【セミナー】

主成分分析

― 社会調査データの多変量解析（2）―

小林 久高
KOBAYASHI Hisataka

1　はじめに

今回は主成分分析について解説する。主成分分析は尺度の構成や変数の分類に役に立つ分析法だ。例えば収入、貯蓄額、保有するその他の資産などから「経済的豊かさ」という尺度を構成したり、さまざまな音楽アーティストへの好き嫌いといった情報から、音楽アーティストを分類したりするのに使える。そのように使い勝手はいいものの実際にその原理について理解するのはなかなか困難だ。そこで今回は図表を大いに用いてその原理を解説することにする。

まず2節で必要となる記述統計学の基礎について解説する。知っていることが多いと思うが、「変数の変換」や「分布のイメージ」についてはここでしっかり把握しておいてほしい。

3節では変数をベクトルとして考える際の基礎的事項を解説する。主成分分析をきちんと理解するためにはこの部分の知識が欠かせない。

4節では最も基本となる2変数の主成分分析について解説する。この解説が実際の主成分分析を知るための基礎的な部分になる。

最後に5節で3変数の主成分分析について解説する。ここまで読めば次元の削減や回転についても理解できるようになるし、より変数の多い分析についてもイメージできるようになる。ではさっそく始めよう。

2　基礎的知識

2.1　変数・値・分布

異なる値をとりうるものを変数という。物の重さや体積は異なる値をとりうるので変数である。身長や体重も異なる値をとりうるので変数である。収入や教育年数や勤続年数なども異なる値をとりうるので変数である。性別も男女という値をとるので変数だ。職業も専門職、ホワイトカラーなどの値をとる変数である。

「変数の分布」は、変数の値全体について、その頻度や割合がどうなっているのかを示す用語だ。たとえば、「このオーケストラには男性が10人、女性が20人いる」という表現や、「体重が50kg未満の人が10人、50kg以上の人が20人」や「収入500万円未満の世帯が50%、500万円以上の世帯が40%」という表現は変数の分布についての表現だ。

変数には量的な変数と質的な変数（カテゴリカルな変数）がある。量的な変数とは、年齢や収入といった変数のことであり、質的な変数とは職業や学歴といった変数である。量的な変数と質的な変数は、平均を出して意味があるかどうかということで区別できる。量的変数である年齢や収入の平均には意味がある。しかし、質的変数である職業に、専門職なら1、ホワイトカラーなら2などと値を与えて平均を出しても意味はない。

量的変数を質的変数にしたり質的変数を量的変数にしたりできる場合がある。たとえば年齢を年少、生産年齢、老齢に区分することは量的変数から質的変数（カテゴリカルな変数）への変換である。質的変数としての学歴も、教育年数からとらえれば量的変数になる。人びとの行政への満足度を調べるとき、「満足、やや満足、やや不満、不満」という形で答えを聞くことがある。この行政への満足度という変数は厳密には質的な変数と考えられるが、それぞれに 4 点から 1 点までを与えて量的変数として分析することもできる。

量的変数の分布について、平均や分散で分布の全体的なありようを示すことがよくある。平均は分布の中心的な傾向を示すものであり、分散は分布のちらばりの傾向を示すものだ。

平均や分散は 1 つの変数の分布の指標である。2 つの変数について、変数と変数との関係の強さを検討することがよくある。たとえば「収入の高い人の方が生活に満足している」ということの検討では、収入という量的変数と生活満足度という量的変数の関係が検討されているのである。2 つの量的変数の関係の強さの指標としてよく利用されるのは相関係数である。

2.2　平均

表 1 は最も基本的な形式で表されたデータであり、各々のケースごとの得点が示されている。このような得点の分布の中心的な傾向は平均 $\bar{x}$ によって表されることが多い。平均は次式で表される。i はケース番号、n はケース数である。

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{n}\left(x_1+x_2+\cdots+x_n\right)$$

表 1　各ケースの得点の表

ケース(i)	得点(x_i)	
1	80	(x_1)
2	70	(x_2)
3	40	(x_3)
4	40	(x_4)
5	20	(x_5)
n	5	
平均	50.0	
分散	480	
標準偏差	21.91	
偏差平方和（変動）	2400	

平均については、これまでもいろいろなところで接してきただろうから説明は不要だろう。表のデータでは平均は下のようになる。

$$\bar{x}=\frac{1}{5}\sum_{i=1}^{5}x_i$$
$$=\frac{1}{5}\left(80+70+40+40+20\right)=50$$

2.3　変動・分散・標準偏差

（1）　変動（偏差平方和）

データの散らばりを示す 1 つの指標は、変動（偏差平方和ともいう）である。変動は次の式で表せる。変動は大文字の S で表すことが多い。x の変動であることを示す場合には S_x または S_{xx} と表記する。

$$S_x(=S_{xx})=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

表 1 のデータの分散は、上で求めた平均＝40 を利用して次のように計算できる。

$$S_x = \sum_{i=1}^{5}(x_i - \bar{x})^2$$
$$= \left\{ \begin{array}{l} (80-50)^2 + (70-50)^2 \\ +(40-50)^2 + (40-50)^2 \\ +(20-50)^2 \end{array} \right\} = 2400$$

変動はケース数の異なるものでの散らばり具合の比較には使えない。というのは同じ散らばり具合の分布でも、ケース数が多くなるとこの値はどんどん大きくなるからである。そういう場合には次の分散が分布の散らばり具合の指標として利用される。

（2） 分散

分散は小文字の s^2 で表されることが多い。x の分散であることを示す場合には s^2_x と表記する。分散は次の式で求められる。n はケース数、i はケース番号であり、$\bar{x}$ は平均だ。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \frac{1}{n}\left\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right\}$$

表 1 のデータの分散は、上で求めた平均＝40 を利用して次のように計算できる。

$$s^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2$$
$$= \frac{1}{5}\left\{ \begin{array}{l} (80-50)^2 + (70-50)^2 \\ +(40-50)^2 + (40-50)^2 \\ +(20-50)^2 \end{array} \right\} = \frac{2400}{5} = 480$$

分散に関連した概念として不偏分散というものがある。不偏分散の定義式は次のものだ。

$$u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

不偏分散は集められたデータを標本として考えるとき、母集団の分散がどうなるかを推定する量である。データ自体の分散を出したければ普通の分散を使い、データから母集団の分散を推定したければ不偏分散を使うといい。

（3） 標準偏差

分散の式は、「各得点と平均のズレの 2 乗の総和」をケース数で割ったものである。したがって、もとの得点が仮に重さ g だとするなら分散の単位は g^2 となる。元の得点が cm で測られた身長なら身長の分散は cm^2 となるのである。平均のほうは元の得点と同じ単位なのだが、分散はその 2 乗を単位とするのである。これではもとの得点の分布との兼ね合いが悪い。たとえば元の分布をグラフで表すとき、平均はそこにすぐさま位置づけられるのに対し、分散は単位が違うので位置づけにくいということになってしまう。そこで、分散の正の平方根をとった標準偏差 s を分布の散らばりの指標とすることがある。

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

表 1 のデータでは標準偏差は次のようになる。

$$s = \sqrt{480} = 21.91$$

2.4 相関係数

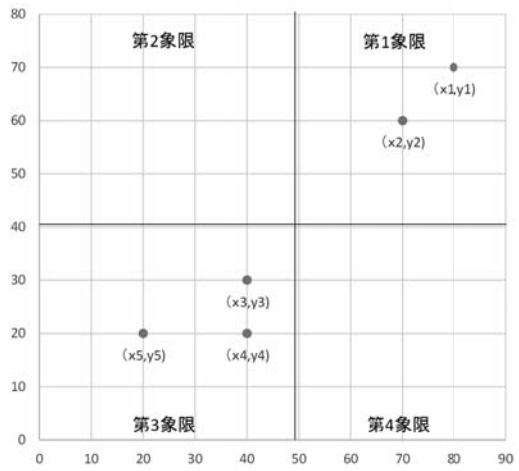
次に 2 つの変数の関係である。表 2 は 2 変数 x と y の得点を示している。x は国語の得点、y は

英語の得点とでも考えればいい。表を見るとおおむねxの得点が高いとyの得点も高いことがわかる。

表 2　変数 x と変数 y の得点

ケース(i)	x	y
1	80	70
2	70	60
3	40	30
4	40	20
5	20	20
n	5	5
平均	50.00	40.00
分散	480.00	440.00
標準偏差	21.91	20.98
偏差平方和（変動）	2400	2200

図 1　変数 x と変数 y の得点



この関連を示したものが図 1 である。中心にある太めの垂直軸と水平軸はそれぞれxの平均とyの平均を表している。

こういった 2 変数間の関係の強さを相関係数 r で表すことがある。相関係数（ピアスンの積率相関係数）は次の式で表される。n はケース数、$\bar{x}$ は x の平均、$\bar{y}$ は y の平均[1]。

$$r=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$=\frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}}=\frac{s_{xy}}{s_x s_y}$$

相関係数の式の分母は x の標準偏差 s_x と y の標準偏差 s_y を掛け合わせたもの、分子は x と y の共分散 s_{xy}（s は小文字）と呼ばれるものだ。だから、相関係数は共分散を 2 つの標準偏差で割ったものということになる。相関係数が x と y についてのものであることを示したいときには r_{xy} と書くのが一般的だ。

相関係数は共変動（偏差積和）と変動（偏差平方和）を使って次のようにも示される（大文字の S であることに注意）。

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{S_{xy}}{\sqrt{S_x S_y}}$$

$$\text{変動（偏差平方和）}: S_x(=S_{xx})=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\text{共変動（偏差積和）}: S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

相関係数の値は、x が大きくなると y も大きくなるという関係があるとき＋1 に近づき、x が大

[1] 相関係数の式に単位を入れて計算するとわかるが、相関係数自体は単位のない数（無名数）となる。

きくなるとyが小さくなるという関係があるとき−1 に近づく。そういった関係があまりないとき、0 に近い値になる。＋1 や−1 に近い値になるとき「強い相関がある」と表現する。

上の例で相関係数を計算すると次のようになる。まず、相関係数の式の分子を計算すると、

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{5}\left\{\begin{array}{l}(80-50)(70-40)\\ +(70-50)(60-40)\\ +(40-50)(30-40)\\ +(40-50)(20-40)\\ +(20-50)(20-40)\end{array}\right\} = \frac{2200}{5} = 440$$

これと表 2 にある標準偏差を使って相関係数を計算すると

$$r = \frac{s_{xy}}{s_x s_y} = \frac{440}{21.909 \times 20.976} = 0.95743\cdots$$

となり、x と y の間には 0.96 程度の正の相関があることがわかる。

2.5 変数と得点の変換

（1） 素点と平均偏差得点

表 3 はある仮想的な調査で得られたデータである。得られたデータは素点の列に並べられている。社会調査の分析ではこのデータを平均偏差得点や標準得点に変換して利用することが多い。

平均偏差得点とは素点から平均を引いた得点だ。たとえば 1 氏の場合、素点の 28 から素点の平均である 39.6 を引いた−11.6 が 1 氏の平均偏差得点になる。すべてのケースの素点を平均偏差得点に変えたとき、平均偏差得点の平均は 0 になるが、分散や標準偏差は変わらない。

表 3 素点・平均偏差得点・標準得点

	年齢		
	素点	平均偏差得点	標準得点
1氏	28	-11.6	-1.36
2氏	35	-4.6	-0.54
3氏	36	-3.6	-0.42
4氏	49	9.4	1.10
5氏	50	10.4	1.22
n	5	5	5
平均	39.6	0.0	0.00
分散	73.0	73.0	1.00
標準偏差	8.5	8.5	1.00
偏差平方和	365.2	365.2	5.00

（2） 標準得点

データ全体の標準偏差が 1(すなわち分散が 1)になるようにさらに値を変換したものが標準得点だ。標準得点は平均偏差得点を標準偏差で割ることによって得られる。1 氏の標準得点は、−11.6 ÷8.5＝−1.36 となる。ここでは平均が 0、分散は 1、標準偏差は 1 となっている [2]。

素点から考えると、標準得点は平均を引き標準偏差で割るという作業で得られる。もとの i さんの得点を x_i で表すと、標準得点 x'_i は次のようになる。

$$x'_i = \frac{x_i - \bar{x}}{s_x} = \frac{x_i - \bar{x}}{\sqrt{s_x^2}}$$

ある変数の個々のケースの値をすべて標準得

[2] 標準偏差は分布全体の性質を示しているのに対し、標準得点は個々のケースの値を示すことに注意すること。

点に変換するということは、その変数自体を標準化することでもある。もとの変数を x で表すと、標準化された変数 x'は次のようになる。

$$x' = \frac{x-\bar{x}}{s_x} = \frac{x-\bar{x}}{\sqrt{s_x^2}}$$

変数を標準得点化すると、新変数の平均は 0、標準偏差は 1、分散は 1 になる。

（3） 偏差値

いわゆる偏差値は、平均を 50、標準偏差を 10 に置き換えて標準化した得点であり、標準得点が −1 のとき偏差値は 40、標準得点が 0 のとき偏差値は 50、標準得点が 1 のとき偏差値は 60、標準得点が 2 のとき偏差値は 70 となる。偏差値によって各人の試験での相対的な位置が把握しやすくなるとともに、さまざまな時に行われるさまざまな科目の試験成績の比較も容易になるということは、学生諸君もよく知っていることだろう。偏差値と標準得点の間には次の式が成り立つ。

$$\text{標準得点} = \frac{\text{偏差値}-50}{10}$$

（4） 変換にともなう平均と分散の変化

素点の平均が $\bar{x}$、分散が s^2 であるとき、平均偏差得点、標準得点、偏差値の平均と分散は表 4 のようになる。

表 4　変換と平均・分散

	平均	分散	標準偏差
素点x	$\bar{x}$	s^2	s
xを平均偏差得点化	0	s^2	s
xを標準得点化	0	1	1
xを偏差値化	50	100	10

（5） 変数の標準化と相関係数

変数 x が平均 $\bar{x}$,分散 s_x^2 の分布を持ち、変数 y が平均 $\bar{y}$,分散 s_y^2 の分布を持っているとし、x を標準得点化して x'を作るとすると、x'と y の相関は、次の x と y の相関と同じになる。

$$r_{x'y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i'-\bar{x}'\right)\left(y_i-\bar{y}\right)}{s_{x'}s_y}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}-\bar{x}'\right)\left(y_i-\bar{y}\right)}{1\cdot s_y}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}-0\right)\left(y_i-\bar{y}\right)}{s_y}$$

$$= \frac{\frac{1}{n}\frac{1}{s_x}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{s_y}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{s_x s_y} = r_{xy}$$

同じことは、両方の変数を標準化した場合にもいえる。

$$r_{x'y'} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i'-\bar{x}'\right)\left(y_i'-\bar{y}'\right)}{s_{x'}s_{y'}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}-\bar{x}'\right)\left(\frac{y_i-\bar{y}'}{s_x}-\bar{y}'\right)}{1\cdot 1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}-0\right)\left(\frac{y_i-\bar{y}'}{s_x}-0\right)$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{s_x s_y} = r_{xy}$$

素点の国語と数学の相関も、標準得点の国語と数学の相関も同じ値なのである。

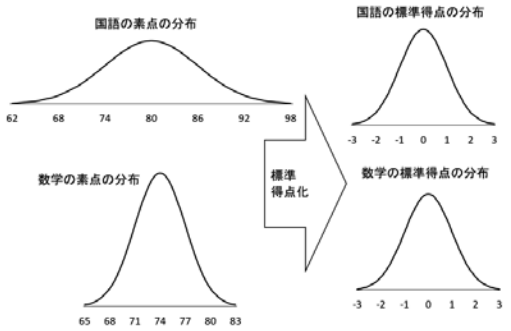
2.6　標準得点化された変数の分布

（1）　標準得点の分布のイメージ

さて、標準得点に注目しよう。今、大量の受験生 1000 人の国語の得点 x と数学の得点 y があったとする。得点は両者ともに平均付近が最も多く、離れるにしたがって少なくなっているとしよう。すなわち、ベル型の分布である。ここで、国語の平均は 80 点で標準偏差は 6 点、数学の平均は 74 点で標準偏差が 3 点とする。国語の分散は 36、数学の分散は9、国語の変動（偏差平方和）は36000、数学の変動は 9000 となっているはずだ。国語と数学の分布は図 2 の左の 2 つで表される。

