データ解析

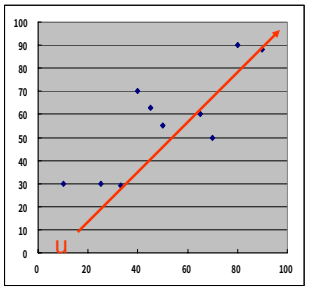
http://coconut.sys.eng.shizuoka.ac.jp/data/

静岡大学工学部
安藤和敏

2005.12.07

主成分分析の考え方

No	x	y
1	70	50
2	80	90
3	33	29
4	50	55
5	90	88
6	25	30
7	45	63
8	65	60
9	10	30
10	40	70



与えられたデータの変数を合成して，データの本質を良く表現する変数（主成分）を見つけ出す．

座標軸の回転

$$a^2 + b^2 = 1$$

を満たすような a, b に対して，

$$u = ax + by$$

という変数の変換を考える．

座標軸の回転

ただし，そのような変換

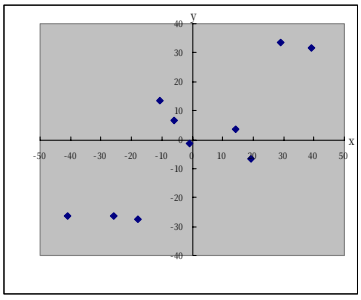
$$u = ax + by$$

の中で，新しい変数 u に関するデータが最も散らばって見えるように，すなわち，u の分散が最も大きくなるように，a と b を選びたい．

Excelで学ぼう

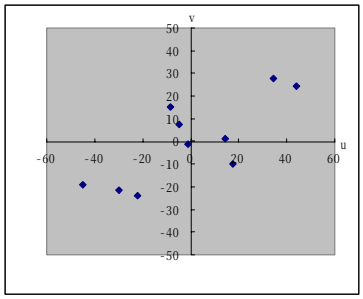
ファイル：第3章/3_1

座標軸の回転（0度）



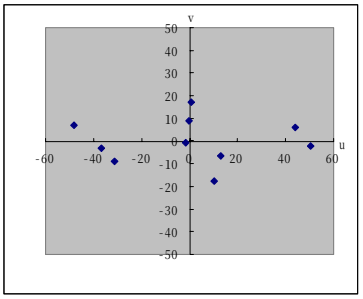
uの分散=575.5

座標軸の回転（10度）



uの分散=716.1

座標軸の回転（41度）



uの分散=941.0

主成分分析のデータ（変数が3個の場合）

No	変数 x	変数 y	変数 z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
⋮	⋮	⋮	⋮
i	x_i	y_i	z_i
⋮	⋮	⋮	⋮
n	x_n	y_n	z_n

合成変数 u

$$a^2+b^2+c^2=1$$

を満たす a, b, c **に対して，**

$$u=ax+by+cz$$

という変数変換を考える．

ただし，ここでも u の分散 s_u^2 **が最大になるように，** a, b, c **を選びたい．**

uの分散

$$s_u^2=\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})^2=\frac{1}{n}\sum_{i=1}^{n}(ax_i+by_i+cz_i-a\bar{x}-b\bar{y}-c\bar{z})^2$$

$$=a^2\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+b^2\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2+c^2\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})^2$$

$$+2ab\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})+2bc\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(z_i-\bar{z})$$

$$+2ca\frac{1}{n}\sum_{i=1}^{n}(z_i-\bar{z})(x_i-\bar{x})$$

$$=a^2s_x^2+b^2s_y^2+c^2s_z^2+abs_{xy}+bcs_{yz}+cas_{zx}$$

uの分散の最大化

条件 $a^2+b^2+c^2=1$

を満足する a,b,c **のうちで，**

$$s_u^2=a^2s_x^2+b^2s_y^2+c^2s_z^2+abs_{xy}+bcs_{yz}+cas_{zx}$$

を最大にするものを求めたい．

ラグランジュの未定係数法

条件 $g(a,b,c)=0$

の下で

$$f(a,b,c)$$

を最大にするa,b,cを求めるためには，

$$L(a,b,c,\lambda)=f(a,b,c)-\lambda g(a,b,c)$$

と置いて，以下の方程式の解を調べればよい．

$$\frac{\partial L}{\partial a}=0,\frac{\partial L}{\partial b}=0,\frac{\partial L}{\partial c}=0,\frac{\partial L}{\partial \lambda}=0.$$

