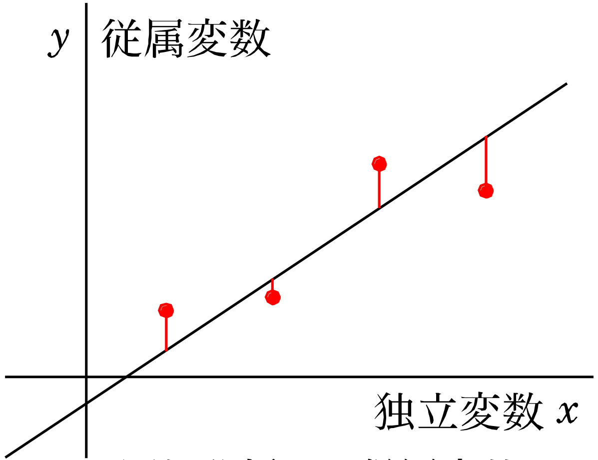
数値計算（７）

最小二乗法

主成分分析

最小二乗法



回帰分析では従属変数と
独立変数の関係を推定

データを直線で近似（回帰直線）
一次回帰曲線

$$y = ax + b$$

データ (x_i, y_i) に対する残差

$$y_i - (ax_i + b)$$

残差平方和 → 最小

$$S = \sum_{i=1}^{N} \{ y_i - (ax_i + b) \}^2$$

２乗和を最小 ⇒ 偏微分＝0

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{N} (y_i - ax_i - b)\, x_i = 0$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{N} (y_i - ax_i - b) = 0$$

$$\begin{cases} \left(\sum_{i=1}^{N} x_i \right) a + \left(\sum_{i=1}^{N} 1 \right) b = \sum_{i=1}^{N} y_i \\ \left(\sum_{i=1}^{N} x_i^2 \right) a + \left(\sum_{i=1}^{N} x_i \right) b = \sum_{i=1}^{N} x_i y_i \end{cases}$$

数値計算

x	0.0	1.0	2.0	3.0	4.0	5.0
y	10.2	12.0	15.7	17.0	20.5	22.4

$$\left(\sum_{i=1}^{6} x_i\right) = 15 \qquad \left(\sum_{i=1}^{6} x_i^2\right) = 55 \qquad \left(\sum_{i=1}^{6} y_i\right) = 97.8 \qquad \left(\sum_{i=1}^{6} x_i y_i\right) = 288.4$$

$$\begin{cases} 15a + 6b = 97.8 \\ 55a + 15b = 288.4 \end{cases} \qquad \begin{cases} a = 2.51 \\ b = 10.03 \end{cases}$$

$$y = 2.51x + 10.03$$

多項式近似（回帰曲線）

$$y = P_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

残差　$y_i - P_n(x_i)$　　　残差平方和　$S = \sum_{i=1}^{N} \{ y_i - P_n(x_i) \}^2$

a_j で偏微分 = 0

$$\frac{\partial S}{\partial a_j} = -2 \sum_{i=1}^{N} (y_i - P_n(x_i))\, x_i{}^j = 0\,, \qquad \sum_{i=1}^{N} P_n(x_i)\, x_i{}^j = \sum_{i=1}^{N} x_i{}^j y_i$$

$$\begin{array}{llll}
j = 0 & \left(\sum_{i=1}^{N} 1\right) a_0 + \left(\sum_{i=1}^{N} x_i\right) a_1 + & \cdots + \left(\sum_{i=1}^{N} x_i{}^n\right) a_n & = \sum_{i=1}^{N} y_i \\
j = 1 & \left(\sum_{i=1}^{N} x_i\right) a_0 + \left(\sum_{i=1}^{N} x_i{}^2\right) a_1 + & \cdots + \left(\sum_{i=1}^{N} x_i{}^{n+1}\right) a_n & = \sum_{i=1}^{N} x_i\, y_i \\
\vdots & & & \vdots \\
j = n & \left(\sum_{i=1}^{N} x_i{}^n\right) a_0 + \left(\sum_{i=1}^{N} x_i{}^{n+1}\right) a_1 + & \cdots + \left(\sum_{i=1}^{N} x_i{}^{2n}\right) a_n & = \sum_{i=1}^{N} x_i{}^n\, y_i
\end{array}$$