ここで国語、数学ともに標準得点化すると、分布は図 2 の右のようになる。それらはともに平均 0、標準偏差 1、分散 1、変動 1000 のベル型の分布である。

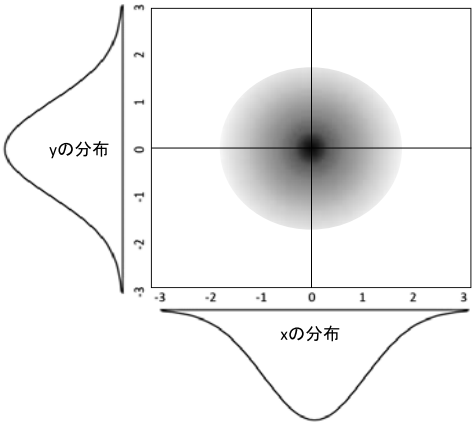
図 2　素点の分布と標準得点化された分布



（2）　標準得点化された 2 変数の分布イメージ

次に、2 変数の関係について。国語 x と数学 y の得点が標準化されており、両者の相関が 0 である場合、x と y の関係は図 3 中央部のような散布図で表される [3]。

図 3　国語と数学の同時分布（相関がない場合）

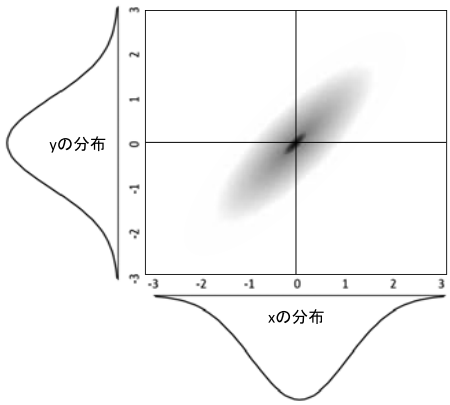


国語数学ともに平均点のケースは (0,0) である図の中心に位置づけられる。この中心に近い部分に多くのケースがあり、その数は周辺に向けて、きれいな円の形でだんだんまばらになっていく。

散布図の左に示されているのは数学 y の分布であり、下に示されているのは国語 x の分布である。これらはそれぞれ y 軸で見たケースの分布と、x 軸で見たケースの分布を示している。標準化されているのでそれらの分布はどちらも平均 0、分散 1 であり、変動はともに 1000 だ。

[3] 中心の図には点が 1000 あるわけではなく濃淡で表されている。したがって、正確には散布図とはいえないがご容赦願いたい。

図 4　国語と数学の同時分布（相関がある場合）



もし、国語と数学の得点の間に相関があるなら、2 変数の関係は図 4 のようなものになる。ここではデータは楕円の形で分布している。ケースは楕円の中心部分では密集し、周辺になるにしたがってまばらになる[4]。

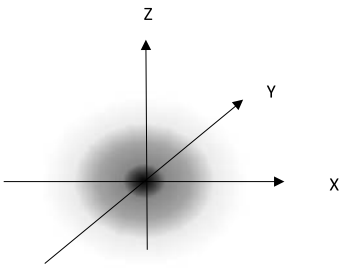
注意してほしいのは、ここで図の左や上にある y や x の分布である。これらは各ケースの点について、x 軸に落としたものの分布と y 軸に落としたもののの分布、つまりデータを横から眺めた場合と下から眺めた場合の分布を意味している。これらの分布は前の図 3 で見た相関がない場合と同じである。それらはともに平均 0、分散＝1（標準偏差 1、変動＝1000）の分布なのである。ここからわかるように、両変数が標準化されていることと、両変数の間に相関があることは別の話なのである。

（3）　変数が 3 つある場合

次に、xy と同様にベル型に分布する英語の得点 z を加え、標準得点化された変数が 3 つある場合について考えよう。このとき、それぞれのケースを表す点は xy 平面ではなく、xyz で構成される空間に位置付けられる。

このとき、もし x と y、y と z、z と x の相関がいずれも 0 であったとするなら、データは(0,0,0)の中心から球の形で広がっていくことになる。すなわちケースを表す点は中心部分で密集しており、周辺部分になるにつれて、どの方向についても等しくまばらになっていくのである（図 5）。

図 5　標準化 3 変数の分布（相関がない場合）

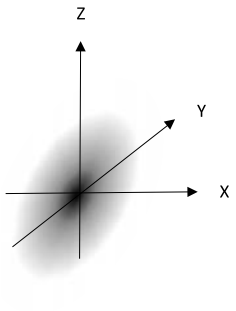


3 つの変数間に相関がある場合はデータは楕円体のような形で分布する。楕円体とはラグビーボールのような形のものだ。この楕円体が xyz 空間に斜めになって浮かんでいるような形でデータが分布しているのである。各ケースを表す点は、ケースの多い楕円体の中心には密集しており、周辺部分ではまばらになっていく（図 6）。

この場合も、2 変数の場合と同様 x,y,z はそれぞれ標準化されているので、x,y,z のどれも平均 0、標準偏差 1（分散 1、変動 1000）である。それらはケースの分布についての、横軸上の分布、縦軸上の分布、高さ軸上のデータの分布に対応したものだ。

[4] 図 4 は、2.4 の相関係数の解説で見た図 1 のような図にデータがたくさんあるものと考えればいい。

図 6　標準化 3 変数の分布（相関がある場合）



2.7　問題と解答

（1）　問題

(a)

・平均、分散、共分散、相関係数、偏差平方和（変動）、偏差積和（共変動）の式をシグマを使って書け。

(b)

・相関係数を分散と共分散を用いて表せ。また、相関係数を変動と共変動を用いて表せ。

(c)

・変数 x1, y1 について、それらを平均偏差得点にしたものをそれぞれ x2, y2 とし、標準得点にしたものをそれぞれ x3, y3 とする。x1 と y1 の相関が 0.8 のとき、x1, x2, x3, y1, y2, y3 の相関を示した相関行列を書け（相関行列とは変数を縦横に並べ、行列の形で相関係数のを示したもの）。

（2）　解答

(a)

平均: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

分散: $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

共分散: $s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$

相関係数: $r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$

偏差平方和（変動）: $S_x = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

偏差積和（共変動）: $S_{xy} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$

(b)

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

（大文字と小文字の違いに注意すること。）

(c)

	x1	x2	x3	y1	y2	y3
x1	1	1	1	0.8	0.8	0.8
x2	1	1	1	0.8	0.8	0.8
x3	1	1	1	0.8	0.8	0.8
y1	0.8	0.8	0.8	1	1	1
y2	0.8	0.8	0.8	1	1	1
y3	0.8	0.8	0.8	1	1	1

3　変数ベクトル

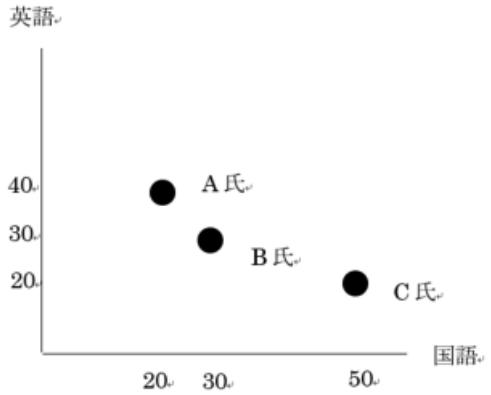
3.1　データの置かれる 2 つの空間

表 5 のようなデータがあったとしよう。こういったデータがある場合、図 7 のように、国語と英語を軸とした座標に A 氏、B 氏、C 氏を位置付けるのが一般的だ。

表 5　データ

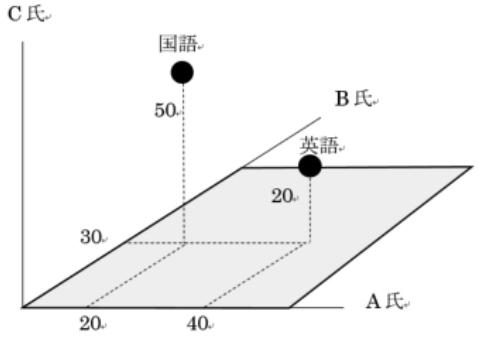
	国語	英語
A氏	20	40
B氏	30	30
C氏	50	20

図 7　ケースがプロットされる空間



しかし、図 8 のように A 氏、B 氏、C 氏を軸とした座標に国語と英語を位置付けることもできる。この節の話はケースが位置付けられる空間の話ではなく、変数が位置付けられる空間の話であることをまず押さえておいてほしい。

図 8　変数がプロットされる空間



3.2　変数ベクトルと平方和・相関係数

（1）　変数ベクトル

主成分分析では回帰分析と同様、ベクトルを用いて図形的にとらえることでイメージがわきやすくなる。ここでは、変数や主成分をベクトルとして考える際の基礎事項について解説する。

表 6　素点と標準得点

	素点		標準得点	
	年齢	収入	年齢	収入
1氏	28	550	-1.36	-1.21
2氏	35	560	-0.54	-1.03
3氏	36	650	-0.42	0.57
4氏	49	630	1.10	0.21
5氏	50	700	1.22	1.46
n	5	5	5	5
平均	39.6	618.0	0.00	0.00
分散	73.0	3176.0	1.00	1.00
標準偏差	8.5	56.4	1.00	1.00
偏差平方和	365.2	15880.0	5.00	5.00
相関係数	0.79		0.79	
偏差積和	1906		3.96	
共分散	381.2		0.79	

表 6 には、1 氏～5 氏の 5 人の年齢と収入についての素点と標準得点、ならびに基本的な統計量が示されている。

この標準得点をベクトルとして表現できることに注目しよう。年齢も収入も、それぞれ 5 次元（ケース数次元）のベクトルとして表現できる。たとえば年齢ベクトルと収入ベクトルは次のようになる（ベクトルはこのように太字で表現される）。

$$\mathbf{x}=\begin{bmatrix}-1.36\\-0.54\\-0.42\\1.10\\1.22\end{bmatrix}\qquad \mathbf{y}=\begin{bmatrix}-1.21\\-1.03\\0.57\\0.21\\1.46\end{bmatrix}$$

これらの変数を表すベクトルを（標準得点の）変数ベクトルと呼ぶことにする。変数ベクトルは n 次元（ケース数次元）のベクトルである。

（2） 変数ベクトルと変動のルート

標準得点の変数ベクトルの長さ（大きさ）|**x**|は、その変数のデータの散らばり具合を示すものだ。それはその変数の変動（偏差平方和）の正の平方根を示している。標準得点化されている変数ではどの変数も同じく標準偏差が 1 なので、どの変数も長さはケース数の正の平方根$\sqrt{n}$ になる。

$$|\mathbf{x}|=\sqrt{S_x}=\sqrt{n}$$

その理由は次の通り。ベクトルの長さは、

$$|\mathbf{x}|=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}$$

標準得点のときには平均が 0 なので、偏差平方和の平方根は、

$$\sqrt{S_x}=\sqrt{\sum\left(x_i-\bar{x}\right)^2}=\sqrt{\sum\left(x_i-0\right)^2}$$
$$=\sqrt{\sum x_i^{\,2}}=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}=|\mathbf{x}|$$

また、標準得点の場合、標準偏差は 1 なので、

$$s=\sqrt{\frac{1}{n}\left(x_1^2+x_2^2+\cdots+x_n^2\right)}$$
$$=\sqrt{\frac{1}{n}}\sqrt{\left(x_1^2+x_2^2+\cdots+x_n^2\right)}$$
$$=\sqrt{\frac{1}{n}}|\mathbf{x}|=1$$

が成り立ち、結局

$$|\mathbf{x}|=\sqrt{n}$$

となるのである。

3.3 2 つの変数ベクトル

（1） 変数ベクトルと相関係数

変数ベクトルは相関係数にも関係している。2 つの標準得点化された変数ベクトルの余弦は、実は相関係数に他ならない。どうしてそんなことになるのかを説明しよう。

相関係数は次のようなものだった。

$$r_{xy}=\frac{S_{xy}}{\sqrt{S_xS_y}}$$

この分子である偏差積和は偏差得点を用いる場合は次のようになる。

$$S_{xy}=\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$$
$$=\sum\left(x_i-0\right)\left(y_i-0\right)=\sum x_iy_i$$

これは、ベクトルの内積の定義と同じである。

$$(\mathbf{x},\mathbf{y})=\sum x_iy_i$$

ここで、相関をベクトルの内積と大きさで表現すると次のようになる。

$$r_{xy}=\frac{S_{xy}}{\sqrt{S_xS_y}}=\frac{\sum x_iy_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}=\frac{(\mathbf{x},\mathbf{y})}{|\mathbf{x}||\mathbf{y}|}$$

この右辺は、高校で習ったベクトル間の角度 θ に関する次の公式の右辺と同一である。

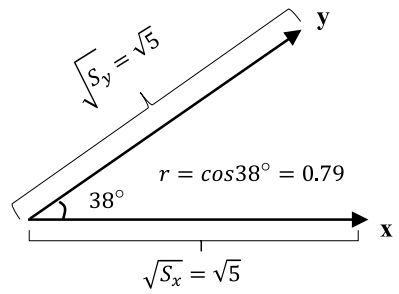
$$\cos\theta=\frac{(\mathbf{x},\mathbf{y})}{|\mathbf{x}||\mathbf{y}|}$$

したがって、2 つの変数ベクトルが作る角度の余弦は相関係数に等しいといえる。

以上のことを図形的に考えてみよう。2 次元ベクトルや 3 次元ベクトルなら図に描けるが、それを超える次元のベクトルは図に描けないと思うかもしれない。しかし、何次元ベクトルでもベクトルが 2 本までだと、それらの関係は平面に描けるし、3 本までだと空間に描ける [5]。

このことを念頭に置き、先の年齢ベクトルと収入ベクトルを図示してみよう。表 6 から、標準化された年齢と収入の変動（偏差平方和）S はともに 5 であること、両者の相関 r は 0.79 であることがわかる。これらをもとに 2 つの変数ベクトルの関係を描くと図 9 のようになる。原点を起点に 2 つの変数ベクトルが伸びていくというイメージをもってほしい。どのような標準得点をもとにした 2 つの変数ベクトルも、このような形で図示できるのである。

図 9　標準化された変数ベクトルと相関係数



相関にかかわるさまざまな問題を感覚的に理解するためには「相関＝ベクトル間の角度の余弦」というとらえ方はとても有効だ。2 つのベクトルが作る角度と相関の値との関係を表にまとめておく（表 7）[6]。ベクトル間の角度 θ は 0° ～180° の範囲であり、負の角度はここでは考えない。

表 7　変数の相関と変数ベクトル間の角度

相関	角度	相関	角度	相関	角度
1.00	0.0	0.30	72.5	-0.40	113.6
0.95	18.2	0.25	75.5	-0.45	116.7
0.90	25.8	0.20	78.5	-0.50	120.0
0.85	31.8	0.15	81.4	-0.55	123.4
0.80	36.9	0.10	84.3	-0.60	126.9
0.75	41.4	0.05	87.1	-0.65	130.5
0.70	45.6	0.00	90.0	-0.70	134.4
0.65	49.5	-0.05	92.9	-0.75	138.6
0.60	53.1	-0.10	95.7	-0.80	143.1
0.55	56.6	-0.15	98.6	-0.85	148.2
0.50	60.0	-0.20	101.5	-0.90	154.2
0.45	63.3	-0.25	104.5	-0.95	161.8
0.40	66.4	-0.30	107.5	-1.00	180.0
0.35	69.5	-0.35	110.5		

