11. 主成分分析

概要: 多くの量的変数が存在するとき、合成数を作成し、少ない合成変数でデータを解釈しようとするものが主成分分析である。主成分分析では、目的変数は無く、第一主成分がその目的変数の役割を演じる。つまり、目的変数のないデータから目的変数を作り出す、という機能を主成分分析は持っているともいえる。合成数である主成分は分散が最大になるという方針で導出する。また、相互作用の強い変数は回帰分析を不安定にする。主成分分析は、これらの相互作用の強い変数を纏めて扱い、上記の不安定性を回避させる。

キーワード: 主成分分析;行列式;固有値;固有ベクトル;回帰;ラグアンジュ法;ラグアンジュ関数。

11.1. 序

あることをいくつかの側面から評価する場合を考える。個々の側面では、それぞれどれが一番いいかは明確に答えが出る。評価項目 1 ではメンバーA が B よりもよく、評価項目 2 では、メンバーA が B に劣る・・・などの場合がある。この場合、どちらが優れていると言えるだろうか？最初から、その複数の評価から総合評価する方法を定義しておけば結果は紛れがない。しかし、それを定義していない場合、どのように評価すれればいいだろうか？この場合、もっともらしい結果を提供するのを示すのが主成分分析である。ここでもっともらしいとは、評価の差が最も明確に出る、という意味である。

ここで断らなければならないのは、結果として、データに差がある場合と、ない場合があるのであって、データにとっては差が明確に出る必要性は必ずしもない、ということである。データがある場合、その順位付けを明確にしたい、という要望に答える解析である、と言える。つまり、データの本質をえぐるという解析ではない。

図 1 は主成分分析で行うことを模式図的に示したものである。

因子として、1 から p までの p 個のものを考える。これは、変数変換し、f_1 から f_p までの、同じ因子数になる。この中で、最初の項 f_1 のみを近似的に考えれば、全体をほぼ考えていることになる、というのが主成分分析である。この場合、p 個の変数を p 個の変数に変換しているだけなので、この部分は数学的に紛れがなく、近似操作は入らない。しかし、変換された p 個の変数の中から、1 個かせいぜい数個のみを選ぶところに近似操作が入る。

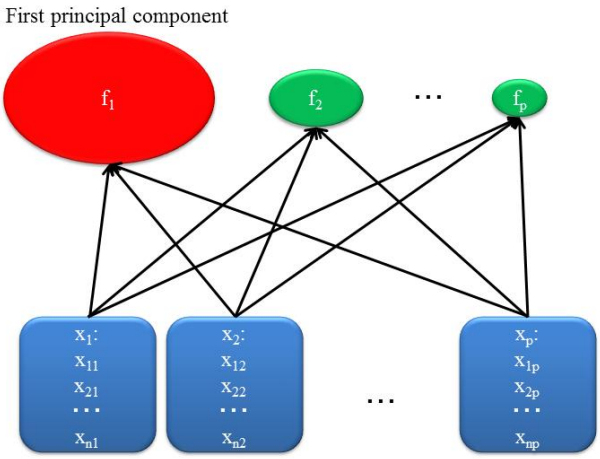

図 1　主成分分析の模式図。 p 個の変数を p 個の別な変数に変換し、その中の 1 つ、または 2 つを選択して解析する

主成分分析はその他の状況にも使われる。先に述べたように、重回帰分析において、説明変数間の相互作用が大きい場合、その重回帰係数は不安定になる。この場合、相互作用の強い変数を主成分分析でまとめる。そして、基本的に一つの変数として扱う。そうすることによって、重回帰分析を安定したものにできる。

ここでの解析では、行列計算を多用する。その行列の演算に関しては別の章で議論し、ここではその結果のみを利用する。

11.2. 二つの変数における主成分分析

ここでは二つの因子に対するデータ $x_j\left(j=1,2\right)$ を考える。データセット数は n とする。

したがって、それぞれのデータは，x_{i1}，$x_{i2}\left(i=1,2,\cdots,n\right)$ のように表現される。その一般的なデータ形式を表 1 に示す。

表 1　主成分分析で扱う二つの因子からなる一般的なデータ形式

ID	x_1	x_1
1	x_{11}	x_{11}
2	x_{21}	x_{12}
…	…	…
i	x_{i1}	x_{i2}
…	…	…
n	x_{n1}	x_{n2}

x_1 の平均と不偏分散を $\bar{x}_1$ および $s_1^{(2)}$ と置く。x_2 の平均と不偏分散をそれぞれ $\bar{x}_2$ および $s_2^{(2)}$ と置く。これらは以下で与えられる。

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1} \tag{1}$$

$$s_1^{(2)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)^2 \tag{2}$$

$$\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2} \tag{3}$$

$$s_2^{(2)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i2} - \bar{x}_2\right)^2 \tag{4}$$

変数 x_1 と x_2 の相関係数は r_{12} と表され、以下で与えられる。

$$r_{12} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left(x_1 - \bar{x}_1\right)\left(x_2 - \bar{x}_2\right)}{\sqrt{s_1^{(2)} s_2^{(2)}}} \tag{5}$$

