

# Rと主成分分析

## 1. 主成分分析とは

観測、実験、調査では、通常個体の属性を複数の項目（変数）に分けて記録する。変数が少ない場合は、簡単なグラフや基本統計量などでデータの構造を明らかにすることができるが、変数が多くなるとデータの構造が複雑になり、解析が難しくなる。一方、変数が多くなると変数の間には相関がある可能性も増える。

主成分分析(principal component analysis)は、多くの変数により記述された量的データの変数間の相関を排除し、できるだけ少ない情報の損失で、少数個の無相関な合成変数に縮約して、分析を行う手法である。主成分分析の手法はホテリング(Hotelling)によって 1933 年頃提案された。

変数が 1 つ、2 つの場合は、棒グラフや散布図でデータの構造を読み取ることが可能であり、主成分分析を行う必要はないが、主成分分析の考え方を説明するため、ここでは 2 変数の場合の例を用いることにする。

たとえば、表 1 の左側に示す 3 つの個体の 2 次元データ ( $x_1, x_2$ ) があるとする。その  $x_1$  を横軸、 $x_2$  を縦軸にした座標系上の散布図を図 1 (a) に示す。図 1 (a) の個体は、座標軸  $x_1$  の値  $x_{1i}$  と座標軸  $x_2$  の値  $x_{2i}$  ( $i=1,2,3$ )、つまり 2 次元で表記しなければならない。しかし、この座標系を図 1 (b) のような  $z_1, z_2$  の座標系に変換すると  $z_1$  変数のみで表すことができる。

表 1 2次元データの例

	$x_1$	$x_2$	変換	$z_1$	$z_2$
個体 1	1	2	⇒	2.236	0.000
個体 2	2	4	⇒	4.472	0.000
個体 3	3	6	⇒	6.708	0.000
分散	1	4	⇒	5.000	0.000
相関係数	1		⇒	0	

図 1 座標の変換

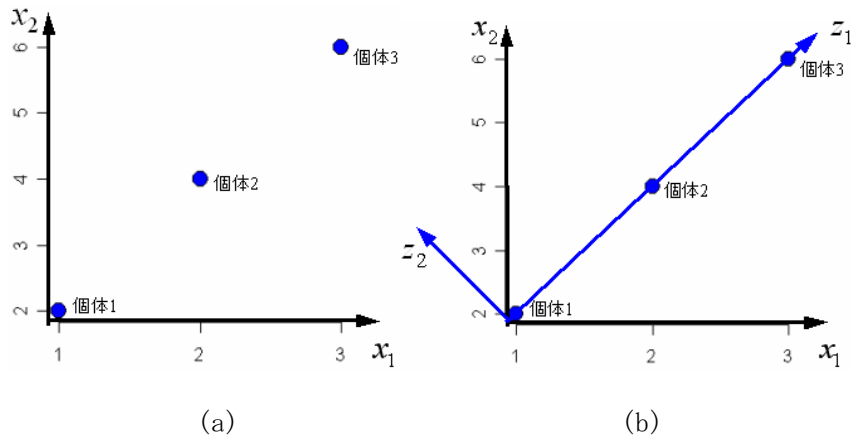


図 1 (b) 中の新しい座標  $z_1$  と  $x_1, x_2$  との関係は次の式(合成変数、あるいは線形結合式と呼ぶ)で表すことができる。

$$z_1 = 0.447x_1 + 0.894x_2$$

この合成変数は、上記の 3 つの個体に限っては情報の損失なしで、表 1 の 2 次元 ( $x_1, x_2$ ) データを 1 次元 ( $z_1$ ) に縮約することができる。これは、 $x_1$  と  $x_2$  の相関関係が非常に強く、その相関係数が 1 であるからである。

上記の合成変数  $z_1$  に個体 1、個体 2、個体 3 の  $x_1, x_2$  の値を代入すると、表 1 の右側に示す  $z_1$  の値が得られる。図 1 (b) に示すように、直線上に全ての点に乗ると、 $z_2$  の値は全てゼロになるので合成変数  $z_2$  は何ら情報を持っておらず、分析には役立たない。ここで言う情報とはデータのバラツキ(分散)を意味し、分散が大きいほど、情報を多く持っていると考えられる。表 1 で分かるように、 $z_1$  の分散は  $x_1, x_2$  の分散より大きい。点が完全に直線に乗る場合は情報の損失がないが、そうではない場合は情報の損失が生じる。その際には、より多くの情報を用いて分析を行うためには他の合成変数(たとえば  $z_2$ ) を付け加えて分析すべきである。

