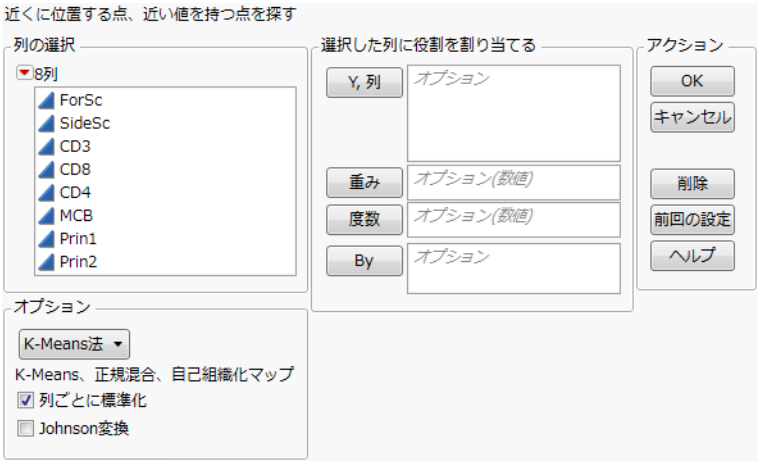


このプロットには、クラスターにあてはめられた正規密度の等高線が表示されます。最初の 3 つの主成分に基づいた場合、1 つのクラスターが他のクラスターとは際立って離れているようです。

「正規混合」クラスター分析プラットフォームの起動

「正規混合」クラスター分析プラットフォームを起動するには、［分析］>［クラスター分析］>［正規混合］を選択します。図 9.4 に示す「クラスター分析」起動ウィンドウは、「Cytometry.jmp」サンプルデータテーブルを使用しています。

図 9.4 正規混合の起動ウィンドウ



Y, 列 クラスター分析の対象となる変数を指定します。

メモ: 正規混合クラスター分析では、数値の列しか使用できません。

重み この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数 この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

By この列の水準に従ってデータがグループ化され、それぞれ個別に分析されます。指定した列の水準ごとに、対応する行が分析されます。分析結果は、個別のレポートにまとめられます。複数の By 変数を指定した場合は、By 変数の水準のすべての組み合わせごとに分析が行われます。

オプション

デフォルトの手法として［正規混合］が選択されていますが、［階層型］または［正規混合］を選択することもできます。［K-Means 法］または［正規混合］を選択し、［OK］をクリックすると、設定パネルが表示されます。「「反復クラスター分析」設定パネル」（177 ページ）を参照してください。

列ごとに標準化 変数の測定尺度が異なる場合に、1 つの変数が結果に強い影響をもつのを防ぐのに用います。たとえば、ある変数が 0 から 1000 までに分布しており、別の変数が 0 から 10 までに分布していたとします。標準化すれば、前者の変数によってクラスター分析の結果がほとんど決まってしまう現象を回避できます。

Johnson 変換 ［列ごとに標準化］を選択した場合は、Y 変数ごとに Johnson 変換が行われます。このとき、Johnson Sb 分布と Johnson Su 分布が検討されます。また、最尤法によって推定されます。『基本的な統計分析』の「一変量の分布」章を参照してください。

Johnson 変換はデータが正規分布に近づき、歪度が小さくなるように、また、外れ値が変換後には分布の中央により近づくように、データの変換を試みます。

メモ：［Johnson 変換］を選択し、［列ごとに標準化］を選択しなかった場合は、すべての Y 変数の値に対して単一の Johnson 変換が行われます。

「反復クラスター分析」レポート

起動ウィンドウで［OK］をクリックすると、以下のものを表示した「反復クラスター分析」レポートウィンドウが開きます。

- 「変換」レポート （［Johnson 変換］オプションを選択した場合にのみ表示されます。）「変換」レポートには、［Y, 列］に指定した変数ごとに、変換の情報が表示されています。Y 変数ごとに、Johnson Sb 分布と Johnson Su 分布のいずれかのうち、あてはまりが良かったほうの $\gamma, \delta, \theta, \sigma$ に対するパラメータ推定値が表示されています。Johnson 分布の詳細については、『基本的な統計分析』の「一変量の分布」章を参照してください。
- モデルのあてはめに関する「設定パネル」。「設定パネル」については、「「反復クラスター分析」設定パネル」（177 ページ）を参照してください。モデルをあてはめるごとに、ウィンドウにレポートが追加されていきます。「「正規混合 クラスター数=<k>」レポート」（179 ページ）を参照してください。

反復クラスター分析のオプション

変換の計算式を保存 （Johnson 変換オプションを選択した場合にのみ表示されます。）変換したデータをデータテーブルの新しい列として保存します。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ 特定のレポートのデータをフィルタリングするためのローカルデータフィルタを表示するか、非表示にします。

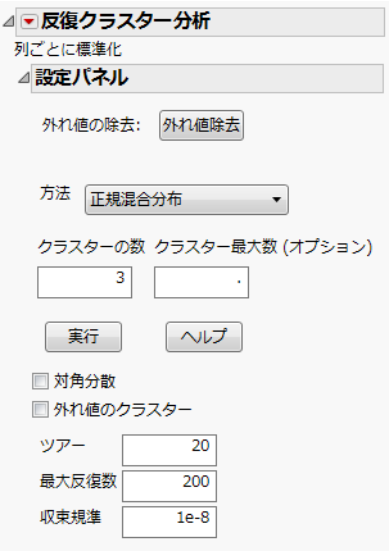
やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

「反復クラスター分析」設定パネル

「CD3」から**「MCB」**までの変数を［**Y, 列**］に指定した「Cytometry.jmp」データテーブルの「設定パネル」を図 9.5 に示します。「設定パネル」を使ってクラスター数を 1 つ指定して 1 回ずつ実行していくことも、［クラスター最大数］オプションを使って、クラスター数の範囲を指定し、まとめて実行することもできます。

図 9.5 正規混合法の設定パネル



正規混合法の設定パネルには、次のようなオプションがあります。

外れ値の除去　多変量の外れ値を検出します。調べたい近傍点の最大数を入力します。[OK] を入力すると、各点から第 1 近傍点への距離、各点から第 2 近傍点への距離、… とグラフが描かれていき、指定した最大数までのグラフが描かれます。

注意 : 近傍点における個数を表すには、「k」という記号が使われることが多いです。一方、k-means 法ではクラスター数が「k」と表されることが多いです。この両者は関係ありません。

「第 k 番目の近傍点への距離」プロットの下に、次のようなオプションが表示されます。

除外のリセット　同じレポート内で新しいクラスター分析を実行する場合に、現在の除外状態を反映させます。たとえば、近傍点プロットや散布図から、データテーブルのいくつかの行に対して、その属性を「除外」にしたとします。同じレポート内の設定パネルで新たなクラスター分析を実行するときに、それらの行を計算から除外してクラスター分析を行うには、このボタンをクリックしてください。

散布図行列　すべての Y 変数に対する散布図行列を描きます。この散布図行列は、現在のレポートとは別のウィンドウに表示されます。

近傍距離の保存　各行から k 番目の近傍点までの距離をデータテーブルの新しい列に保存します。

閉じる　外れ値除去の結果をレポートから削除します。

手法　使用できるクラスター分析の手法は次のとおりです。

K-Means クラスター分析　「K Means クラスター分析」章の「「反復クラスター分析」設定パネル」（163 ページ）で説明しています。

正規混合分布法　この章で説明しています。

ロバスト正規混合 この章で説明しています。「「正規混合 クラスター数=<k>」レポート」（179ページ）を参照してください。

自己組織化マップ 「K Means クラスター分析」章の「自己組織化マップの設定パネル」（168ページ）で説明しています。

クラスターの数 形成されるクラスターの数。

オプション クラスター最大数 形成されるクラスター数の上限値。この値を指定した場合は、[**クラスターの数**]の値とこの値の間にあるすべての整数に対し、分析が行われます。

実行 クラスターをあてはめます。

対角分散 共分散行列の非対角要素をすべて0に固定し、変数間に相関のない多変量正規分布をあてはめます。

メモ: このオプションは、オブザベーション数が列数より少ない場合に、共分散行列が特異になるのを防ぐ目的で使用します。また、変数の数が多い場合に、大規模な共分散行列の計算を回避するためにも使用します。

外れ値のクラスター 通常の正規分布のクラスターには分類されない外れ値を検出する目的で、クラスターを1つ追加します。この外れ値のクラスターには「0」という番号がつけられ、オブザベーションの度数が「クラスター要約」レポートに表示されます。外れ値のクラスターに分類されるオブザベーションの分布は、オブザベーションを囲む超立方体に対して一様であると仮定されます。

ツアー 推定計算を行う回数（ツアーの回数）。推定計算は、異なる初期値を使って行われます。異なる初期値で独立した推定を何回も行うことにより、局所解への収束が軽減されます。

最大反復回数 EMアルゴリズムにおける反復計算の最大反復数です。

収束規準 指定した基準値以下になると、EMアルゴリズムにおける反復計算が停止されます。この基準値には、対数尤度の差が使われています。

「正規混合 クラスター数=<k>」レポート

設定パネルで［実行］をクリックすると、次のレポートが表示されます。

- 「クラスターの比較」レポート 「「クラスターの比較」レポート」（180ページ）を参照してください。
- 1つまたは複数の「正規混合　クラスター数=<k>」レポート。ここで、*k*はクラスターの個数です。「K Means法クラスター数=<k>」レポートは、あてはめを実行する度に表示されます。

「Cytometry.jmp」データテーブルで変数「**CD3**」から「**MCB**」を［**Y, 列**］に指定して実行した結果の「クラスターの比較」レポートと「正規混合 クラスター数=<k>」レポートは、図9.2および「正規混合 クラスター数=6」（174ページ）のようになります。

「クラスターの比較」レポート

「クラスターの比較」レポートには、異なるモデルを比較するための規準統計量が表示されます。それらはBICとAICcの2つです。これらの規準統計量が小さいほど、モデルとして良いことを示唆しています。最良のあてはめは、「最適」列にそれが記されます。なお、［ロバスト正規混合］を選択した場合、「クラスターの比較」レポートにこれらの統計量は表示されません。

「正規混合 クラスター数=<k>」レポート

「正規混合 クラスター数=<k>」レポートには、クラスターごとに要約統計量が表示されます。

- 「クラスター要約」レポートには、クラスター番号、クラスターごとのオブザベーション数（データの行数）、および、割合が表示されます。
- 「クラスター平均」レポートには、クラスターごとに分けて算出された、各変数の平均が表示されます。
- 「クラスター標準偏差」レポートには、クラスターごとに分けて算出された、各変数の標準偏差が表示されます。
- その下の表には、(-1)* 対数尤度、BIC、AICcが表示されます。
- 「正規混合分布の相関」レポートには、クラスターごとの相関行列の推定値が表示されます。

「正規混合 クラスター数=<k>」レポートのオプション

バイプロット　データの主成分のうち最初の2つを軸にして、点とクラスターをプロットします。クラスター中心の周りに円が描かれます。円の大きさはクラスター内のデータ数に比例します。陰影つきの領域は、平均を中心とした50%の等密度面で、そのクラスター内のオブザベーションのうち50%がその領域内に収まることを示しています（Mardia et al., 1980）。プロットの下に、クラスターの色をデータテーブルに保存するオプションが表示されます。固有値は降順に表示されます。