まず最初に、データを以下のように規格化する。

$$u_1 = \frac{x_1 - \bar{x}_1}{\sqrt{s_1^{(2)}}} \tag{6}$$

$$u_2 = \frac{x_2 - \bar{x}_2}{\sqrt{s_2^{(2)}}} \tag{7}$$

これらの変数に関しては以下が成り立つ。

$$\bar{u}_1 = \bar{u}_2 = 0 \tag{8}$$

$$\sum_{i=1}^{n} {u_{i1}}^2 = n - 1 \tag{9}$$

$$\sum_{i=1}^{n} {u_{i2}}^2 = n-1 \tag{10}$$

$$\sum_{i=1}^{n} u_{i1}u_{i2} = (n-1)r_{x_1x_2} \tag{11}$$

ここで、変数の線形和 z を以下のように定義する。

$$z = a_1u_1 + a_2u_2 \tag{12}$$

ここで、z の不偏分散が最大になるように a_1 と a_2 を設定する。不偏分散 s_z^2 は以下のように評価される。

$$\begin{aligned} s_z^{(2)} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \bar{z}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} {z_i}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(a_1u_{i1} + a_2u_{i2}\right)^2 \\ &= \frac{1}{n-1}\left\{a_1^2\sum_{i=1}^{n}u_{i1}^2 + 2a_1a_2\sum_{i=1}^{n}u_{i1}u_{i2} + a_2^2\sum_{i=1}^{n}u_{i2}^2\right\} \\ &= a_1^2 + a_2^2 + 2a_1a_2r_{x_1x_2} \end{aligned} \tag{13}$$

この値は a_1 おおび a_2 が増加すれば、単調に増加する。したがって、この不偏分散に最大値はない。したがって、以下のような a_1 と a_2 に制約条件をつける。

$$a_1^2 + a_2^2 = 1 \tag{14}$$

最終的には以下の問題になる。

係数 a_1 と a_2 を制約条件 Eq. (14)のもとで、線形結合された数の不偏分散を最大にするように定める。

この問題はラグアンジュ法で解くことができる。ラグランジュ関数は以下で与えられる。

$$f(a_1, a_2, \lambda) = a_1^2 + a_2^2 + 2a_1a_2r_{12} - \lambda\left(a_1^2 + a_2^2 - 1\right) \tag{15}$$

Eq. (15) を a_1 について微分し、0 と置くと以下を得る。

$$\frac{\partial f}{\partial a_1}(a_1, a_2, \lambda) = 2a_1 + 2a_2r_{12} - 2\lambda a_1 = 0 \tag{16}$$

これを整理して、

$$a_1 + r_{12}a_2 = \lambda a_1 \tag{17}$$

Eq. (15) を a_2 に関して微分し 0 と置くと、以下を得る。

$$\frac{\partial f}{\partial a_2}(a_1, a_2, \lambda) = 2a_2 + 2a_2r_{12} - 2\lambda a_2 = 0 \tag{18}$$

これを整理して、

$$a_2 + r_{12}a_1 = \lambda a_2 \tag{19}$$

Eqs. (17) および(19) は以下のように行列形式で表現される。

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{20}$$

これは、さらに以下のように表現される。

$$R\boldsymbol{a} = \lambda \boldsymbol{a} \tag{21}$$

ここで、以下である。

$$R = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \tag{22}$$

ここで、この R のことを相関行列と呼ぶ。他のベクトルは

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{23}$$

$$\boldsymbol{a}^T = (a_1, a_2) \tag{24}$$

と定義される。

ここで、λ の意味について考える。 Eq. (24) を Eq. (20),にかけて以下を得る。

$$(a_1, a_2)\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = (a_1, a_2)\lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{25}$$

これを展開する。

$$\begin{aligned} left\ side &= (a_1, a_2)\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= (a_1, a_2)\begin{pmatrix} a_1 + r_{12}a_2 \\ r_{12}a_1 + a_2 \end{pmatrix} \\ &= a_1^2 + r_{12}a_1a_2 + r_{12}a_1a_2 + a_2^2 \\ &= a_1^2 + +a_2^2 + 2r_{12}a_1a_2 \\ &= s_z^{(2)} \end{aligned} \tag{26}$$

$$\begin{aligned} right\ side &= (a_1, a_2)\lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= \lambda\left(a_1^2 + a_2^2\right) \\ &= \lambda \end{aligned} \tag{27}$$

したがって、以下を得る。

$$\lambda = s_z^{(2)} \tag{28}$$

つまり、λ はzの不偏分散に等しい。大きなλ は大きな不偏分散を意味する。このλ のことを固有値と呼ぶ。我々の解析の目的はそれぞれのメンバーの違いを強調することであった。したがって、大きな不偏分散は我々の望んでいたものである。したがって、これは大きなλと呼ぶ。したがって、固有値の中で最大のものを第一固有値、次のものを第二固有値、・・・と呼ぶ。

二つの変数の場合の固有値λ は以下のように解析的に扱うことができる。

Eq. (20)の右辺は以下のように変形できる。

$$\begin{aligned}\begin{pmatrix}1 & r_{12}\\ r_{12} & 1\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix} &= \lambda\begin{pmatrix}a_1\\ a_2\end{pmatrix}\\ &= \lambda\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix}\\ &= \begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix}\end{aligned} \tag{29}$$

