

主成分分析

データ数を n 、変量数を m とする。

$$\text{データを } [X] = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \quad \text{と表す。}$$

$$\text{各変量の平均値ベクトルを } \{\mu\} = \begin{Bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{Bmatrix} \quad \text{とし、}$$

$$\text{また 行列 } [\mu] = \begin{bmatrix} \mu_1 & \cdots & \mu_m \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_m \end{bmatrix} \quad \Bigg\} n \quad \text{とする。}$$

$$\text{分散共分散行列は } [V] = \frac{1}{n-1} ([X] - [\mu])^t ([X] - [\mu])$$

と表現できる。

これは変形すると以下のようになる。

$$[V] = \frac{1}{n-1} ([X]^t [X] - [X]^t [\mu] - [\mu]^t [X] + [\mu]^t [\mu])$$

$$[X]^t [\mu] = \begin{bmatrix} x_{1,1} & \cdots & x_{n,1} \\ \vdots & \ddots & \vdots \\ x_{1,m} & \cdots & x_{n,m} \end{bmatrix} \begin{bmatrix} \mu_1 & \cdots & \mu_m \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_m \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n x_{i,1} \mu_1 & \cdots & \sum_{i=1}^n x_{i,1} \mu_m \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{i,m} \mu_1 & \cdots & \sum_{i=1}^n x_{i,m} \mu_m \end{bmatrix}$$

$$= \begin{bmatrix} n\mu_1\mu_1 & \cdots & n\mu_1\mu_m \\ \vdots & \ddots & \vdots \\ n\mu_m\mu_1 & \cdots & n\mu_m\mu_m \end{bmatrix}$$

$$= n \begin{bmatrix} \mu_1\mu_1 & \cdots & \mu_1\mu_m \\ \vdots & \ddots & \vdots \\ \mu_m\mu_1 & \cdots & \mu_m\mu_m \end{bmatrix}$$

$$= n\{\mu\}\{\mu\}^t$$

となる。したがって、

$$[\mu]^t[X] = (n\{\mu\}\{\mu\}^t)^t = n\{\mu\}\{\mu\}^t$$

また

$$\begin{aligned} [\mu]^t[\mu] &= \begin{bmatrix} \mu_1 & \cdots & \mu_1 \\ \vdots & \ddots & \vdots \\ \mu_m & \cdots & \mu_m \end{bmatrix} \begin{bmatrix} \mu_1 & \cdots & \mu_m \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_m \end{bmatrix} \\ &= \begin{bmatrix} n\mu_1\mu_1 & \cdots & n\mu_1\mu_m \\ \vdots & \ddots & \vdots \\ n\mu_m\mu_1 & \cdots & n\mu_m\mu_m \end{bmatrix} \\ &= n\{\mu\}\{\mu\}^t \end{aligned}$$

であるから、分散共分散行列は 以下のようになる。

$$\begin{aligned} [V] &= \frac{1}{n-1} ([X]^t[X] - [X]^t[\mu] - [\mu]^t[X] + [\mu]^t[\mu]) \\ &= \frac{1}{n-1} ([X]^t[X] - n\{\mu\}\{\mu\}^t - n\{\mu\}\{\mu\}^t + n\{\mu\}\{\mu\}^t) \\ &= \frac{1}{n-1} ([X]^t[X] - n\{\mu\}\{\mu\}^t) \end{aligned}$$

$$[X] \text{ と適当な係数ベクトル } \{a\} = \begin{Bmatrix} a_1 \\ \vdots \\ a_m \end{Bmatrix} \quad (\text{ただし } \{a\}^t\{a\} = 1)$$

との一次結合 $\{y\} = [X]\{a\}$ を考え、 $\{y\}$ の分散が最大に

なるような $\{a\}$ を求める。

$\{y\}$ の平均を μ_y (スカラー) とする。つまり $\mu_y = \{a\}^t\{\mu\}$ 。

$\{y\}$ の分散を v (スカラー) と表示すると、

$$\begin{aligned} v &= \frac{1}{n-1} (\{y\}^t\{y\} - n\mu_y\mu_y) \\ &= \frac{1}{n-1} (\{a\}^t[X]^t[X]\{a\} - n\{a\}^t\{\mu\}\{\mu\}^t\{a\}) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n-1} \{a\}^t ([X]^t[X] - n\{\mu\}\{\mu\}^t) \{a\} \\
&= \{a\}^t[V]\{a\}
\end{aligned}$$

拘束条件 $\{a\}^t\{a\} = 1$ の下で、 $v = \{a\}^t[V]\{a\}$

を最大にする $\{a\}$ を求める問題となるので、L a g r a n g e の
未定乗数法により、

$$h(\{a\}, \lambda) = \{a\}^t[V]\{a\} - \lambda(\{a\}^t\{a\} - 1)$$

を最大化することと同じである。

$$\frac{\partial h(\{a\}, \lambda)}{\partial \{a\}} = 2[V]\{a\} - 2\lambda\{a\} = 0 \quad \text{となり、結局}$$

$[V]\{a\} = \lambda\{a\}$ の固有値問題となる。

一般に 実対称行列 $[X]$ の二次形式について以下が成り立つ。

$$\text{任意の}\{a\} \neq 0\text{について、} \lambda_{min} \leq \frac{\{a\}^t[X]\{a\}}{\{a\}^t\{a\}} \leq \lambda_{max}$$

であるから、最大値を求める場合には、最大の固有値に対する

固有ベクトル $\{a\}$ が解となる。

固有値を最大値 λ_1 から最小値 λ_m 、 $\lambda_{total} = \sum_{k=1}^m \lambda_k$ 、

$$\lambda_c = \sum_{k=1}^c \lambda_k \quad (c \leq m) \text{ として、実際には } \frac{\lambda_c}{\lambda_{total}} \text{ が一定以上}$$

(例えば95%) になる c を求め、対応する固有ベクトル $\{a_1\} \sim \{a_c\}$ を
解とできる。

$\{a_1\} \sim \{a_c\}$ は実対称行列の固有ベクトルなので互いに直交している。

この $\{a_k\} (k = 1 \sim c)$ を主成分ベクトルと呼び、元のデータ

$$[X] = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \quad \text{の行方向の } n \text{ 個のベクトル}$$

$$\begin{Bmatrix} x_{1,1} \\ \vdots \\ x_{1,m} \end{Bmatrix}, \begin{Bmatrix} x_{2,1} \\ \vdots \\ x_{2,m} \end{Bmatrix} \cdots \begin{Bmatrix} x_{n,1} \\ \vdots \\ x_{n,m} \end{Bmatrix} \quad \text{から平均値ベクトル } \{\mu\} = \begin{Bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{Bmatrix} \quad \text{を引き、}$$

$\{a_k\} (k = 1 \sim c)$ への射影としたものを主成分スコアと呼ぶ。

また、主成分 k と元の変量 j との相関係数を 以下で求めたもの

$$r_{j,k} = a_{j,k} \sqrt{\frac{\lambda_k}{v_{j,j}}} \quad j = 1 \sim m, k = 1 \sim c$$

を主成分負荷量という。

但し $a_{j,k}$ は固有ベクトル $\{a_k\} (k = 1 \sim c)$ の第 j 成分($j = 1 \sim m$)

$v_{j,j}$ は分散共分散行列 $[V]$ の (j, j) 成分 ($j = 1 \sim m$) である。

また $\sum_{k=1}^c r_{j,k}^2$ は変量 j の主成分負荷量の寄与率として表示する。