メモ: 起動ウィンドウで［列ごとに標準化］にチェックを入れた場合、バイプロットは相関行列を使用します。［列ごとに標準化］にチェックを入れなかった場合、バイプロットは共分散行列を使用します。

バイプロットオプション　バイプロットの外観をコントロールするためのオプションがあります。

バイプロット線の表示　バイプロット線を表示します。ラベルの付いたバイプロット線は、主成分を基底とした部分空間における共変量の方向を示します。これは、各変数の各主成分に対する関連の度合を示します。

バイプロット線の位置　バイプロット線の位置と半径のスケールを指定できます。デフォルトでは、点(0,0)を出発点とします。プロットでバイプロット線をドラッグして移動するか、またはこのオプションによって出発点の座標を指定できます。半径のスケールオプションを使ってバイプロット線のスケールを調整し、バイプロット線を見やすくすることもできます。

クラスターのマーカー分け　データテーブルの各行に、クラスターに応じたマーカーをつけます。

三次元バイプロット データの3次元バイプロットを表示します。3つ以上の変数がある場合のみ使用できます。

パラレルプロット クラスターごとにパラレルプロットを作成します。オプションで、プロットにおけるデータや平均の表示／非表示を切り替えることができます。詳細については、『グラフ機能』の「パラレルプロット」章を参照してください。

散布図行列 すべての変数に対する散布図行列を描きます。この散布図行列は、現在のレポートとは別のウィンドウに表示されます。

色をテーブルに保存 データテーブルの各行に、クラスターに応じたマーカーをつけます。

クラスターの保存 各行に割り振られたクラスターの番号を含む**「クラスター」**という名前の列を、データテーブルに追加します。正規混合分布法の場合、所属する確率の最も高いクラスターが保存されます。

クラスター計算式の保存 **「クラスター計算式」**という計算式列をデータテーブルに追加します。この計算式は、どのクラスターに属するかを求めるものになっています。（［ロバスト正規混合］を選択した場合は使用できません。）

混合確率の保存 クラスターごとの「確率クラスター **<k>**」という名前の列を追加して、オブザベーションがそのクラスターに属する確率を保存します。

混合計算式の保存 混合確率を計算するために使用する計算式を含む列を、データテーブルに追加します。除外したデータや新たに追加したデータの確率を求めたいときに、これらの計算式を使用します。

距離の計算式 <k> オブザベーションで評価されたクラスター <k> の多変量正規密度関数の推定値。

距離合計 距離合計の計算式列。この列の計算式は、［密度関数の保存］オプションによって作成される**「混合密度」**列の計算式と同じです。

確率の計算式 <k> オブザベーションがクラスター <k> に属する確率。これらの列には、［混合確率の保存］オプションによって作成される**「確率クラスター <k>」**列の値を求める計算式と同じものが保存されます。混合確率を計算する計算式は次のとおりです。混合確率を計算する計算式は次のとおりです。

$$\text{確率の計算式 <k>} = \frac{\text{距離の計算式 <k>}}{\text{距離合計}}$$

密度関数の保存 （［ロバスト正規混合］を選択した場合は使用できません。）正規混合分布の密度関数の推定値を含む**「混合密度」**という列を、データテーブルに追加します。

クラスターのシミュレーション 混合密度に基づいて、乱数データを生成します。正規分布の数値データとそれらが属するクラスターを、新しいデータテーブルに保存します。

削除 クラスター分析のレポートを削除します。

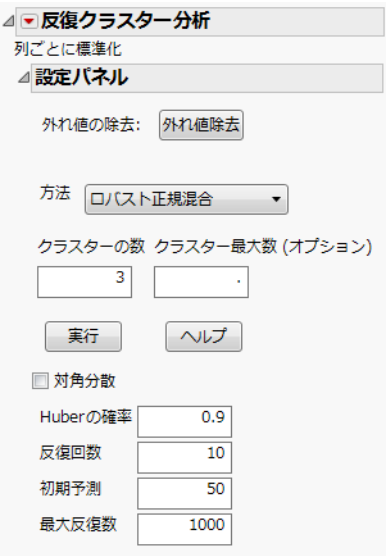
ロバスト正規混合

正規混合分布は外れ値の影響を受けやすいため、JMP では、「ロバスト正規混合」と呼ばれる、外れ値に対して頑健な手法を提供しています。これは、頑健な方法で正規分布のパラメータを推定する手法です。この方法では、混合させる分布に Huber の正規分布を用いて、最尤法で推定します。Huber の正規分布は、正規分布に修正を加え、外れ値に対する頑健性を高めたものです。詳細については、「ロバスト正規混合の詳細」（183 ページ）を参照してください。

ロバスト正規混合分布の設定パネル

「反復クラスター分析」の設定パネルにある「手法」メニューから［ロバスト正規混合］を選択します（図 9.5）。

図 9.6 ロバスト正規混合法の設定パネル



パネルのオプションの一部は、「「反復クラスター分析」設定パネル」（177 ページ）で説明されています。以下、その他のオプションについて説明します。

Huber の確率　0～1 の数値。ロバスト正規混合法は、外れ値に対する重みを小さくすることで、外れ値による影響を抑えます。「Huber の確率」には、外れ値とみなさないデータの割合（重みを小さくしないデータの割合）を指定します。この割合は、多変量正規分布を基準にしたときのものです。1 に近い値を指定すると、重みを**小さくしない**データが増えます。つまり、1 に近い値を指定すると、極端な外れ値だけの重みが小さくなります。逆に、0 に近い値を指定すると、重みが小さくされるデータの割合が増え、それほど外れていないデータでも外れ値とみなされます。詳細については、「ロバスト正規混合の詳細」（183 ページ）を参照してください。

反復回数 独立した推定計算を行う回数。推定計算は、異なる初期値を使って行われます。異なる初期値で独立した推定を何回も行うことにより、局所解への収束が軽減されます。

初期予測 各ツアーにおいて、初期値を求める回数。指定した回数だけ、ランダムなパラメータ初期値が、ツアーの開始時に求められます。

最大反復数 収束までの最大反復数。

ロバスト正規混合のレポート

ロバスト正規混合のレポートは、正規混合のレポートと同じ構造で、同じオプションがあります。レポートには「ロバスト正規混合 クラスター数=<k>」という名前が付きます。「「正規混合 クラスター数=<k>」レポート」（179 ページ）を参照してください。

メモ: ［ロバスト正規混合］を選択した場合、「クラスターの比較」レポートにモデル選択規準の統計量は表示されません。

正規混合法の統計的詳細

正規混合分布の推定には、EM アルゴリズムが使われています。EM アルゴリズムは、Newton-Raphson アルゴリズムよりも安定しています。また、Bayes 法で正規化した EM アルゴリズムを採用しており、共分散行列が特異な場合でも最適化処理がスムーズに行われます。最終的な推定値は、初期値に強く依存します。JMP では、ランダムに選択した点を中心点の初期値として多数のツアーを実行します。

ツアーの回数を増やすと、計算量が増え、特に大規模なデータを処理する場合は計算時間がかかります。そこで、設定パネルでツアーと反復の回数を指定できるようにしています。

ロバスト正規混合の詳細

ロバスト正規混合法の場合、Huber の正規分布の混合分布について最尤法で推定値が計算されます。Huber の正規分布は、正規分布に修正を加え、外れ値に対する頑健性を高めたものです。

Huber 型 Gauss 分布の確率密度関数は、次式 $\Phi_k(x)$ のようになっています。

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

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

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

上記のような確率密度になっているので、kが大きくなるにつれて、$\Phi_k(x)$ は標準正規分布の確率密度関数に基づきます。また、$k \to 0$ の場合、$\Phi_k(x)$ はLaplace分布（二重指数分布）に近づきます。

正則化パラメータkは、P(Normal(x) < k) =「Huberの確率」となるように設定されます。ここで、Normal(x)は多変量正規分布から計算される値です。「Huberの確率」は、分析者によって設定される値で、JMPでのデフォルト値は0.90になっています。

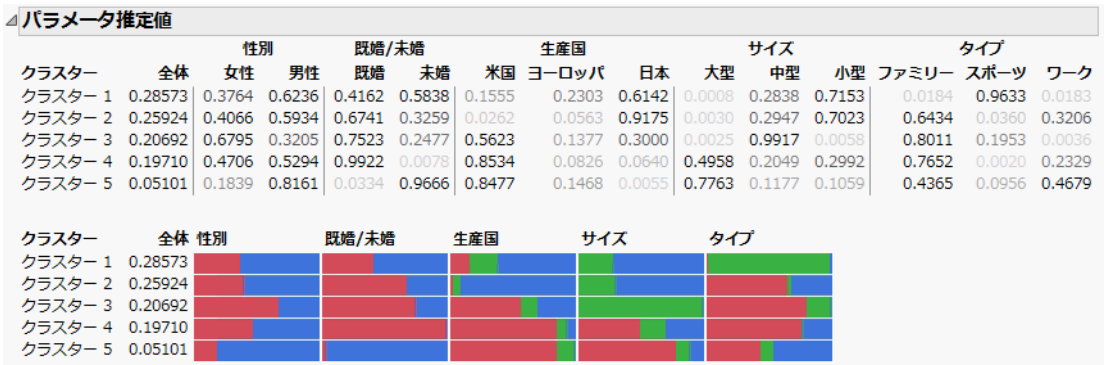
第 10 章

潜在クラス分析

カテゴリカルな変数のデータ行をクラスタリング

潜在クラス分析は、カテゴリカルな変数をもとに、データ行をクラスタリングする手法です。潜在クラス分析では、「分析者から観測されない潜在変数が、グループ変数になっている」と仮定されています。この潜在変数の各水準は、「潜在クラス」と呼ばれています。「潜在クラス分析」プラットフォームは、潜在クラスモデルをあてはめ、各データ行が所属する事後確率の最も高い潜在クラスを特定します。多くの場合、分析者は、潜在クラス分析を実行した後、各クラスの特徴を見て、それらのクラスがどのようになっているかを考察します。

図10.1 潜在クラス分析の例



パラメータ推定値

		性別		既婚/未婚		生産国			サイズ			タイプ		
クラスター	全体	女性	男性	既婚	未婚	米国	ヨーロッパ	日本	大型	中型	小型	ファミリー	スポーツ	ワーク
クラスター 1	0.28573	0.3764	0.6236	0.4162	0.5838	0.1555	0.2303	0.6142	0.0008	0.2838	0.7153	0.0184	0.9633	0.0183
クラスター 2	0.25924	0.4066	0.5934	0.6741	0.3259	0.0262	0.0563	0.9175	0.0030	0.2947	0.7023	0.6434	0.0360	0.3206
クラスター 3	0.20692	0.6795	0.3205	0.7523	0.2477	0.5623	0.1377	0.3000	0.0025	0.9917	0.0058	0.8011	0.1953	0.0036
クラスター 4	0.19710	0.4706	0.5294	0.9922	0.0078	0.8534	0.0826	0.0640	0.4958	0.2049	0.2992	0.7652	0.0020	0.2329
クラスター 5	0.05101	0.1839	0.8161	0.0334	0.9666	0.8477	0.1468	0.0055	0.7763	0.1177	0.1059	0.4365	0.0956	0.4679

「潜在クラス分析」プラットフォームの概要

「潜在クラス分析」プラットフォームは、カテゴリカルなデータに対して潜在クラスモデルをあてはめ、各データ行（各オブザベーション）が所属する事後確率の最も高いクラスターを特定します。潜在クラス分析では、分析者から観測されない**潜在変数**が、グループ変数となっています。このクラスターは、**「潜在クラス」**と呼ばれています。潜在クラス分析の応用例としては、たとえば、リスク行動についての質問紙調査データをクラスタリングする、などが考えられます。