ラグランジュの未定係数法の適用

$$g(a,b,c)=a^2+b^2+c^2-1,$$

$$f(a,b,c)=a^2s_x^2+b^2s_y^2+c^2s_z^2 + abs_{xy}+bcs_{yz}+cas_{zx}$$

$$\therefore L(a,b,c,\lambda)=a^2s_x^2+b^2s_y^2+c^2s_z^2 + abs_{xy}+bcs_{yz}+cas_{zx} - \lambda(a^2+b^2+c^2-1)$$

ラグランジュの未定係数法の適用

$$0=\frac{\partial L}{\partial a}=2s_x^2a+2bs_{xy}+2cs_{zx}-2a\lambda,$$

$$0=\frac{\partial L}{\partial b}=2s_y^2b+2as_{xy}+2cs_{yz}-2b\lambda,$$

$$0=\frac{\partial L}{\partial c}=2s_z^2c+2bs_{yz}+2as_{zx}-2c\lambda,$$

$$0=\frac{\partial L}{\partial \lambda}=a^2+b^2+c^2-1$$

ラグランジュの未定係数法の適用

$$\begin{bmatrix} s_x^2 & s_{xy} & s_{zx} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{zx} & s_{yz} & s_z^2 \end{bmatrix}\begin{bmatrix} a \\ b \\ c \end{bmatrix}=\lambda\begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

$$a^2+b^2+c^2=1$$

主成分の十分条件

$$s_u^2=a^2s_x^2+b^2s_y^2+c^2s_z^2+abs_{xy}+bcs_{yz}+cas_{zx}$$

$$=\begin{bmatrix} a & b & c \end{bmatrix}\begin{bmatrix} s_x^2 & s_{xy} & s_{zx} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{zx} & s_{yz} & s_z^2 \end{bmatrix}\begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

$$=\begin{bmatrix} a & b & c \end{bmatrix}\lambda\begin{bmatrix} a \\ b \\ c \end{bmatrix}=\lambda\left(a^2+b^2+c^2\right)=\lambda.$$

主成分の十分条件

uの分散 s_u^2 の極値を与える (a,b,c) は，

分散共分散行列 $\begin{bmatrix} s_x^2 & s_{xy} & s_{zx} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{zx} & s_{yz} & s_z^2 \end{bmatrix}$ の固有ベクトル

であり，そのとき， $s_u^2=\lambda$ （＝固有ベクトル）となる．

したがって，s_u^2 の最大値を与える(a,b,c)は，分散共分散行列の最大の固有値に属する固有ベクトル（で，長さが1のもの）である．

主成分の解釈

No	国(x)	社(y)	英(z)	理(v)	数(w)
1	66	79	44	79	75
2	75	84	61	76	82
3	59	65	55	58	43
4	74	76	51	69	67
5	86	71	67	74	50
6	87	92	54	71	81
7	73	89	53	71	51
8	79	85	49	52	49
9	87	73	64	82	97
10	88	87	78	70	88

主成分の解釈

$$u = ax + by + cz + dv + ew$$

$$u = 0.298x + 0.133y + 0.237z + 0.303v + 0.863w$$

新変数uは，各教科の得点を適当な正の重みで加え合わせたものである．したがって，uは総合学力を表していると考えられる．

主成分得点

No	国(x)	社(y)	英(z)	理(v)	数(w)	u
1	66	79	44	79	75	129.27
2	75	84	61	76	82	141.77
3	59	65	55	58	43	93.95
4	74	76	51	69	67	122.98
5	86	71	67	74	50	116.52
6	87	92	54	71	81	142.38
7	73	89	53	71	51	111.68
8	79	85	49	52	49	104.50
9	87	73	64	82	97	159.36
10	88	87	78	70	88	153.44

Excelで学ぼう

ファイル：第3章/3_2

本日のまとめ

- **主成分uがどのようにして，得られるかを理解した．（データの分散共分散行列の最大固有値に属する固有ベクトルから得られる．）**
- **主成分得点の計算の仕方を理解した．**
- **Excelを用いて，主成分を計算する方法，主成分得点を計算する方法を理解した．**