したがって、以下を得る。

$$\begin{pmatrix}1-\lambda & r_{12}\\ r_{12} & 1-\lambda\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix} = 0 \tag{30}$$

これは、以下の行列式が 0 の場合のみ解を持つ。すなわち、

$$\begin{vmatrix}1-\lambda & r_{12}\\ r_{12} & 1-\lambda\end{vmatrix} = 0 \tag{31}$$

これは展開すると以下になる。

$$(1-\lambda)^2 - r_{12}^2 = 0 \tag{32}$$

したがって、以下の二つの根を得る。

$$\lambda_1 = 1 + r_{12} \tag{33}$$

$$\lambda_2 = 1 - r_{x1x2} \tag{34}$$

それぞれの固有値λに対して、ベクトル(a_1, a_2)を得ることができる。

$\lambda_1 = 1 + r_{12}$ を Eq. (20)に代入して、

$$a_1 + r_{12}a_2 = (1 + r_{12})a_1 \tag{35}$$

$$a_2 + r_{12}a_1 = (1 + r_{12})a_2 \tag{36}$$

を得る。これら二つの方程式は同じ結果

$$a_1 = a_2 \tag{37}$$

となる。制約条件 Eq. (14)より、

$$2a_1^2 = 1 \tag{38}$$

したがって、以下を得る。

$$a_1 = \pm\frac{1}{\sqrt{2}} \tag{39}$$

この中で、正の係数を選択すると、z_1 は以下となる。

$$z_1 = \frac{1}{\sqrt{2}}u_1 + \frac{1}{\sqrt{2}}u_2 \tag{40}$$

同じように、 $\lambda_2 = 1 - r_{12}$ の場合は以下となる。

$$a_1 + r_{12}a_2 = (1 - r_{12})a_1 \tag{41}$$

$$a_2 + r_{x_1x_2}a_1 = (1 - r_{12})a_2 \tag{42}$$

を得る。これら二つの方程式は以下の同じ結果を導く。

$$a_1 = -a_2 \tag{43}$$

制約条件 Eq. (14)より、

$$2a_1^2 = 1 \tag{44}$$

したがって、以下となる。

$$a_1 = \pm\frac{1}{\sqrt{2}} \tag{45}$$

ここで、正の係数のものを採用すると z_2 は以下となる。

$$z_2 = \frac{1}{\sqrt{2}}u_1 - \frac{1}{\sqrt{2}}u_2 \tag{46}$$

$r_{12} > 0$ の場合、z_1 は第一成分とよばれ、z_2 は第二成分を呼ばれる。 $r_{12} < 0$ の場合は逆である。

$r_{12} > 0$ と仮定して、以上の結果をまとめると以下になる。

第一主成分となる固有値と固有ベクトルは以下で与えられる。

$$\lambda_1 = 1 + r_{12} \tag{47}$$

$$a^{(1)} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \tag{48}$$

第二主成分となる固有値と固有ベクトルは以下で与えられる。

$$\lambda_2 = 1 - r_{12} \tag{49}$$

$$\boldsymbol{a}^{(2)}=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{2}}\end{pmatrix} \tag{50}$$

規格化された線形結合数は以下となる。

第一主成分に対応する変数は以下である。

$$z_i^{(1)}=\frac{1}{\sqrt{2}}u_{i1}+\frac{1}{\sqrt{2}}u_{i2} \tag{51}$$

第二主成分に対応する変数は以下である。

$$z_i^{(2)}=\frac{1}{\sqrt{2}}u_{i1}-\frac{1}{\sqrt{2}}u_{i2} \tag{52}$$

第一主成分の全体に対する比は以下である。

$$\frac{\lambda_i}{s_1^{(2)}+s_2^{(2)}}=\frac{\lambda_i}{1+r_{12}+1-r_{12}}=\frac{\lambda_i}{2} \tag{53}$$

固有値の和は変数の数と等しい。つまり、この場合は

$$\begin{aligned}\lambda_1+\lambda_2&=(1+r_{12})+(1-r_{12})\\&=2\end{aligned} \tag{54}$$

となる。

11.3. 多因子を扱う主成分分析

前節では我々は因子を二つ扱っていた。しかし、実際にはさらに多くの因子を扱う場合が多い。ここでは、我々は解析を拡張し、m 個の因子、つまりデータとしては $x_1,x_2,\cdots,x_m$ を扱う。ここで、where m は 2 以上である。各因子のデータ数はnとする。つまり、我々は$x_{i1},x_{i2},\cdots,x_{im}\ (i=1,2,\cdots,n)$のデータをこの解析では扱う。

これらのデータをまず以下のように規格化する。n

$$u_1=\frac{x_1-\bar{x}_1}{\sqrt{s_1^{(2)}}},u_2=\frac{x_2-\bar{x}_2}{\sqrt{s_2^{(2)}}},\cdots,u_p=\frac{x_p-\bar{x}_p}{\sqrt{s_p^{(2)}}} \tag{55}$$