潜在クラスモデルでは、「観測されるデータは、多項分布の混合分布に従っている」と仮定します。このモデルには、2種類のパラメータの組（γとρ）があります。γパラメータは、各クラスターに属する事前確率を表します。一方、ρパラメータは、該当のクラスターに属すると条件付けられた上で、各観測変数が生じる確率を表します。各潜在クラスがどのような特徴をもつかは、条件付き確率（ρパラメータ）によって解釈されます。

分析結果を意義のあるものとするには、求められたクラスターを適切に解釈する必要があります。潜在クラスの特徴を見て、それらの特徴から各クラスを解釈します。

メモ: 応答列のいずれかに1つでも欠測値がある行は、その行すべてのデータが分析から除外されます。

潜在クラスモデルの詳細については、Collins and Lanza（2010）およびGoodman（1974）を参照してください。

潜在クラス分析の例

この例では、「潜在クラス分析」プラットフォームを使って、米国の高校生を対象に行われた、2005年の調査データを分析します。この調査では、リスク行動に関するさまざまな質問が問われました。

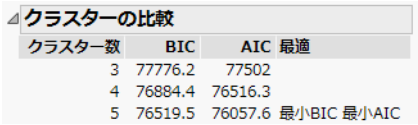
この例では、12個の質問に注目して、生徒をクラスターに分けるために、潜在クラス分析を試してみます。なお、元データは3つ以上の選択肢だったのですが、「Yes」／「No」の2水準データに変換しています。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Health Risk Survey.jmp」を開きます。
2. Health Risk Surveyデータテーブルで、「**潜在クラス分析の起動**」というスクリプトの横にある緑の三角ボタンをクリックします。

   このスクリプトは、「潜在クラス分析」ウィンドウを開き、分析対象の12列を［Y］に指定します。
3. **「最大個数」**の横のボックスに「5」と入力します。

   この指定により、クラスター数が3～5個の潜在クラスモデルがあてはめられます。
4. ［**OK**］をクリックします。

図10.2「クラスター要約」レポート



クラスターの比較

クラスター数	BIC	AIC	最適
3	77776.2	77502	
4	76884.4	76516.3	
5	76519.5	76057.6	最小BIC 最小AIC

「潜在クラス分析」アウトラインには、「クラスターの比較」レポートと3個の独立した「潜在クラスモデル」レポートが含まれます。「潜在クラスモデル」レポートには、クラスター数が3個、4個、および5個のモデルが表示されています。「クラスターの比較」レポートは、5個のクラスターのあるモデルのBICとAICが最小で、3つのうちで最適なモデルであることを示しています。これが分析するモデルとなります。

5. 「潜在クラスモデル（クラスター数: 5個）」レポートで、パラメータ推定値の下の棒グラフを確認します。次のような検証が行われます。
   - クラスター1は、ほぼすべてのリスク行動について「No」と回答している。
   - クラスター2は、13歳未満で行ったリスク行動について「Yes」と回答した数が多い。
   - クラスター3は、「過去30日間で飲酒運転した」と「過去30日間で5杯以上アルコールを飲んだ」に「Yes」と回答した数が多い。
   - クラスター4は、13歳未満で行ったリスク行動以外のほとんどのリスク行動について「Yes」と回答した数が多い。
   - クラスター5は、ほとんどのリスク行動について「Yes」と回答した数が最も多い。

   これらの情報を使って、クラスターにわかりやすい名前をつけましょう。
6. 「潜在クラスモデル（クラスター数: 5個）」レポートの横の赤い三角ボタンをクリックし、［**クラスター名の変更**］を選択します。
   - クラスター1に「低リスク」と入力します。
   - クラスター2に「早期リスクテイカー」と入力します。
   - クラスター3に「飲酒」と入力します。
   - クラスター4に「後期高リスク」と入力します。
   - クラスター5に「高リスク」と入力します。
7. ［**OK**］をクリックします。
8. JMPの警告ウィンドウで［**OK**］をクリックします。

メモ: 新しいクラスター名はスクリプトに保存されません。

図10.3「パラメータ推定値」レポートの一部

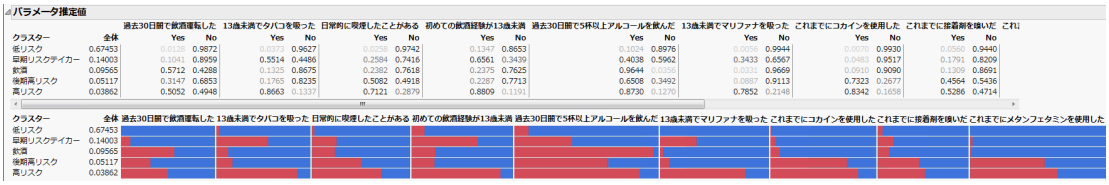
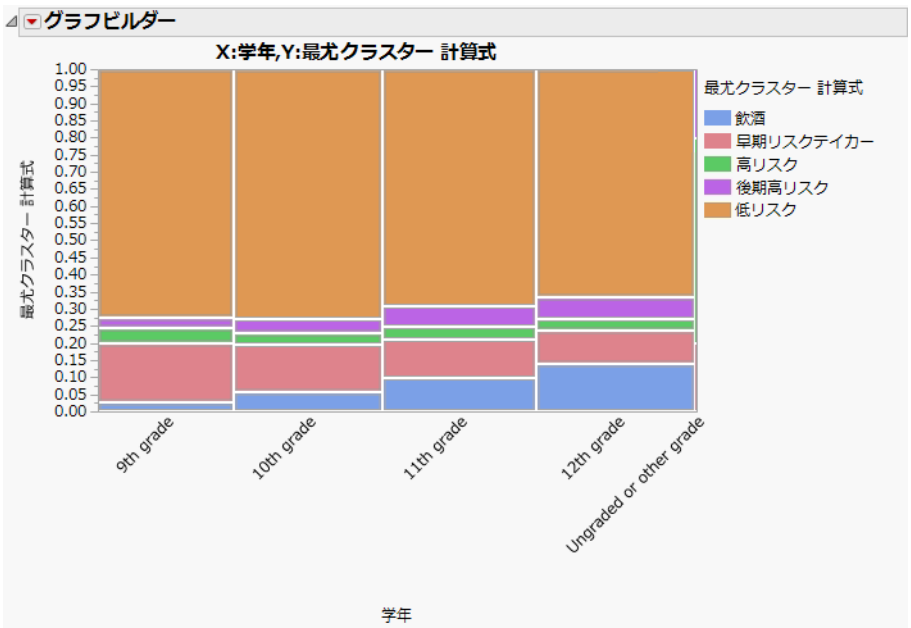


図10.3は、最初の8個の変数のパラメータ推定値を示しています。レポートには、新しいクラスター名が表示されています。

次に、生徒の「学年」ごとに、どのクラスターが多いかを比較してみましょう。

9. 「潜在クラス分析（クラスター数:5個）」レポートの横の赤い三角ボタンをクリックし、**[混合計算式とクラスター計算式を保存]**を選択します。
10. **[グラフ] > [グラフビルダー]**を選びます。
11. **「学年」**を**[X]**に入力します。
12. **「最尤クラスター 計算式」**を**[Y]**に入力します。
13. 「モザイク」アイコンを選択します。
14. **[終了]**をクリックします。

図10.4 クラスターメンバーと「学年」の水準のモザイク図

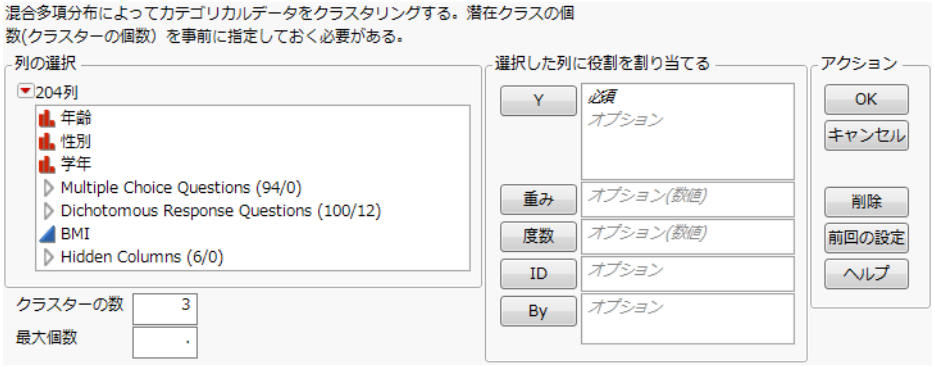


ほとんどの回答者が「低リスク」のクラスターに含まれていることがわかります。また、「飲酒」という名前のクラスは、学年が上がるにつれて回答者が多くなっています。

「潜在クラス分析」プラットフォームの起動

［分析］>［クラスター分析］>［潜在クラス分析］を選択すると、「潜在クラス分析」プラットフォームが起動します。

図10.5「潜在クラス分析」起動ウィンドウ



「潜在クラス分析」プラットフォームの起動ウィンドウには、次のようなオプションがあります。

Y　分析対象となる複数のカテゴリカルな応答列を指定します。

重み　この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数　この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

ID　個々の回答者を識別するための列。このIDは、一部の出力テーブルで使用されます。

By　ここで指定した変数の水準ごとに、分析が実行され、レポートが作成されます。複数のBy変数を指定した場合は、By変数の水準のすべての組み合わせごとに分析が行われます。

クラスターの数　モデルで仮定するクラスター数（クラスターの個数）を指定します。

最大個数　クラスター数に対する上限を指定します。この数が「クラスターの数」に指定した値よりも大きい場合、「クラスターの数」から「最大個数」までのすべての整数をクラスター数とするモデルがあてはめられます。これらの結果は、1つのウィンドウにまとめて表示されます。

起動ウィンドウで［OK］をクリックすると、「潜在クラス分析」レポートが表示されます。

「潜在クラス分析」レポート

デフォルトでは、「潜在クラス分析」レポートには「クラスターの比較」レポートと、指定されたクラスター数ごとの「潜在クラスモデル」レポートが表示されます。

「クラスターの比較」レポート

「クラスターの比較」レポートには、異なるモデルを比較するための規準統計量が表示されます。適合度統計量はBICとAICです。よくあてはまっているほど、これらの規準値は小さくなります。最良のあてはめは、**「最適」**列にそれが記されます。

「潜在クラスモデル（クラスター数: <k>個）」レポート

「潜在クラスモデル（クラスター数: <k>個）」レポートには、以下のような結果およびアウトラインが含まれます。

- 「モデルの要約」（190ページ）
- 「パラメータ推定値」（190ページ）
- 「転置したパラメータ推定値」（191ページ）
- 「効果の大きさ」（191ページ）
- 「多次元尺度構成プロット」（191ページ）
- 「混合確率」（192ページ）

モデルの要約

各「潜在クラスモデル」レポートの最上部には、それぞれ該当する個数のクラスターによるモデルの要約が表示されます。モデルの概要には、(-1)*対数尤度、パラメータ数、BIC、AICが含まれます。これらはすべてモデルのあてはまりの良さを決定するのに使用できます。(-1)*対数尤度、AICc、BICの値が小さいほど、良いモデルであることを示唆しています。詳細は、『基本的な回帰モデル』の付録「統計的詳細」を参照してください。「パラメータ数」は、潜在クラスモデルの固有パラメータ数です。詳細は、「「潜在クラス分析」プラットフォームの統計的詳細」（194ページ）を参照してください。