上記の具体的な式を一般化すると  $z_1 = a_1x_1 + a_2x_2$  となる。式の中の  $x_1, x_2$  は用いるデータであり、係数  $a_1, a_2$  は用いた既知のデータから求める。主成分分析では、合成変数(線形結合式)の係数を主成分と呼ぶ。主成分分析での主な問題は、いかにデータを縮約する係数(主成分)を求めるかである。

通常用いる主成分分析は、データの分散が最大になるように線形結合式の係数を求める方法と相関が最大になるように線形結合の係数を求める方法がある。

すでに説明した 2 変数の線形結合式を一般化するため、 $m$  個の個体、 $n$  個の変数 ( $x_1, x_2, \dots, x_n$ ) により構成された表 2 のようなデータセットがあり、これを  $X_{m \times n}$  で表す。

表2 データ ( $X_{m \times n}$  と記する)

	$x_1$	$x_2$	...	$x_j$	...	$x_n$
個体 1	$x_{11}$	$x_{12}$	...	$x_{1j}$	...	$x_{1n}$
個体 2	$x_{21}$	$x_{22}$	...	$x_{2j}$	...	$x_{2n}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
個体 $i$	$x_{i1}$	$x_{i2}$	⋮	$x_{ij}$	⋮	$x_{in}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
個体 $m$	$x_{m1}$	$x_{m2}$	...	$x_{mj}$	...	$x_{mn}$

この  $n$  次元のデータをより低い  $k$  ( $k \leq n$ ) 次元に縮約する線形結合の一般式を次に示す。

$$\begin{aligned}
 z_1 &= a_{11}x_1 + a_{21}x_2 + \dots + a_{n1}x_n \\
 z_2 &= a_{12}x_1 + a_{22}x_2 + \dots + a_{n2}x_n \\
 &\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
 z_j &= a_{1j}x_1 + a_{2j}x_2 + \dots + a_{nj}x_n \\
 &\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
 z_k &= a_{1k}x_1 + a_{2k}x_2 + \dots + a_{nk}x_n
 \end{aligned}$$

この線形結合式  $z_1$  に用いる係数を第 1 主成分、 $z_2$  に用いた係数を第 2 主成分と呼ぶ。説明の便利のため、係数データを表 3 のように並べ、これを  $A_{n \times k}$  で示す。

表3 係数データ ( $A_{n \times k}$  と記する)

	主 成 分					
	第 1	第 2	...	第 $j$	...	第 $k$
$x_1$ の係数	$a_{11}$	$a_{12}$	...	$a_{1j}$	...	$a_{1k}$
$x_2$ の係数	$a_{21}$	$a_{22}$	...	$a_{2j}$	...	$a_{2k}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$x_i$ の係数	$a_{i1}$	$a_{i2}$	⋮	$a_{ij}$	⋮	$a_{ik}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$x_n$ の係数	$a_{n1}$	$a_{n2}$	...	$a_{nj}$	...	$a_{nk}$

データ  $X_{m \times n}$  と係数  $A_{n \times k}$  を線形結合式で求めた値  $z_j$  を主成分得点と呼ぶ。主成分得点のデータ ( $Z_{m \times k}$ ) を表 4 に示す。主成分得点  $Z_{m \times k}$  とデータ  $X_{m \times n}$ 、係数  $A_{n \times k}$  との関係は、 $Z_{m \times k} = X_{m \times n} A_{n \times k}$  (行列の演算) となる。

表4 主成分得点 ( $Z_{m \times k}$  と記する)

