第13章
主成分分析

『データ分析入門』（第 7 版）の第 13 章（266 ～ 278 頁）については、本冊子をご覧ください。

2009年 3 月
慶應義塾大学出版会

第13章
主成分分析

本章の概要

本章では、３つ以上の変量を同時に扱い、その関係を見通しよく整理して情報を引き出すための方法である主成分分析を学ぶ。まず、第８章で学んだ散布図の考え方を拡張し、多変量データの散らばり方を表示する３次元散布図の取り扱い方を学習する。続いて、そうした多次元空間内でのデータの散らばりについての情報を効率的に要約する方法である主成分分析を学習する。その分析のメカニズムを正確に理解するためには、数学的説明によらなければならないが、ここではグラフ表示による、視覚的で直感的な理解をめざす。

1　3次元散布図

1.1　3次元散布図

主成分分析の準備として、まず３次元データの取り扱いを JMP でどのように行うのかを修得しよう。そのために、まず**３次元散布図**（*three-dimensional plot*）の書き方について学習する。アメリカ50州の各州で起きた犯罪についてのデータである、“DAsample”フォルダ内の“crime”ファイルを用いて３次元プロットを行う。なお、データテーブルの変量名については表13.1に示す。

表13.1：“crime”ファイルにおける変量名

変量名
州名
殺人
婦女暴行
強奪
暴行
強盗
窃盗
自動車盗
地域区分

操作

1. “JMP スターター”の“カテゴリをクリック”の欄から“グラフ”を選択し、“三次元散布図”をクリックする。

2. “データファイルを開く”ウィンドウ内の“DAsample”にある“crime”を開く。

3. “グラフ”メニューから“三次元散布図”を選択し、そこの“列の選択”欄から“殺人”、“婦女暴行”、“自動車盗”の３つのデータを順に“Y，列”欄にクリックしながら選択すると、選択された3つの変量による三次元散布図が描かれる（図13.1）。なお、離れた位置にある２つ以上の変量を選択するときに Ctrl キーを押しながらクリックするとよい。

図13.1：“回転プロット”の列の選択画面

4. “OK”ボタンをクリックすると図13.2のような３次元散布図となる。

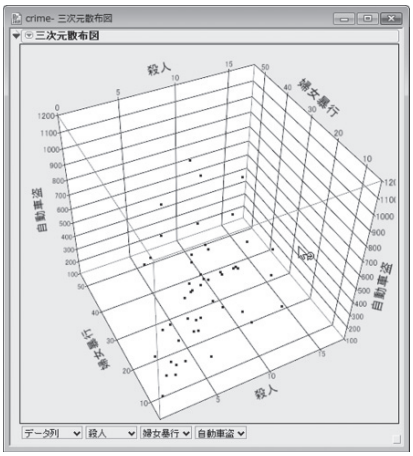


図13.2： ３次元散布図の画面

5. 回転のさせ方

キーボードの十字キーで回転させることができる。また、画面の中の散布図を直接ドラッグすることによっても回転させることができる。

6. プロットされた各点がどの州かを知る方法

“crime”の“州名”にはラベルが指定されているので、図13.2のプロットの各点にマウスのカーソルを合わせると、それぞれの州の名前が表示される。

練習問題

軸をいろいろと回転させ、特徴的な位置にある点をクリックして、データの3次元の散らばりの様子をじっくりと観察しよう。

1.2 確率楕円体

三次元散布図でも、二次元の場合（第8章参照）と同じように95%の確率でデータが分布する範囲を図で示すことができる。それを**確率楕円体**と呼ぶ。これを使って、データが主に分布する範囲や外れ値を空間内で発見することができる。

確率楕円体は、個々の変量の値に対しても、変量としてあらかじめ作成された区分ごとに対しても、指定した変量全体を考慮しても作ることもできるが、まずは、指定した変量全体を考慮した確率楕円体をつくってみるほうが、分布の広がり具合や全体的な外れ値が見いだせて、データのばらつき具合を知るための役に立つ。

もし、データ内にいくつかのサブグループがあれば、そのサブグループごとの分類を示す変量をつくり、サブグループごとの確率楕円体を描くこともできる。これは、それぞれの小グループが、空間の中でどのような位置関係になっているかを示す有力な手がかりを与えてくれる。

操作

1. “カテゴリをクリック”のウィンドウの中から“ファイル”を選択し、さらに“データテーブルを開く”をクリックして、“DAsample”内の“crime”ファイルを開く。

2. “カテゴリをクリック”のウィンドウの中から“グラフ”を選択し、さらに“三次元散布図”をクリックする。

3. Y列に“殺人”から“自動車盗”までを選択し、“OK”ボタンをクリックする。

4. 三次元散布図の横の赤い三角ボタンを押し、“確率楕円体”を選択すると、データ全体の値に対し

て95%確率楕円体の描画を行う。これで、データ全体の傾向と外れ値が見いだせる。

5. 次に、サブグループ（このデータの場合はアメリカの地域区分）ごとの分布の傾向と相違を見るために、サブグループを示す変量を使って、それぞれのサブグループごとの確率楕円体を描く。ウィンドウ内の“二択”の次の“列の値ごとに行う”」を選択し、サブグループとしての“地域区分”をクリックし、OK ボタンを押す。すると、図13.3のように、サブグループごとの確率楕円体が現れるので、よく観察してみよう。