これらの平均と不偏分散はそれぞれ 0 および 1 となる。

これから、以下の線形結合した変数を構成する。

$$z=a_1u_1+a_2u_2+\cdots+a_mu_m \tag{56}$$

この線形結合された変数に対応する不偏分散は以下で与えられる。

$$
\begin{aligned}
s_z^{(2)} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i-\bar{z}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} z_i{}^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(a_1u_{i1}+a_2u_{i2}+\cdots+a_mu_{im}\right)^2 \\
&= a_1^2+a_2^2+\cdots+a_m^2 \\
&\quad 2a_1a_2r_{12}+2a_1a_3r_{13}+\cdots+2a_1a_mr_{1m} \\
&\quad +2a_2a_3r_{23}+2a_2a_4r_{24}+\cdots+2a_2a_mr_{2m} \\
&\quad +2a_3a_4r_{34}+2a_3a_5r_{35}+\cdots+2a_3a_mr_{2m} \\
&\quad \cdots \\
&\quad +2a_{m-1}a_mr_{m-1,m}
\end{aligned}
\tag{57}
$$

ここで、係数に以下の制約条件をつける。

$$
a_1^2+a_2^2+\cdots+a_m^2=1 \tag{58}
$$

ラグランジュ関数は以下となる。

$$
\begin{aligned}
f &= a_1^2+a_2^2+\cdots+a_m^2 \\
&\quad 2a_1a_2r_{12}+2a_1a_3r_{13}+\cdots+2a_1a_mr_{1m} \\
&\quad +2a_2a_3r_{23}+2a_2a_4r_{24}+\cdots+2a_2a_mr_{2m} \\
&\quad +2a_3a_4r_{34}+2a_3a_5r_{35}+\cdots+2a_3a_mr_{2m} \\
&\quad \cdots \\
&\quad +2a_{m-1}a_mr_{m-1,m} \\
&\quad -\lambda\left(a_1^2+a_2^2+\cdots+a_m^2-1\right)
\end{aligned}
\tag{59}
$$

Eq. (59)を a_1 で偏微分して 0 と置き、以下を得る。

$$
\frac{\partial f}{\partial a_1}=2a_1+2a_2r_{12}+2a_3r_{13}+\cdots+2a_mr_{1m}-2\lambda a_1=0 \tag{60}
$$

Eq. (59) を a_2 で偏微分して 0 と置き、以下を得る。

$$
\frac{\partial f}{\partial a_2}=2a_1r_{12}+2a_2+2a_3r_{23}+2a_4r_{24}+\cdots+2a_mr_{2m}-2\lambda a_2=0 \tag{61}
$$

他の係数でも同様の解析を繰り返し、それらを纏めると以下を得る。

$$
\begin{pmatrix}
1 & r_{12} & \cdots & r_{1m} \\
r_{21} & 1 & \cdots & r_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
r_{m1} & r_{m2} & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
a_1 \\ a_2 \\ \vdots \\ a_m
\end{pmatrix}
=\lambda
\begin{pmatrix}
a_1 \\ a_2 \\ \vdots \\ a_m
\end{pmatrix}
\tag{62}
$$

ここで、

$$R=\begin{pmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{pmatrix} \tag{63}$$

のことを相関行列と呼ぶ。これは前節の 2 行列のものを拡張したものである。

これは、以下の行列式が 0 になる場合に解を持つ。

$$\begin{vmatrix} 1-\lambda & r_{12} & \cdots & r_{1m} \\ r_{21} & 1-\lambda & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1-\lambda \end{vmatrix}=0 \tag{64}$$

これから、固有値と固有ベクトルを得ることができる。それは一般的には解析的に求めることができず、数値的に得ることができる。それを数値的に得る方法を行列演算の章で示す。それを数値的に得たとしてここでは解析をすすめていく。ここで、

$$\lambda_1>\lambda_2>\cdots>\lambda_m \tag{65}$$

とする。λ_i は $i-th$ 番目の主成分である。

主成分に対応する行列式は以下となる。

$$\begin{pmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}=\lambda_i\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \tag{66}$$

これからは、固有ベクトルを得ることができるが、我々はその絶対値を得ることはできないが、その比率は一般的なものを得ることができる。したがって、a_i の値は a_1 の値を使って以下のように表現できる。

$$a_i=\rho_i a_1 \tag{67}$$

我々は ρ_i を Eq. (66)から評価できる。しかし、a_1 は決定されない。そこで、制約条件 Eq. (56)から、a_1 を以下のように設定する。

$$\begin{aligned} a_1^2+\rho_2^2a_1^2+\cdots+\rho_m^2a_m^2 &=\left(1+\rho_2^2+\cdots+\rho_m^2\right)a_1^2 \\ &=1 \end{aligned} \tag{68}$$

正の a_1 を仮定すると、以下を得る。

$$a_1=\frac{1}{\sqrt{1+\rho_2^2+\cdots+\rho_m^2}} \tag{69}$$

したがって、固有ベクトルは以下となる。

$$z=\frac{1}{\sqrt{1+\rho_2^2+\cdots+\rho_m^2}}\left(u_1+\rho_2u_2+\cdots+\rho_mu_m\right) \tag{70}$$