[5] 3 次元空間の原点から別々の方向に延びる 2 本の線に 1 枚のガラスをぺたりと貼るとき、そのガラス面は 2 次元になっている。3 次元空間の 2 本の線は 2 次元で表現できるということだ。3 本の線があるときにはこんなことはできない。線の本数が問題なのである。

[6] 相関から角度を求めるには cos の逆関数 arccos を用いる。ラジアンではなく度で計算するためには、エクセルでは DEGREES(ACOS(相関))とすればいい。

表を見るとわかるように、相関が 1 のとき 2 つの変数ベクトルは同じ方向を向いており、相関が −1 のとき逆方向を向いている。相関が 0 であるとき、2 つの変数ベクトルは直交する。

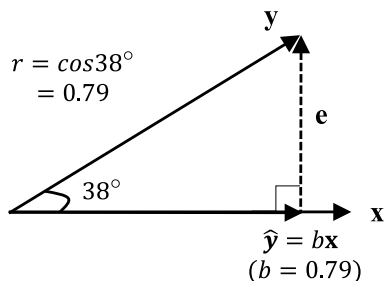
（2） 回帰モデル

ここで表 6 のデータを用いて年齢から収入を予測するということを考えよう [7]。年齢と所得を標準得点に変換した場合、回帰の予測式は次のようになる。

$$\hat{y}_i = 0.79 x_i$$

これをベクトルを示す図で表したものが図 10 である。

図 10　回帰分析の図形表現



x は年齢ベクトル、**y** は実際の収入のベクトル、$\hat{\mathbf{y}}$ は収入の予測値のベクトルである。$\hat{\mathbf{y}}$ は原点を起点とし、**y** の先端から **x** ベクトル方向に垂線を下ろした足を終点とするベクトルだ。回帰式をベクトルの式とその要素からなる式で表すと次のようになる。

$$\hat{\mathbf{y}} = 0.79\mathbf{x}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \end{bmatrix} = 0.79 \begin{bmatrix} -1.36 \\ -0.54 \\ -0.42 \\ 1.10 \\ 1.22 \end{bmatrix}$$

回帰式をベクトルで表すと次のようになる。

$$\mathbf{y} = b\mathbf{x} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$$

$$\begin{bmatrix} -1.21 \\ -1.03 \\ 0.57 \\ 0.21 \\ 1.46 \end{bmatrix} = 0.79 \begin{bmatrix} -1.36 \\ -0.54 \\ -0.42 \\ 1.10 \\ 1.22 \end{bmatrix} + \begin{bmatrix} -0.13 \\ -0.60 \\ 0.90 \\ -0.66 \\ 0.49 \end{bmatrix}$$

$$= \begin{bmatrix} -1.07 \\ -0.43 \\ -0.33 \\ 0.87 \\ 0.96 \end{bmatrix} + \begin{bmatrix} -0.13 \\ -0.60 \\ 0.90 \\ -0.66 \\ 0.49 \end{bmatrix}$$

（3） y の変動の分解

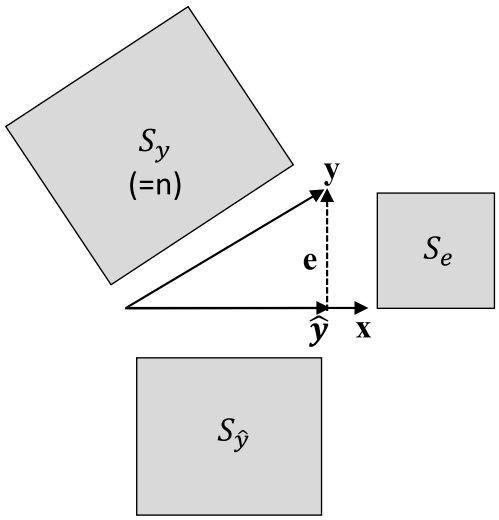
さて、**y**, $\hat{\mathbf{y}}$, **e** については、次の図のような関係が成り立っていることがわかる（図 11）。ここで、変数 y の偏差平方和（変動）S_y（=n）は、変数ベクトル **y** の長さの 2 乗、$|\mathbf{y}|^2$ になる。それは面積 S_y として把握できる。

同様に、S_e や S_R も面積として把握できる。ここで、ピタゴラスの定理より、$S_y = S_{\hat{y}} + S_e$ が成立している。y の変動 S_y は、変動 $S_{\hat{y}}$ と変動 S_e にきれいに分解されるのである。

[7] 回帰分析については、小林・山本（2020）を参照されたい。

$$S_y = S_R + S_e$$

図 11　y の変動の分解



このことをベクトルの要素からも確認しておこう。

$$\mathbf{y} = \begin{bmatrix} -1.21 \\ -1.03 \\ 0.57 \\ 0.21 \\ 1.46 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} -1.07 \\ -0.43 \\ -0.33 \\ 0.87 \\ 0.96 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} -0.13 \\ -0.60 \\ 0.90 \\ -0.66 \\ 0.49 \end{bmatrix}$$

$$S_y = |\mathbf{y}|^2 = \left\{ \begin{matrix} (-1.21)^2 + (-1.03)^2 \\ +0.57^2 + 0.21^2 \\ +1.46^2 \end{matrix} \right\} = 5$$

$$S_{\hat{y}} = |\hat{\mathbf{y}}|^2 = \left\{ \begin{matrix} (-1.07)^2 + (-0.43)^2 \\ +(-0.33)^2 + 0.87^2 \\ +0.96^2 \end{matrix} \right\} = 3.13$$

$$S_e = |\mathbf{e}|^2 = \left\{ \begin{matrix} (-0.13)^2 + (-0.60)^2 \\ +0.90^2 + (-0.66)^2 \\ +0.49^2 \end{matrix} \right\} = 1.87$$

$$S_{\hat{y}} + S_e = 3.13 + 1.87 = 5 = S_y$$

（4） 決定係数

S_yは y の変動そのものであり、$S_{\hat{y}}$は y の変動S_yのうち x 方向で説明される部分と考えられる。またS_eは、y の変動S_yのうち x 方向では説明できない部分と考えられる。そこで次の「決定係数 R^2（R2 乗）」を y の変動が x でどの程度説明できるのかということに関する指標とすることができる。

$$R^2 = \frac{S_{\hat{y}}}{S_y} = \frac{3.13}{5} = 0.626$$

決定係数について次の式が成り立つ。

$$R^2 = \frac{S_{\hat{y}}}{S_y} = \left(\frac{|\hat{\mathbf{y}}|}{|\mathbf{y}|} \right)^2 = \cos^2\theta = r^2$$

すなわち、決定係数は相関係数の 2 乗である。今回のデータで見ても、相関係数（0.79）の 2 乗は決定係数（0.626）になっている[8]。

[8] 決定係数は「y の分散の内で x で説明される分散の割合」と考えてもよい。というのは次の式が成り立つからだ。

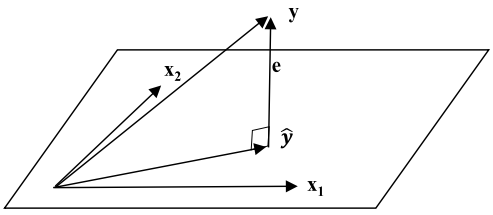
$$R^2 = S_{\hat{y}} / S_y = n s_{\hat{y}}^2 / n s_y^2 = s_{\hat{y}}^2 / s_y^2$$

3.4　3 つの変数ベクトル

（1）　重回帰モデル

3 つの変数、x_1, x_2, y があるとして、x_1 と x_2 から y を説明することを考えよう。これは独立変数が x_1 と x_2、従属変数が y の重回帰モデルである。このモデルは、$\mathbf{x}_1$ と $\mathbf{x}_2$ という 2 つのベクトルで、$\hat{\boldsymbol{y}}$（**y** ベクトルの予測値のベクトル）を導き出すモデルであり、これらのベクトルの関係は、図 12 のように表現できる。

図 12　重回帰モデル



（2）　y の変動の分解

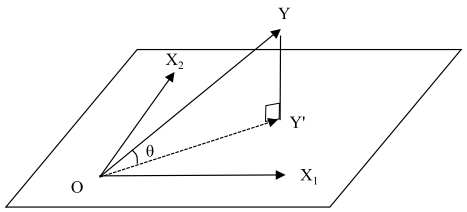
ここで、$\hat{\boldsymbol{y}}$ベクトルは、原点 O を起点にし、**y** ベクトルの頂点から $\mathbf{x}_1$ と $\mathbf{x}_2$ という 2 つのベクトルで作られる平面に下ろした垂線の足を終点にするベクトルである。真上から光を当てたときの **y** ベクトルの影のベクトルといってもいい。回帰式が意味することは $\mathbf{x_1x_2}$ 平面（これを予測平面という）で **y** ベクトルを説明するということである。

$\mathbf{x_1x_2}$ 平面で説明できるのはもちろん **y** ベクトルのすべてではない。説明できるのは **y** ベクトルの影である$\hat{\boldsymbol{y}}$ベクトルである。$\hat{\boldsymbol{y}}$ベクトルは$b_1\boldsymbol{x}_1 + b_2\boldsymbol{x}_2$という形で表現できるから、$\hat{\boldsymbol{y}}$ベクトルは $\mathbf{x_1x_2}$ 平面で完全に説明できる。

（3）　y の変動の分解と決定係数

図 13 において、y の変動は$|OY|^2$ である。この $|OY|^2$ は$\hat{y}$の変動$|OY'|^2$ と e の変動$|Y'Y|^2$ にきれいに分割できる [9]。

図 13　重相関係数と決定係数



$|OY'|^2$ は（$\mathbf{x_1x_2}$ 平面に$|OY'|$があるので）　$\mathbf{x}_1$ と $\mathbf{x}_2$ で説明できるが、$|YY'|^2$ は説明できない。したがって、$|OY|^2$ に占める$|OY'|^2$ の割合は、x_1、x_2 を説明変数としたモデルで、y の変動がどの程度説明できるかを意味することになる。これが決定係数だ [10]。

ところで、$\hat{\boldsymbol{y}}$ベクトルと **y** ベクトルの作る角度 θ の余弦は、説明変数 x_1、x_2 を用いた場合の y にたいする重相関係数とよばれる。ここで$|OY'|$ は $|OY|\cos\theta$ と表現できることに注目しよう。すると、$|OY|^2$ に占める$|OY'|^2$ の割合は、$|OY|^2\cos^2\theta / |OY|^2 = \cos^2\theta$ となる。重回帰モデルでは決定係数は重相関係数の 2 乗なのである。

---

[9] y の分散は、$\hat{y}$の分散と e の分散に分割できるといってもいい。

[10] y の分散の内 x_1、x_2 で説明できる部分といってもいい。

3.5　変数の単位ベクトル

（1）　変数ベクトルの単位ベクトル化

これまで標準得点化した変数ベクトルについて、基礎的な事項を解説してきたのだが、ベクトルを図で示す際には実はやっかいな問題が存在する。それは「標準得点ベクトルの長さ＝標準得点の標準偏差＝1」とはならず、ベクトルの長さは$\sqrt{n}$になるということである。これは図を用いて解説する際の大きな問題である。そこで、元の標準得点のベクトルを$\sqrt{n}$で割って変数の単位ベクトル化を試みよう。

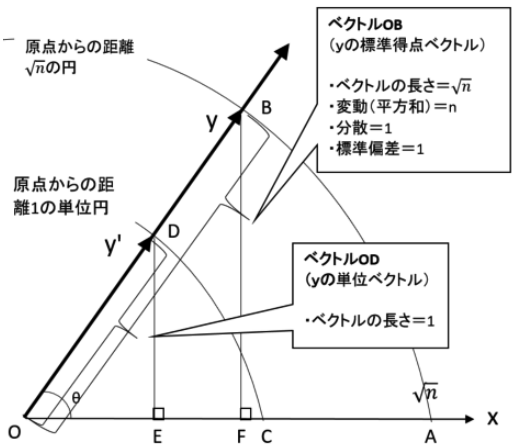
表 8 は標準得点ベクトルとそれを単位ベクトル化したものを示したものである。表の網掛け部分が太線で囲まれた部分に対応しているので、単位ベクトルの図を書くとベクトルの長さがそのまま標準得点の標準偏差に対応することになる。

表 8　標準得点のベクトルと単位ベクトル

	標準得点		単位ベクトル	
	年齢	収入	年齢	収入
1氏	-1.36	-1.21	-0.61	-0.54
2氏	-0.54	-1.03	-0.24	-0.46
3氏	-0.42	0.57	-0.19	0.25
4氏	1.10	0.21	0.49	0.10
5氏	1.22	1.46	0.54	0.65
n	5	5	5	5
平均	0.00	0.00	0.00	0.00
標準偏差	1.00	1.00	0.45	0.45
分散	1.00	1.00	0.20	0.20
偏差平方和	5.00	5.00	1.00	1.00
ベクトル長さ	2.24	2.24	1.00	1.00
長さの2乗	5.00	5.00	1.00	1.00
相関係数	0.79		0.79	

このあたりの関係は少しややこしいので絵で解説したものが図 14 である。原点からベクトルの終点までの長さを$\sqrt{n}$にしたものが標準得点ベクトル、1 にしたものが単位ベクトルである。単位ベクトル化してもベクトルの方向は変わらない。

図 14　変数ベクトルの単位ベクトル化



（2）単位ベクトル化の効用

図 14 からは次のことがわかる。

たとえば、標準得点ベクトル **y** を図に描くと、そのベクトルの長さ|OB|は$\sqrt{n}$になるが、単位ベクトル **y'**を図に描くと、ベクトルの長さ|OD|は y の標準偏差に等しい 1 になる。

また、標準得点ベクトル **y** の長さを一辺とする正方形を描くと、その面積は$|OB|^2=n$ となるが、単位ベクトル **y'**をもとに正方形を描くと、その面積$|OD|^2$ は y の分散に等しい 1 になる。

標準得点 **y** と **x** の相関は次のようになる。

$$r_{xy}=\cos\theta=\frac{|OB|}{|OF|}$$

一方、単位ベクトル **y'**では、

$$r_{xy'}=\cos\theta=\frac{|OE|}{|OD|}=\frac{|OE|}{1}=|OE|$$

と簡単に表せる。

このように、図を用いて種々の事柄を考える際には、変数を単位ベクトル化するととても便利なのである。

図 15　単位ベクトル化の効用

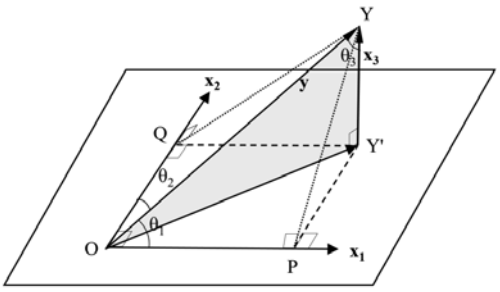


図 15 を使ってこの便利さについてもう少し解説しよう。O から Y に向かう **y** ベクトルの分散について述べるとき、標準得点ベクトルでは「y の分散は図の|OY|² を n で割ったものである」というかなりややこしい表現になる。しかし単位ベクトル化しておくと、「y の分散は図の|OY|² だ」というように簡単だ。