パラメータ推定値

「パラメータ推定値」レポートには、表とグラフが表示されます。これらの各行は、クラスターに対応しています。先頭の表には、結果が数値で表示されます。続くグラフでは、結果がシェアチャートで描かれています。

表とグラフの「全体」列には、各クラスターに属する事前確率が示されています。（これらはγパラメータです。「「潜在クラス分析」プラットフォームの統計的詳細」（194ページ）を参照してください。）

先頭の表における各列には、「潜在クラス分析」起動ウィンドウで指定したY列ごとに、パラメータ推定値が表示されています。この表の列は、Y変数の水準ごとに1つの列になっています。表のセルに表示されている数値は、表の行で示されているクラスターに属しているという条件のもとで、Y変数が該当の水準となる確率（ρパラメータ）を示しています。

次に示されるグラフは、前述の条件付き確率を**シェアチャート**で描いたものです。該当のクラスターに属するという条件のもとでの確率を横につなげた棒グラフになっています。棒での表示順序は、データ値の表示順序に従っています。

ヒント:「パラメータ推定値」レポートの表やグラフにおいて、1行または複数行を選択すると、対応するクラスターに割り振られたデータ行も選択されます。

転置したパラメータ推定値

「転置したパラメータ推定値」レポートの表は、「パラメータ推定値」レポートの表を転置したものです。この表では、クラスターが表の列になっています。該当するクラスターに属するという条件のもとでの確率が表示されています。

メモ:「全体」列の推定値は、転置したテーブルには含まれません。

効果の大きさ

「効果の大きさ」(effect size) の表には、各Y列とクラスタとの関係の大きさを示す指標が計算されます。この表の統計量は、Y列と割り振られたクラスターとの分割表から、各セルの期待度数からの乖離を求めて算出されています。期待度数は、各クラスターの標本サイズに、Y列の各水準の条件付き確率を掛けて求められています。

この分割表分析では、Y列ごとにPearsonのカイ2乗 (X^2) が計算されます。nを標本サイズとすると、「効果の大きさ」は次のように求められます。

$$\text{効果の大きさ} = \sqrt{\frac{X^2}{n}}$$

「尤度比 対数価値」の値は、-log10(p_{LR}) です。ここで、p_{LR}は分割表に対する尤度比検定のp値です。p値が0.01のときに、対数価値は2となります。

ヒント:「効果の大きさ」テーブルで行を選択すると、それらの行に対応するデータテーブルの列も選択されます。

多次元尺度構成プロット

多次元尺度構成プロットは、クラスターの類似性を2次元で表したものです。プロットの各点は、各クラスターを表しています。距離が近いクラスターほど、類似性が高いことを示しています。このプロットは、ρパラメータの相違性行列から作成されます。多次元尺度構成プロットの詳細については、『消費者調査』の「多次元尺度構成」章を参照してください。

混合確率

「混合確率」の表には、各クラスターに属する事後確率が表示されています。また、「最尤クラスター」列の値は、事後確率が最も高いクラスターです。

メモ: Y 列のいずれかに 1 つでも欠測値がある行は分析から除外され、また、「混合確率」の表にも表示されません。

「潜在クラス分析」プラットフォームのオプション

「潜在クラス分析」のオプション

「潜在クラス分析」の赤い三角ボタンのメニューには、次のようなオプションがあります。

新しいクラスター数 クラスター数を変えて別の分析を実行できます。新しい分析レポートが現在のレポートに追加されます。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

潜在クラスモデルのオプション

「潜在クラスモデル（クラスター数 : <k>）」の赤い三角ボタンのメニューには、次のようなオプションがあります。

クラスターによる色分け 事後確率が最も高いクラスターに合わせて、データテーブルの行を色分けします。例については、「「潜在クラス分析」プラットフォームの別例」（193 ページ）を参照してください。

混合計算式とクラスター計算式を保存 各クラスターについての事後確率と、最も確率の高いクラスターを特定する計算式の列を、データテーブルに保存します。

クラスター計算式だけを保存 事後確率の最も高いクラスターを特定する計算式の列を、データテーブルに保存します。

JMP PRO **確率の計算式を発行** 確率の計算式を作成し、それを「計算式デポ」レポート内の計算式列スクリプトとして保存します。「計算式デポ」レポートが開いていない場合は、このオプションを選択した時点でレポートが作成されます。『予測および発展的なモデル』の「計算式デポ」章を参照してください。

混合確率の保存 「混合確率」表の値を、データテーブルの対応する行に保存します。

クラスターだけを保存 事後確率の最も高いクラスターの列を、データテーブルに保存します。この列に計算式は含まれません。

クラスター名の変更 レポート内のクラスターにわかりやすい名前をつけることができます。

メモ: 新しいクラスター名は、レポートに乱数シード値を指定していない限りスクリプトには保存されません。乱数シード値は、スクリプトからレポートを起動した場合にのみ設定できます。

あてはめの削除 所定のクラスタリングレポートをレポートウィンドウから削除します。

「潜在クラス分析」プラットフォームの別例

クラスターメンバーの確率をプロットする

ここでは、車の所有者とメーカーに関する調査データである「Car Poll.jmp」データテーブルを使用します。車の所有者を3つのクラスターに分類し、各クラスターに属する確率を視覚的に表示するプロットを作成します。クラスターが3つの場合は、三角図を使うとわかりやすいです。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Car Poll.jmp」を開きます。
2. ［**分析**］>［**クラスター分析**］>［**潜在クラス分析**］を選択します。
3. 「**年齢**」以外の列をすべて選択し、［**Y**］をクリックします。
4. ［**OK**］をクリックします。
5. 「潜在クラスモデル（クラスター数: 3個）」レポートの横の赤い三角ボタンをクリックし、［**クラスターによる色分け**］を選択します。
6. 「潜在クラスモデル（クラスター数: 3個）」レポートの横の赤い三角ボタンをクリックし、［**混合確率の保存**］を選択します。
7. Car Pollのデータテーブルで、列リストから「**LCA クラスター確率**」列グループを選択します。
8. ［**グラフ**］>［**三角図**］を選択します。
9. ［**X, プロット**］をクリックします。
10. ［**OK**］をクリックします。

図10.6 クラスターメンバーの確率を示す三角図

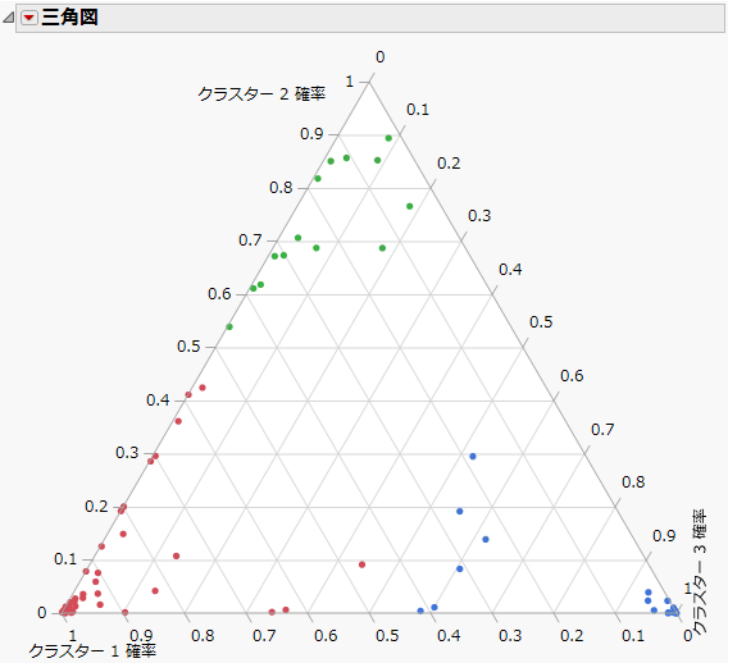


図10.6の三角図は、各回答者が各クラスターに属する事後確率を示しています。ほとんどの事後確率が頂点近くに分布しています。これは、多くの回答者がいずれか1つだけのクラスターに属する事後確率が高く、それ以外の2つのクラスターに属する事後確率が低いことを示しています。ただし、プロットの中央付近にもいくつかの点があり、これらの回答者はどのクラスターに属する確率も特に高くないことを示しています。このような回答者がいるということは、さらに検討したり、クラスター数を増やしたりする必要があるかもしれません。

メモ: 乱数シードが設定されていないため、実際の結果がこの例とは異なる場合があります。

「潜在クラス分析」プラットフォームの統計的詳細

ここでは、「潜在クラス分析」プラットフォームであてはめられる潜在クラスモデルについて説明します。潜在クラスモデルの詳細については、Collins and Lanza (2010) および Agresti (2002) を参照してください。

メモ: 「テキストエクスプローラ」プラットフォームでも潜在クラス分析が実行できますが、そこで使われているアルゴリズムは文書単語行列の特定の構造を利用しています。そのため、「テキストエクスプローラ」プラットフォームでの潜在クラス分析の結果は、「潜在クラス分析」プラットフォームでの結果と正確には一致しません。

$j = 1, ..., J$を観測された応答の列とします。これらは、「潜在クラス分析」起動ウィンドウで［Y］に指定した列です。列jの水準数をR_jとします。

これらJ変数から構成されるベクトルには、全部で$W = R_1*...*R_J$個のパターンがあります。ある回答者の応答は、これらJ変数に対する応答値によって定義されます。この応答値は、$\mathbf{y} = (y_1, ..., y_j)$という長さ$J$の行ベクトルで表されます。いま、$\mathbf{Y}$を、$W$個の応答パターンを回答者に関して縦に並べた、$W \times J$の行列とします。そして、ある行でパターンが$\mathbf{y}_w$となる確率を$\Pr(\mathbf{y}_w)$と表します。この確率の全パターンでの合計は1となります。

$$\sum_{w=1}^{W} Pr(y_w) = 1$$

次の表記を使用します。

- Cは、潜在クラスモデル内のクラスターの数です。
- γ_cは、クラスターcに属する事前確率です。（γ_cは潜在クラスの普及率です。）これらのパラメータの合計は1となります。
- $r_{j,k}$は、j番目の応答のk番目の水準です。
- $\rho_{j,k|c}$は、クラスcに属するという条件のもとで、列jの応答が$r_{j,k}$になる確率です。（$\rho_{j,k|c}$は項目応答の確率です。）クラスター c と応答変数jの各組み合わせ内において、$\rho_{j,k|c}$を合計したものは1となります。
- $I(y_j = r_{j,k})$は指示関数で、j番目の応答y_jが第k水準である場合は1、そうでない場合は0となる関数です。

潜在クラスモデルでは、特定の応答ベクトル$\mathbf{y} = (y_1, ..., y_j)$が観測される確率が、各潜在クラス$C$において、応答変数ごとに独立に条件付き確率を掛け合わせ、それらを合計したものと仮定されています。

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

この式は、「潜在クラス分析」の赤い三角ボタンから［混合計算式とクラスター計算式の保存］オプションを選択すると保存される**「確率計算式 クラスター」**の分母となります。一方、**「確率計算式 クラスター」**は、Pr(クラスター = c | $\mathbf{y}$) で、Pr($\mathbf{y}$, クラスター = c) / Pr($\mathbf{y}$) です。