数値計算 $y = P_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$

x	0.0	1.0	2.0	3.0	4.0	5.0
y	10.2	14.0	15.7	16.0	18.5	22.4

$$\left(\sum_{i=1}^{6} 1\right) = 6, \quad \left(\sum_{i=1}^{6} x_i\right) = 15, \quad \left(\sum_{i=1}^{6} x_i^2\right) = 55, \quad \left(\sum_{i=1}^{6} x_i^3\right) = 255,$$

$$\left(\sum_{i=1}^{6} x_i^4\right) = 979, \quad \left(\sum_{i=1}^{6} x_i^5\right) = 4425, \quad \left(\sum_{i=1}^{6} x_i^6\right) = 20515,$$

$$\sum_{i=1}^{6} y_i = 96.8, \quad \sum_{i=1}^{6} x_i y_i = 279.4, \quad \sum_{i=1}^{6} x_i^2 y_i = 1076.8, \quad \sum_{i=1}^{6} x_i^3 y_i = 4555.6$$

$$\begin{cases} 6a_0 + 15a_1 + 55a_2 + 255a_3 = 96.8 \\ 15a_0 + 55a_1 + 255a_2 + 979a_3 = 279.4 \\ 55a_0 + 255a_1 + 979a_2 + 4425a_3 = 1076.8 \\ 255a_0 + 979a_1 + 4425a_2 + 20515a_3 = 4555.6 \end{cases}$$

多項式近似　行列版

$$\begin{pmatrix} \sum_{i=1}^{N} 1 & \sum_{i=1}^{N} x_i & \cdots & \sum_{i=1}^{N} x_i{}^n \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i{}^2 & \cdots & \sum_{i=1}^{N} x_i{}^{n+1} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{N} x_i{}^n & \sum_{i=1}^{N} x_i{}^{n+1} & \cdots & \sum_{i=1}^{N} x_i{}^{2n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i\, y_i \\ \\ \sum_{i=1}^{N} x_i{}^n\, y_i \end{pmatrix}$$

ここで,

$$\boldsymbol{A} = \begin{pmatrix} 1 & x_1 & x_1{}^2 & \cdots & x_1{}^n \\ 1 & x_2 & x_2{}^2 & \cdots & x_2{}^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N & x_N{}^2 & \cdots & x_N{}^n \end{pmatrix}$$

とすると, $\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{a} = \boldsymbol{A}^t\boldsymbol{y}$

線形最小二乗法

$$y = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_n f_n(x) = \sum_{k=1}^{n} a_k f_k(x)$$

残差平方和 $S = \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{n} a_k f_k(x_i) \right)^2$

$$\frac{\partial S}{\partial a_j} = -2 \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{n} a_k f_k(x_i) \right) f_j(x_i) = 0 \,, \qquad \sum_{k=1}^{n} \sum_{i=1}^{N} f_j(x_i) f_k(x_i) a_k = \sum_{i=1}^{N} f_j(x_i)\, y_i$$

ここで,

$$\boldsymbol{A} = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_N) & f_2(x_N) & \cdots & f_n(x_N) \end{pmatrix}$$ とすると, $\boldsymbol{A}^t \boldsymbol{A} \boldsymbol{a} = \boldsymbol{A}^t \boldsymbol{y}$

数値計算 $f_1(x)=1,\quad f_2(x)=x,\quad f_3(x)=\sin(\frac{2\pi}{5}x)$

x	0.0	1.0	2.0	3.0	4.0	5.0
y	10.2	14.0	15.7	16.0	18.5	22.4

$$\boldsymbol{A}=\begin{pmatrix} 1 & 0 & 0.00 \\ 1 & 1 & 0.95 \\ 1 & 2 & 0.58 \\ 1 & 3 & -0.58 \\ 1 & 4 & -0.95 \\ 1 & 5 & 0.00 \end{pmatrix} \qquad \boldsymbol{y}=\begin{pmatrix} 10.2 \\ 14.0 \\ 15.7 \\ 16.0 \\ 18.5 \\ 22.4 \end{pmatrix}$$