また、図 15 での次のような展開もわかりやすいものになる（**x1, x2,** は平面で直交するベクトル、**x3** は平面と直交し上部に向かうベクトル）。

$$
\begin{aligned}
&|OY|^2 = |OY'|^2 + |Y'Y|^2 \\
&= |OP|^2 + |OQ|^2 + |Y'Y|^2 \\
&= \cos^2\theta_1 + \cos^2\theta_2 + \cos^2\theta_3 \\
&= r_{x_1 y}^2 + r_{x_2 y}^2 + r_{x_3 y}^2
\end{aligned}
$$

これは、x_1, x_2, x_3 に相関がないとき、y の分散が y と x_1, x_2, x_3 との相関で表現できることを表しているのだが、図との対応がはっきりしていて、直感的な理解が容易である。

もし **y** を単位ベクトル化せず、標準得点ベクトルのまま同じことを示そうとすると、議論は少し面倒になる。そこではまず|OY|² が変動を表し、それが n であることを示したうえで、分散が|OY|²/n であると述べる必要がある。そして式の展開は次のようになされる。

$$
\begin{aligned}
&\frac{1}{n}|OY|^2 = \frac{1}{n}\left\{|OY'|^2 + |Y'Y|^2\right\} \\
&= \frac{1}{n}\left\{|OP|^2 + |OQ|^2 + |Y'Y|^2\right\} \\
&= \frac{1}{n}\left\{\left(|OY|\cos\theta_1\right)^2 + \left(|OY|\cos\theta_2\right)^2 + \left(|OY|\cos\theta_3\right)^2\right\} \\
&= \frac{1}{n}\left\{n\cos^2\theta_1 + n\cos^2\theta_2 + n\cos^2\theta_3\right\} \\
&= \cos^2\theta_1 + \cos^2\theta_2 + \cos^2\theta_3 \\
&= r_{x_1 y}^2 + r_{x_2 y}^2 + r_{x_3 y}^2
\end{aligned}
$$

この展開は、式では理解できるが、図と対応させて理解が容易になるようにはなっていない。

そういうわけで、以下の主成分分析の解説では、変数ベクトルを単位ベクトル化した図を使うことが多くなる。ここでは単位ベクトルという用語と意味を覚えておいてほしい。

3.6　問題と解答

（1）　問題

100 人を対象とした調査で、年齢、教育年数、収入のデータを得、それぞれの得点を標準化して x, y, z という変数にした。これらについて以下に答えよ。

(a)

x, y, z の平均と分散を求めよ。

(b)

x, y, z をベクトルと考えた場合、それぞれのベクトルの長さを求めよ。

(c)

x ベクトル、y ベクトル、z ベクトルそれぞれの要素の数（次元数）を述べよ。

(d)

x, y, z についてどの変数のペアも相関が 1 や－1 ではない場合、3 つの変数ベクトルを空間に位置付けるためには、最低何次元が必要か。

(e)

3 つの変数ベクトルをさらに単位ベクトル化した。この操作は具体的にはどうなされるか。

(f)

単位ベクトル化された x ベクトル、y ベクトル、z ベクトルの長さはどうなるか。

（2） 解答

(a)

どれも平均は 0、分散は 1。

(b)

どれも長さは 10。

(c)

どれも 100 の要素からなる（100 次元）。

(d)

3 次元。

(e)

各ベクトルを 10 で割る。

(f)

どれも長さは 1。

4　2 変数で考える主成分分析

4.1　分散と主成分軸

（1） 全変動と全分散

データの各変数を標準化して分散や変動（偏差平方和）を算出すると、どの変数においても「分散＝1」になり、「変動＝ケース数」になる。変動がケース数になるのは、

$$s_x^2 = \frac{1}{n} S_x$$

をもとに、

$$S_x = n s_x^2 = n \times 1 = n$$

となるからである。

このように各変数が標準得点化されている状態において、考えられている変数の変動の合計を全変動、分散の合計を全分散ということにしよう。100 人から獲得されたデータについて、各変数が標準化されている場合、1 つの変数を考慮するときデータの全変動は 100。2 つの変数を考慮するときデータの全変動は 200、3 つの変数を考慮するときデータの全変動は 300 となる。

全分散については、変数 1 つで 1、変数 2 つで 2、変数 3 つで 3 である。

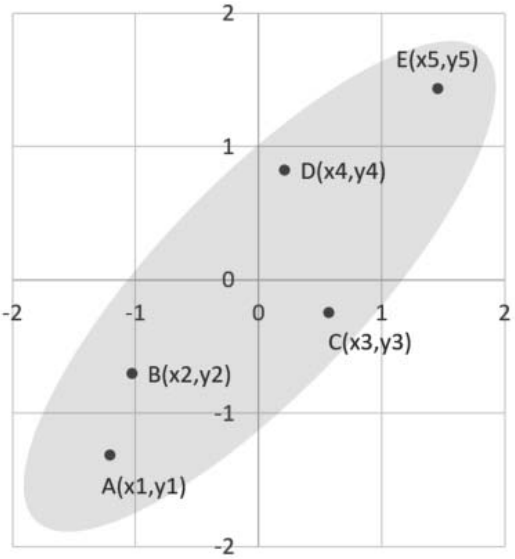
全変動は考慮されている変数の数×ケース数に一致し、全分散は考慮されている変数の数に一致する。年齢と教育年数と所得という 3 つの変数を同時に考え、それらの変数がすべて標準化されている場合、ケースが 100 あればデータの全変動は 300、全分散は 3 である。

ここで全変動や全分散についてのきちんとしたイメージを得るために、表 9 のようなケース数 5 のデータをもとに解説を続けよう。このデータの標準得点をプロットしたものが図 16 である。

表 9　データ

	素点		標準得点	
	収入	資産	収入	資産
A氏	550	300	-1.21	-1.31
B氏	560	500	-1.03	-0.70
C氏	650	650	0.57	-0.24
D氏	630	1000	0.21	0.82
E氏	700	1200	1.46	1.43
n	5	5	5	5
平均	618	730	0.00	0.00
分散	3176	107600	1.00	1.00
標準偏差	56.4	328.0	1.00	1.00
偏差平方和	15880	538000	5.00	5.00
相関係数	0.885		0.885	
偏差積和	81800		4	
共分散	16360.0		0.9	

図 16　収入と資産の分布



ここで、x の変動は次の式で表される。

$$S_x = \left\{ \begin{matrix} \left(x_1-\bar{x}\right)^2+\left(x_2-\bar{x}\right)^2+\left(x_3-\bar{x}\right)^2 \\ +\left(x_4-\bar{x}\right)^2+\left(x_5-\bar{x}\right)^2 \end{matrix} \right\}$$

$$= \left\{ \begin{matrix} \left(x_1-0\right)^2+\left(x_2-0\right)^2+\left(x_3-0\right)^2 \\ +\left(x_4-0\right)^2+\left(x_5-0\right)^2 \end{matrix} \right\}$$

$$= x_1^2+x_2^2+x_3^2+x_4^2+x_5^2$$

同様に y の変動は、

$$S_y = y_1^2+y_2^2+y_3^2+y_4^2+y_5^2$$

したがって、全変動は

$$S_x+S_y = \left\{ \begin{matrix} \left(x_1^2+x_2^2+x_3^2+x_4^2+x_5^2\right) \\ +\left(y_1^2+y_2^2+y_3^2+y_4^2+y_5^2\right) \end{matrix} \right\}$$

$$= \left\{ \begin{matrix} \left(x_1^2+y_1^2\right)+\left(x_2^2+y_2^2\right)+\left(x_3^2+y_3^2\right) \\ +\left(x_4^2+y_4^2\right)+\left(x_5^2+y_5^2\right) \end{matrix} \right\}$$

ここでピタゴラスの定理より、

$$= |OA|^2+|OB|^2+|OC|^2+|OD|^2+|OE|^2$$

となって、全変動は原点からそれぞれのケースを表す点までの距離の 2 乗の総和であることがわかる。

また、全分散については、

$$s_x^2+s_y^2 = \left\{ \begin{matrix} \frac{1}{5}\left(x_1^2+x_2^2+x_3^2+x_4^2+x_5^2\right) \\ +\frac{1}{5}\left(y_1^2+y_2^2+y_3^2+y_4^2+y_5^2\right) \end{matrix} \right\}$$

$$= \frac{1}{5}\left\{ \begin{matrix} \left(x_1^2+y_1^2\right)+\left(x_2^2+y_2^2\right)+\left(x_3^2+y_3^2\right) \\ +\left(x_4^2+y_4^2\right)+\left(x_5^2+y_5^2\right) \end{matrix} \right\}$$

$$= \frac{1}{5}\left(|OA|^2+|OB|^2+|OC|^2+|OD|^2+|OE|^2\right)$$

となり、「原点からそれぞれのケースを表す点までの距離の 2 乗の総和」をケース数で割ったもの

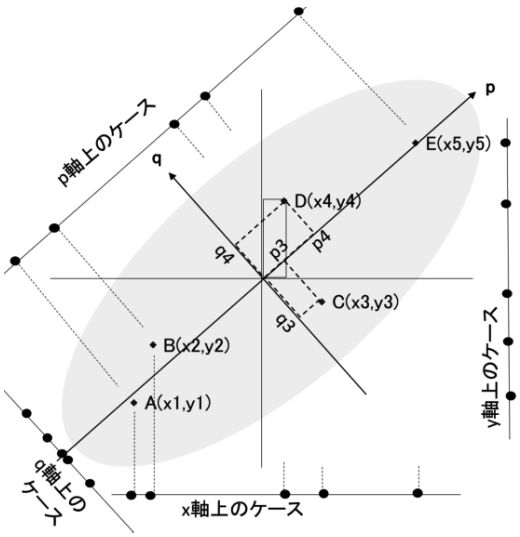
となる[11]。

全変動も全分散も、取り上げられている変数を全体として見たとき、データがどれぐらい散らばっているのかを示すものだ。それらは原点から各ケースを表す点までの距離をもとに考えられている。このことは 3 変数やそれ以上の変数の全変動や全分散についても同様である[12]。

（2） 主成分軸

図 16 のようにデータが分布している場合、最初の x, y 座標軸を用いるよりも、図 17 の p, q のような座標軸を用いたほうがデータ全体の散らばりがうまく説明できる。p はデータ全体の散らばりをかなり多く説明し、残った部分が q によって説明されると考えることができるのである。

図 17　主成分軸



これはデータの全分散が p 軸方向の分散と、それに直交する q 軸方向の分散に分割されることを意味する。全分散は変わらないが分散の割合が異なる 2 軸が導かれるのである。

式で見ていこう。A,B,C,D,E の座標は次のように表されていた。

$$A(x_1,y_1)\quad B(x_2,y_2)\quad C(x_3,y_3)$$
$$D(x_4,y_4)\quad E(x_5,y_5)$$

同じ A,B,C,D,E の座標を pq 座標で次のように表そう。

$$A(p_1,q_1)\quad B(p_2,q_2)\quad C(p_3,q_3)$$
$$D(p_4,q_4)\quad E(p_5,q_5)$$

ここで、x,y,の分散は次のようになる。

$$s_x^2=\frac{1}{5}\left(x_1^2+x_2^2+x_3^2+x_4^2+x_5^2\right)=1$$
$$s_y^2+\frac{1}{5}\left(y_1^2+y_2^2+y_3^2+y_4^2+y_5^2\right)=1$$

すなわち、どちらの分散も 1 であり、全分散は 2 だ。

一方 p,q 軸から見た分散は次のようになる。

$$s_p^2=\frac{1}{5}\left(p_1^2+p_2^2+p_3^2+p_4^2+p_5^2\right)>1$$
$$s_q^2=\frac{1}{5}\left(q_1^2+q_2^2+q_3^2+q_4^2+q_5^2\right)<1$$

すなわち、p 軸での分散は q 軸での分散よりも大きい。ただし、次に示すようにこの場合も、全分散はやはり 2 である。

---

[11] したがって、全分散は「原点からケースまでの距離の 2 乗の平均的な値」ということになる。

[12] 3 変数の場合は平面ではなく空間で考えればいい。

$$
\begin{aligned}
& s_p^2 + s_q^2 \\
&= \left\{ \begin{array}{l} \frac{1}{5}\left(p_1^2 + p_2^2 + p_3^2 + p_4^2 + p_5^2\right) \\ + \frac{1}{5}\left(q_1^2 + q_2^2 + q_3^2 + q_4^2 + q_5^2\right) \end{array} \right\} \\
&= \frac{1}{5}\left\{ \begin{array}{l} \left(p_1^2 + q_1^2\right) + \left(p_2^2 + q_2^2\right) + \left(p_3^2 + q_3^2\right) \\ + \left(p_4^2 + q_4^2\right) + \left(p_5^2 + q_5^2\right) \end{array} \right\} \\
&= \frac{1}{5}\left(|OA|^2 + |OB|^2 + |OC|^2 + |OD|^2 + |OE|^2\right) \\
&= \frac{1}{5}\left\{ \begin{array}{l} \left(x_1^2 + y_1^2\right) + \left(x_2^2 + y_2^2\right) + \left(x_3^2 + y_3^2\right) \\ + \left(x_4^2 + y_4^2\right) + \left(x_5^2 + y_5^2\right) \end{array} \right\} \\
&= \left\{ \begin{array}{l} \frac{1}{5}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2\right) \\ + \frac{1}{5}\left(y_1^2 + y_2^2 + y_3^2 + y_4^2 + y_5^2\right) \end{array} \right\} \\
&= s_x^2 + s_y^2 = 2
\end{aligned}
$$

元のデータの全分散は、p, q の 2 軸の設定によって、p 軸での分散と q 軸での分散に再分割されるのである。

主成分分析はこのようにデータの全分散を x, y 軸よりもふさわしい別の軸で説明しようとする方法である。そこではデータの全分散をできるだけ多く説明する第 1 主成分軸（p 軸）をまず導き出し、次に第 1 主成分軸に直交し、残りの分散を説明するような第 2 主成分軸（q 軸）を導き出すことがめざされる。

4.2　寄与率・主成分得点・主成分負荷量

（1）　主成分軸の求め方

図 17 の p 軸や q 軸を求めるには、「変数群の相関行列の固有値と固有ベクトルのセットを求める」という数学的な操作が必要になる。固有値 λ と固有ベクトル $\mathbf{e}$ とは、相関行列を $\mathbf{A}$ とするとき、$\mathbf{Ae} = \lambda\mathbf{e}$ を満たす $\boldsymbol{\lambda}$（スカラー）と $\mathbf{e}$（ベクトル）のことだ。

固有値と固有ベクトルのセットとは、「固有値 λ_1 と固有ベクトル $\mathbf{e_1}$ のセット」、「固有値 λ_2 と固有ベクトル $\mathbf{e_2}$ のセット」…といった意味だ。変数が 2 つならこのセットは 2 つ、変数が 3 つならこのセットは 3 つになる。行列とベクトルの形で表現すると次のようになる。

$$
\mathbf{Ae} = \lambda\mathbf{e}
$$

$$
\begin{bmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{bmatrix} \begin{bmatrix} e1_1 \\ e1_2 \end{bmatrix} = \lambda_1 \begin{bmatrix} e1_1 \\ e1_2 \end{bmatrix}
$$

$$
\begin{bmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{bmatrix} \begin{bmatrix} e2_1 \\ e2_2 \end{bmatrix} = \lambda_2 \begin{bmatrix} e2_1 \\ e2_2 \end{bmatrix}
$$

固有値と固有ベクトルのセットは、固有値のもっとも大きいものから順に算出される [13]。また、固有ベクトルの長さは 1 に調整されて算出される。

表 9 のデータにおいて、相関係数は 0.885 だった。これをもとに固有値と固有ベクトルを求めると次のような 2 セットになる。

$$
\begin{bmatrix} 1 & 0.885 \\ 0.885 & 1 \end{bmatrix} \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} = 1.885 \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}
$$