	主 成 分 得 点 ( $Z$ )					
	$z_1$ 第 1	$z_2$ 第 2	...	$z_j$ 第 $j$	...	$z_k$ 第 $k$

個体 1	$z_{11}$	$z_{12}$	...	$z_{1j}$	...	$z_{1k}$
個体 2	$z_{21}$	$z_{22}$	...	$z_{2j}$	...	$z_{2k}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
個体 $i$	$z_{i1}$	$z_{i2}$	⋮	$z_{ij}$	⋮	$z_{ik}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
個体 $m$	$z_{m1}$	$z_{m2}$	...	$z_{mj}$	...	$z_{mk}$

分散(相関)を最大にする方法で主成分を求めることは、データの分散共分散行列(相関係数行列)の固有値と固有ベクトルを求める問題に帰する。その固有ベクトルが主成分であり、固有値の大小がそれに対応する固有ベクトル(主成分)に含まれる情報の多少を決める。

データ  $X_{m \times n}$  の分散共分散行列(あるいは相関係数行列)の固有値は通常  $\lambda_i$  で表記する。固有値を求めるアルゴリズムは、固有値を大小の降順に並べて返す。

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_j \geq \dots \geq \lambda_k \geq 0$$

固有値は主成分得点の標準偏差の 2 乗に等しい。この値が大きい主成分(固有ベクトル)ほど、元のデータの情報を多く含んでいる。最も大きい固有値に対応する主成分を第 1 主成分、その次に大きい固有値に対応する主成分を第 2 主成分と呼ぶ。

ある主成分にデータ全体の情報がどれくらい含まれているかは、その主成分に対応する固有値(標準偏差)が固有値全体(標準偏差全体)の中でどれだけの割合を占めているかで説明できる。各固有値が固有値全体に占める割合を寄与率、その寄与率を累積したものを累積寄与率と呼ぶ。

主成分分析を行う際には、寄与率が大きい少数個の主成分を用いる。つまり、 $q$  個の主成分にデータ全体の何割の情報が含まれているかは、第  $q$  主成分までの累積寄与率を用いて説明する。

## 2. R の主成分分析関数

R には主成分分析を行う関数 **prcomp** と **princomp** がある。ここでは、prcomp についてデータを用いて説明する。princomp の基本的な使用方法と機能は prcomp とほぼ同じである。

主成分分析の手法を用いて、データを縮約した場合、データの構造の再現性について考察するため、人工データを用いることにする。

たとえば、図 2 のように 2 次元平面状に、2 つの同心円周上に点が散布されているとする。内側の点 1~16 は半径が 5cm である円周上に等間隔に、点 A、B、C、D、E は半径が 10cm である円弧上に等間隔に並んでいる。A、B、C、D、E を観測点とし、各点 1~16 までの直線距離を測ると表 5 のようなデータが得られる(丘本 1992)。このデータを丘本の円周データと呼ぶことにする。

図 2 丘本の円周データ図

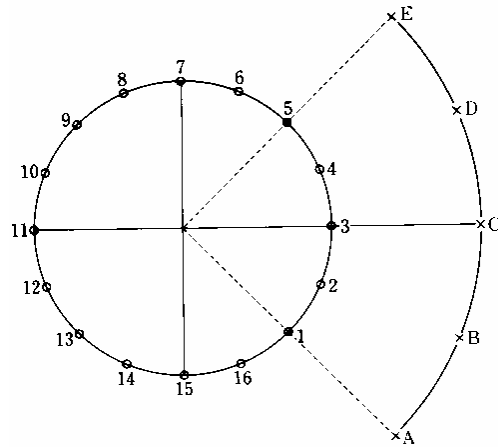


表 5 丘本の円周データ

個体	A	B	C	D	E
1	50	57	74	94	112
2	57	50	57	74	94
3	74	57	50	57	74
4	94	74	57	50	57
5	112	94	74	57	50
6	128	112	94	74	57
7	140	128	112	94	74
8	147	140	128	112	94
9	150	147	140	128	112
10	147	150	147	140	128
11	140	147	150	147	140
12	128	140	147	150	147
13	112	128	140	147	150
14	94	112	128	140	147
15	74	94	112	128	140
16	57	74	94	112	128