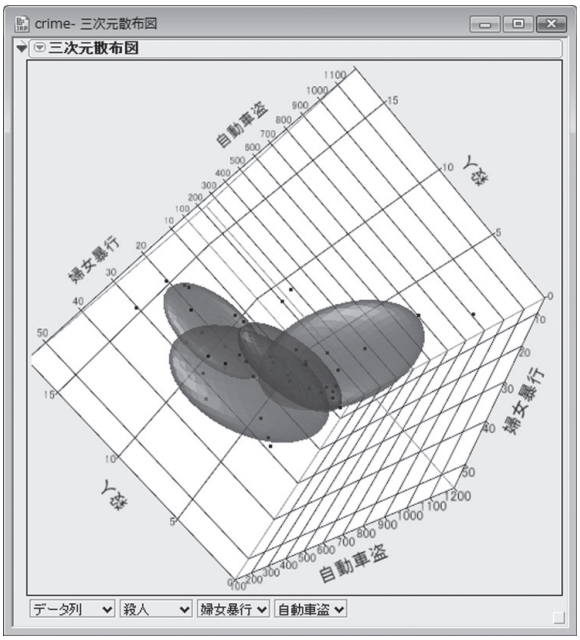


図13.3：サブグループごとの確率楕円体

三次元散布図をリポート作成などのために、より見やすくする工夫を考えよう。三次元散布図作成の操作をしている時に、ツールの中の左端の矢印ボタンを押し、図の中で右クリック（Macintosh の場合は control キーを押しながらクリック）し、“設定”をクリックすると、散布図中のそれぞれの値のプロットに対して、サイズや形などを設定することができる。この操作によって、より見やすい散布図を作成することができ、よりよいリポートの作成が可能となる。

2 主成分分析

2.1 主成分分析とは

たとえば、「身長」と「体重」という 2 つのデータから、人間の体つきについて分析することを考えよ

う。身長と体重には正の相関があることが知られている。身長の高い人は概して体重も重く、身長の小さい人は体重も軽いことが普通だからである。第８章で学んだ散布図で表現すると、図13.4のようになる。

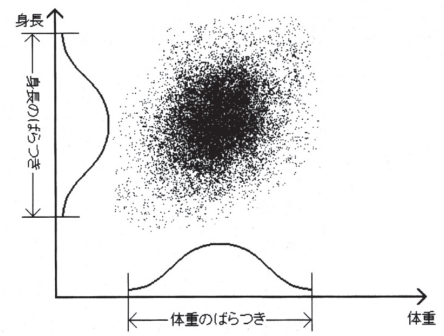


図13.4：「身長」と「体重」の散布図

図13.4は、身長・体重ともにそれぞれが独自のばらつきを持って散らばっていることを示している。しかし、両者の間に正の相関があるということは、身長・体重の両方とも、体格が大きい人のとる値が大きいことを意味する。この場合、身長と体重という２つの変量がそれぞれ持っている「人間の体つき」に関する情報の内容がかなり重なりあっているため、身長と体重を別々に取り上げて分析すると、両方の変量とも体格の大きさだけを主に表現してしまい、肥満型・やせ型などといった人間の体つきについてのほかの種類の情報は、ここから読み取りにくい。これではデータが持つ多様な情報を一度に十分表現できず、分析の効率がわるそうである。

それならば、ばらつきを身長方向や体重方向で考えるのではなく、図13.5に示した、A－B と C－D の方向で考えることにしたらどうであろうか。A－B 方向は、図13.4の散布図でプロットされた点が形づくる楕円の最も長い方向を示しており、ここに投影するとき、各点のばらつきは最も大きくなる。この方向の意味を考えると、A の向きは身長・体重がともに大きいことを、B の向きは身長・体重がともに小さいことを表している。すなわち A－B 方向は、いわば「体格」を示す方向であるといえる。

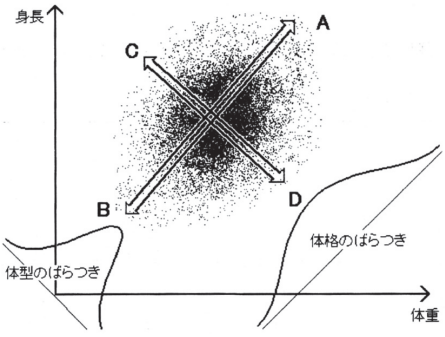


図13.5：主成分によるばらつき

一方、A－B 方向で表せない情報は、それと直交する C－D 方向が示す情報である。これは、C の向きに行くほど、身長が高い割に体重が軽いというやせ型であることを、逆に D の向きは体重が重い割には身長が小さいという肥満型であることを示している。つまり、C－D は「体型」を示す方向であるといえる。

