多変量解析　～主成分分析～

１．主成分解析とは

２．適用例と解析の目的

３．解析の流れ

４．変数が2個の場合の主成分分析

５．変数がp個の場合の主成分分析

６．行列とベクトルによる表現

第9章　P.132~150

2006/05/26　神津　健太

主成分分析とは

- 多くの量的変数が存在する場合に、
  それらの間の相関構造を考慮して、
  低い次元の合成変数（主成分）に変換し、
  データが有している情報をより解釈しやすくするための方法。

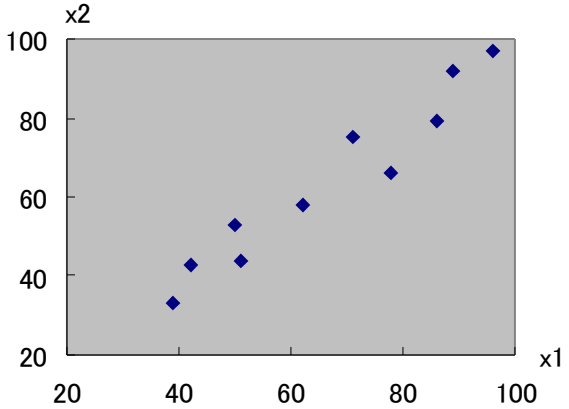
- 「相関係数行列から出発する方法」と
  「分散共分散行列から出発する方法」の
  ２種類がるが、今回は前者だけを説明する。

適用例と解析の目的(1)