寄与率は以下となる。

$$\frac{\lambda_i}{m} \tag{71}$$

主成分分析においては、最大 2 個までの因子を考える。それは、図式的に表してイメージできるのが 2 次元までであることによる。2 個までの因子に限定することの有効性は対応する寄与率

$$\frac{\lambda_1 + \lambda_2}{m} \tag{72}$$

で評価できる。この値がいくら以上でなければならない、ということを物理的または数学的に決定することはできない。通常は 0.8 以上というふうに扱われている。

この結果は通常の 2 次元プロットされる。

それぞれの固有ベクトル $j = 1, 2, \cdots, m$ は以下のような近似的な係数で表現される。

$$\begin{aligned} &\left(a_1^{(1)}, a_2^{(1)}, a_3^{(1)} \cdots, a_m^{(1)}\right) \\ &\left(a_1^{(2)}, a_2^{(2)}, a_3^{(2)} \cdots, a_m^{(2)}\right) \\ &\cdots \\ &\left(a_1^{(m)}, a_2^{(m)}, a_3^{(m)} \cdots, a_m^{(m)}\right) \end{aligned} \tag{73}$$

つまり、線形合成数として、m 因子の和から、m 個の線形結合数

$$z^{(1)} = a_1^{(1)} u_1 + a_2^{(1)} u_2 + \cdots + a_m^{(1)} u_m \tag{74}$$

$$z^{(2)} = a_1^{(2)} u_1 + a_2^{(2)} u_2 + \cdots + a_m^{(2)} u_m \tag{75}$$

・・・

$$z^{(m)} = a_1^{(m)} u_1 + a_2^{(m)} u_2 + \cdots + a_m^{(m)} u_m \tag{76}$$

が解析から出てくる。この中で、第一主成分と第二主成分のみ

$$z^{(1)} = a_1^{(1)} u_1 + a_2^{(1)} u_2 + \cdots + a_m^{(1)} u_m \tag{77}$$

$$z^{(2)} = a_1^{(2)} u_1 + a_2^{(2)} u_2 + \cdots + a_m^{(2)} u_m \tag{78}$$

を考え、各メンバーは以下のように近似的に表現さる。

$$\left(z_i^{(1)}, z_i^{(2)}\right) \tag{79}$$

ここまで、第一主成分、第二主成分を等価に扱ってきた。しかし、それらに重みをつけることを考える。ここで、固有値は不変分散を表現していたことを思い出す。したがって、第一主成分には$\sqrt{\lambda_1}$ を第二主成分には $\sqrt{\lambda_2}$ を重みとして掛けることを提案する。

したがって、二人のメンバーi と j の距離d_{ij}は以下となる。

$$d_{ij}=\sqrt{\left(\sqrt{\lambda_1}z_i^{(1)}-\sqrt{\lambda_1}z_j^{(1)}\right)^2+\left(\sqrt{\lambda_2}z_i^{(2)}-\sqrt{\lambda_2}z_j^{(2)}\right)^2} \tag{80}$$

11.4. 主成分分析の例(成績評価)

主成分分析を成績評価の例を取り解析する。表 2 に示す成績表を得たとする。

表 2　各メンバーの成績

Member ID	Japanese x1	English x2	Mathematics x3	Science x4
1	86	79	67	68
2	71	75	78	84
3	42	43	39	44
4	62	58	98	95
5	96	97	61	63
6	39	33	45	50
7	50	53	64	72
8	78	66	52	47
9	51	44	76	72
10	89	92	93	91
Mean	66.4	64.0	67.3	68.6
Stdev.	20.5	21.6	19.4	18.0

表 3　各メンバーの規格化値と第一、第二主成分

Member ID	Japanese u1	English u2	Mathematics u3	Science u4	First component	Second component
1	0.954	0.696	-0.015	-0.033	0.796	0.857
2	0.224	0.510	0.552	0.857	1.073	-0.348
3	-1.188	-0.974	-1.461	-1.368	-2.493	0.321
4	-0.214	-0.278	1.585	1.469	1.283	-1.765
5	1.441	1.531	-0.325	-0.312	1.165	1.802
6	-1.334	-1.438	-1.151	-1.035	-2.479	-0.297
7	-0.798	-0.510	-0.170	0.189	-0.643	-0.678
8	0.565	0.093	-0.790	-1.202	-0.671	1.342
9	-0.750	-0.928	0.449	0.189	-0.518	-1.148
10	1.100	1.299	1.327	1.246	2.488	-0.086

この規格化値を表 3 に示す。対応する相関行列は以下となる。

$$R=\begin{pmatrix} 1 & 0.967 & 0.376 & 0.311 \\ 0.967 & 1 & 0.415 & 0.398 \\ 0.376 & 0.415 & 1 & 0.972 \\ 0.311 & 0.398 & 0.972 & 1 \end{pmatrix} \tag{81}$$