この例は、人工データであるためやや面白味に欠けるが、主成分分析による縮約されたデータから、元のデータ構造の再現性を説明するには都合が良い。

ここでは、5つの変数により記述されている円周上の16個の点、円弧上の5個の点に関するデータを主成分分析で2変数に縮約した場合、円周、円弧上の点の相対位置がどの程度まで再現できるかを考察することを主な目的とする。まず次のようにデータオブジェクトを作成する。

```
>temp<-c(50,57,74,94,112,128,140,147,150,147,140,128,112,94,74,57,57,50,57,74,94,112,128,
140,147,150,147,140,128,112,94,74,74,57,50,57,74,94,112,128,140,147,150,147,140,128,112,9
4,94,74,57,50,57,74,94,112,128,140,147,150,147,140,128,112,112,94,74,57,50,57,74,94,112,12
8,140,147,150,147,140,128)
>okamoto<-matrix(temp,16,5,byrow=F)
>colnames(okamoto)<-c("A","B","C","D","E")
```

主成分分析の関数 `prcomp` の使用方法は簡単で、用いる引数も少ない。デフォルトでは、分散共分散を用いる主成分になっている。作成したデータ `okamoto` の主成分の要約を次に示す。

```
>oka.pc<-prcomp(okamoto)
```

```
>summary(oka.pc)
```

```
Importance of components:
      PC1      PC2      PC3      PC4      PC5
Standard deviation 68.002 40.815 3.83644 1.30318 0.30578
Proportion of Variance 0.733 0.264 0.00233 0.00027 0.00001
Cumulative Proportion 0.733 0.997 0.99972 0.99999 1.00000
```

ソフトによっては、主成分分析の結果として固有値を返すが `prcomp` ではその代わりに標準偏差 (`standard deviation`) を返す。標準偏差は、各主成分得点の標準偏差で、固有値の正の平方根に等しい。元の変数が  $n$  個あると、非ゼロである固有値および主成分(固有ベクトル)は  $k$  ( $k \leq n$ ) 個求まる。主成分分析の目的は、できるだけ少ない主成分に、もとの変数の情報を吸収することにある。実際主成分分析を行う際には、用いる第  $q$  番目の主成分までに、もとの変数の情報がどれだけ吸収できたかが問題となる。そのために、寄与率と累積寄与率に関する情報が必要となる。

返された結果の中の `Proportion of Variance` は、各標準偏差が標準偏差の合計に占める割合(寄与率)で、`Cumulative Proportion` は累積寄与率である。返された結果から分かるように第 2 主成分までの累積寄与率が 0.997(99.7%)で、5次元データの情報のほとんどが、第 1、2 主成分に縮約されている。

関数 `prcomp` で計算された主成分(固有ベクトル)は `$rotation` に、列を単位に記録する。主成分は、変数の相互関係を分析するのに用いる。丘本円周データの主成分を次に示す。

```
>oka.pc$rotation
```

```
      PC1      PC2      PC3      PC4      PC5
A 0.3648367 6.201120e-01 0.5837693 -3.397958e-01 -0.1615786
B 0.4805596 3.397958e-01 -0.1641399 6.201120e-01 0.4920575
C 0.5214531 6.070177e-17 -0.5143375 3.293264e-17 -0.6808404
D 0.4805596 -3.397958e-01 -0.1641399 -6.201120e-01 0.4920575
E 0.3648367 -6.201120e-01 0.5837693 3.397958e-01 -0.1615786
```

主成分分析では、情報がより多く含まれている上位のいくつかの主成分を用いる。分析にはこれらの棒グラフや散布図が多く用いられている。

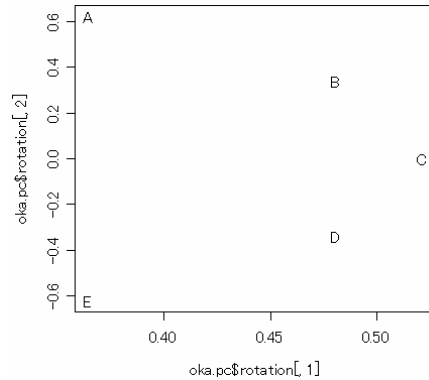
円周データでは第 2 主成分までの累積寄与率が 99.7%に上るので、図 3 に第 1、2 主成分の散布図のみを示す。

```
>plot(oka.pc$rotation[,1],oka.pc$rotation[,2],type="n")
```

```
>text(oka.pc$rotation[,1],oka.pc$rotation[,2],colnames(okamoto))
```