[13] 固有値 λ と固有ベクトル $\mathbf{e}$ は行列式を用いて、$|\boldsymbol{A} - \boldsymbol{E\lambda}| = 0$ の解として求められる。計算はやっかいなのでネット上の計算サイトを利用すればいい。

$$\begin{bmatrix} 1 & 0.885 \\ 0.885 & 1 \end{bmatrix}\begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix} = 0.115\begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix}$$

すなわち、第 1 主成分の固有値は 1.885、固有ベクトルは（0.707, 0.707）、第 2 主成分の固有値は 0.115、固有ベクトルは（-0.707, 0.707）である。

固有ベクトルとはそれぞれの直交する主成分軸に対応するものである[14]。また、固有値 λ_1, λ_2…はそれぞれの主成分軸から見たデータの分散を表し、λ_1, λ_2…の順に小さくなっていく。

全分散は変数軸から見ても主成分軸から見ても同じだった。したがって、全固有値の和＝全分散が成立する。変数が標準化されているとき、それぞれの変数の分散は 1 だから、結局、全分散＝全固有値の和＝変数数ということになる。2 変数を扱う場合には全分散も固有値の和も 2 になり、3 変数を扱う場合には全分散も固有値の和も 3 になるのである。

ここでのデータでは、第 1 主成分に対応する固有値は 1.885、第 2 主成分に対応する固有値は 0.115 になる。1.885＋0.115＝2 となるのは、変数が 2 つであり全分散が 2 であるからだ。

（2） 主成分得点

各ケースのデータを主成分軸上で測ったものが各ケースの主成分得点である。図 17 でいうと、D 氏の第 1 主成分についての主成分得点とは p 軸を物差しとした D 氏の位置であり p4 がその値になる。また、D 氏の第 2 主成分についての主成分得点とは、q 軸を物差しとした D 氏の位置であり q4 がその値になる[15]。

表 10　主成分得点

	収入x	資産y	主成分得点		標準化された主成分得点	
	標準得点		第1主成分 p	第2主成分 q	第1主成分 p	第2主成分 q
A氏	-1.207	-1.311	-1.780	-0.074	-1.296	-0.218
B氏	-1.029	-0.701	-1.224	0.232	-0.891	0.685
C氏	0.568	-0.244	0.229	-0.574	0.167	-1.692
D氏	0.213	0.823	0.733	0.431	0.534	1.272
E氏	1.455	1.433	2.042	-0.016	1.487	-0.047
平均	0.00	0.00	0.00	0.00	0.00	0.00
分散	1.00	1.00	1.89	0.11	1.00	1.00
標準偏差	1.00	1.00	1.37	0.34	1.00	1.00
平方和	5.00	5.00	9.43	0.57	5.00	5.00
相関	0.88		0.00		0.00	
n	5		5		5	

A 氏から E 氏までの第 1 主成分 p と第 2 主成分 q についての主成分得点をまとめると表 10 中央の「主成分得点」の 2 列のようになる。2 つの主成分得点の平均は 0 である。分散はそれぞれの固有値 1.89 と 0.11 に対応しており、両者を足すと 2 になる。

主成分得点も通常の変数と同様、標準得点化されて利用されることが多い。その値は表 10 の「標準化された主成分得点」の 2 列に記されている。

---

[14] 主成分軸は固有ベクトル方向の軸である。ベクトル（a_1, a_2）は、空間内の直線の式としては、$y = \frac{a_2}{a_1}x$ となる。

[15] 第 1 の固有ベクトルが **e1**=（a1, a2）、第 2 の固有ベクトルが **e2**=（b1, b2）であり、あるケースの x と y についての標準得点が（x, y）だとすると、そのケースの第 1 主成分得点は a1x＋a2y、第 2 主成分得点は b1x＋b2y である。ここの例での D 氏の収入の得点は 0.213、資産の得点は 0.823 であった。また、固有ベクトルは **e1**=（0.707, 0.707）、**e2**=（0.707, -0.707）だった。したがって、D 氏の第 1 主成分得点は、(0.707)(0.213)＋(0.707)(0.823)＝0.733 となり、D 氏の第 2 主成分得点は、(0.707)(0.213)＋(-0.707)(0.832)＝0.431 となる。3 変数以上の場合も同様、固有ベクトルと標準得点から主成分得点が求められる。なお、このように主成分得点が産出されるので、主成分得点は変数の重み付き合成得点といわれる。

こちらの分散はそれぞれ 1 であり、相関は 0 だ[16]。

（3） 負荷量

表 10 を見るとイメージがわくと思うが、主成分得点で表された列は新しくできた変数と考えられる。すなわち、収入 x、資産 y に追加して、第 1 主成分 p、第 2 主成分 q という新たな変数が作られたのである。主成分 p と主成分 q の相関は 0 である（表 10 参照）。

これらの主成分と各変数の相関は負荷量（主成分負荷量・因子負荷量）といわれるものだ。それは表 11 の「主成分分析結果の表」の網掛け部分に示されている（「主成分分析結果の表」は主成分分析で最も重要な表である）。相関の値は変数が標準化されても変わらないので、この負荷量は標準化された主成分と各変数の相関でもある。

表 11　主成分分析結果の表

	第1主成分 p	第2主成分 q	共通性
収入	0.971	-0.240	1.000
資産	0.971	0.240	1.000
固有値	1.885	0.115	
寄与率	0.942	0.058	

表からは第 1 主成分と収入・資産の相関は大きいこと、第 2 主成分と収入・資産の相関は大きくはないことがわかる。

実際にどのような関係があるのかを図で示しておこう。第 1 主成分と収入・資産の関係を図で表すと図 18、図 19 のようになる。これらの図から、収入も資産も第 1 主成分と高い相関を持つことがわかる。

---

[16] 分散ではなく不偏分散（の正の平方根）を用いて標準得点の計算をすると、表 10 とやや異なる下の表のような結果になる。汎用の統計ソフトでは不偏分散が標準得点化において用いられることが多いので、SPSS を使った分析でもこうなるはずである。2 つの表の値の違いは、今回の例において n が 5 と小さいために生じる。n が大きい通常の社会調査データでは分散と不偏分散の値はほぼ同じになるので、結果はほとんどかわらない。なお、SPSS で算出される主成分得点は標準化された主成分得点である。

	収入x 資産y		主成分得点		標準化された主成分得点	
	標準得点		第1主成分	第2主成分	第1主成分	第2主成分
A氏	-1.079	-1.172	-1.592	-0.066	-1.160	-0.195
B氏	-0.921	-0.627	-1.094	0.207	-0.797	0.611
C氏	0.508	-0.218	0.205	-0.513	0.149	-1.513
D氏	0.190	0.736	0.655	0.386	0.477	1.139
E氏	1.301	1.282	1.826	-0.014	1.330	-0.041
平均	0.00	0.00	0.00	0.00	0.00	0.00
不偏分散	1.00	1.00	1.88	0.11	1.00	1.00
標準偏差	1.00	1.00	1.37	0.34	1.00	1.00
平方和	4.00	4.00	7.54	0.46	4.00	4.00
相関	0.88		0.00		0.00	
n	5		5		5	

（標準得点化に不偏分散を用いた場合）

図 18　第 1 主成分と収入の関係

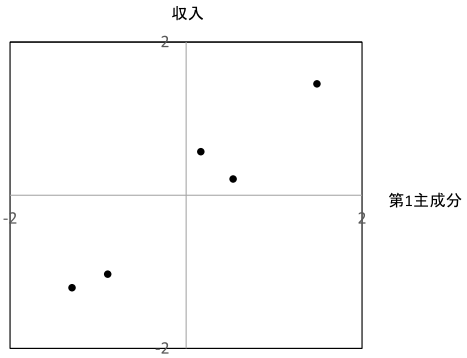


図 19　第 1 主成分と資産の関係

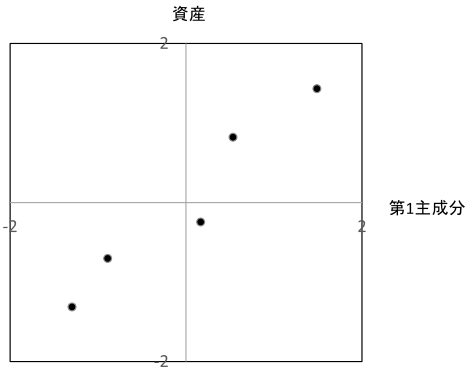


図 20　第 2 主成分と収入の関係

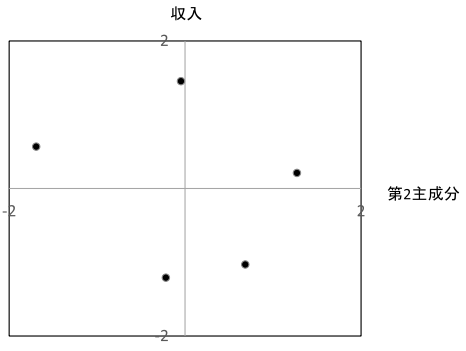
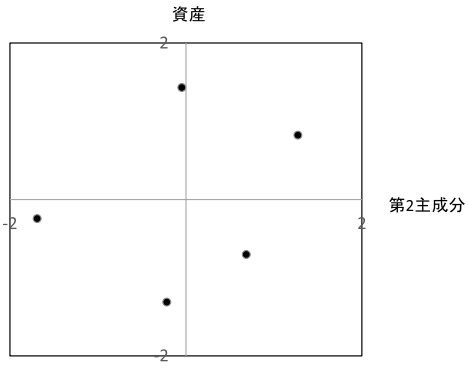


図 21　第 2 主成分と資産の関係



第 2 主成分と収入・資産の関係を示したのが図 20 と図 21 である。これらの図からは収入・資産ともに第 2 主成分との相関は小さいことがわかる。

（4） 固有値と寄与率

さて、負荷量の掲載されていた「主成分分析結果の表」（表 11）には負荷量以外に、各主成分の固有値や寄与率、各変数の共通性が掲載されている。順番に解説しよう。

全分散に占めるその主成分で説明できる分散の割合をその主成分の寄与率という。主成分の分散はその固有値に等しいので、「当該主成分に対応する固有値÷全固有値の和」でこの寄与率が算出できる。各変数の分散は 1 だから、全分散＝変数数＝全固有値の和が成立し、「当該主成分に対応する固有値÷変数数」でも寄与率が出る。

「寄与率」という考え方は、回帰分析の「決定係数」の考え方に似たところがある。決定係数は「従属変数の分散が独立変数全体でどの程度説明できるか」を示す概念であった。それに対して主成分分析における寄与率は「変数全体の分散の総和（全分散）がその主成分でどの程度説明できるか」を示す概念なのである。ここでのデータで

は表 12 のようになる。

表 12 固有値と寄与率

	第1主成分 p	第2主成分 q	計
固有値	1.885	0.115	2.00
寄与率	0.942	0.058	1.00

（5） 共通性

固有値や寄与率が各主成分にかかわる概念であるのに対し、共通性は各変数にかかわる概念である。共通性とは各変数（収入・資産）の分散がとりあげられた主成分でどの程度説明されるのかを表したものである。

表 11 ではこの値はすべて 1 になっている。だから値が 1 のこのような概念は不要と感じるかもしれない。しかし、「とりあげる主成分」がいつも変数の数（この場合は 2）とは限らない。次にこのあたりのことを解説しよう。

4.3 主成分の解釈と尺度構成

いくつかの変数を主成分分析にかけて、今回のデータのように 1 つの主成分の寄与率がとても高かったとしよう。そんなときに通常提示されるのは次のような省略された表だ（表 13）。

表 13 省略された主成分分析結果の表

	第1主成分 p	共通性
収入	0.971	0.943
資産	0.971	0.943
固有値	1.885	
寄与率	0.942	

これが示しているのは収入の分散の 94.3%が第 1 主成分で説明でき、資産の分散もまた 94.3%が第 1 主成分で説明できるということである。変数の数ととりあげる主成分の数が同じとき、共通性は各変数でかならず 1 になるが、取り上げる主成分の数が変数の数より少ないとき 1 より小さくなる。

ところで表 13 で示されているような負荷量を見て、読者はこの第 1 主成分がどんな概念を表すものと考えるだろうか。主成分分析において、各主成分の意味する概念が何なのかは分析者自らが解釈しなければならない問題である。

今回のデータから得られた第 1 主成分では、収入と資産についての負荷量が高い。分析者はそれを見てたとえば「この主成分は経済的豊かさを表す」といった解釈をする。そしてそこからさらに第 1 主成分の主成分得点（表 10 参照）をその人の経済的豊かさの値とみなしたりする。そしてさらにこの主成分得点を以後の分析で用いたりするのである。これは「経済的豊かさ」についての尺度を構成していることに他ならない。主成分分析はこのように変数をとりまとめた新たな尺度を構成するためによく用いられる。

4.4 2 変数の分散と負荷量・固有値・共通性

（1） 2 変数の単位ベクトル

さて、表 11 や表 13 のような「主成分分析結果の表」を見た読者は不思議なことに気づくかもしれない。それは次のことだ。表 11 について、(1) 負荷量を縦に見るとき、第 1 主成分の収入の負荷量と資産の負荷量の 2 乗和は第 1 主成分の固有値になり（$0.971^2+0.971^2=1.885$）、第 2 主成分でもそうなること（$(-0.240)^2+0.240^2+=0.115$）、(2) 表の負荷量を横に見るとき、収入の第 1 主成

分負荷量と第2主成分負荷量の2乗和は共通性になり（$0.971^2+(-0.240)^2=1.000$）、資産でもそうなること（$0.971^2+0.240^2=1.000$）である。つまり負荷量の縦の2乗和は固有値になり、横の2乗和は共通性になっているのである。同じことは表 13 でもいえる（$0.971^2=0.943$）。

これらのことは変数や主成分を単位ベクトル化して図形でとらえることで理解できる。

図 22　2 変数の主成分分析の図形的表現

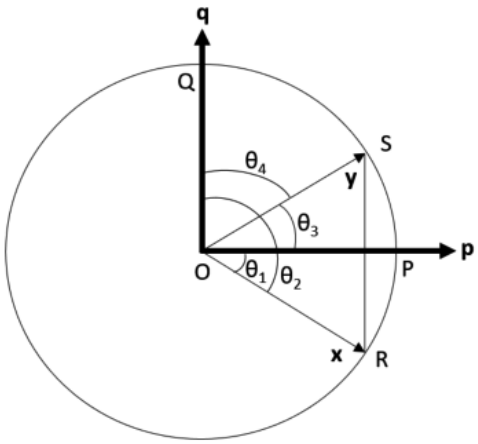


図 22 には 2 つの変数ベクトル **x**, **y**、2 つの主成分ベクトル **p**, **q**、ならびに半径 1 の円が描かれている。変数ベクトルは各ケースの値を要素とするベクトルであり、主成分ベクトルは各ケースの主成分得点を要素とするベクトルである（表 10 参照）。変数ベクトルは単位ベクトル化されているので、図の **x**, **y** ベクトルはどちらも円の中心を始点とし円周を終点とする長さ 1 のベクトルとなる [17]。

∠ROS の余弦は x と y の相関を表している。また、∠POQ は 90 度だが、これは p と q という 2 つの主成分の相関が 0 であるため、ベクトル **p**, **q** が直交することによる [18]。

ここで角度について、次のシンボルを使おう。

$$\theta_1 = \angle ROP$$
$$\theta_2 = \angle ROQ$$

$$\theta_3 = \angle SOP$$
$$\theta_4 = \angle SOQ$$

このとき、相関とベクトル間の角度との関係から、x, y と p, q について次の式が成り立つ。

$$r_{x,p} = \cos\theta_1$$
$$r_{x,q} = \cos\theta_2$$

$$r_{y,p} = \cos\theta_3$$
$$r_{y,q} = \cos\theta_4$$

これらは x と y の p と q についての負荷量でもある。

次に R と S の **pq** ベクトル軸での座標について考えよう（θ はここでは $0°\leqq\theta\leqq180°$ とし、負の θ は考えない）[19]。