対応する固有ベクトルは以下となる。

$$\lambda_1=2.721,\lambda_2=1.222,\lambda_3=0.052,\lambda_4=0.005 \tag{82}$$

まmた、対応する線形結合数は以下となる。

$$z_1=0.487u_1+0.511u_2+0.508u_3+0.493u_4 \tag{83}$$

$$z_2=0.527u_1+0.474u_2-0.481u_3-0.516u_4 \tag{84}$$

$$z_3=-0.499u_1+0.539u_2-0.504u_3+0.455u_4 \tag{85}$$

$$z_4=0.485u_1-0.474u_2-0.506u_3+0.533u_4 \tag{86}$$

各固有値の寄与率は以下となる。

$$\frac{\lambda_1}{4}=0.680,\frac{\lambda_2}{4}=0.306,\frac{\lambda_3}{4}=0.013,\frac{\lambda_4}{4}=0.001 \tag{87}$$

第二成分までの寄与率の和は 98% であり、80%以上となっている。したがって、この二つの成分に注目する。第一、第二成分を表 3 に表記している。

係数はその因子、つまりここの場合は教科に対する点数である。その係数から以下を得る。

$$\begin{aligned} \text{Japanese:}&(0.487,0.527) \\ \text{English:}&(0.511,0.474) \\ \text{Mathematics:}&(0.508,-0.481) \\ \text{Science:}&(0.493,-0.516) \end{aligned} \tag{88}$$

上の例では第一成分、第二成分を等価に扱っている。重みをつけるとすれば、その固有値のルートをかけるのが尤もらしいであろう。それは、この場合 $\sqrt{\lambda_1}=\sqrt{2.721}$ を第一成分にかけ、 $\sqrt{\lambda_2}=\sqrt{1.222}$ を第二成分にかけるこれから以下を得る。

$$\begin{aligned}\text{Japanese:}&(0.803,0.583)\\ \text{English:}&(0.843,0.524)\\ \text{Mathematics:}&(0.838,-0.532)\\ \text{Science:}&(0.813,-0.570)\end{aligned} \tag{89}$$

次に各生徒のことを考える。各生徒の成績もそれぞれの教科の第一因子、第二因子にそれぞれの規格化した成績をかけたものとなる。したがって、それぞれは、表 4 に示すのものとなる。この場合も、第一成分、第二成分を等価にあつかっており、同じように重みづけをしたものも同じ表に示している。

上の重みづけした結果を 2 次元プロットしたものを図 2 に示す。

最初は第一成分のみをみる。No.10 が圧倒的に成績がよく、つづいて No.4 および No.5 のグループ、最後に No.3 および No.6 のグループになる。

次に、詳細な情報を得るために、縦方向もみる。強かとしては、英語、国語があるから、縦方向は文系か理系かをあらわし、上のほうにあると文系、下の方にあると理系、と定性的に考えることができる。No.4 と 5 は同じような第一成分の値を持っていたが、第二成分は大変異なる。すなわち、No.4 は理系で、No5 は文系である、といえる。 No. 10 はその片寄りがなく、オールラウンダーである。.

表 4　各メンバーの第一、第二主成分およびそれらの重みづけされたもの

Member ID	First component	Second component	Weighted first component	Weighted second component
1	0.796	0.857	1.313	0.948
2	1.073	-0.348	1.770	-0.385
3	-2.493	0.321	-4.113	0.355
4	1.283	-1.765	2.116	-1.951
5	1.165	1.802	1.922	1.992
6	-2.479	-0.297	-4.090	-0.328
7	-0.643	-0.678	-1.060	-0.750
8	-0.671	1.342	-1.107	1.483
9	-0.518	-1.148	-0.854	-1.270
10	2.488	-0.086	4.104	-0.095