この図 3 を図 2 と比べると、主成分の散布図円弧が若干変形されたものの、図 3 を、横軸を軸とし 180 度回転すると A、B、C、D、E の相対位置は元の点の位置とほぼ一致し、元のデータの構造が再現されていると言える。

図 3 第 1、2 主成分の散布図



主成分が、線形結合式の合成変数を求める係数であるので、用いたデータと主成分の結果で主成分得点を求めることは可能であるが、通常の主成分分析のソフトパッケージは、主成分得点を返す機能を揃えていない。関数 `prcomp` では、主成分の得点は、`$x` に列単位で記録する。データ `okamoto` の主成分得点を次に示す。主成分得点は、個体のデータ構造に関する情報を縮約したものである。

```
>oka.pc$x
      PC1      PC2      PC3      PC4      PC5
[1,] -85.348592 -5.101938e+01 -2.879591 -1.876805e+00 -0.13974065
[2,] -91.201607 -3.109924e+01  3.874481 -2.310244e+00 -0.07364076
[3,] -100.751885 -1.232348e-14  7.364914  1.332268e-15  0.25640318
[4,] -91.201607  3.109924e+01  3.874481  2.310244e+00 -0.07364076
[5,] -85.348592  5.101938e+01 -2.879591  1.876805e+00 -0.13974065
[6,] -29.708700  5.694019e+01 -5.484544 -5.612449e-01 -0.25084539
[7,]   7.557885  5.248045e+01 -3.722346 -1.342714e+00  0.52231640
[8,]  40.168493  4.238022e+01 -1.114172 -6.480407e-01  0.02797145
[9,]  65.140371  3.002037e+01  1.197715 -1.130112e+00 -0.21794250
[10,]  80.741814  1.518008e+01  2.724255 -2.550000e-01  0.29651471
[11,]  86.052595  8.548717e-15  3.443529 -7.549517e-15 -0.58566967
[12,]  80.741814 -1.518008e+01  2.724255  2.550000e-01  0.29651471
[13,]  65.140371 -3.002037e+01  1.197715  1.130112e+00 -0.21794250
[14,]  40.168493 -4.238022e+01 -1.114172  6.480407e-01  0.02797145
[15,]   7.557885 -5.248045e+01 -3.722346  1.342714e+00  0.52231640
[16,] -29.708700 -5.694019e+01 -5.484544  5.612449e-01 -0.25084539
```

さて、各個体（図 2 に示された 16 個の点）の構造を合成変数で縮約した主成分得点で再現するため、次のように第 2 主成分までの主成分得点の散布図を作成してみる。

```
>plot(oka.pc$x[,1],oka.pc$x[,2],type="n")
>text(oka.pc$x[,1],oka.pc$x[,2],1:16)
```

図 4 第 1, 2 主成分得点散布図

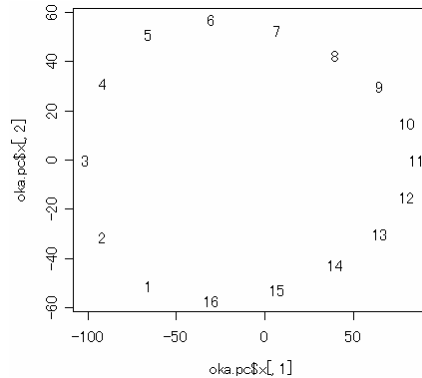


図4で分かるように、縦軸を軸とし180度回転すると円形が若干変形されたものの、16個の点の相対位置は図2と変わらない。

ここでは5変数(次元)のデータを2変数に縮約している。縮約した2次元のデータは、多少歪みはあるものの、元のデータ構造をある程度再現している。

主成分分析は、変数が多いとき情報の損失を最小限に押さえながら少ない合成変数に縮約する方法であるため、元のデータ構造が100%正確に再現できないことと、再現されているのは元の変数、または個体間の相対的な関係に過ぎないことを強調しておきたい。