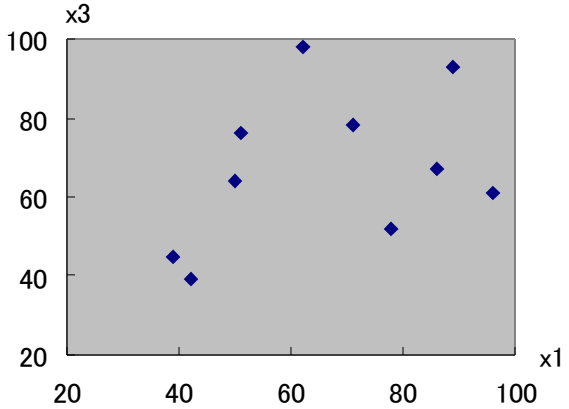
(P.132~)　表1. 試験の成績のデータ

生徒No.	国語 x1	英語 x2	数学 x3	理科 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

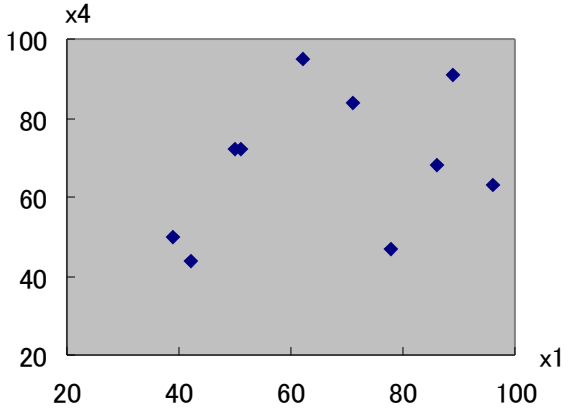
それぞれの科目を量的変数と考える。 変数の個数は $p = 4$ である。



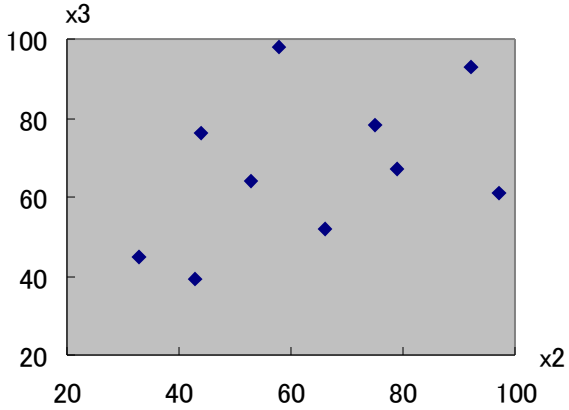
x1 と x2 の散布図



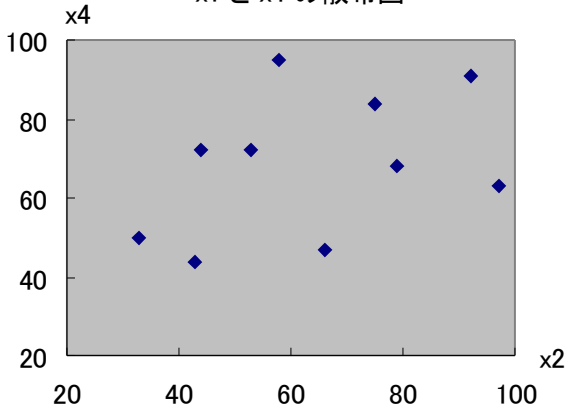
x1 と x3 の散布図



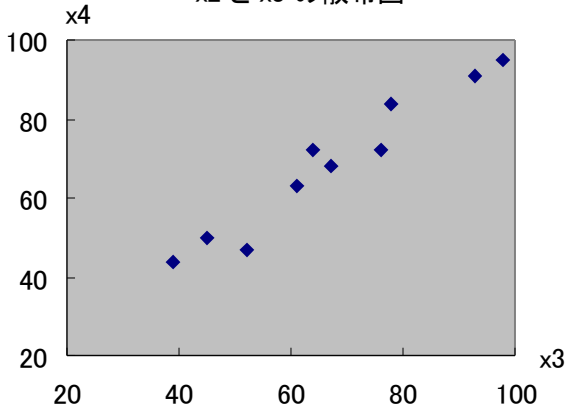
x1 と x4 の散布図



x2 と x3 の散布図



x2 と x4 の散布図



x3 と x4 の散布図

適用例と解析の目的(2)

それぞれの相関係数を求めると、　（相関係数の求め方はP.13を参照）

$$r_{x_1x_2} = 0.967 \qquad r_{x_1x_3} = 0.376$$

$$r_{x_1x_4} = 0.311 \qquad r_{x_2x_3} = 0.415$$

$$r_{x_2x_4} = 0.398 \qquad r_{x_3x_4} = 0.972$$

各変数間の相関係数はすべて正となり、

「国語と英語」「数学と理科」の相関係数は高い値となっている。

つまり、データにはなんらかの相関構造があると考えられる。

・「より低い次元でデータのばらつきを解釈できないか」

・「そのためにはどのように合成変数(主成分)を構成すればよいか」

・「それぞれの主成分の説明力はどれくらいか」

・「科目や生徒をどのように分類できるか」

などを検討したい。

主成分分析の流れ

（1）主成分の導出

相関係数行列 R の第1固有値（最大固有値）λ1 に対応する固有ベクトルから第1主成分 z1 を求める。次にRの第2固有値λ2 に対応する固有ベクトルから第2主成分 z2 を求める。同様にして、第 k 主成分 を求める。(k = 3,4,･･･,p)

（2）寄与率および累積寄与率

それぞれの主成分の寄与率および累積寄与率を求める。「固有値が1以上」ないしは「累積寄与率80%を超える」を目安として主成分を選択する。

（3）因子負荷量

因子負荷量を求める。固有ベクトルや因子負荷量の値を参考にして、選択した各主成分の意味について考察する。また、因子負荷量を散布図にプロットし、変数の分類を行う。

（4）主成分得点

主成分得点を散布図にプロットし、サンプルの特徴付けや分類を行う。

変数が2個の場合の主成分分析（1）

（1）主成分の導出　(P.134~)

変数がx1,x2の2つで、サンプルサイズがnとする。

変数x1,x2を標準化　$u_1 = \dfrac{x_1 - \bar{x}_1}{s_1}$,　$u_2 = \dfrac{x_2 - \bar{x}_2}{s_2}$

また、次のこともにも注意しておく。

$$\sum_{i=1}^{n} {u_{i1}}^2 = \sum_{i=1}^{n} {u_{i2}}^2 = n - 1$$

$$\sum_{i=1}^{n} u_{i1} u_{i2} = (n-1) r_{x_1 x_2}$$

第1主成分z1を　$z_1 = a_1 u_1 + a_2 u_2$ とおく。

$\bar{u}_1 = \bar{u}_2 = 0$ だから $\bar{z}_1 = 0$ である。

目的は、データの情報をできるだけ多く有するようにz1を定めることである。
（つまり、係数a1とa2をデータから定めることである。）

変数が2個の場合の主成分分析（2）

「z1がもとのデータの情報をできるだけ多く有する」

→「データの全体のバラツキをできるだけz1のバラツキに反映させる」と考える。

z1の分散 $$V_{z_1} = \frac{1}{n-1}\sum_{i=1}^{n}(z_{i1} - \bar{z}_1)^2 = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}^{\ 2}$$