この図内の点のばらつきについての情報を表現するためには、「身長」と「体重」で考えても、「体格」と「体型」で考えても、同じ散布図が再現できるという点で情報量は同じといえる。それならば、部分的に情報が重なって相関のある「身長」と「体重」よりも、ばらつきの情報をうまく分けて考えられる「体格」と「体型」のほうが、データに含まれている情報を解釈するときにたくさんのことがわかりやすく表現される。また、ばらつきの最大方向である「体格」には、人間の体つきについての多くの情報が集約されていると考えられ、効率的でもある。

ここで見た「体格」と「体型」のように、多種のデータによってできた散布図の空間の中で、まずデータのばらつきが最大の方向（**軸** *axis* と呼ばれることもある）を見つけ、それと直交しながら、ばらつきがそのつぎに大きい方向を順次見つけてゆくことによって、データの持つ情報を効率的に記述し、理解することができる。このような軸を見つける分析手法が**主成分分析**（*principal component analysis*：略して PCA）である。分析の結果として見い出されたそれぞれの軸のことを**主成分**（*principal component*）と呼ぶ。抽出された主成分は、ばらつきが最大の方向を示す主成分から順に、それぞれ第 1 主成分、第 2 主成分……と呼ばれる。

理屈の上では、主成分は分析に使われた変量の数だけ見い出すことができる。しかし、見い出されたすべての主成分が、先の例の「体格」や「体型」ように解釈可能で、意味を持つとは限らない。複雑にからみ合った関係を持つ多くの変量を同時に処理する場合には、少数の主成分だけを抽出することによって、散布図の中の点のばらつきが表現している情報を明確に理解できることが多い。すなわち、主成分分析は、多くの変量に重複して含まれる情報を整理し、主成分という互いに直交した独立な、つまり数学的に相互に無関係な新しい軸を作って表現することによって、点のばらつきの中に潜む情報を理解しやすくするための手法なのである。

2.2 主成分分析の実行と結果

JMP で主成分分析を行うには、JMP スターターの“カテゴリをクリック”の中から“多変量”を選択し、“主成分分析”をクリックする。次に、“列の選択”欄から主成分分析を行うためのデータ（今回の場合は“殺人”、“婦女暴行”、“自動車盗”の 3 つのデータ）を順に“Y，列”にクリックしながら選択し“OK”ボタンをクリックすると、主成分分析の結果が表示される（図13.6）。

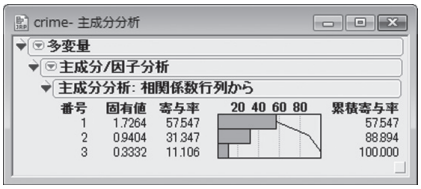


図13.6：主成分分析の結果を示した画面（1）

次に、図13.6の状態から、“主成分 / 因子分析”の左にある赤い三角マークをクリックし、“回転プロット”オプションを選ぶと、主成分分析の結果をもとにした新たな三次元散布図が描かれる（図13.7）。この図には、中央にもとのデータの３つの軸（“殺人”、“婦女暴行”、“自動車盗”）が描かれている。なお、この図では、３つの軸があるうち、どの軸がどのデータを表しているか判断が付きにくいかもしれないが、描かれている軸にマウスカーソルを合わせると、“殺人”、“婦女暴行”、“自動車盗”と、それぞれの軸がどのデータを表しているのか表示されるので、確認してほしい。

ここで、再び軸をいろいろと回転させ、主成分が点の散らばりの中をどんな方向に通っているかを観察しよう。2.1節では、「体格」「体型」として説明したが、この犯罪データでそれにあたるものは何と考えればよいであろうか。

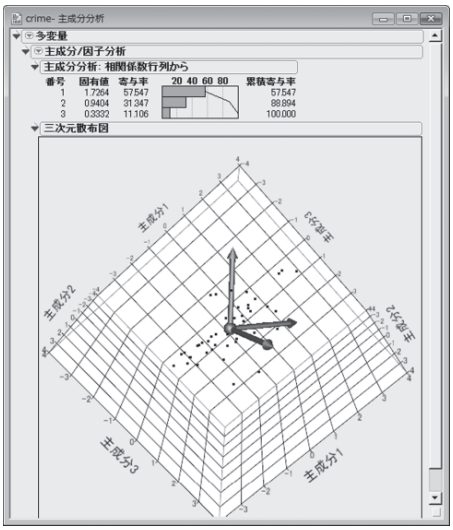


図13.7：主成分分析の結果を示した画面（2）

主成分分析によって行われることは、上で見たように「データの散らばりから抽出できる互いに関係のある情報を分類して独立した何種類かのものに集約し、表現する」ことだけである。コンピュータの分析出力から何らかの結論がすぐに得られるわけではなく、出力結果として得られた「集約された結果の情報（＝主成分）」を解釈するという作業が欠かせない。そして、この「解釈」という作業にあたっては、分析者自身のセンスが非常に重要になってくる。したがって、主成分分析を使いこなすには、分析するデータに関する分析者の事前知識が必要になる。