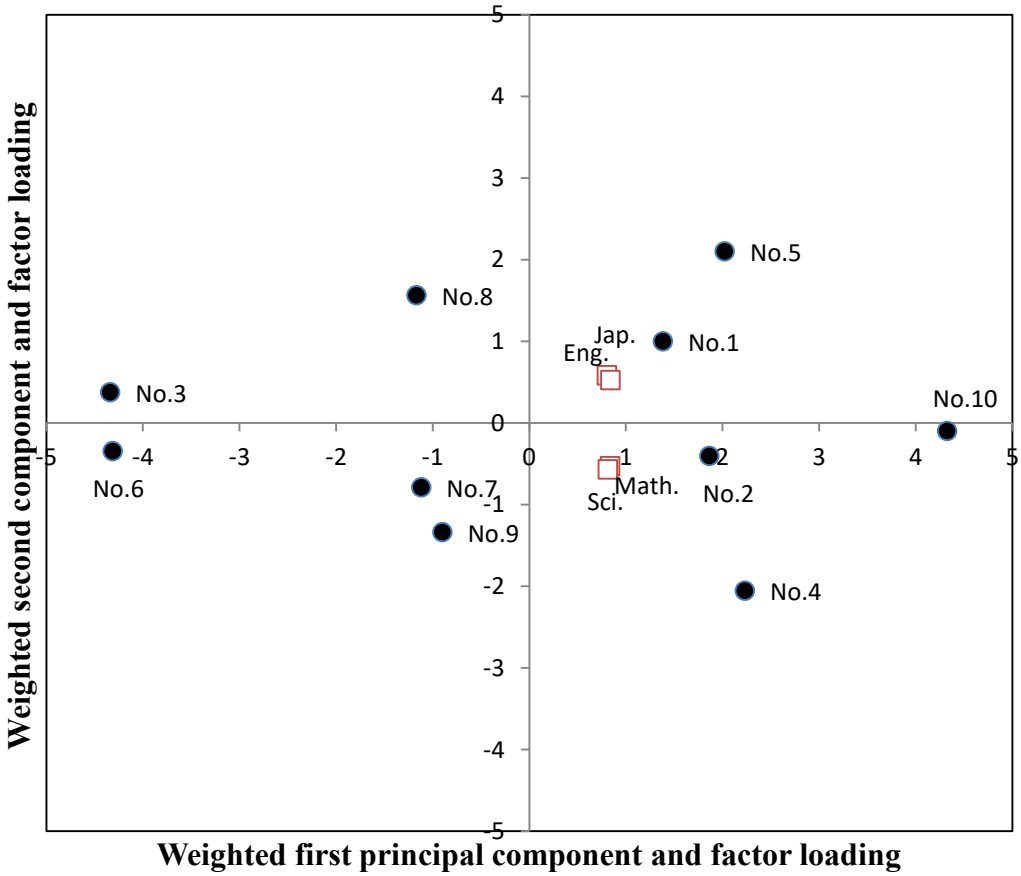

図 2　重みづけされた 2 次元プロット

これらのメンバー間の距離を表 5 に示す。距離が 1.5 以下のものに黄色の網かけをしている。距離は上三角と、下三角領域で対称なので下三角領域のみに網掛けを施している。No. 1 と No .2, No. 5、No.3 と No.6、No.7 と No.9 が似ていることを距離の判定から評価することができる。.

表 5　それぞれのメンバーの距離

メンバー	No.1	No.2	No.3	No.4	No.5	No.6	No.7	No.8	No.9	No.10
No.1		1.5	5.8	3.2	1.3	5.9	3.1	2.6	3.3	3.1
No.2	1.5		6.3	1.7	2.5	6.2	3.0	3.6	2.9	2.5
No.3	5.8	6.3		7.0	6.6	0.7	3.4	3.4	3.8	8.7
No.4	3.2	1.7	7.0		4.2	6.8	3.6	5.0	3.2	2.9
No.5	1.3	2.5	6.6	4.2		6.8	4.3	3.2	4.5	3.2
No.6	5.9	6.2	0.7	6.8	6.8		3.2	3.7	3.6	8.6
No.7	3.1	3.0	3.4	3.6	4.3	3.2		2.4	0.6	5.5
No.8	2.6	3.6	3.4	5.0	3.2	3.7	2.4		2.9	5.7
No.9	3.3	2.9	3.8	3.2	4.5	3.6	0.6	2.9		5.4
No.10	3.1	2.5	8.7	2.9	3.2	8.6	5.5	5.7	5.4	

11.5. 主成分分析を利用した重回帰分析

重回帰分析においては、説明変数間の相互作用が大きいと解析が不安定になるという問題があった。この場合、主成分分析が利用できる。

変数 x_1 と x_2 の相互作用が大きいとする。この場合の規格化した変数から線形結合は以下となる。

$$z = a_1 u_1 + a_2 u_2 \tag{90}$$

ここで、u_1 and u_2 は変数 x_1 と x_2 の規格化値である。この合成数の第一主成分は以下となる。

$$z_i^{(1)} = \frac{1}{\sqrt{2}} u_{i1} + \frac{1}{\sqrt{2}} u_{i2} \tag{91}$$

この場合の寄与率は以下で評価される。

$$\frac{\lambda_1}{2} \tag{92}$$

これが 0.8 以上である必要がある。 そうであると、仮定して以下の解析を進める。

この規格化された第一主成分合成数の平均は 0 であり、分散 $\sigma_{z1}^{(2)}$ は以下である。

$$\begin{aligned} \sigma_{z1}^{(2)} &= \frac{1}{n-1} \sum_{i=1}^{n} {z_{i1}}^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{1}{\sqrt{2}} u_{i1} + \frac{1}{\sqrt{2}} u_{i2} \right)^2 \\ &= \frac{1}{2} \frac{1}{n-1} \sum_{i=1}^{n} \left({u_{i1}}^2 + {u_{i2}}^2 + 2 u_{i1} u_{i2} \right) \\ &= 1 + r_{12} \end{aligned} \tag{93}$$

この変数と目的変数の相関係数は以下となる。

$$
\begin{aligned}
r_{yz1} &= \frac{1}{n-1}\sum_{i=1}^{n}\frac{v_i z_i^{(1)}}{\sqrt{\sigma_{z1}^{(2)}}} \\
&= \frac{1}{\sqrt{1+r_{12}}}\frac{1}{n-1}\sum_{i=1}^{n} v_i\left(\frac{1}{\sqrt{2}}u_{i1}+\frac{1}{\sqrt{2}}u_{i2}\right) \\
&= \frac{1}{\sqrt{2(1+r_{12})}}\left(r_{y1}+r_{y2}\right)
\end{aligned}
\tag{94}
$$

r_{12} が 1 の場合、それは、それぞれの相関係数の平均になる。つまり、二つの変数は 1 つの変数になる。

したがって、この二つの変数のみを考えた場合の回帰直線は以下となる。

$$
\begin{aligned}
v &= r_{yz1}z^{(1)} \\
&= \frac{1}{\sqrt{2(1+r_{12})}}\left(r_{y1}+r_{y2}\right)\left(\frac{1}{\sqrt{2}}u_1+\frac{1}{\sqrt{2}}u_2\right) \\
&= \frac{r_{y1}+r_{y2}}{2\sqrt{1+r_{12}}}\left(u_1+u_2\right)
\end{aligned}
\tag{95}
$$