が最大となるようなa1とa2を求める。 $$V_{z_1} = \frac{1}{n-1}\sum_{i=1}^{n} z_{i1}^{\ 2} = \frac{1}{n-1}\sum_{i=1}^{n}(a_1u_{i1} + a_2u_{i2})^2$$

$$= \frac{1}{n-1}\{a_1^{\ 2}\sum_{i=1}^{n} u_{i1}^{\ 2} + 2a_1a_2\sum_{i=1}^{n} u_{i1}u_{i2} + a_2^{\ 2}\sum_{i=1}^{n} u_{i2}^{\ 2}\}$$

$$= a_1^{\ 2} + a_2^{\ 2} + 2r_{x_1x_2}a_1a_2$$

となるので、$a_1^{\ 2} + a_2^{\ 2} = 1$ の値は（a1,a2）の値が大きくなればいくらでも大きくなる。

そこで　　の制約条件を設けた上で、Vz1の最大化を考える。

変数が2個の場合の主成分分析（3）

制約付きの最大化問題を求めるために、ラグランジュの未定乗数法を用いる。

未定乗数λを用いて

$$f(a_1, a_2, \lambda) = a_1^2 + a_2^2 + 2r_{x_1x_2}a_1a_2 - \lambda(a_1^2 + a_2^2 + 1)$$

とおき、a1,a2のそれぞれで微分（偏微分）してゼロとおく。

$$2a_1 + 2r_{x_1x_2}a_2 - 2\lambda a_1 = 0$$

$$2r_{x_1x_2}a_1 + 2a_2 - 2\lambda a_2 = 0$$

両辺を2で割って行列の形に表現

$$\begin{bmatrix} 1 & r_{x_1x_2} \\ r_{x_1x_2} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

相関係数行列**R**

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 \end{bmatrix}'$$

$$\mathrm{R}\,\mathbf{a} = \lambda\mathbf{a}$$

λが行列Rの固有値であり、求めるべき[a1,a2]は　　固有ベクトルで あることを示している。

変数が2個の場合の主成分分析（4）

両辺に左からベクトル[a1,a2]をかけてみると

$$a_1^{\,2} + a_2^{\,2} + 2r_{x_1x_2}a_1a_2 = \lambda\,(a_1^{\,2} + a_2^{\,2})$$

$$V_{z_1} = \lambda$$

以上より、Vz1を最大化させることは

「相関係数行列Rの固有値問題を解いて、最大固有値λ1に対応する（長さ1の） 固有ベクトル **a** を求めれば（つまり、R **a** =λ1 **a**）、それがVz1の最大値を与える[a1,a2]であり、Vz1の最大値はλ1となる。」

という手続きで実行される。

変数が2個の場合の主成分分析（5）

第1主成分だけでデータの情報を十分説明できないとき

→ 第1主成分に含まれない情報を追加するために第2主成分を導入する

第2主成分 z2 を $z_2 = b_1u_1 + b_2u_2$ とおき、z1 と無相関となるように定める。

まず、z1,z2 の相関係数 $r_{x_1x_2}$ の分子を考えると、

$$\sum_{i=1}^{n}(z_{i1}-\bar{z}_1)(z_{i2}-\bar{z}_2)=\sum_{i=1}^{n}z_{i1}z_{i2}$$

$$=\sum_{i=1}^{n}(a_1u_{i1}+a_2u_{i2})(b_1u_{i1}+b_2u_{i2})$$

$$=a_1b_1\sum_{i=1}^{n}{u_{i1}}^2+a_1b_2\sum_{i=1}^{n}u_{i1}u_{i2}+a_2b_1\sum_{i=1}^{n}u_{i1}u_{i2}+a_2b_2\sum_{i=1}^{n}{u_{i2}}^2$$

$$=(n-1)\{a_1b_1+r_{x_1x_2}a_1b_2+r_{x_1x_2}a_2b_1+a_2b_2\}$$

$$=(n-1)\mathbf{a}'\mathrm{R}\mathbf{b}$$

$$=(n\text{-}1)\lambda_1\mathbf{a}'\mathbf{b} \qquad \leftarrow (R\mathbf{a}=\lambda\mathbf{a}\text{ より }\mathbf{a}'R=\lambda\mathbf{a}')$$