では次に、解釈という作業を行うために必要となる、主成分分析によって得られた各主成分についての詳細な結果の見方を説明していく。その前に、図13.7の状態から、“主成分 / 因子分析”の左にある赤い三角マークをクリックし、“固有ベクトル”オプションを選び、固有ベクトルを表示させておく（図13.8）。この固有ベクトルについても他の項目と一緒に、後で説明していく。

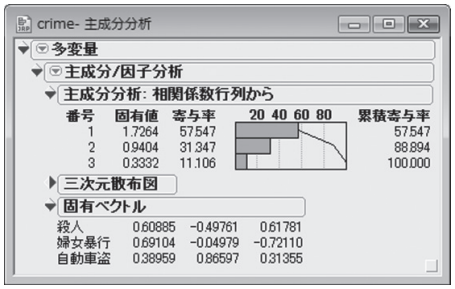


図13.8：主成分分析の結果を示した画面（3）

ここで、各主成分についての分析結果の詳しい見方について、重要なことを以下に示す。

固有値（*eigenvalue*）：
データの全情報量（データのすべてのばらつき）のうち、それぞれ主成分がどのくらいの分量の情報を表現しているかを示す[1)]。第1主成分がもっとも大きな値をとり、以下だんだん小さくなる。

寄与率（*contribution ratio*）：
各主成分がそれぞれ受けもって表現している情報量（固有値に表現されている）を比率に直して表現したものである。たとえば、図13.8では、第1主成分の値が57.547であるが、これは第1主成分が全情報量の57.547％を集約して表現していることを意味している。

累積寄与率（*cumulative contribution ratio*）：
寄与率（パーセント）の第1主成分からの累積値。たとえば、図13.8では第2主成分における値が88.894であるが、これは第2主成分までででデータの全情報量の88.894％を表すことができることを意味している。

固有ベクトル（*eigenvector*）：
その主成分がなにを表しているかを解釈するときの手がかりであり、新しい方向である主成分ともとの変量との相関係数である。これを**主成分負荷量**（*principal component loading*）という。

これまで見てきたように、主成分は、分析対象となるすべての変量がつくる空間に、データのばらつき方を手がかりとした計算によって、それぞれが直交するように新たに引かれた軸である。つまり、主成分は、ばらつきの情報を集約して効率的に担うことができるように、計算によって作り出されたものである。空間内のx軸からz軸までの軸には、もともと“殺人”、“婦女暴行”および“自動車盗”という意味があったわけだが、先の「体格」と「体型」の例でもわかる通り、計算によって新たにつくられた主成分軸にもそれぞれなにかの意味があるはずである。その意味は、主成分を示す新たな3つの軸が、分析対象とな

[1)] 固有値と固有ベクトルの正確な意味については、線形代数を学ぶとよく理解できるが、本書では述べない。

ったもとの変量を示す３つの軸とどれほど似た方向を指しているかを見ることによって考えてゆける。その方向の似かより具合を相関係数の数値として示すのが、この主成分負荷量である。これらの値からそれぞれの主成分と関係の深い変量が持つ共通な性質は何であるかを考えて、主成分の意味を浮かび上がらせるようにしよう。

2.3 主成分の解釈

アメリカの各州の犯罪データ“crime”で行った主成分分析の結果である図13.8の固有ベクトルから、それぞれの主成分が持つ意味について考えてみよう。

第１主成分

すべての犯罪について正の値をとっているところから考えて、全体的な犯罪の起こりやすさを表す主成分であると解釈できる。

第２主成分

“自動車盗”という物めあての犯罪が正の値の負荷を、“殺人”、“婦女暴行”といった人に対する凶悪犯罪が負の値の負荷をとっているところから、各州で物めあての犯罪と人に対する凶悪犯罪のどちらが特徴的に多いのか、どちらに片寄っているのかを表している主成分だと解釈できる。

第３主成分

固有値や寄与率から判断すると、第３主成分は担っている情報量が小さい。累積寄与率を見ても、第２主成分までで、全情報量の９割近くを表現できているので、第３主成分は解釈しないことする。

上で述べたように、主成分を解釈するときには、たとえば第２主成分までの累積寄与率が十分に高いならば、データの持つばらつきは第１・第２主成分の２つだけで十分に説明できると判断して、これら２つの主成分の解釈のみを行えばよい。累積寄与率の値によっては、第３主成分以下の解釈が必要になることもある。