---

[17] 変数と主成分についての混乱しやすい概念を整理しておく。変数（x）に対応する概念が主成分（p）である。変数軸（x 軸）に対応する概念が主成分軸（p 軸）である。ケースの変数の値（x_i）に対応するのがケースの主成分得点（p_i)である。ケースの変数の値を並べた変数ベクトル（**x**）に対応するのがケースの主成分得点を並べた主成分ベクトル（**p**）である。

[18] 図 25 の **pq** 軸の意味は図 20 の p,q の主成分軸とは異なることに注意してほしい。図 20 の主成分軸はケースを位置付ける空間の軸であるのに対し、ここでの軸は主成分ベクトル **p**,**q** の方向を表す軸である。ざっくりいえば、図 17 はケースを位置付けた図 7 に対応し、図 22 は変数を位置付けた図 8 に対応する。

[19] この座標はケース数次元（この場合は 5 次元）のベクトルを、変数数次元（この場合は 2 次元）の平面に写した場合の平面の座標ということになる。

R の座標はまず次のように表せる。

$$R(\cos\theta_1, -\sin\theta_1)$$

ここで

$$-\sin\theta_1 = \cos(90^\circ + \theta_1) = \cos\theta_2$$

したがって R の座標は、

$$R(\cos\theta_1, \cos\theta_2)$$

また、S の座標は次のように表せる。

$$S(\cos\theta_3, \sin\theta_3)$$

ここで

$$\sin\theta_3 = \sin(90^\circ - \theta_4) = \cos\theta_4$$

だから、S の座標は

$$S(\cos\theta_3, \cos\theta_4)$$

結局、次のものが R と S の座標となる。

$$R(\cos\theta_1, \cos\theta_2) \qquad S(\cos\theta_3, \cos\theta_4)$$

このことと上で述べたことより、R と S の座標を負荷量で表すと次のようになる。

$$R(r_{x,p}, r_{x,q}) \qquad S(r_{y,p}, r_{y,q})$$

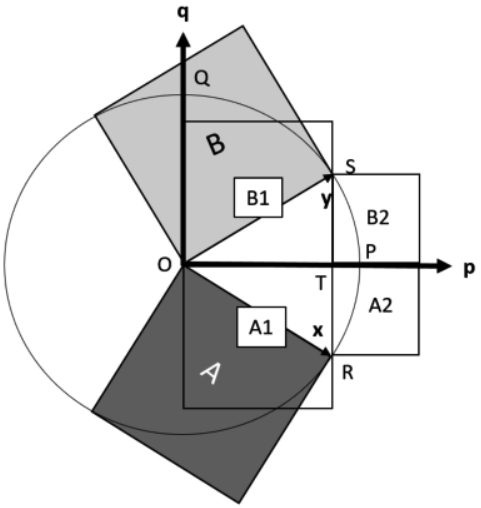
（2） 変数の分散の分解

以上の知識をもとに、分散の分解と負荷量平方和について同時に考えていこう。図 23 には A, A1, A2 ならびに B, B1, B2 の計 6 つの正方形が描かれている。これらの記号をその面積も表すものとする。

濃い灰色の正方形 A は x の分散を表し、薄い灰色の正方形 B は y の分散を表す。それぞれの変数ベクトルは単位ベクトル化されているので、A と B の面積はともに 1 であり全分散はそれらを合わせた 2 になる。

A1 ならびに B1 は OT を一辺とする正方形であり、A2 は TQ を一辺とする正方形、B2 は TS を一辺とする正方形だ。

図 23　分散の分解図



R と S の座標は次のものだった。

$$R(r_{x,p}, r_{x,q}) \qquad S(r_{y,p}, r_{y,q})$$

したがって、A1, A2, B1, B2 の面積は次のように表せる。

$$A1 = r_{x,p}^2 \qquad A2 = r_{x,q}^2$$

$$B1 = r_{y,p}^2 \qquad B2 = r_{y,q}^2$$

A と A1,A2 や B と B1,B2 についてはピタゴラスの定理より次のことがいえる。

$$A = A1 + A2 = 1$$

$$B = B1 + B2 = 1$$

すなわち、x の分散 A は **p** 方向の分散 A1 と、**q** 方向の分散 A2 に分割され、y の分散 B も **p** 方向

の分散 B1 と **q** 方向の分散 B2 に分割されるのである。

これを負荷量の式で表すと次のようになる。

$$A = A1 + A2 = r_{x,p}^2 + r_{x,q}^2 = 1$$

$$B = B1 + B2 = r_{y,p}^2 + r_{y,q}^2 = 1$$

（3） 固有値、寄与率と負荷量平方和

ここまでの議論をもとにすると p の分散 λ_1 と q の分散 λ_2 はそれぞれ次のように負荷量平方和で表されることがわかる。

$$p: \quad \lambda_1 = A1 + B1 = r_{x,p}^2 + r_{y,p}^2$$
$$q: \quad \lambda_2 = A2 + B2 = r_{x,q}^2 + r_{y,q}^2$$

すなわち、「主成分分析結果の表」の縦の負荷量平方和はそれぞれの主成分の固有値に等しい。

また、各主成分の寄与率はそれぞれ、

$$p: \quad \frac{A1 + B1}{A + B} = \frac{A1 + B1}{2} = \frac{\lambda_1}{2}$$

$$q: \quad \frac{A2 + B2}{A + B} = \frac{A2 + B2}{2} = \frac{\lambda_2}{2}$$

であり、それぞれの固有値を全分散（＝変数数）で割ったものとなる。

（4） 共通性と負荷量平方和

共通性については次のようになる。まず、第 1 主成分 p と第 2 主成分 q がとりあげられている場合、

$$x: \quad A1 + A2 = r_{x,p}^2 + r_{x,p}^2 = 1 (= A)$$
$$y: \quad B1 + B2 = r_{y,q}^2 + r_{y,q}^2 = 1 (= B)$$

となり、主成分分析表の横の平方和が共通性になることがわかる。これは必ず 1 になる。

次に、第 1 主成分 p のみがとりあげられている場合の x と y それぞれの共通性は、

$$x: \quad A1 = r_{x,p}^2$$
$$y: \quad B1 = r_{y,q}^2$$

となり、こちらは通常 1 より小さい値になる。

以上が、図を用いた固有値、寄与率、共通性の解釈であり、「主成分分析結果の表」の負荷量の縦の平方和が固有値になり、横の平方和が共通性になることについての解説である。

4.5　この節の議論の展開

この節では主成分分析の最も基本的な部分の解説をしたが、話の展開はやや複雑である。ここでこれまでの流れについてまとめておく。

（1）2 変数 5 ケースのデータがあるとした。

（2）このデータについて、xy 軸の空間に 5 ケースを位置付け、xy 軸よりも全分散をうまく説明できる p 軸と q 軸を描いた。

（3）各ケースの p 軸から見た得点を第 1 主成分得点とし、q 軸から見た得点を第 2 主成分得点とした。

（4）各ケースの x の得点を要素とする x ベクトル、y の得点を要素とする y ベクトル、第 1 主成分得点を要素とする第 1 主成分ベクトル、第 2 主成分得点を要素とする第2主成分ベクトルを考えた。

（5）これらのベクトルはそれぞれ 5 つの要素をもつ 5 次元ベクトルであるが、それらを 2 次元平面に写して図を描いた。図に登場する 2 つの軸は 2 つの主成分ベクトル方向の軸だ。

（6）この図をもとに、変数の分散と全分散、主成分分析における負荷量、固有値、寄与率、共通性などを解説した。

4.6 問題と解答

（1） 問題

(a)

次の主成分分析の結果の表を見て、収入と第 1 主成分の相関の値を述べよ。また a～e の値を自分で計算して求めよ。

	第1主成分	第2主成分	共通性
収入	0.971	-0.240	a
資産	0.971	0.240	b
固有値	c	d	
寄与率	0.942	e	

(b)

収入と資産という 2 変数についてたずねた 100 人のデータを主成分分析にかけた。このとき次のことがどうなるかを答えよ。

・相関行列から得られる固有値の数
・相関行列から得られた固有ベクトルの数
・得られた固有ベクトルの次元数
・変数ベクトルの個数
・変数ベクトルの次元数
・主成分ベクトルの個数
・主成分ベクトルの次元数

・2 つの変数ベクトルを図で表現するために必要となる空間の次元数
・2 つの主成分ベクトルを図で表現するために必要となる空間の次元数
・2 つの変数ベクトルと 2 つの主成分ベクトルを 1 つの図で表現するために必要な空間の次元数

・各ケースをプロットするために用いられる主成分軸と固有ベクトルの関係
・相関行列の固有ベクトルと主成分ベクトルは同じものか。

(c)

図 23 が表 11 の主成分分析の結果にもとづいているとする場合、∠SOP と∠SOQ の角度を求めよ（3.3(1)の表 7 およびその表についての注を見てエクセルで計算するといい)。

（2） 解答

(a)

・収入と第 1 主成分の相関は 0.971。
・a=1, b=1, c=1.885, d=0.115, e=0.058

(b)

・相関行列から得られる固有値の個数は 2 つ。
・相関行列から得られる固有ベクトルの数は 2。
・得られる固有ベクトルの次元数はそれぞれ 2。
・変数ベクトルの個数は、収入ベクトルと資産ベクトルの 2 つ。
・変数ベクトルの次元数はケースが 100 なのでそれぞれ 100。
・主成分ベクトルの個数は第 1 主成分と第 2 主成分の 2 つ。
・主成分ベクトルの次元数はケースが 100 なのでそれぞれ 100。

・2 つの変数ベクトルを図で表現するために必要となる空間の次元数はベクトルが 2 つなので 2。
・2 つの主成分ベクトルを図で表現するために必要となる空間の次元数はベクトルが 2 つなので 2。
・2 つの変数ベクトルと 2 つの主成分ベクトルを図で表現するために必要な空間の次元数は、すべてのベクトルが同一平面にあるので 2。

・各ケースをプロットするために用いられる主成分軸は 2 つの固有ベクトル方向の軸である。

・相関行列の固有ベクトルは主成分ベクトルとは異なる。前者はケースを位置付ける軸になり、後者は主成分得点を要素とするベクトルである。2 つのものが異なるのはケースを位置付ける x 軸とケースの値を要素とする変数 x のベクトルが異なるのと同様だ。固有ベクトルでは次元数＝変数数となり、主成分ベクトルでは次元数＝ケース数となる。ただし後者は変数数次元の空間に描くことができる。

(c)

$$\angle SOP = \arccos\left(0.971\right) = 14^{\circ}$$

$$\angle SOQ = \arccos\left(0.240\right) = 76^{\circ}$$

・arccos は cos の逆関数である。下の関係が成り立つ。

$$\cos\left(14^{\circ}\right) = 0.971$$

$$\cos\left(76^{\circ}\right) = 0.240$$

・arccos はエクセルでは acos 関数を使う。「=DEGREES(ACOS(0.971))」とすれば、「度」で結果が表示される。

5　3 変数で考える主成分分析

5.1　データ・分析結果・基本的原理

（1） 3 変数のデータと分析結果

ここからは次のような 3 変数、5 ケースの仮想的データに基づいてさらに解説をすすめる（表 14）。これは「この 1 週間にテレビで次のような番組を何回見ましたか」という問いに対する A 氏から E 氏の答えだと考えればいいだろう。

表 14　テレビ視聴についての仮想的データ

	音楽 x	スポーツ y	ニュース z
A氏	5	3	6
B氏	4	4	6
C氏	3	2	4
D氏	3	2	8
E氏	1	1	5

このデータをもとに相関行列を求めたものが、表 15 である。

表 15　相関行列

	音楽	スポーツ	ニュース
音楽x	1.000	0.828	0.250
スポーツy	0.828	1.000	0.207
ニュースz	0.250	0.207	1.000

この相関行列をもとに、固有値と固有ベクトルを求めるという数学的操作をする。すなわち、x,y,z の相関行列 **A** に対して、

$$\mathbf{Ae} = \lambda\mathbf{e}$$

が成り立つような固有値 λ と固有ベクトル **e** を 3 セット求めるのである。具体的にいえば、次の式が成り立つような固有値と固有ベクトルを 3 セット求めるわけである。

$$\begin{bmatrix} 1 & 0.828 & 0.250 \\ 0.828 & 1 & 0.207 \\ 0.250 & 0.207 & 1 \end{bmatrix} \begin{bmatrix} e1_1 \\ e1_2 \\ e1_3 \end{bmatrix} = \lambda_1 \begin{bmatrix} e1_1 \\ e1_2 \\ e1_3 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0.828 & 0.250 \\ 0.828 & 1 & 0.207 \\ 0.250 & 0.207 & 1 \end{bmatrix} \begin{bmatrix} e2_1 \\ e2_2 \\ e2_3 \end{bmatrix} = \lambda_2 \begin{bmatrix} e2_1 \\ e2_2 \\ e2_3 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0.828 & 0.250 \\ 0.828 & 1 & 0.207 \\ 0.250 & 0.207 & 1 \end{bmatrix} \begin{bmatrix} e3_1 \\ e3_2 \\ e3_3 \end{bmatrix} = \lambda_3 \begin{bmatrix} e3_1 \\ e3_2 \\ e3_3 \end{bmatrix}$$

このような式を解き、さらにいろいろ計算すると、表 16 の「主成分分析結果の表」が得られる[20]。固有値の欄にある数値は左から λ_1、λ_2、λ_3 の値である。固有値の合計は 3 になるが、これは変数が 3 つあるからである。中央の灰色部分は主成分負荷量だ。

表 16　主成分分析結果の表

	主成分			共通性
	1(p)	2(q)	3(v)	
音楽	0.937	-0.190	-0.294	1.000
スポーツ	0.925	-0.244	0.290	1.000
ニュース	0.453	0.891	0.016	1.000
固有値	1.94	0.89	0.17	
寄与率	64.64	29.67	5.70	
累積寄与率	64.64	94.30	100.00	

このとき A 氏から E 氏のそれぞれの主成分についての主成分得点は表 17 のようになっている（値は標準得点化されている）。

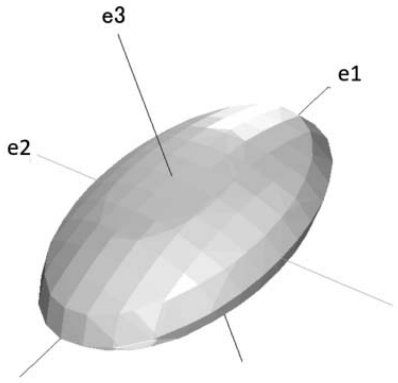
表 17　主成分得点

	主成分		
	1(p)	2(q)	3(v)
A氏	0.869	-0.269	-1.185
B氏	0.962	-0.365	1.464
C氏	-0.516	-1.090	-0.479
D氏	0.114	1.610	-0.220
E氏	-1.429	0.114	0.420

（2）　次元の削減と共通性の変化

上で 3 変数を主成分分析にかけると 3 つの固有値が算出されたのだが、その中の 1 つは 0 に近かった。こんな場合、0 に近い方向の主成分軸でのデータの分散はとても小さいといえる。このデータ分布のイメージは、3 次元空間に浮かぶ「お盆」である。縦横は広がりがあるのだが上下のひろがりはとても小さい（図 24）[21]。

図 24　3 次元空間に浮かぶ平べったいデータ



[20] 以下では統計ソフト SPSS を用いて分析しているので不偏分散が利用されている。

[21] 図 24 はケースの分布の図であり、変数ベクトルの空間の図ではない。前に見た図 17 のような「ケースの分布」の図なのである。

こんなときにはデータの全分散を説明するために 3 次元は必要なく、「お盆」が表現できる 2 次元で十分と考えられる。主成分分析は、このように、データの全分散を少数の主成分で説明するためにしばしば用いられる。