この解析を複数の因子に拡張できる。もしも m 変数の相互作用が強ければ、以下の構成をすればいい。

$$
z = \sum_{i=1}^{m} a_i^{(1)} u_i \tag{96}
$$

その後に、通常の重回帰分析をしていけばいい。

11.6. まとめ

この章のまとめを行う。

データとして

$$
x_{ij}\left(j=1,2,\cdots,p;i=1,2,\cdots,n\right)
$$

があるとする。この規格化値を以下のように求める。

$$
u_{ij}\left(j=1,2,\cdots,p;i=1,2,\cdots,n\right)
$$

そして、この線形結合した合成数の分散が最大になるように要請し以下を得る。

$$
\begin{pmatrix}
1 & r_{12} & \cdots & r_{1m} \\
r_{21} & 1 & \cdots & r_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
r_{m1} & r_{m2} & \cdots & 1
\end{pmatrix}
\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}
= \lambda
\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}
$$

ここで、 r_{ij} は因子 i と j の間の相関係数である。これから、p 個の固有値と p 個固有ベクトルを以下のように得る。

$$\lambda_1; a^{(1)} = \left(k_1^{(1)}, k_2^{(1)}, \cdots, k_p^{(1)}\right)$$

$$\lambda_2; a^{(2)} = \left(k_1^{(2)}, k_2^{(2)}, \cdots, k_p^{(2)}\right)$$

・・・

$$\lambda_p; a^{(p)} = \left(k_1^{(p)}, k_2^{(p)}, \cdots, k_p^{(p)}\right)$$

各整形結合データは以下のように表現される。

$$z_i^{(1)} = k_1^{(1)} u_{i1} + k_2^{(1)} u_{i2} + \cdots + k_p^{(1)} u_{ip}$$

$$z_i^{(2)} = k_1^{(2)} u_{i1} + k_2^{(2)} u_{i2} + \cdots + k_p^{(2)} u_{ip}$$

・・・

$$z_i^{(p)} = k_1^{(p)} u_{i1} + k_2^{(p)} u_{i2} + \cdots + k_p^{(p)} u_{ip}$$

我々は、この中で最初の 2 成分のみが有効で、あとは近似的に無視できるとする。その精度は寄与率

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_1 + \lambda_2}{p}$$

から評価できる。それが通常は 0.8 以上であることを要請する。すると、各メンバーは二つの要素

$$\left(z_i^{(1)}, z_i^{(2)}\right)$$

で表現される。一方、各因子はその係数から、以下であると同定される。

$$Item\,1: \left(k_1^{(1)}, k_1^{(2)}\right)$$

$$Item\,2: \left(k_2^{(1)}, k_2^{(2)}\right)$$

…

$$Item\ p: \left(k_p^{(1)}, k_p^{(2)}\right)$$

それぞれのメンバーのスコアを固有値で重みづけし、

$$\left(\sqrt{\lambda_1} z_i^{(1)}, \sqrt{\lambda_2} z_i^{(2)}\right)$$

として 2 次元平面上にプロットする。

それぞれのメンバーの近さは以下の距離で評価する。

$$d_{zij} = \sqrt{\left(\sqrt{\lambda_1} z_i^{(1)} - \sqrt{\lambda_1} z_j^{(1)}\right)^2 + \left(\sqrt{\lambda_2} z_i^{(2)} - \sqrt{\lambda_2} z_j^{(2)}\right)^2}$$

それぞれに因子の近さはとし、

$$\left(\sqrt{\lambda_1} k_j^{(1)}, \sqrt{\lambda_2} k_j^{(2)}\right)$$

その間の距離で評価する。

$$d_{kij} = \sqrt{\left(\sqrt{\lambda_1} k_i^{(1)} - \sqrt{\lambda_1} k_j^{(1)}\right)^2 + \left(\sqrt{\lambda_2} k_i^{(2)} - \sqrt{\lambda_2} k_j^{(2)}\right)^2}$$

それぞれの因子とメンバーの近さは以下で評価する。

$$d_{zkij} = \sqrt{\left(\sqrt{\lambda_1} z_i^{(1)} - \sqrt{\lambda_1} k_j^{(1)}\right)^2 + \left(\sqrt{\lambda_2} z_i^{(2)} - \sqrt{\lambda_2} k_j^{(2)}\right)^2}$$

ある組の変数の相互作用が強い場合は、それらを纏めて一つの変数

$$z = \sum_{i=1}^{m} a_i^{(1)} u_i$$

とし、他の変数と併せて重回帰分析すればいい。