主成分の意味を解釈する際に注意すべきことは、主成分どうしは互いに直角に交わる（直交する）方向を向くように計算されて作られているという点である。つまり、主成分どうしは互いに無相関・無関係なのだから、このことを解釈にも反映させなければならない。すなわち、ある主成分の意味するものは、他の主成分が意味するものとは性質が違っているということである。たとえば、第１主成分を「重さ」を表すものと解釈したならば、第２主成分を「軽さ」を表すものと解釈するのは誤りである。なぜならば、もし第１主成分が重さを表すものであるならば、「軽い」ということは符号の違い（「重い」が正ならば「軽い」は負）で表されてしまい、第１主成分だけで説明がついてしまうからである。したがって、この場合、第２主成分は「軽い－重い」とは無関係なものに解釈しなければならない。たとえば「高い－低い」などのように、第１主成分とは無関係な要素に注目して解釈を考える必要がある。

2.4 軸の回転（Varimax 回転）

前節で見たように、各主成分の解釈は、分析に用いた変量と抽出された主成分との関係を手がかりにして行われる。ところが、主成分分析を実際に使ってみたとき、1つの変量が多くの主成分と多少の相関関係を持つという結果が出る場合が多い。これは、その変量が解釈の上で多義的な意味をあわせ持っているためであると理解することもできるが、ふつう、個々の変量の意味は比較的単純であると想定しておいた方が分析上、都合がよいことが多い。そうでなければ、各々の主成分の意味を考える作業が相当困難なものになってしまうからである。

このようなときには、それぞれの変量がいずれか1つの主成分とのみ関わる**単純構造**（*simple structure*）を持つように、抽出された軸を回転させることによって、軸の意味を考えやすくすることがある。このためには、まずデータから情報として重要なのは何番目の主成分まで解釈すれば十分なのかということを決め（**主成分数の決定**）、次いで、そこで抽出されたデータのばらつきに関する情報の総量は変えずに、主成分の方向を**回転**（*rotation*）により調整するという手続きが用いられる。結果として、全体としてできるだけ単純構造に近づけた、新しい主成分負荷量が得られるようにする方法である。ここでは、「主成分どうしが直交する」という性質はそのまま保持される。

本書では、主成分分析と似た分析手法である、**因子分析**（*factor analysis*）で用いられる軸の回転を主成分分析にも適用して分析する方法を紹介する[2)]。軸の回転をする方法にはいくつかの方法があるが、ここでは **Varimax 回転**（*Varimax rotation*）の実行方法について解説する。

操作

1. まず、“主成分 / 因子分析”の左にある赤い三角マークをクリックし、“主成分の回転”オプションを選ぶ。すると、図13.9の画面が現れる。

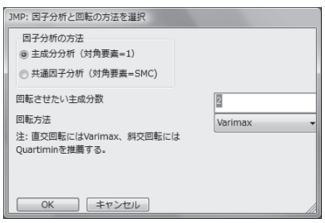


図13.9：主成分数の選択画面

2. 今回は主成分分析において、累積寄与率の値から第2主成分までで9割近くが説明できていたので、回転する主成分数を2つとし、ウィンドウに“2”と入力して“OK”ボタンをクリックすると、図13.10が現れる。

[2)] 主成分分析の結果に対して軸の回転を適用している実例については、たとえば奥野忠一・久米均・芳賀敏郎、吉澤正『多変量解析法』日科技連、1971、p.211参照。

この操作以後、**因子**（*factor*）という見慣れない用語が出てくるが、ここでは先に学習した「主成分」とまったく同じと思ってよい。

3. 図13.10の分析結果にある“回転前の因子パターン”は回転前の主成分負荷量（固有ベクトル）、“回転後の因子パターン”は回転後の主成分負荷量のことと考えてよい。新しい主成分が得られたのであるから、それぞれがどのような主成分であるか解釈しなおす必要がある。軸の解釈の方法は、先の場合と同じである。回転前に比べて、主成分の意味や解釈のしやすさがより明確になっているかどうか確認してみよう。