このことを表 16 の主成分分析の結果に適用してみよう。データは 3 変数のものであるから主成分は 3 つ導ける。しかし、2 つの主成分だけでデータの全分散の 94.3%が説明できる。それゆえ、2 つの主成分だけ取り上げるのが妥当だということになる（表 18）[22]。

表 18　削減された主成分分析結果の表

	主成分		共通性
	1(p)	2(q)	
音楽	0.937	-0.190	0.913
スポーツ	0.925	-0.244	0.916
ニュース	0.453	0.891	1.000
固有値	1.94	0.89	
寄与率	64.64	29.67	
累積寄与率	64.64	94.30	

表 18 を表 16 と比べて見るとわかるが、共通性の部分だけ値が異なっている。2 つの主成分しか取り上げられていないので、それぞれの変数の分散は完全には説明されないのである。

次元が削減されても主成分得点は変わらない。表 19 と表 17 を比べてみるといい。

表 19　次元が削減されたときの主成分得点

	主成分	
	1(p)	2(q)
A氏	0.869	-0.269
B氏	0.962	-0.365
C氏	-0.516	-1.090
D氏	0.114	1.610
E氏	-1.429	0.114

5.2　3 変数の分散と負荷量・固有値・共通性

（1）　3 変数の単位ベクトル

3 変数の場合も 2 変数の場合と同様、主成分分析結果の表の負荷量の縦の平方和は固有値、横の平方和は共通性になる。なぜそうなるかは 2 変数の場合と同様である。図をもとに解説していこう。

図 25 には半径 1 の球が描かれており、中心を起点に直交する第 1 主成分ベクトル **p**、第 2 主成分ベクトル **q**、第 3 主成分ベクトル **v** が示されている[23]。**p** ベクトルと **q** ベクトルは球の中心を通る平面を構成している。

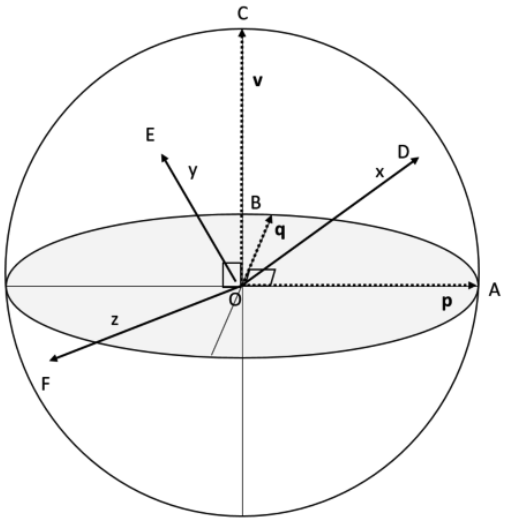
この空間に 3 つの変数ベクトル **x**, **y**, **z** も存在している。どれも長さが 1 に調整された単位ベクトルなので、各ベクトルは球の中心を始点とし球面のどこかを終点とするベクトルである（**z** ベクトルは平面より下方にある球面に向かっている）。

x, y, z それぞれの分散は、それぞれのベクトルの長さの 2 乗、$|OD|^2$, $|OE|^2$, $|OF|^2$ で表現できる。それらはどれも 1 であり、全分散は 3 である。

[22] いくつ主成分を取り上げるかということに関しては「固有値 1 以上を基準」になされることが多いが、別にそうすべき強い理由はない。その理由をしいてあげれば、元の変数の分散（＝1）よりも大きい分散をもつ主成分を取り上げるべきといった程度のことだ。

[23] 図 25 は一般的な表現でありここでのデータの数値に対応するものではない。データの数値に対応するものは後の部分で提示する。

図 25　3 変数の主成分分析の図形的表現



ここで **x** ベクトルに注目すると図 26 のようになる。この図では、角度を表す θ は次のように使われている [24]。

$$\theta_1 = \angle DOP$$
$$\theta_2 = \angle DOQ$$
$$\theta_3 = \angle ODR$$

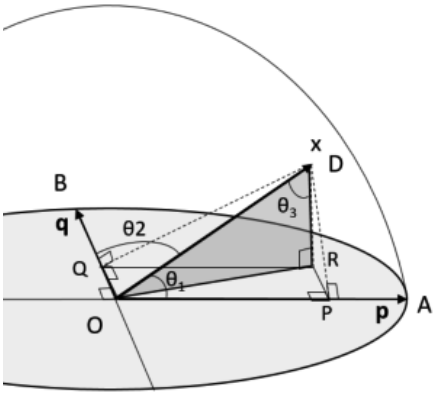
このとき、**p**, **q**, **v** という直交軸でできる空間において、D の座標が次のようになることをまず確認しよう [25]。

$$D\left(\cos\theta_1, \cos\theta_2 \cos\theta_3\right)$$

これは次のようにも表せる。

$$D\left(r_{x,p}, r_{x,q}, r_{x,v}\right)$$

図 26　x の分散の分解



（2）　変数の分散の分解

図 26 より x の分散|OD|² は|RD|² と|OR|² に分解され、さらに|OR|² は|OP|² と|OQ|² に分解されることがわかる。結局次のことが成立する。

$$\left|OD\right|^2 = \left|OP\right|^2 + \left|OQ\right|^2 + \left|RD\right|^2$$

|OP|² は x の **p** 方向の分散の量、|OQ|² は x の **q** 方向の分散の量、|RD|² は x の **v** 方向（天頂方向）の分散の量ということになる。x の分散は 1 なので、|OP|²、|OQ|²、|RD|² はそれぞれ x の **p**, **q**, **v** 方向の分散の割合と表現してもいい。

（3）　共通性と負荷量の平方和

ここで、次のことが成り立っていることに注目しよう。

---

[24] ∠ODR は O から天頂に向かうベクトルと OX ベクトルの作る角に等しい。

[25] x と p の相関（負荷量）は∠DOP の余弦であり∠ROP の余弦ではない。

$$|OP|^2 = \cos^2\theta_1 = r_{x,p}^2$$
$$|OQ|^2 = \cos^2\theta_2 = r_{x,q}^2$$
$$|RD|^2 = \cos^2\theta_3 = r_{x,v}^2$$

これらの相関はxと各主成分との相関（＝負荷量）である。したがって、主成分分析の表の負荷量の横の2乗和は、とりあげられた主成分でその変数の分散が説明される割合になる。この割合が共通性である。3主成分の内、とりあげられる主成分がpとqならば、xの共通性は次のようになる。

$$x:\ |OP|^2 + |OQ|^2 = r_{x,p}^2 + r_{x,q}^2 \quad (\leq 1)$$

以上のことは、変数yやzでも同じであり、結局、第1主成分と第2主成分をとりあげた場合の共通性は、x, y, z で次のようになる。

$$x:\ r_{x,p}^2 + r_{x,q}^2 \quad (\leq 1)$$
$$y:\ r_{y,p}^2 + r_{y,q}^2 \quad (\leq 1)$$
$$z:\ r_{z,p}^2 + r_{z,q}^2 \quad (\leq 1)$$

（4） 固有値と負荷量の平方和

主成分分析の表の負荷量の縦の2乗和が固有値になることについても考え方は同じである。図27はその解説のためのものであり、**pqv** 空間にある **x,y,z** の各変数ベクトルを表している（いずれも長さは1）。
p 方向の x, y, z, の分散は、それぞれ次のように表せる。

$$|OK|^2 = r_{x,p}^2$$
$$|OL|^2 = r_{y,p}^2$$
$$|OM|^2 = r_{z,p}^2$$

そして、これらを足しあげると **p** 方向の分散の総量ということになり、それがpに対応する固有値に一致するのである。

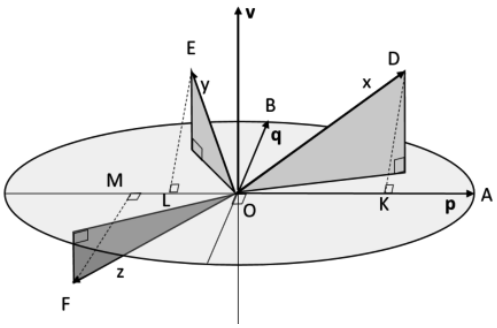
$$\lambda_1 = |OK|^2 + |OL|^2 + |OM|^2 = r_{x,p}^2 + r_{y,p}^2 + r_{z,p}^2$$

このことは、qやvでも同じである。それぞれの主成分で説明できる分散は次のようになる。

$$p:\ \lambda_1 = r_{x,p}^2 + r_{y,p}^2 + r_{z,p}^2$$
$$q:\ \lambda_2 = r_{x,q}^2 + r_{y,q}^2 + r_{z,q}^2$$
$$v:\ \lambda_3 = r_{x,v}^2 + r_{y,v}^2 + r_{z,v}^2$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 3$$

図 27　p 方向の分散の解説図



5.3　単純構造と回転

（1） 主成分の解釈と単純構造

さて、前に収入と財産の2変数をもとにした第1主成分の意味について、われわれの分析者はそれを「経済的豊かさ」を表すと解釈した。では今回の2つの主成分についてはどう解釈できるだろうか（表 20）。おそらく、この解釈はなかなか難しいだろう。

ではどんな場合に主成分の解釈が容易になるかというと、全体の負荷量が表 21 のようになっている場合だ。

表 20　主成分分析結果の表（表 18 再掲）

	主成分		共通性
	1(p)	2(q)	
音楽	0.937	-0.190	0.913
スポーツ	0.925	-0.244	0.916
ニュース	0.453	0.891	1.000
固有値	1.94	0.89	
寄与率	64.64	29.67	
累積寄与率	64.64	94.30	

表 21　単純構造

	主成分		
	1	2	3
変数1	0.90	0	0
変数2	0.90	0	0
変数3	0	0.90	0
変数4	0	0.90	0
変数5	0	0	0.90
変数6	0	0	0.90

ここでは各変数は1つの主成分だけに強くかかわり、その他の主成分とは関係しない。このような負荷量の付置を「単純構造」というが、単純構造になっている場合、第 1 主成分が何を意味するかを解釈するには変数 1 と変数 2 だけに注目すればいい。同様に、第 2 主成分や第 3 主成分についても、特定の変数だけに注目してその主成分の意味を解釈をすればいいのである。

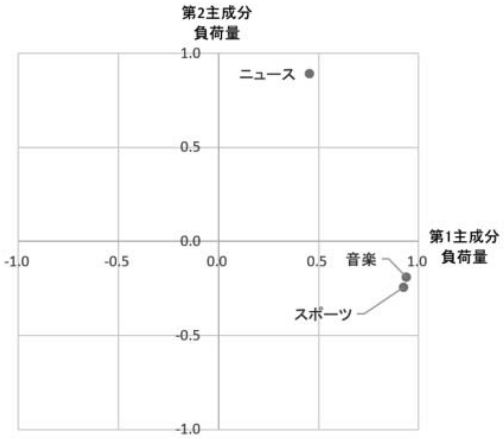
（2） 負荷量のプロット

ではもとの因子負荷量の表をなんとかして単純構造の表にできないか。このために考えられたのが軸の回転ということである。3 変数 2 主成分のわれわれのデータをもとに解説していこう。

まず、表 20 の負荷量を図にプロットしてみる。その結果描けるのが図 28 である。横軸と縦軸はそれぞれ第 1 主成分負荷量と第 2 主成分負荷量の値を示しており、そこに音楽、スポーツ、ニュースといった各変数が位置付けられる。

図を見ると、音楽とスポーツが第 1 主成分に高く負荷しているが、ニュースもそこそこ第 1 主成分に負荷しているので解釈しにくいことがわかる。

図 28　主成分負荷量のプロット図

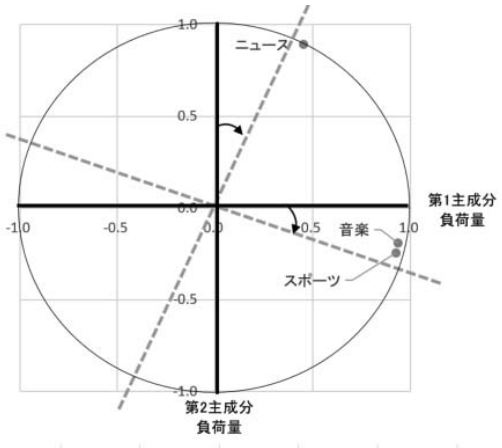


（3） 軸の回転

ここで行われるのが負荷量の軸の回転である。音楽、スポーツ、ニュースの負荷量の点は動かさず、直交性を保ちつつ矢印方向に向けて 2 つの軸を回転させると、元の軸は図 29 の点線の 2 軸になる[26]。

[26] 回転の方法はいろいろあるが、軸（主成分ベクトル）の直交性を保ったまま回転させる方法の代表としてバリマックス回転がある。軸の回転に直交性を要求しない斜交回転というものもある。

図 29　回転前の各変数の負荷量のプロット



（4）　回転後の負荷量のプロット

回転した軸をあらたな負荷量の軸として書き直したものが図 30 である。また、この結果を表にしたものが表 22 である。

図 30　回転後の各変数の負荷量のプロット

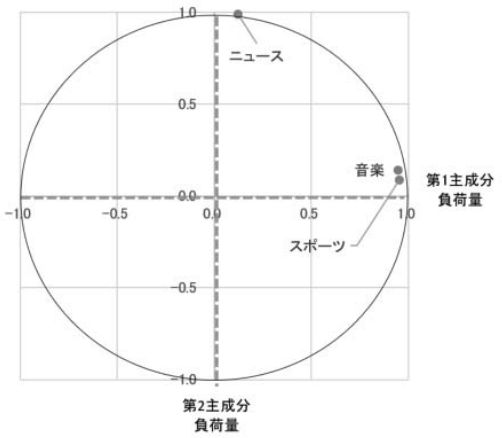


表 22　　回転後の主成分分析結果の表

	主成分		共通性
	1(p)	2(q)	
音楽	0.945	0.142	0.913
スポーツ	0.953	0.088	0.916
ニュース	0.120	0.993	1.000
負荷量平方和	1.82	1.01	
寄与率	60.52	33.78	
累積寄与率	60.52	94.30	

回転によって音楽やスポーツは第 1 主成分への負荷量が高く、第 2 主成分への負荷量は低くなっており、スポーツは第 2 主成分への負荷量が高く、第 1 主成分への負荷量が低くなっていることがわかる。すなわち、単純構造に近づいているのである。それゆえ主成分の解釈が容易になる。

以上が回転についてのよくある解説なのだが、これだけではここで行われていることの意味を理解できない読者も多いと思う。そもそもプロットされた負荷量の点はどのような意味をもつのか。ここでの軸はそもそも回転させていいような軸なのか、回転させた後の座標を新しい負荷量と考えていいのか、そういった疑問は生じて当然であるし生じないのはむしろおかしい。そこで以下では今回のデータをもとにもう少し詳しく解説することにしよう。

5.4　回転の意味と尺度構成

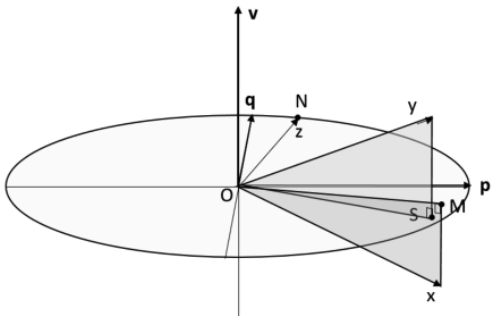
（1）　変数と主成分のベクトルの全体像

まず、今回のデータの主成分分析結果についてもう一度見ておこう（表 23）。この表から変数ベクトルや主成分ベクトルを空間的に表したものが図 31 である。

表 23　主成分分析結果の表（表 16 再掲）

	主成分			共通性
	1(p)	2(q)	3(v)	
音楽	0.937	-0.190	-0.294	1.000
スポーツ	0.925	-0.244	0.290	1.000
ニュース	0.453	0.891	0.016	1.000
固有値	1.94	0.89	0.17	
寄与率	64.64	29.67	5.70	
累積寄与率	64.64	94.30	100.00	