潜在クラスモデルのこれらのパラメータ（γおよびρ）は、反復法の一種である EM（Expectation-Maximization）法を用いて推定されます。潜在クラスモデルの固有パラメータ数は次のように定義されます。

$$(C + 1) + C \sum_{j=1}^{J} (R_j - 1)$$

第 11 章

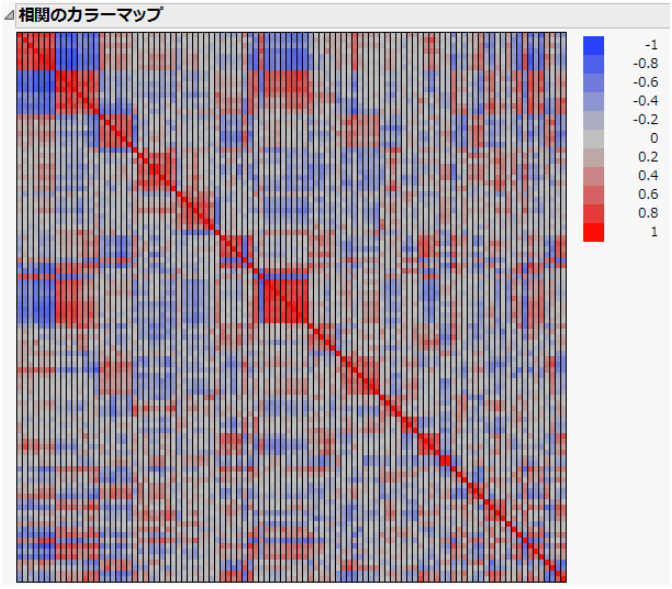
変数のクラスタリング

似通った変数をクラスターに分類

「変数のクラスタリング」プラットフォームでは、変数に対するクラスタリングを行います。変数のクラスタリングにより、少数の主成分にまとめたり、代表的とみなされる変数を選び出したりできます。「変数のクラスタリング」の各クラスターは、主成分（そのクラスターに属する変数の線形結合）です。また、各クラスターは、クラスター内で最も代表的とみなされる変数で特徴付けることもできます。

［変数のクラスタリング］プラットフォームは、データの次元を減らすための手法として使えます。変数の個数が多いような状況において、変動のかなりの部分を、クラスター成分、または、クラスター内の最も代表的な変数によって説明できることがあります。これらの新しい成分は、予測などのモデルに流用できるでしょう。通常、求められたクラスター成分は、すべての変数から求められる主成分よりも解釈が容易です。

図11.1 変数の相関マップの例



「変数のクラスタリング」プラットフォームの概要

主成分分析では、分析対象となっているすべての変数の線形結合によって主成分を求めます。それとは対照的に、［変数のクラスタリング］プラットフォームは、クラスター内の変数（類似している変数）だけの線形結合によって成分を求めます。この分析では、すべての変数が、クラスターのいずれか1つに分類されます。各クラスターにおける、そのクラスターに属する変数の第1主成分を、**クラスター成分**（cluster component）を呼びます。この第1主成分（クラスター成分）は、そのクラスターに属する変数がもつ変動を最も説明する線形結合になっています。

［変数のクラスタリング］プラットフォームは、データの次元を減らすための手法として使えます。変数の個数が多いような状況において、変動のかなりの部分を、クラスター成分、または、クラスター内の最も代表的な変数によって説明できることがあります。これらの新しい成分は、予測などのモデルに流用できるでしょう。通常、求められたクラスター成分は、すべての変数から求められる主成分よりも解釈が容易です。

すべての変数に対して主成分分析を行った場合は、その主成分は互いに直交します。しかし、「変数のクラスタリング」におけるクラスター成分は、いくつかの変数に分割して求められるので、互いに直交しません。

変数の数が多い場合、「変数のクラスタリング」プラットフォームでは、特異値分解に基づいたアルゴリズムを使用して計算時間を短縮します。詳細については、「統計的詳細」の付録の「「線形 横長データ」の手法と特異値分解」（210ページ）を参照してください。

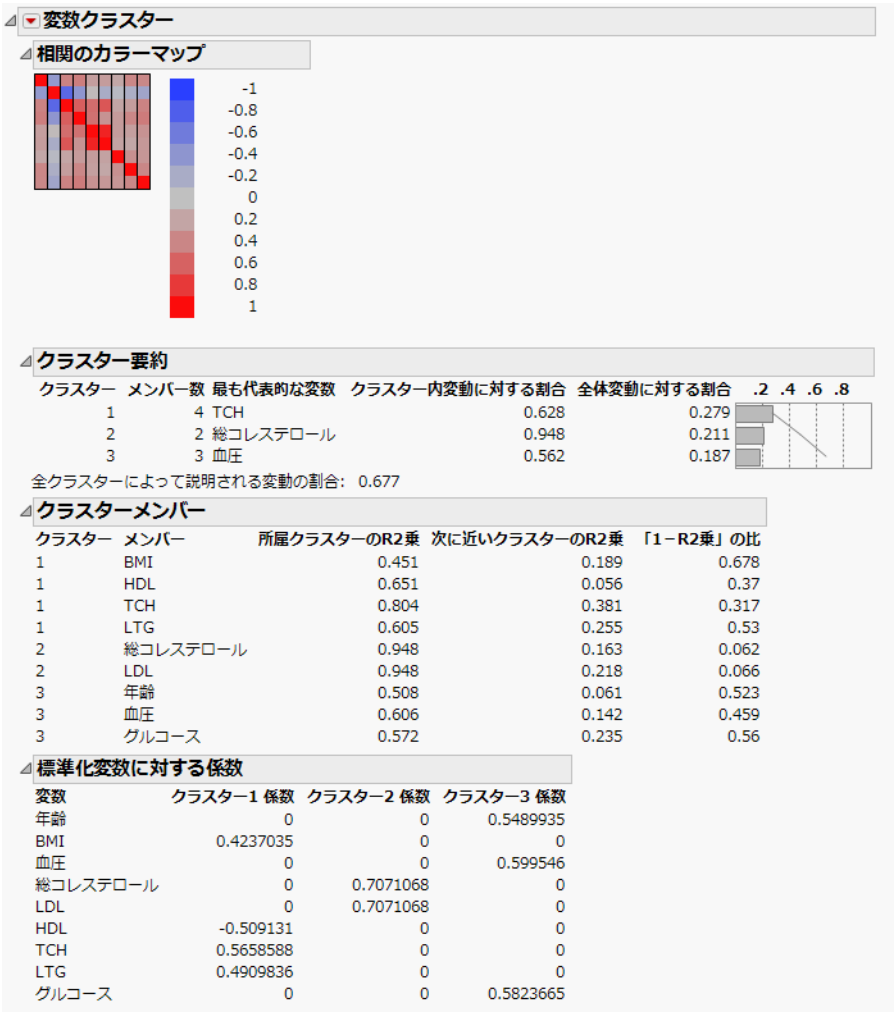
「変数のクラスタリング」プラットフォームの例

「Diabetes.jmp」サンプルデータテーブルには、病症の進行をモデル化するのに使われる10個の説明変数が含まれています。この例では、それらの連続尺度の変数をクラスタリングします。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Diabetes.jmp」を開きます。
2. ［**分析**］>［**クラスター分析**］>［**変数のクラスタリング**］を選択します。
3. 「**性別**」を除く「**年齢**」から「**グルコース**」までの列（「**年齢**」、「**BMI**」、「**血圧**」、「**総コレステロール**」、「**LDL**」、「**HDL**」、「**TCH**」、「**LTG**」、「**グルコース**」）を選択し、［**Y, 列**］をクリックします。

   ［変数のクラスタリング］では数値の連続変数を使用する必要があるため、「**性別**」の列は含めることができません。
4. ［**OK**］をクリックします。

図11.2「Diabetes.jmp」の「変数のクラスタリング」レポート



変数クラスター

相関のカラーマップ

-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

クラスター要約

クラスター	メンバー数	最も代表的な変数	クラスター内変動に対する割合	全体変動に対する割合
1	4	TCH	0.628	0.279
2	2	総コレステロール	0.948	0.211
3	3	血圧	0.562	0.187

全クラスターによって説明される変動の割合: 0.677

クラスターメンバー

クラスター	メンバー	所属クラスターのR2乗	次に近いクラスターのR2乗	「1－R2乗」の比
1	BMI	0.451	0.189	0.678
1	HDL	0.651	0.056	0.37
1	TCH	0.804	0.381	0.317
1	LTG	0.605	0.255	0.53
2	総コレステロール	0.948	0.163	0.062
2	LDL	0.948	0.218	0.066
3	年齢	0.508	0.061	0.523
3	血圧	0.606	0.142	0.459
3	グルコース	0.572	0.235	0.56

標準化変数に対する係数

変数	クラスター1 係数	クラスター2 係数	クラスター3 係数
年齢	0	0	0.5489935
BMI	0.4237035	0	0
血圧	0	0	0.599546
総コレステロール	0	0.7071068	0
LDL	0	0.7071068	0
HDL	-0.509131	0	0
TCH	0.5658588	0	0
LTG	0.4909836	0	0
グルコース	0	0	0.5823665

図11.2は、「変数クラスタリング」レポートです。「クラスター要約」レポートを見ると、変数は3つのクラスターにグループ化されています。

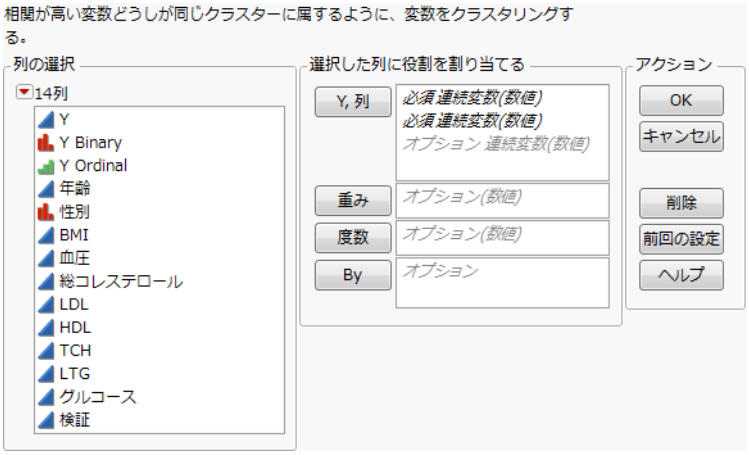
- 「クラスターメンバー」レポートを見ると、クラスター1は「**BMI**」、「**HDL**」、「**TCH**」、「**LTG**」で構成されています。「クラスター要約」レポートを見ると、クラスター1の最も代表的な変数は「**TCH**」で、これがクラスター1の変動の62.8%を説明していることがわかります。
- クラスター2は、「**総コレステロール**」と「**LDL**」の2つだけで構成されています。「クラスター要約」レポートを見ると、クラスター2の最も代表的な変数は「**総コレステロール**」で、これがクラスター2の変動の94.8%を説明していることがわかります。

- クラスター 3 は、**「年齢」**、**「血圧」**、**「グルコース」**で構成されています。「クラスター要約」レポートを見ると、クラスター 3 の最も代表的な変数は**「血圧」**で、これがクラスター 3 の変動の 56.2% を説明していることがわかります。