となる。

変数が2個の場合の主成分分析（6）

相関係数 $r_{x_1x_2} = 0$ となる条件は

$\mathbf{a}'\mathrm{R}\mathbf{b} = 0$　　または　　$\mathbf{a}'\mathbf{b} = a_1b_1 + a_2b_2 = 0$

このことから、第2主成分は

$$V_{z_2} = \frac{1}{n-1}\sum_{i=1}^{n}(z_{i2} - \bar{z}_2)^2 = \frac{1}{n-1}\sum_{i=1}^{n} {z_{i2}}^2$$

$$= {b_1}^2 + {b_2}^2 + 2r_{x_1x_2}b_1b_2$$

第1主成分のときと同様に、 ${b_1}^2 + {b_2}^2 = 1$ の制約条件、
そして相関係数がゼロになる条件のもとで最大化する。

この場合もラグランジュの未定乗数法を用いるが、
相関係数がゼロになる条件あるので、λとηの2つの乗数を用いる。

変数が2個の場合の主成分分析（7）

λとηの2つの乗数を用いると

$$f(b_1, b_2, \lambda, \eta) = b_1^2 + b_2^2 + 2r_{x_1x_2}b_1b_2 - \lambda(b_1^2 + b_2^2 - 1) - \eta(a_1b_1 + a_2b_2)$$

とおき、b1,b2のそれぞれで微分（偏微分）してゼロとおく。

$$2b_1 + 2r_{x_1x_2}b_2 - 2\lambda b_1 - \eta a_1 = 0$$

$$2r_{x_1x_2}b_1 + 2b_2 - 2\lambda b_2 - \eta a_2 = 0$$

それぞれを2で割って行列で表現

$$\begin{bmatrix} 1 & r_{x_1x_2} \\ r_{x_1x_2} & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \lambda \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \frac{\eta}{2} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad \rightarrow \quad R\mathbf{b} = \lambda \mathbf{b} + \frac{\eta}{2}\mathbf{a}$$

ここで両辺に**a'**ををかけ、相関係数がゼロになる条件を考えると
η＝0となる。よって式は

$$R\mathbf{b} = \lambda \mathbf{b}$$

変数が2個の場合の主成分分析（8）

$R\mathbf{b} = \lambda \mathbf{b}$ より、第2主成分の係数（b1,b2）もRの固有ベクトルである。

Vz1のときと同じ理由でVz2の最大値もRの最大固有値となる。

しかし、Vz2の最大化において、Rの最大固有値λ1に対応する固有ベクトル**b**は　制約条件：$\mathbf{a}'\mathbf{b} = 0$ を満たさない。

そこで、Vz2の最大値はRの2番目に大きな固有値λ2となる。

第2主成分z2の構成には、λ2に対応する（長さ1の）固有ベクトル**b**を用いる。この**b**は制約条件を満たす。

（対象行列の固有値はすべて実数であり、異なる固有値に対応する

固有ベクトルは直交する。）

変数が2個の場合の主成分分析（9）

（2）寄与率および累積寄与率　（P.139~）

2つの主成分z1とz2は、それぞれVz1,Vz2を最大とするように求めた。

そのVz1,Vz2の最大値はλ1,λ2だったので、寄与率は以下のように定義する。

$$\text{第１主成分の寄与率} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

$$\text{第２主成分の寄与率} = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

次に、累積寄与率は以下のように定義する。

$$\text{第１主成分の累積寄与率} = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

$$\text{第２主成分の累積寄与率} = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2} = 1$$

相関係数 r の値が大きいほど、第1主成分の寄与率は1に近づく。
r＝0のときは、どちらの寄与率も0．5になる。

変数が2個の場合の主成分分析（10）

（3）因子負荷量と主成分の解釈　（P.140~）

主成分z1,z2と、もとの変数x1,x2との相関係数を考える。

→ これらの相関係数を因子負荷量という。

これらは主成分z1,z2と、もとの変数を標準化した変数u1,u2との相関係数に等しい。

因子負荷量と固有値、固有ベクトルの間には以下のような関係がある。

$$r_{z_1x_1} = \sqrt{\lambda_1}a_1 \qquad r_{z_1x_2} = \sqrt{\lambda_1}a_2$$

$$r_{z_2x_1} = \sqrt{\lambda_2}b_1 \qquad r_{z_2x_2} = \sqrt{\lambda_2}b_2$$

因子負荷量と固有ベクトルは主成分に対して同じ情報を与える。

第1主成分z1については、[$r_{z_1x_1}, r_{z_1x_2}$]によってもとの変数との関わり具合を考察してその解釈を与えるが、これは[a1,a2]を見ることと同じである。

