主成分分析と合成変数

多変量解析では，複数個の変数の重み付き合計値を用いる。解析の目的により各種の重み付けを行う。p 個の変数 x_1，x_2，$\cdots$，x_p が，重み w_1，w_2，$\cdots$，w_p で重み付けされた f を，合成変数と呼ぶ。

$$f = w_1\, x_1 + w_2\, x_2 + \cdots + w_p\, x_p \tag{1}$$

ここで，x_i はそれぞれ標準化されており，それぞれの変数が n ケースについて測定されているとき，$n \times p$ の大きさのデータ行列を $\boldsymbol{X}$ で表すことにする。

$$\underset{n\times p}{\boldsymbol{X}} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & x_{ij} & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{2}$$

変数 x_j の平均値 $E(x_j)$ と分散 $V(x_j)$ はそれぞれ定義により，

$$E(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0 \tag{3}$$

$$V(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \tag{4}$$

となる。

また，変数 x_j と x_k の共分散 Cov_{jk} と相関係数 r_{jk} は，

$$Cov_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\, x_{ik} \tag{5}$$

$$r_{jk} = \frac{Cov(x_j,\ x_k)}{\sqrt{V(x_j)\ V(x_k)}} = Cov(x_j,\ x_k) \tag{6}$$

となる。相関係数行列 $\boldsymbol{R}$ は，データ行列 $\boldsymbol{X}$ を用いて，

$$\underset{p\times p}{\boldsymbol{R}} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & r_{ij} & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \tag{7}$$

のように表される（$\boldsymbol{X}'$ は転置行列を表す）。

重みベクトルを $\boldsymbol{w}' = (w_1,\ w_2,\ \cdots,\ w_p)$ として，

$$\boldsymbol{f} = \boldsymbol{X}\boldsymbol{w} \tag{8}$$

で表すことにする。

合成変数 f の平均値と分散は，

$$\begin{aligned} E(f) &= \frac{1}{n}\sum_{i=1}^{n} f_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} w_j\, x_{ij}\right) \\ &= \sum_{j=1}^{p} w_j \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\right) \\ &= 0 \end{aligned} \tag{9}$$

$$\begin{aligned} V(f) &= \frac{1}{n}\sum_{i=1}^{n} f_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} w_j\, x_{ij}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} w_j\, x_{ij}\right)\left(\sum_{k=1}^{p} w_k\, x_{ik}\right) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p} w_j\, w_k \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\right) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p} w_j\, w_k\, r_{jk} \\ &= \boldsymbol{w}'\boldsymbol{R}\boldsymbol{w} \end{aligned} \tag{10}$$

となる。

標準化された合成変数ベクトル $\boldsymbol{g}$ は，

$$\boldsymbol{g} = \frac{\boldsymbol{f}}{\sqrt{V(f)}} = \frac{\boldsymbol{X}\boldsymbol{w}}{\sqrt{\boldsymbol{w}'\boldsymbol{R}\boldsymbol{w}}} \tag{11}$$

となり，合成変数 g と，もとの変数 x_j との相関係数は，

$$a_j = r(g,\ x_j) = r(f,\ x_j) = \frac{1}{n}\sum_{i=1}^{n} g_i\, x_{ij} \tag{12}$$

となるので，全ての変数との相関係数ベクトルは，

$$\begin{aligned} \boldsymbol{a} &= \frac{1}{n}\boldsymbol{X}'\boldsymbol{g} \\ &= \frac{1}{n}\boldsymbol{X}'\frac{\boldsymbol{X}\boldsymbol{w}}{\sqrt{\boldsymbol{w}'\boldsymbol{R}\boldsymbol{w}}} \\ &= \frac{\boldsymbol{R}\boldsymbol{w}}{\sqrt{\boldsymbol{w}'\boldsymbol{R}\boldsymbol{w}}} \end{aligned} \tag{13}$$

ここで，データ行列から標準化された合成変数を直接求めることのできる重みベクトル（**標準重みベクトル**）を $\boldsymbol{w}_s$ とする。すなわち，

$$\boldsymbol{g} = \boldsymbol{X}\boldsymbol{w}_s \tag{14}$$

とすると，（11）式から，$\boldsymbol{w}$ が既知であるときは，

$$\boldsymbol{w}_s = \frac{\boldsymbol{w}}{\sqrt{\boldsymbol{w}'\boldsymbol{R}\boldsymbol{w}}} \tag{15}$$

となる。

また，$\boldsymbol{a}$ が既知のときは，（14）式の両辺に左から $\frac{1}{n}\boldsymbol{X}'$ をかけて，

$$\begin{aligned} \boldsymbol{g} &= \boldsymbol{X}\boldsymbol{w}_s \\ \frac{1}{n}\boldsymbol{X}'\boldsymbol{g} &= \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{w}_s \\ \boldsymbol{a} &= \boldsymbol{R}\boldsymbol{w}_s \end{aligned} \tag{16}$$

さらに，両辺に $\boldsymbol{R}$ の逆行列 $\boldsymbol{R}^{-1}$ を 左からかけることにより，

$$\boldsymbol{R}^{-1}\boldsymbol{a} = \boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{w}_s = \boldsymbol{w}_s \tag{17}$$

が得られる。

相関係数行列 $\boldsymbol{R}$ の固有値を λ，固有ベクトルを $\boldsymbol{u}$ とすると，（18）式に示すような性質がある。

$$\boldsymbol{R}\boldsymbol{u} = \lambda\boldsymbol{u}, \qquad \lambda = \boldsymbol{u}'\boldsymbol{R}\boldsymbol{u}, \qquad \boldsymbol{u}'\,\boldsymbol{u} = 1 \tag{18}$$

これらと前述の関連式を組み合わすと，主成分分析の場合には以下のような関連があることを示せる。

$$V(f) \equiv \lambda \qquad \text{固有値} \tag{19}$$

$$\boldsymbol{w} \equiv \boldsymbol{u} \qquad \text{固有ベクトル} \tag{20}$$

$$\boldsymbol{a} = \frac{\boldsymbol{R}\boldsymbol{w}}{\sqrt{\boldsymbol{w}'\boldsymbol{R}\boldsymbol{w}}} = \frac{\lambda\boldsymbol{w}}{\sqrt{\lambda}} = \sqrt{\lambda}\boldsymbol{w} \qquad \text{主成分負荷量} \tag{21}$$

$$\boldsymbol{w}_s = \frac{\boldsymbol{w}}{\sqrt{\lambda}} \qquad \text{主成分得点係数} \tag{22}$$