「変数のクラスタリング」プラットフォームの起動

「変数のクラスタリング」プラットフォームを起動するには、**［分析］＞［クラスター分析］＞［変数のクラスタリング］**を選択します。図 11.3 は、「Diabetes.jmp」テーブルの「変数のクラスタリング」起動ウィンドウです。

図 11.3 「変数のクラスタリング」起動ダイアログ



Y, 列　クラスタリング対象の変数。変数は数値の連続変数でなければなりません。

重み　この役割を割り当てた列の数値は、分析において各行の重みとして使用されます。

度数　この役割を割り当てた列の数値は、分析において各行の度数として使用されます。

By　この列の水準に従ってデータがグループ化され、それぞれ個別に分析されます。指定した列の水準ごとに、対応する行が分析されます。分析結果は、個別のレポートにまとめられます。複数の By 変数を指定した場合は、By 変数の水準のすべての組み合わせごとに分析が行われます。

「変数のクラスタリング」レポート

デフォルトでは、「変数のクラスタリング」レポートに次のものが表示されます。

- 「相関のカラーマップ」（201 ページ）
- 「クラスター要約」（201 ページ）
- 「クラスターメンバー」（202 ページ）

- 「標準化変数に対する係数」（202 ページ）

ヒント：「変数のクラスタリング」レポートの表でも、行を選択すると、データテーブル内の対応する列も選択されます。また、データテーブル内の対応する列の選択を解除するには、Ctrl キーを押しながら表の行をクリックしてください。

相関のカラーマップ

「相関のカラーマップ」レポートには、変数間の相関が色で示されます。変数は、「クラスターメンバー」レポートにリストされている順番に並べられています。これにより、「相関のカラーマップ」では同じクラスターに所属する変数が隣接して表示されます。「相関のカラーマップの例」（203 ページ）を参照してください。

ヒント：カラーマップのセルにカーソルを置くと、そのセルに対応する 2 つの変数の名前とその相関係数が表示されます。

同じクラスターに所属する変数どうしの相関係数は、異なるクラスターに属する変数との相関係数よりも、その絶対値が大きい傾向にあります（セルは濃い赤または濃い青になる傾向にあります）。そのため、「変数のクラスタリング」で並べられた「相関のカラーマップ」では、対角線上あたりのセルが濃くなる傾向にあります。

相関係数は、リストワイズ法によって計算されます。リストワイズ法では、いずれかの変数に 1 つでも欠測値がある行は計算から除外されます。リストワイズ法の詳細については、「多変量の相関」章の「推定法について」（43 ページ）を参照してください。

クラスター要約

「クラスター要約」レポートには、次の情報が表示されます。

クラスター　クラスターの ID。

メンバー数　クラスター内のメンバーの数。

最も代表的な変数　クラスター成分との相関係数の 2 乗が最も大きいクラスター変数。

クラスター内変動に対する割合　クラスター内の変動の中で、第 1 主成分によって説明される変動の割合。クラスター内の変数が 1 つだけの場合、この割合は 1 です。この統計量は、すべての変数ではなく、クラスター内の変数だけをベースにしています。

全体変動に対する割合　各クラスター成分によって説明される、全体変動に対する割合。これは、各クラスター内の変数だけを使って第 1 主成分を計算するのと同じです。

この表の下には、すべてのクラスター成分によって説明される全体変動の割合が示されます。

クラスターメンバー

「クラスターメンバー」レポートには、次の情報が表示されます。

クラスター　クラスターの ID。

メンバー　クラスター内の変数。

所属クラスターの R2 乗　変数と、その変数が属するクラスターのクラスター成分との相関係数を 2 乗したもの。

次に近いクラスターの R2 乗　変数と、次に近いクラスターのクラスター成分との相関係数を 2 乗したもの。「次に近いクラスター」とは、変数とクラスター成分との相関係数を 2 乗したものが 2 番目に大きいクラスターのことを指します。

「1 - R2 乗」の比　変数が属するクラスターと、次に近いクラスターとの相対的な近さを示す指標です。「1 − R2 乗」の比は次のように定義されます。

(1 - 所属クラスターの R2 乗)/(1 - 次に近いクラスターの R2 乗)

標準化変数に対する係数

クラスター成分を求めるための係数を示す「標準化変数に対する係数」レポートを表示します。これらの係数は、各クラスター内の第 1 主成分の固有ベクトルです。

「変数のクラスタリング」プラットフォームのオプション

「変数クラスター」の赤い三角ボタンをクリックすると、次のようなオプションが表示されます。

相関のカラーマップ　「相関のカラーマップ」プロットの表示／非表示を切り替えます。「相関のカラーマップ」（201 ページ）を参照してください。

クラスター要約　「クラスター要約」レポートの表示／非表示を切り替えます。「クラスター要約」（201 ページ）を参照してください。

クラスターメンバー　「クラスターメンバー」レポートの表示／非表示を切り替えます。「クラスターメンバー」（202 ページ）を参照してください。

クラスター成分の係数　「標準化変数に対する係数」レポートの表示／非表示を切り替えます。「標準化変数に対する係数」（202 ページ）を参照してください。

クラスター成分の保存　**「クラスター <i> 成分」**という列をデータテーブルに保存します。各列に含まれる計算式は、（中心化・尺度化されていない）元データからクラスター成分を求めるものです。

モデルのあてはめを起動 各クラスターの「最も代表的な変数」が「モデル効果の構成」リストに入力された状態の「モデルの指定」ウィンドウを開きます。「最も代表的な変数」に基づいてモデルを構成する場合は、このオプションを使用します。

ヒント: クラスター成分を回帰モデルなどに利用する場合には、まず［**クラスター成分の保存**］オプションを選択して、クラスター成分をデータテーブルに保存してください。次に、「モデルのあてはめ」の「モデル効果の構成」リストなどで、それらのクラスター成分を指定してください。

以下のオプションについて詳しくは、『JMP の使用法』の「JMP のレポート」章を参照してください。

ローカルデータフィルタ 現在のレポートにおいてデータをフィルタリングするローカルデータフィルタを表示するか、非表示にします。

やり直し 分析を繰り返すオプションや、やり直すオプションを含みます。また、［自動再計算］オプションを選択すると、このオプションに対応しているプラットフォームにおいて、データテーブルに加えた変更が、該当するレポートに即座に反映されるようになります。

スクリプトの保存 レポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。

By グループのスクリプトを保存 By 変数の全水準に対するレポートを再現するためのスクリプトを保存するオプションが、保存先ごとに用意されています。起動ウィンドウで By 変数を指定した場合のみ使用可能です。

「変数のクラスタリング」プラットフォームの別例

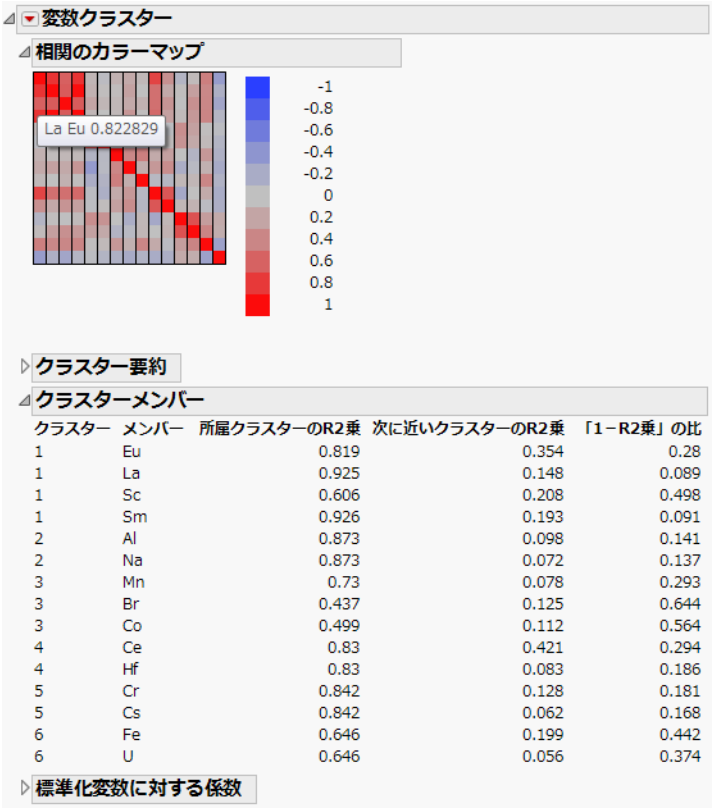
相関のカラーマップの例

この例では、「相関のカラーマップ」を見てみます。

1. ［**ヘルプ**］>［**サンプルデータライブラリ**］を選択し、「Cherts.jmp」を開きます。
2. ［**分析**］>［**クラスター分析**］>［**変数のクラスタリング**］を選択します。
3. 連続尺度の列をすべて選択し、［**Y, 列**］をクリックします。
4. ［**OK**］をクリックします。
5. 「クラスター要約」レポートと「標準化変数に対する係数」レポートを閉じます。
6. 相関のカラーマップの 2 行目、1 列目の上にカーソルを置きます。

   このセルに対応する変数が「**La**」と「**Eu**」で、それらの相関が 0.822829 であることがわかります。

図 11.4「Cherts.jmp」の相関カラーマップ



「クラスターメンバー」レポートを見ると、クラスター 1 には 4 つの変数があります。「相関のカラーマップ」では、これら 4 つの変数に対応する左上の 4x4 セルの正方形が、強い相関があることを示すパターンになっています。また、カラーマップを見ると、クラスター 2、4、5 の変数間にもやや強い相関があります。さらに、クラスター 6 の 2 つの変数に対応しているカラーマップ右下の 2x2 セルの正方形は、負の相関を示すパターンになっています。詳細については、「相関のカラーマップ」（201 ページ）を参照してください。

「変数のクラスタリング」プラットフォームの次元削減の例

この例では、「変数のクラスタリング」プラットフォームで説明変数の次元を少なくします。「Penta.jmp」サンプルデータテーブルには、応答変数「log RAI」の予測に使用する 15 個の変数が含まれています。変数のクラスタリングを使用して、この数を減らしましょう。