変数が2個の場合の主成分分析(11)

(4)主成分得点　(P.141~)

個々のサンプルに対して各変数の値を標準化し、

$$z_1 = a_1u_1 + a_2u_2$$

に代入して得られた値を、第1成分の主成分得点という。

$z_2 = b_1u_1 + b_2u_2$ に代入すれば、第2主成分の主成分得点となる。

各主成分に対して、主成分得点はサンプルの個数だけ計算することができる。

これらを散布図にプロットし、各主成分に与えた意味付けを考慮しながら

サンプルの特徴付けや分類などを試みる。

変数がp個の場合の主成分分析（1） (P.142~)

変数が3個以上になっても考え方は同様である。

変数の標準化

$$u_1 = \frac{x_1 - \bar{x}_1}{s_1} \quad , \quad u_2 = \frac{x_2 - \bar{x}_2}{s_2} \quad \cdots \quad u_p = \frac{x_p - \bar{x}_p}{s_p}$$

第1主成分

$$z_1 = a_1 u_1 + a_2 u_2 + \quad \cdots \quad + a_p u_p$$

相関係数行列

$$R = \begin{bmatrix} 1 & r_{x_1 x_2} & \cdots & r_{x_1 x_p} \\ r_{x_2 x_1} & 1 & & r_{x_2 x_p} \\ \vdots & & \ddots & \vdots \\ r_{x_p x_1} & r_{x_p x_2} & \cdots & 1 \end{bmatrix}$$

変数がp個の場合の主成分分析（2）

第k主成分の寄与率

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

第k主成分までの累積寄与率

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

「固有値が１以上」または「累積寄与率が80%を超える」という目安で主成分を選択。

第１主成分の因子負荷量 （第２成分以降も同様）

$$r_{z_1 x_1} = \sqrt{\lambda_1} a_1 \quad , \quad r_{z_1 x_2} = \sqrt{\lambda_1} a_2 \quad , \quad \cdots \quad r_{z_1 x_p} = \sqrt{\lambda_1} a_p$$

変数がp個の場合の主成分分析（3）

例として、表1のデータに主成分分析を適用する。

1．データの平均と標準偏差を求める。

$\bar{x}_1 = 66.4$　　$\bar{x}_1 = 64.0$　　$\bar{x}_1 = 67.3$　　$\bar{x}_1 = 68.6$

$s_1 = 20.5$　　$s_1 = 21.6$　　$s_1 = 19.4$　　$s_1 = 18.0$

2．相関係数行列Rを求める。

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

3．固有値と固有ベクトルを求める。（解析ソフトを使用した）

$\lambda_1 = 2.721$　　$\mathbf{a} = [0.487, 0.511, 0.508, 0.493]'$

$\lambda_2 = 1.222$　　$\mathbf{b} = [0.527, 0.474, -0.481, -0.516]'$

$\lambda_3 = 0.052$　　$\mathbf{c} = [-0.499, 0.539, -0.504, 0.455]'$

$\lambda_4 = 0.005$　　$\mathbf{d} = [0.485, -0.474, -0.506, 0.533]'$

変数がp個の場合の主成分分析（4）

4．求めた固有ベクトルから、4つの主成分を求める。

$$z_1 = 0.487u_1 + 0.511u_2 + 0.508u_3 + 0.493u_4$$
$$z_2 = 0.527u_1 + 0.474u_2 - 0.481u_3 - 0.516u_4$$
$$z_3 = -0.499u_1 + 0.539u_2 - 0.504u_3 + 0.455u_4$$
$$z_4 = 0.485u_1 - 0.474u_2 - 0.506u_3 + 0.533u_4$$