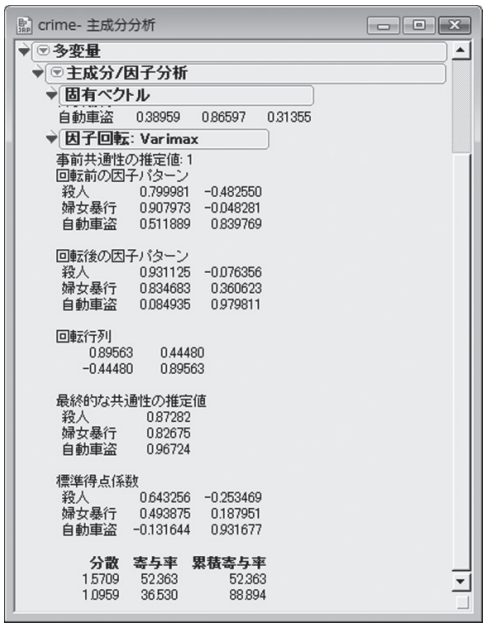


図13.10：軸の回転後の結果を示した画面

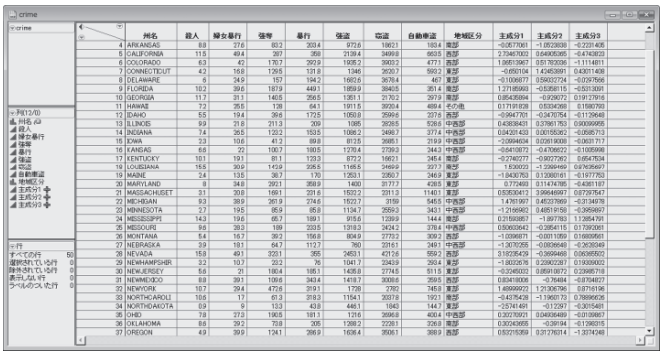
2.5　主成分得点とその保存方法

主成分得点（*principal component score*）とは、散布図の中の各々の点が、抽出された各主成分の軸上でとる値のことである。主成分得点は、それぞれの軸において、値が平均０、標準偏差１となるように標準化されている。本章2.1節の例でいえば、各人の「体格」や「体型」についての標準化された得点ということになる。主成分得点は、効率的に情報が集約された主成分の各軸の上で、データの個々の標本に対して与えられた得点であるから、有用な新しい変量として、それ自体が分析対象となる。また、保存した主成分得点を散布図に描いたり、回帰分析など他の手法でさらに分析するといったことも少なくない。

操作

Varimax 回転前の主成分得点を保存するには、主成分分析の操作後に“主成分 / 因子分析”の左にある赤い三角マークをクリックし、“主成分の保存”を選ぶ。そして、出てきたウィンドウに保存したい主成分の数を入力すると、図13.11に示すようにデータテーブルの最後の列に各主成分の主成分得点が保存される。保存された主成分得点には第 1 主成分の得点に“主成分 1”、第 2 主成分に“主成分 2”……などという変量名が自動的につけられる。

Varimax 回転後の主成分得点を保存するには、回転の操作後に“主成分 / 因子分析”の左にある赤い三角マークをクリックし、“回転後の成分を保存”を選択すればよい。保存された回転後の主成分得点には、第 1 主成分の得点に“因子 1”、第 2 主成分の得点に“因子 2”……などという変量名が自動的につけられる。



	州名	殺人	婦女暴行	強奪	暴行	強盗	窃盗	自動車盗	地域区分	主成分1	主成分2	主成分3
4	ARKANSAS	8.8	27.6	83.2	203.4	972.6	1862.1	1853.4	南部	−0.0577061	−1.0523838	−0.2231405
5	CALIFORNIA	11.5	49.4	287	358	2139.4	3499.8	663.5	西部	2.7346700	0.6409305	−0.4743822
6	COLORADO	6.3	42	170.7	292.9	1935.2	3903.2	477.1	西部	1.9951997	0.8178336	−0.1114811
7	CONNECTICUT	4.2	16.8	129.5	131.8	1346	2620.7	593.2	東部	−0.650104	1.4245899	0.4301408
8	DELAWARE	6	24.9	157	194.2	1682.6	2679.4	467	東部	−0.1006677	0.5902274	−0.0297566
9	FLORIDA	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4	南部	1.2718569	−0.6358115	−0.5213091
10	GEORGIA	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9	南部	0.8543594	−0.0226072	0.1917916
11	HAWAII	7.3	25.5	128	64.1	1911.5	3920.4	489.4	その他	0.1711938	0.5334288	0.1580793
12	IDAHO	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6	西部	−0.9947701	−0.2470754	−0.1139648
13	ILLINOIS	9.9	21.8	211.3	209	1085	2828.5	528.6	中西部	0.4838431	0.3796175	0.9009995
14	INDIANA	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4	中西部	0.0402143	0.0010502	−0.0585713
15	IOWA	2.3	10.6	41.2	89.8	812.5	2685.1	219.9	中西部	−2.0994634	0.0291908	−0.0631717
16	KANSAS	6.6	22	100.7	180.5	1270.4	2739.3	244.3	中西部	−0.5410072	−0.4706922	−0.1005998
17	KENTUCKY	10.1	19.1	81.1	123.3	872.2	1662.1	245.4	南部	−0.2746277	−0.9077262	0.6547634
18	LOUISIANA	15.5	30.9	142.9	335.5	1165.5	3469.9	337.7	南部	1.530023	−1.2994969	0.9763697
19	MAINE	2.4	13.5	38.7	170	1253.1	2351.7	246.9	東部	−1.8430753	0.1336016	−0.1977753
20	MARYLAND	8	34.8	292.1	358.9	1400	3177.7	428.5	東部	0.772469	0.1147485	−0.4381187
21	MASSACHUSET	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1	東部	0.8353041	2.3664697	0.8739754
22	MICHIGAN	9.3	38.9	261.9	274.6	1522.7	3159	545.5	中西部	1.4761997	0.4523769	−0.2134978
23	MINNESOTA	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1	中西部	−1.2166982	0.4851958	−0.3958697
24	MISSISSIPPI	14.3	19.6	65.7	189.1	915.6	1239.9	144.4	南部	0.2156936	−1.8977883	1.1265479
25	MISSOURI	9.6	28.3	189	233.5	1318.3	2424.2	378.4	中西部	0.9606642	−0.3854115	0.1739061
26	MONTANA	5.4	16.7	39.2	156.8	804.9	2773.2	309.2	西部	−1.0396871	−0.0011059	0.1693561
27	NEBRASKA	3.9	18.1	64.7	112.7	760	2316.1	249.1	中西部	−1.3070255	−0.0839648	−0.2630349
28	NEVADA	15.8	49.1	323.1	355	2453.1	4212.6	559.2	西部	3.1923549	−0.9994468	0.8638502
29	NEWHAMPSHIR	2.2	10.7	23.2	76	1041.7	2343.9	293.4	東部	−1.9033576	0.2390238?	0.1930902
30	NEWJERSEY	5.6	21	180.4	185.1	1435.8	2774.5	511.5	東部	−0.2245022	0.8591972	0.2398578
31	NEWMEXICO	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5	西部	0.8034106	−0.76484	−0.8704827
32	NEWYORK	10.7	29.4	472.6	319.1	1728	2782	745.8	東部	1.4899922	1.2130679?	0.8716196
33	NORTHCAROLI	13.9	17	61.3	318.3	1154.1	2037.8	192.1	南部	−0.4437428	−1.1960773	0.7889636
34	NORTHDAKOTA	0.9	9	13.3	43.8	446.1	1843	144.7	東部	−2.5741491	−0.12297	−0.3015481
35	OHIO	7.8	27.3	190.5	181.1	1216	2696.8	453.1	中西部	0.3027893	0.0400949	−0.0109867
36	OKLAHOMA	8.6	29.2	73.8	205	1288.2	2228.1	326.8	南部	0.3024365	−0.39194	−0.1286015
37	OREGON	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9	西部	0.5321539?	0.2176314	−1.3374248