図 31　主成分ベクトルと変数ベクトル



図では直交する **p**, **q**, **v** のベクトルで構成される空間に、**x**（音楽）、**y**（スポーツ）、**z**（ニュース）のベクトルが位置付けられている。変数ベクトルはすべて長さ 1 の単位ベクトルである。

表 23 から、第 1 主成分 p は、音楽 x、スポーツ y、ニュース z のいずれとも正の相関関係にあることがわかる。したがって図 31 では、**x**, **y**, **z** ベクトルはいずれも **p** と 0～90 度の関係にある。

表 23 の第 2 主成分 q は、音楽 x とスポーツ y と負に、ニュース z と正に相関している。したがって図 31 では、**x**, **y** ベクトルと **q** の作る角度はいずれも 90～180 度の範囲にあり、**z** ベクトルと **q** の作る角度は 0～90 度の範囲にある。

表 23 の第 3 主成分 v は、音楽 x とは負に相関し、スポーツ y とは正に相関している。したがって図 31 では、**x** は空間下部に向かい、**y** は空間上部に向かう。第 3 主成分 v とニュース z の相関はほぼ 0 なので **z** ベクトルは **v** とほぼ直交し、**pq** 平面にあるということになる[27]。これら全体のことを考慮して図 31 は描かれたわけである。

（2） 因子負荷量のプロットの意味

ところで、図 31 の **x**, **y**, **z** ベクトルの先端の座標を X, Y, Z で表すとき、それらの座標は次のようになる（このことは 5.2(1)で述べた）。

$$X\left(r_{x,p}, r_{x,q}, r_{x,v}\right)$$
$$Y\left(r_{y,p}, r_{y,q}, r_{y,v}\right)$$
$$Z\left(r_{z,p}, r_{z,q}, r_{z,v}\right)$$

図 31 には M, S, N という点が小さな黒丸で表現されている。これらは **x**, **y**, **z** ベクトルの先端から **pq** 平面に下した垂線の足をそれぞれ示している。これら M, S, N の **pq** 平面上の座標は、次のようになる。

$$M\left(r_{x,p}, r_{x,q}\right)$$
$$S\left(r_{y,p}, r_{y,q}\right)$$
$$N\left(r_{z,p}, r_{z,q}\right)$$

以上より、図 28 の因子負荷量のプロット平面と図 31 の関係が見えてくる。実は、図 28 の平面は図 31 の平面と同じものなのである。図 28 にある変数の因子負荷量を示す軸は主成分ベクトルの軸と考えることができる。各変数と主成分の相関が負荷量であるから、図 28 にプロットされた各点は、図 31 でいえば各変数ベクトルの先端から平面に下した垂線の足 M, S, N である。

[27] このことは表 18 においてニュースの共通性が 1 であることからも確かめられる。

（3）　単純構造に向けての回転

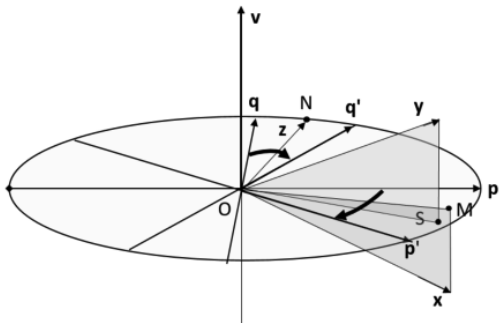
単純構造に向けての軸の回転とは、簡単にいえば、空間にある変数ベクトルはそのままにして、取り上げられた主成分で構成される平面の円盤を回転させることだ（図 32）[28]。
回転後の **p'q'v** 空間における X, Y, Z 座標は次のようになる。

$$X\left(r_{x,p'}, r_{x,q'}, r_{x,v}\right)$$
$$Y\left(r_{y,p'}, r_{y,q'}, r_{y,v}\right)$$
$$Z\left(r_{z,p'}, r_{z,q'}, r_{z,v}\right)$$

したがって、平面上の M, S, N の座標は次のようになる。

$$M\left(r_{x,p'}, r_{x,q'}\right)$$
$$S\left(r_{y,p'}, r_{y,q'}\right)$$
$$N\left(r_{z,p'}, r_{z,q'}\right)$$

図 32　回転のベクトル的表現



変数と主成分の相関は負荷量である。したがって回転後の図 30 の負荷量の 2 つの軸は、回転後の主成分ベクトルの方向をそれぞれ表しており、プロットされている点は新しい主成分ベクトル軸での M, S, N の座標ということになる。

（4）　回転前後の変化

以上が回転についてのより詳しい解説である。この解説をもとに、回転前後でどのような変化が生じるかを見てみよう。表 24 には回転前後の主成分分析結果の表がともに示されている。ここから次のことがわかる。

表 24　回転前後の主成分分析結果の表

		主成分		共通性
		1(p)	2(q)	
回転前	音楽	0.937	-0.190	0.913
	スポーツ	0.925	-0.244	0.916
	ニュース	0.453	0.891	1.000
	固有値	1.94	0.89	
	寄与率	64.64	29.67	
	累積寄与率	64.64	94.30	
		主成分		共通性
		1(p)	2(q)	
回転後	音楽	0.945	0.142	0.913
	スポーツ	0.953	0.088	0.916
	ニュース	0.120	0.993	1.000
	負荷量平方和	1.82	1.01	
	寄与率	60.52	33.78	
	累積寄与率	60.52	94.30	

(1) 主成分負荷量の付置が回転後、単純構造に近づいている。単純構造をめざしているのだから当然それに近づくことになる。スポーツを例として変化を図で解釈すると、回転前のスポーツのベクトル **y** と第 2 主成分のベクトル **q** の角度が 90 度以上（104 度）あったものが、回転によって両ベクトル間の角度が 90 度近く（85 度）になったということになる [29]。

[28] 軸の回転で問題とされるのはケースを位置付ける主成分軸ではなく主成分ベクトル方向の軸である。「主成分軸」という用語が使われている場合、どちらの意味で用いられているかについては注意が必要だ。

[29] 3.3(１)の表 7 参照。これらの角度は図 29・図 30 の「負荷量の軸」と「負荷量の点と原点を結ぶ線」が平面上で作る角

(2) 各変数の共通性の値は回転後も変わらない。図形的にいうと平面にある2主成分を回転させただけだから、各変数についての2つの主成分を合わせた分散の説明割合は変化しない。説明されないのはどの変数においても垂直な**v**方向の分散であり、そこには変化はないのだから、残された**pq**平面で説明できる分散の割合は変化しないのである。

(3) 回転前の固有値に対応するところに回転後は「負荷量平方和」が示されている。負荷量平方和とは、回転後の主成分方向のデータの分散のことだ[30]。この値は回転前の固有値と異なっており、寄与率や累積寄与率の値も変わっている。

回転前の2主成分で説明されていた分散はそれぞれの固有値に対応しており、全分散3のうちの1.94（第1主成分）と0.89（第2主成分）だった。それらは全分散の64.64%と29.67%を説明するものだ。回転によって全分散3の内、第1主成分では1.82、第2主成分では1.01が説明されることになった。それらは全分散の60.52%と33.78%である。これらの変化は各変数ベクトルに対する2つの主成分ベクトル **p**, **q** の方向が変わったことによって生じている。

2主成分を合わせた分散の総量は回転前後で変化がない（1.94+0.89=2.83, 1.82+1.01=2.83）。したがって、累積寄与率にも変化はない（94.3%）。2主成分で平面が構成されているとき、全分散の内この平面で説明される分散の量や割合は回転後も変化しない。全分散について説明されないのは垂直の**v**方向の分散であり、ここには変化はないので、残された**pq**平面で説明できる分散の量には変化はないのである。

（5） 主成分の解釈と尺度構成

さて、回転後の負荷量をもとにすると、それぞれの主成分についての解釈が容易になる。たとえば、回転後の第1主成分は音楽やスポーツに高く相関していることから「感情充足的な視聴」を意味しており、第2主成分はニュースに高く相関していることから「情報獲得的な視聴」を意味している、などといった解釈が可能になるのである。

こういった解釈をもとに、さらにここから「感情充足視聴尺度」や「情報獲得視聴尺度」というものも考えることができる。すなわち、回転後の第1主成分得点を「感情充足視聴尺度」の得点とし、第2主成分得点を「情報獲得視聴尺度」の得点とするのである。回転前後で主成分得点は異なる（表 25）。

表 25　回転前後の主成分得点の表

	回転前		回転後	
	主成分		主成分	
	1	2	1	2
A氏	0.869	-0.269	0.908	0.046
B氏	0.962	-0.365	1.029	-0.013
C氏	-0.516	-1.090	-0.111	-1.201
D氏	0.114	1.610	-0.445	1.552
E氏	-1.429	0.114	-1.381	-0.383

主成分分析を用いた尺度構成にはいくつかのバリエーションがある。たとえば多変数を主成分分析にかけて第1主成分に高く負荷している諸変数をとりあげ、それらについて再度主成分分析に

度ではなく、図 32にあるような「平面にある主成分軸」と「空間に向かう変数ベクトル」の作る角度だということに注意すること。

[30] 固有値という用語はもとの相関行列に基づくものだから厳密にいうと回転後の表には記載できない。そういうわけでここには負荷量平方和と記されている。主成分方向の分散の量という意味は同じである。

かけて、その第 1 主成分得点を 1 つの尺度にするといったことはよくなされる。

また、ここでの例のように 1 つの主成分分析をもとに複数の尺度を構成することもある。このときの主成分分析は、回転を伴う場合も伴わない場合もある。回転しない場合は尺度間の相関は 0 になる。回転する場合も通常主成分ベクトル間の直交性を保ったままの回転が一般的であり（バリマックス回転）、そのようなときには尺度間の相関は 0 になる [31]。

最後に注意点を 1 つ。主成分分析によって作られた尺度は、分析に用いられた全変数に関連している。たとえば上の例のような場合、感情充足視聴尺度には音楽とスポーツの得点だけが関連しているわけではなく、音楽、スポーツ、ニュースの得点すべてが関連しているのである。ただし、主成分得点を算出する際の重みは変数ごとに異なる。感情充足視聴尺度では、音楽とスポーツの重みは大きくニュースの重みは小さいのだ。情報獲得視聴尺度も同様である。この尺度はすべての変数に関連しているのだが、ニュースに重みを置いて構成されているのである。

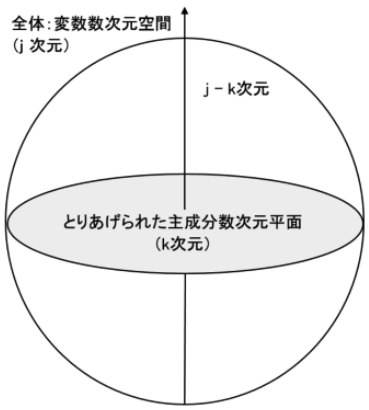
5.5　変数や主成分がさらに多い場合

以上で 3 変数の主成分分析についての解説は終わりである。解説は最も単純な条件の下での例をもとになされてきた。すなわち、たかだか 5 人から得られた 3 変数のデータから 3 主成分を導き、それを 2 主成分に縮減してデータの分布を考えるという例を用いた解説である。このような例を用いたのは、ここまで簡単にしてはじめて主成分分析が行っていることを絵に描いて表現することができるからである。

実際の分析においては、ケースの数はずっと多く、変数やとりあげる主成分の数も今回の例よりも多いのが普通である。このような場合はこれまで解説してきた原理を次のように拡大解釈して理解すればいい。

一般に、j 個（j＞4）の変数について k 個（k＞3）の主成分で見ていくような場合（j＞k）、j 次元に存在するデータを k 次元平面で考えていくということになる。これを図に描いたのが図 33 である。

図 33　多変数の主成分分析での次元縮減イメージ



図にある平面は、分析でとりあげる主成分で構成される k 次元平面であり、とりあげた主成分では説明できない部分が j−k 次元の上部下部の空間になる。

たとえば、500 人から得られた 10 変数のデータがあるとする。これを主成分分析にかけ、10 変数を 4 主成分から解釈するといった場合、それぞれ 500 の要素からなる 10 の変数ベクトルを 10 次

[31] 相互に相関しない尺度を構成する必要がある場合、この特徴は役に立つ。

元空間に描き、それらの影を 4 次元平面に映して考えるということになる。説明されないのは 6(＝10－4）次元の部分である。

このような空間や平面はもちろん絵に描くことはできない。それは実際の空間ではなく数学的にのみ想定できる空間である。しかしながら考え方は同じである。基本的なところがきちんと理解できていれば、多変数、多主成分の分析についてもイメージがわくはずである。

5.6 問題と解答

（1） 問題

(a)

・計算して下の主成分分析の表を完成させよ。

・ニュースの共通性が 1 であることの意味を述べよ。

・ニュースのベクトルと主成分ベクトルの位置関係について述べよ。

	主成分		共通性
	1	2	
音楽	0.937	-0.190	a
スポーツ	0.925	-0.244	b
ニュース	0.453	0.891	1.000
固有値	c	d	
寄与率	e	f	
累積寄与率	g	h	

(b)

・上の表は主成分分析の回転前の表である。回転後（直交回転）、a〜h の値はどうなるか。それぞれ、増加する、減少する、変化しないから適切なものを選べ（回転後の固有値は負荷量平方和と考えること）。

（2） 解答

(a)

・表 18 参照。

・ニュースの分散を説明するには第 1 主成分と第 2 主成分をとりあげるだけで十分であり、それ以外の要素である第 3 主成分は説明に不要である。

・ニュースのベクトルは 2 つの主成分ベクトルで構成される平面上にある。

(b)

・増加する：d, f

・減少する：c, e, g

・変化しない：a, b, h

6　おわりに

社会調査から得られたデータに対してなされる多変量解析の代表的なものに重回帰分析と主成分分析がある。前者は多数の変数から 1 つの変数を説明するために利用され、後者は多数の変数をまとめて尺度を構成したり、多数の変数を分類したりするのによく用いられる。それゆえ、2 つの分析法はまったく異なるもののように見える。

しかしながら、その方法の内実を探ると、それらには同じような発想があることがわかる。すなわち、データの分散の分解である。重回帰分析ではある変数の分散を別の変数の分散に分解しようとしており、主成分分析は全変数の分散、すなわち全分散を分解しようとしているのである。

背後にこういった発想があることを知ることは、重要であるとともに楽しいことでもある。こういった楽しみを感じつつ、さらなる分析法を理解し、その方法を使って実際に社会を分析し、さまざまなことを発見し、そこでまた楽しみを感じるといった方向に、読者が進んでいくことを期待している。楽しくなけりゃ学問じゃないのである。

文献

小林久高, 2018a「母集団・標本・確率変数」『同志社社会学研究』22.
————, 2018b「離散型確率変数とその分布」『同志社社会学研究』22.
————, 2018c「連続型確率変数とその分布」『同志社社会学研究』22.
————, 2019a「統計的仮説検定の原理と実際」『同志社社会学研究』23.
————, 2019b「統計的推定の原理と実際」『同志社社会学研究』23.
小林久高・山本圭三, 2020「線形回帰と相関：社会調査データの多変量解析 (1)」『同志社社会学研究』24:.55-118.
永田靖・棟近雅彦, 2001『多変量解析法入門』サイエンス社.