主成分分析を行うとき、いくつの合成変数(主成分)を用いて分析するのが適切であるかに関しては明確な決まりがない。

分散共分散行列を用いる場合は、一般的には累積寄与率70%~80%を大まかな目安とし、累積寄与率がこれを超える主成分まで用いて分析をすることが多い。

相関行列を用いた主成分分析の場合は、固有値の値が1前後になる主成分まで用いるのが1つの目安である。

関数 `prcomp` には引数 `scale` があり、データの標準化(データのスケールを統一)が必要なときは `scale=TRUE` を指定する。デフォルトには `scale=FALSE` になっている。`scale=TRUE` にすると元のデータの相関行列を用いた主成分に等しい。これは関数 `princomp` の引数 `cor = TRUE` の結果と対応する。

関数 `prcomp` を用いた主成分分析では、関数 `predict` を用いて、あるデータに基づいて作成した合成変数に、新しいデータを当てることが出来る。その書き方を次に示す。

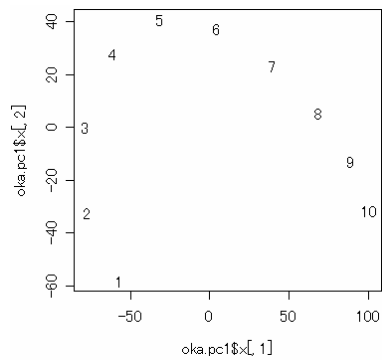
```
predict(object,newdata,..)
```

たとえば、丘本の円周データの個体1~10を用いて、合成変数を作成し、残りの11~16の個体を、求めた合成関数に当てた2次元の主成分得点の散布図を図5に示す。

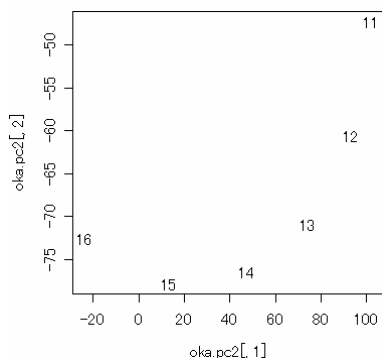
```
>oka.pc1<-prcomp(okamoto[1:10,])
>oka.pc2<-predict(oka.pc1,okamoto[11:16,])
> plot(oka.pc1$x[,1],oka.pc1$x[,2],type="n")
> text(oka.pc1$x[,1],oka.pc1$x[,2],1:10)
> plot(oka.pc2[,1],oka.pc2[,2],type="n")
```

```
> text(oka.pc2[,1],oka.pc2[,2],11:16)
```

図 5 predict の結果の散布図



(a)

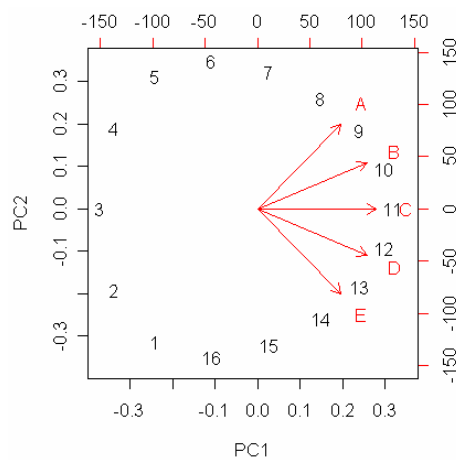


(b)

関数 `prcomp` の結果については、主成分と主成分得点を同一の画面上で散布図を作成する **biplot** 関数を用いることができる。関数 `biplot` を用いた結果を次に示す。

```
>biplot(oka.pc)
```

図 6 関数 `biplot` による散布図



主成分分析で求まる主成分および主成分得点の正負の符号は、固有値及び固有ベクトルを求めるときにアルゴリズムが異なると逆になる場合がある。つまり個体の散布図を描いたときに、異

なるアルゴリズムによる結果の上下が逆になったり、左右が逆になったりすることがある。主成分分析で行う分析は、変数間、個体間の絶対的關係ではなく、相対的關係であるので、分析には問題がない。

**引用文献：**

丘本正(1991)多変量解析の諸方法のモデル再現性に関する数値実験、行動計量学、Vol. 18、No. 2、P. 47-56.