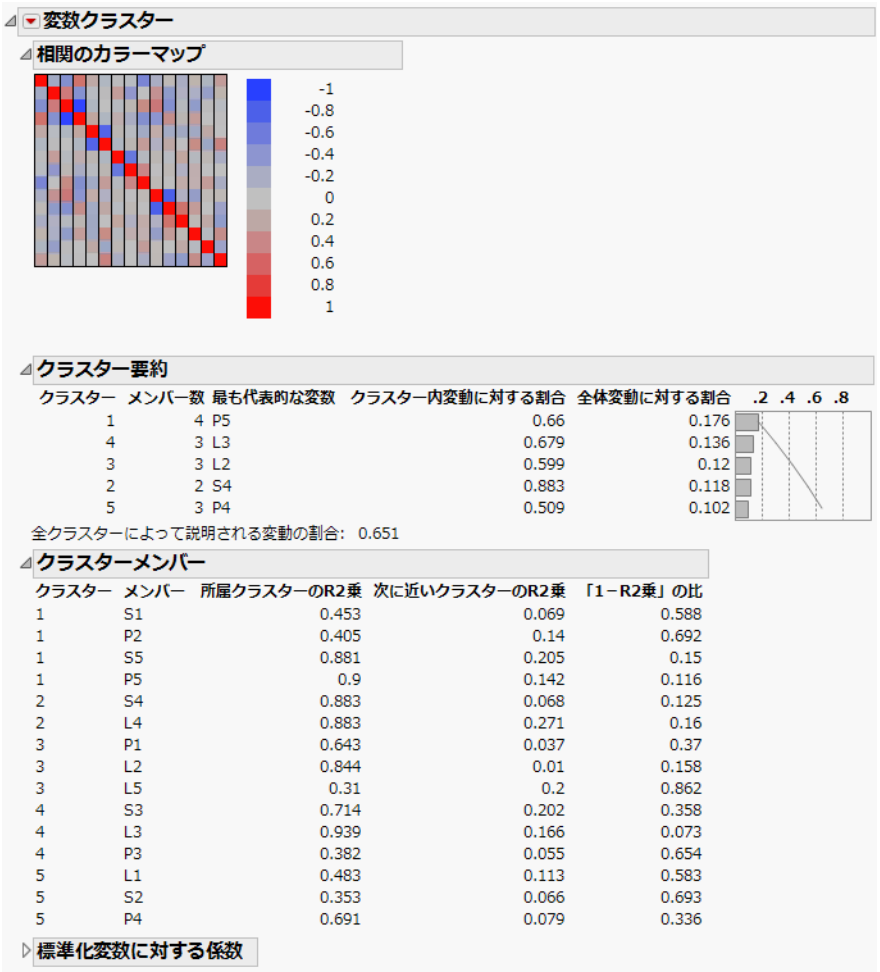
変数のクラスタリング

1. ［ヘルプ］＞［サンプルデータライブラリ］を選択し、「Penta.jmp」を開きます。
2. ［分析］＞［クラスター分析］＞［変数のクラスタリング］を選択します。
3. 「logRAI」を除くすべての連続変数を選択し、［Y, 列］をクリックします。

4. ［OK］をクリックします。
5. 「変数クラスター」の赤い三角ボタンをクリックし、［**クラスター成分の保存**］を選択します。

   3つの計算式の列がデータテーブルに追加されます。

図11.5「Penta.jmp」の「変数のクラスタリング」レポート



「クラスター要約」レポートと「クラスターメンバー」レポートから、変数が5つのグループにクラスタリングされ、5つのクラスター成分があることがわかります。

モデルのあてはめ

次に、次の2つのモデルをあてはめて、「**logRAI**」を予測します。

- すべての変数を説明変数として使用したモデル

- クラスター成分を説明変数として使用したモデル

1. 「変数クラスター」の赤い三角ボタンをクリックし、［**モデルのあてはめを起動**］を選択します。
2. 「**logRAI**」を選択し、［**Y**］をクリックします。

   最も代表的な変数の5つのクラスターが、「モデル効果の構成」リストに入力されていることに注目してください。ただし、この例では、まず、すべての変数を用いています。
3. 「**S1**」から「**P5**」までのすべての連続変数を選択し、［**追加**］をクリックします。

   「**Obs Name**」を含めないように注意してください。
4. 「**ダイアログを開いたままにする**」の横のボックスにチェックを入れます。
5. ［**実行**］をクリックします。

図11.6 すべての連続尺度の説明変数を使ったモデルの「最小2乗法によるあてはめ」レポート

応答 log RAI

効果の要約

あてはめの要約

R2乗	0.929316
自由度調整R2乗	0.853582
誤差の標準偏差(RMSE)	0.331225
Yの平均	0.734333
オブザベーション(または重みの合計)	30

分散分析

要因	自由度	平方和	平均平方	F値
モデル	15	20.193596	1.34624	12.2709
誤差	14	1.535941	0.10971	p値(Prob>F)
全体(修正済み)	29	21.729537		<.0001*

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob>|t|)
切片	-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

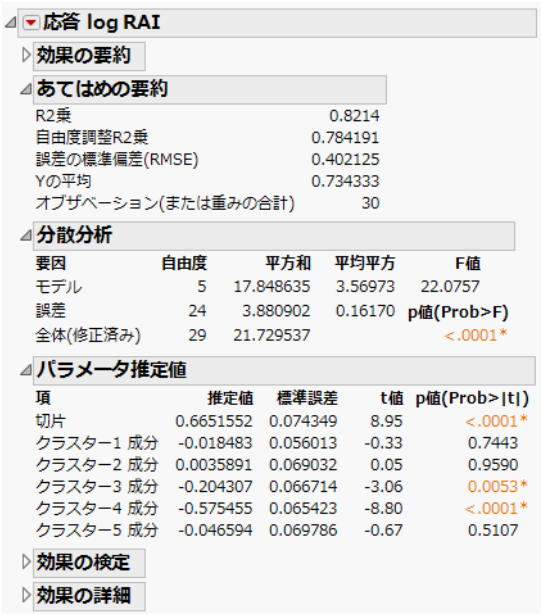
効果の検定

効果の詳細

6. 「モデルのあてはめ」ウィンドウで、「モデル効果の構成」リスト内のすべての変数を選択して、［**削除**］をクリックします。
7. 5つのクラスター成分を選択して［**追加**］をクリックします。

8. ［実行］をクリックします。

図11.7 クラスター成分を説明変数として使ったモデルの「最小2乗法によるあてはめ」レポート



応答 log RAI

効果の要約

あてはめの要約

R2乗	0.8214
自由度調整R2乗	0.784191
誤差の標準偏差(RMSE)	0.402125
Yの平均	0.734333
オブザベーション(または重みの合計)	30

分散分析

要因	自由度	平方和	平均平方	F値
モデル	5	17.848635	3.56973	22.0757
誤差	24	3.880902	0.16170	p値(Prob>F)
全体(修正済み)	29	21.729537		<.0001*

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob>|t|)
切片	0.6651552	0.074349	8.95	<.0001*
クラスター1 成分	-0.018483	0.056013	-0.33	0.7443
クラスター2 成分	0.0035891	0.069032	0.05	0.9590
クラスター3 成分	-0.204307	0.066714	-3.06	0.0053*
クラスター4 成分	-0.575455	0.065423	-8.80	<.0001*
クラスター5 成分	-0.046594	0.069786	-0.67	0.5107

効果の検定

効果の詳細

5つのクラスター成分だけを説明変数として含めたモデルでは、自由度調整R2乗が0.784で、応答内のかなりの変動を説明しています。15個すべての説明変数を使用したモデルの自由度調整R2乗は、それより少しだけ高い0.853です（図11.6）。

「変数のクラスタリング」プラットフォームの統計的詳細

ここでは、「変数のクラスタリング」プラットフォームの統計的詳細について説明します。

変数クラスターのアルゴリズム

クラスタリングのアルゴリズムは、反復的に変数のクラスターを分割し、それ以上の分割が不可能となるまで変数をクラスターに割り当てます。最初のクラスターはすべての変数で構成されます。このアルゴリズムはSASが開発し、PROC VARCLUS（SAS Institute Inc., 2011）にインプリメントされています。

メモ: このアルゴリズムでは、［Y, 列］に指定されたすべての変数に関して欠測値がないオブザベーションだけを使用します。

アルゴリズムの反復ステップは次のとおりです。

1. すべてのクラスターについて、次を実行します。
   a. 各クラスターにおいて、そのクラスターに属する変数だけを対象に主成分分析を行います。
   b. すべてのクラスターの2番目の固有値が1より小さい場合、アルゴリズムを終了します。
2. 2番目の固有値が最も大きい（かつ1より大きい）クラスターを、次のようにして新しい2つのクラスターに分けます。
   a. オーソブリク（orthoblique）回転を用いて、現在のクラスター内の変数の主成分を回転させます。
   b. クラスターを1つ定義して、回転後の第1主成分との相関のR2乗が第2主成分との相関のR2乗よりも強い現在のクラスター内の変数を構成します。
   c. もう1つのクラスターを定義して、元のクラスター内の残りの変数を構成します。これらは第2主成分との相関がより強い変数です。
   d. この2つの新しいクラスターの主成分を計算します。
3. データセット内に、別のクラスターに割り当てるべき変数がないかどうかをテストします。変数ごとに、次を実行します。
   a. 各クラスターの第1主成分との相関のR2乗を計算します。
   b. 変数を、相関のR2乗が最も大きいクラスター内に含めます。

メモ: オーソブリク（orthoblique）回転の直交回転には、生のコーティマックス（quartmax）回転を用いています。オーソブリク回転については、Harris and Kaiser（1964）を参照してください。

付録 A

統計的詳細

多変量分析

この付録では、「線形 横長データ」手法での特異値分解について説明します。また、多変量検定、正確な F 統計量および近似の F 統計量に使用される計算式についても詳しく説明します。

「線形 横長データ」の手法と特異値分解

「クラスター分析」、「主成分分析」、および「判別分析」の各プラットフォームにある「線形 横長データ」の手法では、数千個あるいは数百万個の変数を持つデータセットの分析ができます。多変量分析の多くでは、共分散行列やその逆行列を求めます。変数の個数が多いデータで多変量分析を行うと、共分散行列が巨大になり、その計算や、その逆行列の計算が難しくなり、多大な計算コストがかかってしまいます。

n 個の行と p 個の列から成るデータがあるとしましょう。共分散行列のランクは、最大でも n と p のうち小さい方と等しくなります。横長のデータセットでは、p が n より大きいのが普通です。このような場合、共分散行列の逆行列における0以外の固有値は、多くても n 個となります。「線形 横長データ」手法は、特異値分解とともにこの性質を利用して、効率的な計算を行います。「特異値分解のアルゴリズム」（212ページ）を参照してください。

特異値分解

特異値分解（Singular Value Decomposition; SVD）は、行列による線形変換を、回転、尺度変換、もう一つの回転で表したものです。特異値分解は、$n \times p$ の行列 $\mathbf{X}$ を次のように分解します。

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

r を $\mathbf{X}$ のランクとします。また、$r \times r$ の単位行列を $\mathbf{I}_r$ とします。

行列 $\mathbf{U}$、$Diag(\mathbf{\Lambda})$、および $\mathbf{V}$ には、次のような性質があります。

$\mathbf{U}$ は、$\mathbf{U}'\mathbf{U} = \mathbf{I}_r$ を満たす $n \times r$ の準直交行列

$\mathbf{V}$ は、$\mathbf{V}'\mathbf{V} = \mathbf{I}_r$ を満たす $p \times r$ の準直交行列

$Diag(\mathbf{\Lambda})$ は、列ベクトル $\mathbf{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_r)'$ を対角要素に持つ $r \times r$ の対角行列。ここで、$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$

λ_i は、$\mathbf{X}$ の0以外の**特異値**です。

特異値分解は、正方行列のスペクトル分解（固有値分解）と次のような関係があります。

- λ_i の平方は、$\mathbf{X}'\mathbf{X}$ の0以外の固有値です。
- $\mathbf{V}$ の r 列は、$\mathbf{X}'\mathbf{X}$ の固有ベクトルです。

メモ: 行列 $\mathbf{U}$、$\mathbf{V}$、および特異値に関し、行列の次元に応じていくつかの異なる表現方法が文献で紹介されています。しかし、それらには、$\mathbf{X}$ のランクまでの特異値分解に関しては根本的な違いはありません。

詳細については、Press et al.（1998, Section 2.6）を参照してください。

特異値分解と共分散行列

ここでは、特異値分解によって共分散行列の固有ベクトルと固有値を求める方法について説明します。対象の共分散行列が巨大であるがランクが小さい場合には、データを特異値分解したほうが、共分散行列を固有値分解するよりもずっと効率的です。

nをオブザベーション数、pを変数の個数とします。データの$n \times p$行列を$\mathbf{X}$とします。

特異値分解は通常、標準化されたデータに適用されます。データの標準化とは、データから平均を引き、それを標準偏差で割る変換を指します。標準化したデータの共分散行列は、$\mathbf{X}$の相関行列となります。標準化したデータ値のn x p行列をXsとすると、相関行列は次のように求められます。