5．固有値より、各主成分の寄与率を求める

第１主成分の寄与率 $= \dfrac{\lambda_1}{p} = 0.680$

第２主成分の寄与率 $= \dfrac{\lambda_2}{p} = 0.306$

第３主成分の寄与率 $= \dfrac{\lambda_3}{p} = 0.013$

第４主成分の寄与率 $= \dfrac{\lambda_4}{p} = 0.001$

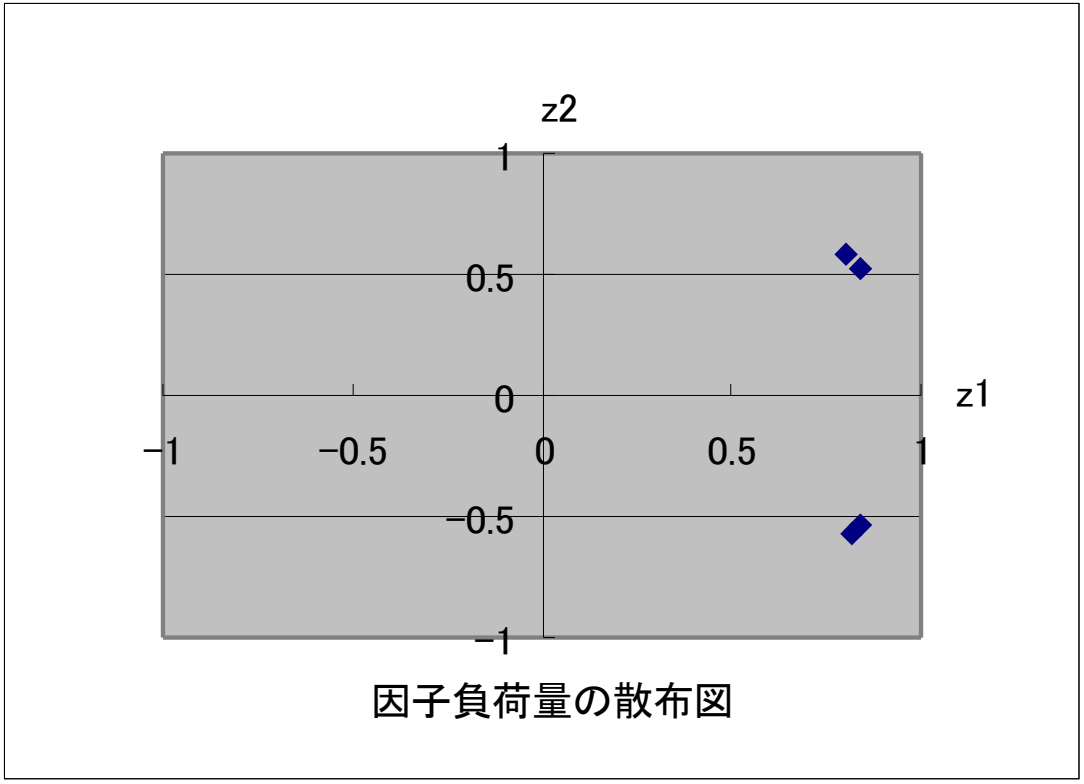
第2主成分までの累積寄与率は
0.680+0.306=0.986なので、

第2主成分までで十分

変数がp個の場合の主成分分析(4)

表2. 因子負荷量

	国語x1	英語x2	数学x3	理科x4
z1	0.804	0.842	0.838	0.814
z2	0.583	0.524	-0.531	-0.570
z3	-0.114	0.123	-0.115	0.104
z4	0.035	-0.034	-0.036	0.038