$\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{a}=\boldsymbol{A}^t\boldsymbol{y}$ を解くと，

$$\begin{cases} a_0 = 10.01 \\ a_1 = 2.45 \\ a_2 = 1.59 \end{cases}$$

よって，

$$y = 10.01 + 2.45x + 1.59\sin(\frac{2\pi}{5}x)$$

多次元の場合

独立変数 $x_1,\ x_2,\ \cdots, x_n$ と従属変数 y

j 番目のデータを $(x_{j1}, x_{j2},\ \cdots, x_{jn}, y_j)$ と表すと $(1 \leq j \leq N)$

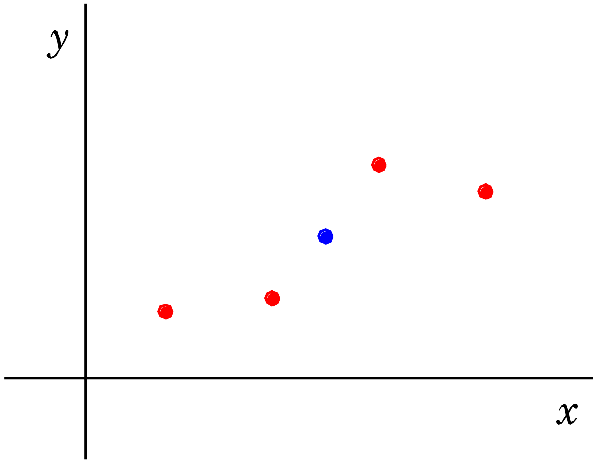
$x_1 = f_1(\xi),\ x_2 = f_2(\xi),\ \cdots, x_n = f_n(\xi)$ と考える

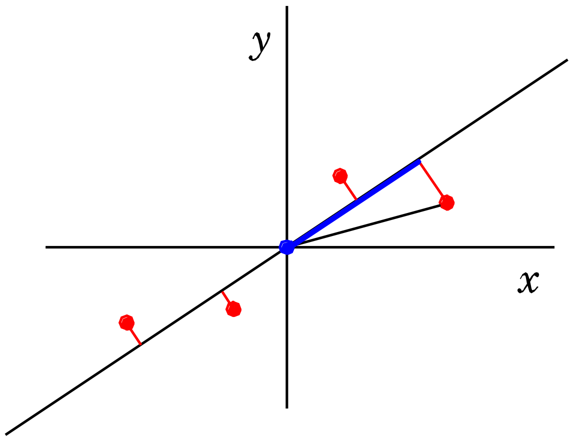
$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = \sum_{k=1}^{n} a_k x_k$$

$f_k(\xi)$?　　計算に必要なのは，データ $f_k(\xi_j) = x_{jk}$

$$\boldsymbol{A} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix} \qquad \boldsymbol{A}^t\boldsymbol{A}\boldsymbol{a} = \boldsymbol{A}^t\boldsymbol{y}$$

主成分分析





生データ (x^*_i, y^*_i)

データの平均 $(\overline{x^*}, \overline{y^*})$

$$x_i = x^*_i - \overline{x^*}$$
$$y_i = y^*_i - \overline{y^*}$$

$$\boldsymbol{x}_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

データを最も良く表す直線

$$y = \frac{b}{a} x$$

データ (x_i, y_i) に対する距離最小

$$\Downarrow$$

$\boldsymbol{x}_i$ の直線に対する**射影長**最大

$$\boldsymbol{u} = \begin{pmatrix} a \\ b \end{pmatrix} \qquad (a^2 + b^2 = 1)$$

射影長は $\boldsymbol{x}_i{}^t \boldsymbol{u}$

射影長の平方和 $= \sum_{i=1}^{N} (\boldsymbol{x}_i{}^t \boldsymbol{u})^2$

先ほどと同様に

$$\boldsymbol{A} = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_N & y_N \end{pmatrix} \quad \text{とおくと,} \quad \boldsymbol{A}\boldsymbol{u} = \begin{pmatrix} \boldsymbol{x}_1{}^t \boldsymbol{u} \\ \boldsymbol{x}_2{}^t \boldsymbol{u} \\ \vdots \\ \boldsymbol{x}_N{}^t \boldsymbol{u} \end{pmatrix}$$

$$(\boldsymbol{A}\boldsymbol{u})^t \boldsymbol{A}\boldsymbol{u} = (\boldsymbol{x}_1{}^t \boldsymbol{u} \quad \boldsymbol{x}_2{}^t \boldsymbol{u} \quad \cdots \quad \boldsymbol{x}_N{}^t \boldsymbol{u}) \begin{pmatrix} \boldsymbol{x}_1{}^t \boldsymbol{u} \\ \boldsymbol{x}_2{}^t \boldsymbol{u} \\ \vdots \\ \boldsymbol{x}_N{}^t \boldsymbol{u} \end{pmatrix} = \sum_{i=1}^{N} (\boldsymbol{x}_i^t \boldsymbol{u})^2$$

$$\parallel$$

$\boldsymbol{u}^t \boldsymbol{A}^t \boldsymbol{A}\boldsymbol{u}$ ←これを最大にすれば良い

$\boldsymbol{u}^t(\boldsymbol{A}^t \boldsymbol{A}\boldsymbol{u})$ は， $\boldsymbol{u}$ と $\boldsymbol{A}^t \boldsymbol{A}\boldsymbol{u}$ の内積