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

$\mathbf{X}_s$の特異値分解によって、$\mathbf{X}_s'\mathbf{X}_s$の固有ベクトルと固有値を求めることができます。この性質を利用すれば、行列$\mathbf{X}$が非常に横長の場合（列数が膨大な場合）、または縦長の場合（行数が膨大な場合）に、固有ベクトルや固有値を効率的に計算できます。この計算方法が、横長データに対する主成分分析の基本になります。「主成分分析」章の「「主成分分析」レポート」（54ページ）を参照してください。

特異値分解および共分散行列の逆行列

一部の多変量分析では、共分散行列の逆行列を計算する必要があります。ここでは、特異値分解によって共分散行列の逆行列を計算する方法について説明します。

標準化したデータ行列を$\mathbf{X}_s$とし、また、$\mathbf{S} = \mathbf{X}_s'\mathbf{X}_s$とします。特異値分解された行列は、$\mathbf{S}$と次のような関係になっています。

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

$\mathbf{S}$がフルランクの場合、$\mathbf{V}$は$p \times p$の直交行列で、$\mathbf{S}^{-1}$は次のように表されます。

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

$\mathbf{S}$ がフルランクではない場合、$Diag(\mathbf{\Lambda})^{-1}$ の部分が、$Diag(\Lambda)$ の対角要素を逆数にした一般化逆行列 $Diag(\mathbf{\Lambda})^{-}$ に置き換えられます。これにより、$\mathbf{S}$の一般化逆行列は次のように定義されます。

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

この一般化逆行列は、特異値分解だけによって求めることができます。

横長データの線形判別分析における特異値分解の適用については、「線形判別分析」章の「横長データに対する線形判別法」（98ページ）を参照してください。

特異値分解のアルゴリズム

JMPの多変量分析プラットフォームで使われている特異値分解のアルゴリズムは、Golub and Kahan（1965）が提唱したものに従っています。Golub and Kahan（1965）のアルゴリズムは、2つのステップから成ります。最初のステップでは、行列**M**を2重対角行列**J**に変形します。2番目のステップで、**J**の特異値を計算します。この特異値は、元の行列**M**の特異値と同じです。行列**M**の列は、通常、計算における列の効果を均一化するために標準化されます。Golub and Kahan（1965）の手法は、計算が効率的です。

付録 B

参考文献

Agresti, A. (2002), *Categorical Data Analysis*, Second Edition, New York: John Wiley and Sons, Inc.

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

Ballard, D.H., (1981), "Generalizing the Hough Transform to Detect Arbitrary Shapes," *Pattern Recognition*, 13:2, 111-122.

Bartlett, M.S. (1937), "Properties of sufficiency and statistical tests," *Proceedings of the Royal Society of London Series A*, 160, 268–282.

Bartlett, M.S. (1954), "A Note on the Multiplying Factors for Various Chi Square Approximations," *Journal of the Royal Statistical Society, 16 (Series B)*, 296–298.

Boulesteix, A.-L. and Strimmer, K. (2007), "Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data," *Briefings in Bioinformatics*, 8(1), 32–44.

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

Cox, I. and Gaudard, M. (2013), *Discovering Partial Least Squares with JMP*, Cary NC: SAS Institute Inc.

Cronbach, L.J. (1951), "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, 16, 297–334.

De Jong, S. (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 251–263.

Denham, M.C. (1997), "Prediction Intervals in Partial Least Squares," *Journal of Chemometrics*, 11, 39–52.

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

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., and Wold, S. (2006), *Multi- and Megavariate Data Analysis Basic Principles and Applications (Part I)*, Chapter 4, Umetrics.

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

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

Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951a), "Sur La Liaison et la Division des Points d'un Ensemble Fini," *Colloquium Mathematicae*, 2, 282–285.

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

Golub, G.H., Kahan, W. (1965), "Calculating the singular values and pseudo-inverse of a matrix," *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis* 2:2, 205–224.

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

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

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

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

Hand, D, Mannila, H, and Smyth, P. (2001), Principles of Data Mining, MIT Press.

Harris, C.W. and Kaiser, H.F. (1964), "Oblique Factor Analytic Solutions by Orthogonal Transformation," *Psychometrika*, 32, 363–379.

Hartigan, J.A. (1981), "Consistence of Single Linkage for High–Density Clusters," *Journal of the American Statistical Association*, 76, 388–394.

Hastie, T., Tibshirani, R., and Friedman, J.H.(2009), *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, New York: Springer Science and Business Media.

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

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

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

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

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

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

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

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

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

Kohonen, T. (1990), "The Self-Organizing Map," *Proceedings of the IEEE*, 78:9, 1464-1480.

Lindberg, W., Persson, J.-A., and Wold, S. (1983), "Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate," *Analytical Chemistry*, 55, 643–648.

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

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

McLachlan, G.J. and Krishnan, T. (1997), *The EM Algorithm and Extensions,* New York: John Wiley and Sons.

McQuitty, L.L. (1957), "Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies," *Educational and Psychological Measurement*, 17, 207–229.

Milligan, G.W. (1980), "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika*, 45, 325–342.

Nelson, Philip R.C., Taylor, Paul A., MacGregor, John F. (1996), "Missing Data Methods in PCA and PLS: Score calculations with incomplete observations," *Chemometrics and Intelligent Laboratory Systems*, 35, 45–65.

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

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

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

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

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

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

Schafer, J. and Strimmer, K. (2005), "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics", *Statistical Applications in Genetics and Molecular Biology*, 4, 1175–1189.

Sneath, P.H.A. (1957) "The Application of Computers to Taxonomy," *Journal of General Microbiology*,17, 201–226.

Sokal, R.R. and Michener, C.D. (1958), "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Science Bulletin*, 38, 1409–1438.

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

Tracy, N.D., Young, J.C., Mason, R.R. (1992), "Multivariate Control Charts for Individual Observations," *Journal of Quality Technology*, 24, 88–95.

Umetrics (1995), *Multivariate Analysis (3-day course)*, Winchester, MA.

White, K.P., Jr., Kundu, B., and Mastrangelo, C.M., (2008), "Classification of Defect Clusters on Semiconductor Wafers Via the Hough Transform," *IEEE Transactions on Semiconductor Manufacturing*, 21:2, 272-278.

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

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

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

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

索引
多変量分析

数字
95% 二変量正規確率楕円 39

B
Baltic.jmp 107
By 変数 52

C
Cronbach の α 係数 41–43
Cronbach の α 係数
　統計的詳細 48

D
Danger.jmp 42

E
EM アルゴリズム 156
E 行列 101

H
Hoeffding の D 統計量 38, 45
H 行列 101

J
JMP の計算で使用される式 210

K
Kendall の順位相関係数（τ）44
Kendall の順位相関係数（τ）38
k-means クラスター分析プラットフォーム
　自己組織化マップ 167
k-means 法 156

L
L 行列 101

M
Mahalanobis の距離 40, 46
MDS プロット 191
M 行列 101

P
PCA（主成分分析）50
Pearson 相関 34, 44
PLS 105–128
　統計的詳細 126–129
PLS 回帰プラットフォーム
　検証 111
PLS における欠測値の補完 112
p 値のカラーマップ 36

R
ROC 曲線 86

S
Solubility.jmp 51
Spearman の順位相関係数（ρ）44
Spearman の順位相関係数（ρ）37

T
T^2 統計量 41

W-Z
Ward 法 137, 154

ア
値が近い行をグループにまとめる
　クラスター分析プラットフォームを参照
アルゴリズム 210

イ
一変量 32
因子分析 50
因子分析、概要 50
因子分析プラットフォーム
　By変数 52

カ
各グループに属する確率を表示 86
各グループ平均への距離を表示 85
確率楕円 39
カラーマップ 144

キ
幾何級数 144
期待値最大化アルゴリズム 156
逆相関表 33
距離行列の保存 145
距離グラフ 144
距離スケール 144
近似F検定 102

ク
クラスター階層の保存 145
クラスター数の選択規準 144
クラスターの色分け 143
クラスターのシミュレーション 181
クラスターの保存 145, 167, 181
クラスターのマーカー分け 143
クラスター分析の履歴 143
クラスター分析プラットフォーム 131
　k-means法 156–181
　階層型 132–154
　概要 132, 156, 172
　起動 136
　手法の比較 132, 156, 172
　正規混合分布法 171–184
グループ内共分散行列の表示 84
グループの追加 84
グループ平均の表示 84
群平均法 137, 154

ケ
計算についての詳細 210
計算方法 210
欠測値 34
欠測値の補完、PLS 112

コ
効果の大きさ、潜在クラス分析 191
高次元プロットの次数を落とす 50
項目の信頼性 41–43
固有値分解 50, 54
固有ベクトル 56
混合確率 192
混合確率の保存 181
混合計算式の保存 181

サ
最短距離法 137, 154
最長距離法 138, 154
三次元正準プロット 84
三次元バイプロット 166, 181
散布図行列 31
散布図行列 38, 85

シ
次元 50
自己組織化マップ 167
質問票調査の分析 41–43
ジャックナイフ法による距離 41
集塊クラスター分析 132
重心 40
重心法 137, 154
樹形図 131–132, 141
樹形図データの保存 145
樹形図のスケールコマンド 144
樹形図の表示 144
主成分の回転 62

主成分分析 50
信頼性分析 41–43
　生存時間分析プラットフォームも参照

ス

スクリープロット 59
スコアプロット 60

セ

正規50%等高線の表示 87
正規確率楕円 39
星座樹形図 145
正準スコアの保存 88
正準の詳細を表示 87
積率相関 34, 44
線形結合 50

ソ

相関行列 31
相関のカラーマップ 36
相関の逆行列 33, 46
相関のクラスタリング 36
相関の表示 39
相関の棒グラフ 34
その他 39

タ

対比M行列 101
楕円色 39
楕円内を塗る 39
楕円の信頼率 39
楕円の透明度 39
多項混合分布 186
多変量の相関 29–30
多変量の相関プラットフォーム 209
　主成分分析 50
多変量の外れ値 40
多変量平均 40

チ

直線のあてはめ 39

テ

データの標準化 139
転置したパラメータ推定値 191
点の表示 39, 87

ト

等間隔 144
統計的詳細 210

ニ

二変量正規確率楕円 39

ノ

ノンバラメトリック相関係数 37
ノンパラメトリックな連関の測度 37
ノンパラメトリック密度 40

ハ

バイプロット 166, 180
バイプロットオプション 166, 181
バイプロット線 61
パイプロット線の位置 87
バイプロット線の表示 87
外れ値の距離プロット 46
外れ値分析 40
パラレルプロット 167, 181
判別分析、PLS 130
凡例 144

ヒ

ヒストグラムの表示 39
表示順序の保存 145

フ

負荷量プロット 60
プロット点の色分け 87

ヘ

ペアごとの相関係数 32
ペアごとの相関係数表 34
平均の信頼限界楕円の表示 87

変数間クラスター 144
偏相関 33
偏相関表 33

ミ

密度関数の保存 181

モ

最も近いクラスターの計算式を保存 145

ユ

有意確率 34