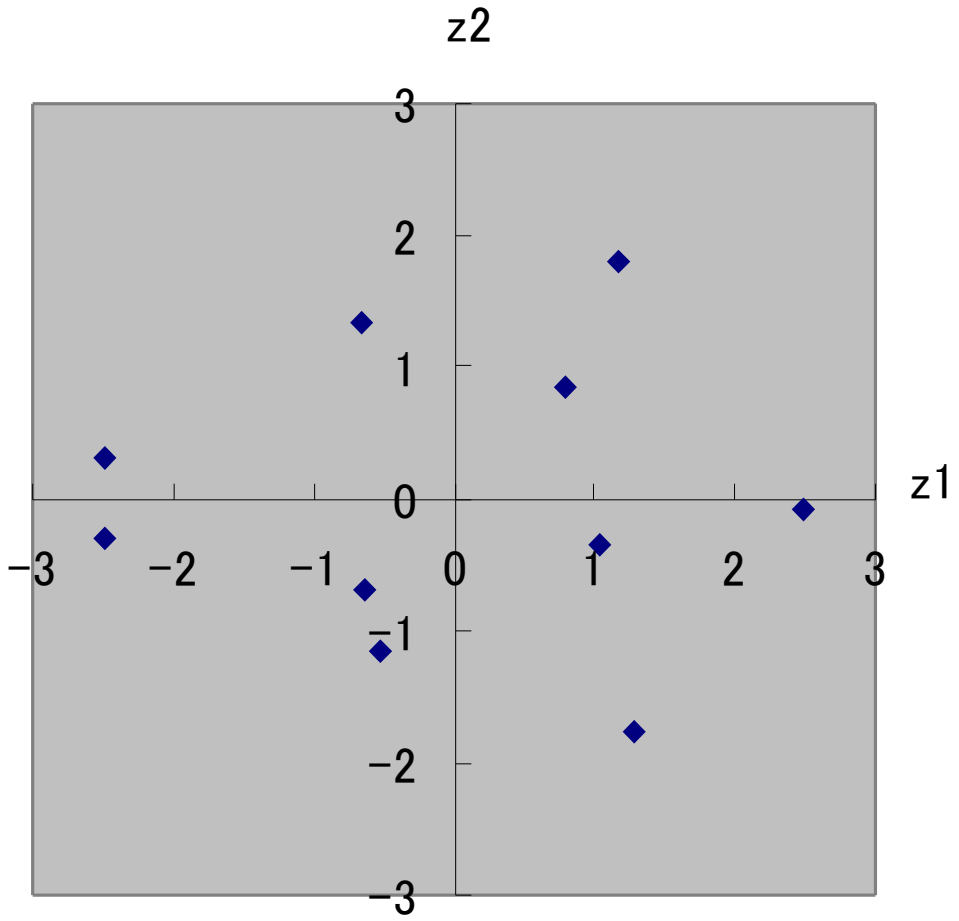
因子負荷量の散布図

z1:総合的な学力

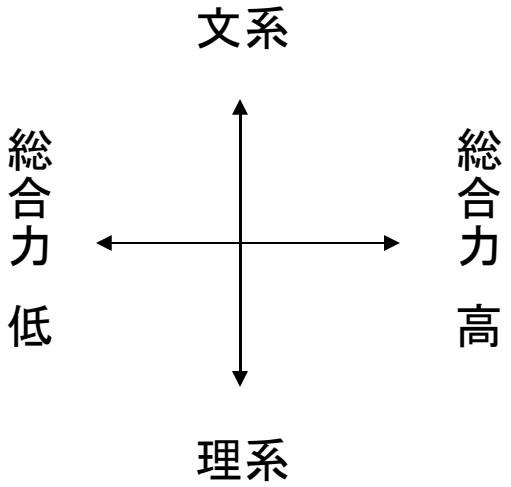
z2:文系・理系の違い

表3. 標準化した値と主成分得点

生徒No.	標準化した値				主成分得点	
	u1	u2	u3	u4	z1	z2
1	0.956	0.694	-0.015	-0.033	0.796	0.857
2	0.244	0.509	0.552	0.856	1.027	-0.348
3	-1.190	-0.972	-1.459	-1.367	-2.491	0.319
4	-0.215	-0.278	1.582	1.467	1.280	-1.763
5	1.444	1.528	-0.325	-0.311	1.166	1.802
6	-1.337	-1.435	-1.149	-1.033	-2.477	-0.299
7	-0.800	-0.509	-0.170	0.189	-0.643	-0.679
8	0.566	0.093	-0.789	-1.200	-0.669	1.341
9	-0.751	-0.926	0.448	0.189	-0.518	-1.148
10	1.102	1.296	1.325	1.244	2.485	-0.084



主成分得点の散布図



行列とベクトルによる表現（1）

（1）主成分の導出　（P.146~）

変数の個数p　　サンプルサイズn

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_p \end{bmatrix}, \qquad \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \cdot \\ \cdot \\ u_{ip} \end{bmatrix} \qquad (i = 1,2,\cdots,n)$$

第1主成分

$$\begin{aligned} z_1 &= a_1 u_{i1} + a_2 u_{i2} + \quad \cdots \quad + a_p u_{ip} \\ &= \mathbf{a}'\mathbf{u}_i \\ &= \mathbf{u}_i'\mathbf{a} \end{aligned}$$

行列とベクトルによる表現（2）

z1の分散

$$
\begin{aligned}
V_{z_1} &= \frac{1}{n-1}\sum_{i=1}^{n} {z_{i1}}^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{a}'\mathbf{u}_i)(\mathbf{u}_i'\mathbf{a}) = \mathbf{a}'\left(\frac{1}{n-1}\sum_{i=1}^{n} \mathbf{u}_i\mathbf{u}_i'\right)\mathbf{a} \\
&= \mathbf{a}'R\mathbf{a}
\end{aligned}
$$