図13.11：保存された主成分得点

練習問題

今回取り上げた例題では、“殺人”、“婦女暴行”、“自動車盗”の 3 つの変量を利用して主成分分析を行ったが、本来、主成分分析は多数の変量を使って分析することが普通である。ここでは、“crime”のデータテーブルにある州名・地域区分以外の 7 つの変量をすべて投入して、主成分分析と軸の回転を行い、結果の主成分得点を保存してみよう。

1. 犯罪データのデータテーブル“crime”をアクティブウィンドウにして、以後の操作で誤って破壊しないように“crime.JMP”という名前をつけて、“MyData”フォルダに保存しなさい。

2. “グラフ”メニューから“三次元散布図”を選択し、選択画面の左側に表示される“殺人”から“自動車盗”までの 7 つの変量をすべて選択して“Y, 列”ボタンをクリックし、7 つの変量による 3 次元プロットを出力してみよう。

3. 4 つ以上の変量を選択して 3 次元散布図を出力すると、選択した変量のうち、はじめの 3 つの

変量の３次元散布図が出力される。３次元散布図に別の変量を表示したいときは、図13.2の軸の記号をいくつかクリックして新たな変量を表示し、グラフの表示が変わることを確認しよう。画面下にある“殺人”、“婦女暴行”、“強奪”などのボタンをクリックしてみるとよい。別の変量を選びたい時は変量名と軸名をクリックすればよい。

4. ３次元散布図を少しずつ回転させて、データのばらつき具合を確認しよう。

5. “JMP スターター”の“多変量”カテゴリから“主成分分析”を選び、同じく７つの変量をすべて選択し、“OK”ボタンを押すと、７つの変量を用いた主成分分析の計算結果を表示される。

6. 本章の2.3節で説明した方法で、主成分の意味を解釈してみよう。第１主成分と第２主成分はそれぞれ何を意味する軸であると考えられるか。主成分の意味が解釈しにくいときは、主成分負荷量の絶対値が比較的大きい値を取るいくつかの変量に特に注目して考えるとよい。

7. ここで、寄与率が大きい２つの主成分に注目して、Varimax 回転による軸の回転を行い、6. と同様に結果を解釈してみよう。Varimax 回転を行うには、“主成分 / 因子分析”の左にある赤い三角マークをクリックし、“主成分の回転”オプションを選べばよい。

8. Varimax 回転により単純構造に近づけることによって、軸の解釈は簡単になっただろうか。考察しなさい。

9. Varimax 回転後の主成分得点を新しいデータテーブルに出力してみよう。

10. 出力された主成分得点である“回転成分１”と“回転成分２”の分布をそれぞれヒストグラムに出力して、観察しよう。また、２つの成分の散布図を作成してみよう。

（ア）出力された平均と標準偏差の値から、主成分得点が標準化されていることを確認しなさい。

（イ）グラフに現わされた主成分得点はどのように分布しているか。8. で行った軸の解釈とあわせて結果を考察しなさい。

（ウ）２つの回転後の主成分得点の散布図に州名のラベルを表示して、どの州の犯罪の傾向が似ているか、結果を考察し、この分析で得た他の情報とあわせて結論を述べなさい。

主成分得点どうしのグラフは、先の操作10. のようにいちいち主成分得点を外部のファイルに出力してからグラフを描くよりも簡単に出力することができる。外部のファイルに出力するのは、計算された主成分得点を使って、データの分類などのさまざまな深い分析ができるからである。しかし、主成分得点どうしのグラフを描くだけなら、簡単に行うことができる。ここでは、その方法を簡単に紹介しよう。