$\boldsymbol{A}^t \boldsymbol{A}$ により向きが変わらないベクトル（固有ベクトル）で，

変換倍率（固有値）が最大のものを選べば良いのでは？

$\boldsymbol{A}^t \boldsymbol{A}$ は分散共分散行列であり，固有値はすべて非負である

数値計算

x	0.0	1.0	2.0	3.0	4.0	5.0
y	10.2	12.0	15.7	17.0	20.5	22.4

$$\overline{x} = 2.5$$
$$\overline{y} = 16.3$$

$$\boldsymbol{A} = \begin{pmatrix} -2.5 & -6.1 \\ -1.5 & -4.3 \\ -0.5 & -0.6 \\ 0.5 & 0.7 \\ 1.5 & 4.2 \\ 2.5 & 6.1 \end{pmatrix}$$

$$\boldsymbol{A}^t\boldsymbol{A} = \begin{pmatrix} 17.5 & 43.9 \\ 43.9 & 111.4 \end{pmatrix}$$

$\boldsymbol{A}^t\boldsymbol{A}$　固有値，固有ベクトルは，

$$\det(\boldsymbol{A}^t\boldsymbol{A} - \lambda \boldsymbol{E}) = 0$$

$(17.5-\lambda)(111.4-\lambda) - 43.9^2 = 0$ を解いて

$\lambda_1 = 128.73$　　$\lambda_1 = 0.17$

$$\boldsymbol{u}_1 = \begin{pmatrix} 0.37 \\ 0.93 \end{pmatrix} \qquad \boldsymbol{u}_2 = \begin{pmatrix} -0.93 \\ 0.37 \end{pmatrix}$$

$$y = \frac{0.93}{0.37}(x - 2.5) + 16.3$$

$$y = 2.51x + 9.67$$

多次元のとき

$$\boldsymbol{A} = \begin{array}{c} \longleftarrow \quad n \text{ 次元} \quad \longrightarrow \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix} \end{array} \quad \begin{array}{c} \uparrow \\ N \text{ データ} \\ \downarrow \end{array}$$

$\boldsymbol{A}^t\boldsymbol{A}$ の最大の固有値に対応する固有ベクトル $\boldsymbol{u}_1$

$$\boldsymbol{z}_1 = \boldsymbol{A}\boldsymbol{u}_1 \qquad \underbrace{\begin{pmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{N1} \end{pmatrix}}_{\text{主成分 } \boldsymbol{z}_1} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix} \underbrace{\begin{pmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{n1} \end{pmatrix}}_{\text{固有ベクトル } \boldsymbol{u}_1}$$

n次元 $\longrightarrow$ 1 次元

$\boldsymbol{A}^t\boldsymbol{A}$ の大きい方から1～k番目の固有値に対応する固有ベクトル

$\boldsymbol{u}_1 \ \boldsymbol{u}_2 \ \cdots \ \boldsymbol{u}_\mathrm{k}$

$$\boldsymbol{Z}_k = \boldsymbol{A}\boldsymbol{U}_k \qquad \begin{array}{c} \text{第1主成分 } \boldsymbol{z}_1 \quad \text{第}k\text{主成分 } \boldsymbol{z}_k \\ \downarrow \qquad\qquad \downarrow \\ \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{Nk} \end{pmatrix} \end{array} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{pmatrix} \begin{array}{c} \boldsymbol{u}_1 \qquad\qquad \boldsymbol{u}_\mathrm{k} \\ \downarrow \qquad\qquad \downarrow \\ \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1k} \\ u_{21} & u_{22} & \cdots & u_{2k} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{nk} \end{pmatrix} \end{array}$$

n 次元　→　k 次元　（次元削減，次元圧縮）

$$\text{寄与率} = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i} \qquad\qquad \text{累積寄与率} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$