ここでRは相関係数行列である。

最大化の制約条件

$$
{a_1}^2 + {a_2}^2 + \cdots + {a_p}^2 = \mathbf{a}'\mathbf{a} = 1
$$

行列とベクトルによる表現（３）

ラグランジュの未定乗数法

$$f(\mathbf{a},\lambda) = \mathbf{a}'R\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1)$$

これをベクトル**a**により微分して0とおく。（P.39参照）

$$\frac{\partial f}{\partial \mathbf{a}} = 2R\mathbf{a} - 2\lambda\mathbf{a} = 0$$

$$R\mathbf{a} = \lambda\mathbf{a}$$

$$\mathbf{a}'R\mathbf{a} = \lambda\mathbf{a}'\mathbf{a} \qquad \rightarrow \qquad V_{z_1} = \lambda$$

第１主成分の係数は、固有値問題を解いて、

最大固有値に対応する（長さ1の）固有ベクトルを求めればよい

。

行列とベクトルによる表現（４）

第２主成分の導出

$$z_2 = b_1 u_1 + b_2 u_2 + \quad \cdots \quad + b_p u_p$$

$\mathbf{b} = [b_1, b_2, \quad \cdots \quad , b_p]'$ とすると

$$z_2 = \mathbf{b}'\mathbf{u}_i = \mathbf{u}_i'\mathbf{b}$$

z1のときと同様に、z2の分散は

$$V_{z_2} = \frac{1}{n-1}\sum_{i=1}^{n} {z_{i2}}^2 = \mathbf{b}'R\mathbf{b}$$

制約条件

$${b_1}^2 + {b_2}^2 + \cdots + {b_p}^2 = \mathbf{b}'\mathbf{b} = 1$$

行列とベクトルによる表現（５）

z1,z2が無相関である制約条件の設定

$$\sum_{i=1}^{n} z_{i1}z_{i2} = \sum_{i=1}^{n} (\mathbf{a}'\mathbf{u}_i)(\mathbf{u}_i'\mathbf{b}) = (n-1)\mathbf{a}'\left(\frac{1}{n-1}\sum_{i=1}^{n}\mathbf{u}_i\mathbf{u}_i'\right)\mathbf{b}$$

$$= (n-1)\mathbf{a}'R\mathbf{b} = (n-1)\lambda_1\mathbf{a}'\mathbf{b}$$

$$\downarrow$$

$\mathbf{a}'\mathrm{R}\mathbf{b} = 0$　　または　　$\mathbf{a}'\mathbf{b} = a_1b_1 + a_2b_2 = 0$

ラグランジュの未定乗数法

$$f(b,\lambda,\eta) = \mathbf{b}'R\mathbf{b} - \lambda(\mathbf{b}\mathbf{b}' - 1) - \eta\,\mathbf{a}'\mathbf{b}$$

$$\downarrow$$

$$2\mathbf{a}'R\mathbf{b} - 2\lambda\,\mathbf{a}'\mathbf{b} - \eta\,\mathbf{a}'\mathbf{a} = 0 \qquad \rightarrow \qquad \eta = 0$$

$$\therefore \qquad R\mathbf{b} = \lambda\,\mathbf{b}$$

行列とベクトルによる表現（6）

（2）相関係数行列　（P.150~）

Rについての固有値問題を解いて、

$$R\mathbf{a}_1 = \lambda \mathbf{a}_1$$
$$R\mathbf{a}_2 = \lambda \mathbf{a}_2$$
$$\vdots$$
$$R\mathbf{a}_p = \lambda \mathbf{a}_p$$

$\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_p$は固有ベクトル

Rは対象行列なのでスペクトル分解が成り立つ（P.38）

$$R = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \lambda_2 \mathbf{a}_2 \mathbf{a}_2' + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}_p'$$

第2主成分までの累積寄与率が1に近いとすると、

$$R \fallingdotseq \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \lambda_2 \mathbf{a}_2 \mathbf{a}_2'$$

主成分分析はスペクトル分解に基づいた相関係数行列Rの近似である。

近似がうまくいくためには、いくつかの大きな固有値が存在し、その他の固有値がゼロに近ければよい。つまり、そのような相関構造が変数間にあればいい。

逆に、特徴のある相関構造がなく、各変数が無相関に近い場合は、近似は成立せず、データの情報を小数個の主成分に縮約できない。