操作

1. “カテゴリをクリック”の中から“多変量”を選択し、“主成分分析”をクリックする。

2. “殺人”から“自動車盗”までを Y 列に選択し、“OK”ボタンを押す。すると、各主成分ごとの寄

与率と累積寄与率がグラフによって表される。

3. “主成分 / 因子分析”の横の赤い三角ボタンをクリックし、“主成分の回転”をクリックする。

4. もう一度“主成分 / 因子分析”の横の赤い三角ボタンをクリックし、“スコアプロット”をクリックする。すると図13.12のような図が出力される。

単に主成分得点間の関係を見て、結果を考察したいだけならば、このやり方のほうが簡単に主成分得点どうしの関係を見ることができる。

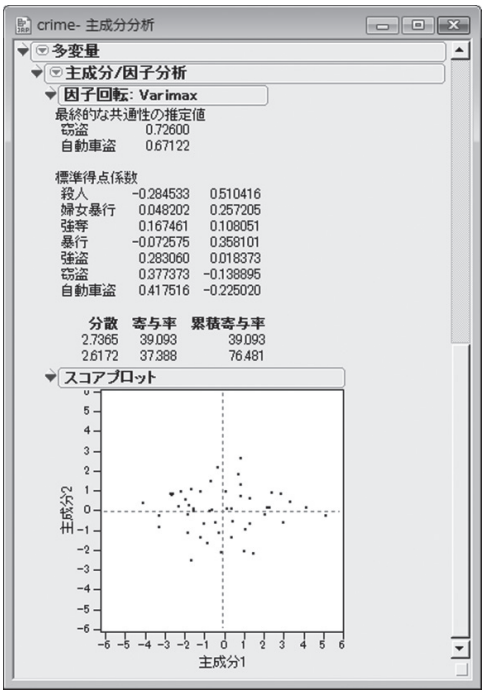


図13.12：スコアプロットの画面

軸の回転を行った後の主成分についても、図13.7のように元の変量の軸と主成分の関係を探るグラフを描きたい場合、次のように操作することによって、グラフを描くことができる。

操作

1. “カテゴリをクリック”の中から“グラフ”を選択し，“三次元散布図”をクリックする。

2. “殺人”から“自動車盗”までを“Y，列”に選択し、“OK”ボタンを押す。すると、各データ列を

軸とした三次元散布図が描かれる。

3. “三次元散布図”の横の赤い三角ボタンをクリックし、“主成分分析”をクリックする。すると、各データ列を軸とした三次元散布図から、各主成分を軸とした三次元散布図へと描きなおされる。

4. 次に、もう一度“三次元散布図”の横の赤い三角ボタンをクリックし、“成分の回転”をクリックすると、図13.9と同じ“主成分数の選択”の画面が出るので、今回は回転させたい主成分数を３とする。“OK”ボタンを押すと，図13.13のように回転後の各主成分を軸とした図が出力される。

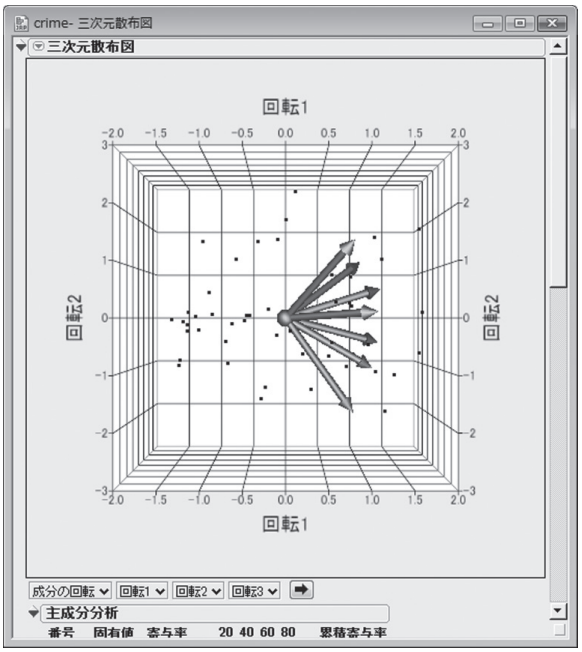


図13.13：回転後の成分の三次元散布図

このとき、グラフ内に描かれている矢印は図13.7と同様に、もとのデータの軸を表すものであるので、回転後の主成分の解釈に役立てることができる。なお、4. の手順はグラフを作るためのものであるので、各主成分の固有ベクトルについては表示させることができないので注意してほしい。