V-3. 分散共分散行列の活用

V-3-1. 分散共分散行列

V-3-1-1 分散共分散行列の計算

データが得られた時、私たちが最初にすることはデータシートを作ることです。そして、これらを使って、基礎的な統計値たとえば、平均、最大値、最小値を求めたり、データを分布の仕方を見ます。それに加えて、この段階で、得られたデータの相互の関係を確認しておくべきです。なぜならば、データの各要因間にたがいに相関があると（多重共線性）相関のある要因同士が引っ張りあって、統計的なパラメータの推定値が不安定になるからです。そういう場合は、相関のある要因を足し合わせたりして一つの要因に求めるとか、代表的要因一つに絞り込む。あるいは、交差項を作って、要因同士の組み合わせによって生まれる効果を取り分けるとか、何らかの対応をすることが求められます。基礎統計量の計算に続いて、要因間の関係を確かめておくことは、統計解析に先立ってやるべき基本的な作業です。そのために行われるのが、分散・共分散行列あるいは相関行列の作成です。どちらも、多変量解析のベースとなる主成分分析の基礎データになります。
そのためのデータシートを作るフォーマットを自分で作っている分析者もいますが、多くの場合、データシートはエクセルのような既存のソフトウェアーを使って作ります。表 41 はその一例です。下の列に、合計や平均、分散などを計算しておきます。

表 37. データシートの例

	要因			
標本番号	A	B	…	P
1	d_{11}	d_{12}	…	d_{1p}
2	d_{21}	d_{22}	…	d_{2p}
⋮	⋮	⋮	⋱	⋮
n	d_{n1}	d_{n2}	…	d_{np}
平均	$\bar{d}_1 = \frac{1}{n}\sum_{i=1}^{n} d_{i1}$	$\bar{d}_2 = \frac{1}{n}\sum_{i=1}^{n} d_{i2}$	…	$\bar{d}_p = \frac{1}{n}\sum_{i=1}^{n} d_{ip}$

次に、データを平均値からの距離（$c_{ij} = d_{ij} - \bar{d}_j$）にした表を作ります。表 42 がその例です。

表 38. 平均値からの距離として標準化したデータシート

	要因			
標本番号.	A	B	…	P
1	c_{11}	c_{12}	…	c_{1p}
2	c_{21}	c_{22}	…	c_{2p}
⋮	⋮	⋮	⋱	⋮
n	c_{n1}	c_{n2}	…	c_{np}

表 38 から次のような行列が作れます。表 38 のようにそれぞれの標本を縦に並べた行列を転置行列と考えるところがポイントです。

$$\boldsymbol{c}^T = (\boldsymbol{c_1} \quad \boldsymbol{c_2} \quad \cdots \quad \boldsymbol{c_p}) = \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{p1} \\ c_{12} & c_{22} & \cdots & c_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1n} & c_{2n} & \cdots & c_{pn} \end{pmatrix}$$

$$\boldsymbol{c} = \begin{pmatrix} \boldsymbol{c_1} \\ \boldsymbol{c_2} \\ \vdots \\ \boldsymbol{c_p} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pn} \end{pmatrix}$$

この２つの行列を次のように掛け合わせます。これを分散・共分散行列といいます。

$$\boldsymbol{cc}^T = \begin{pmatrix} \boldsymbol{c_1} \\ \boldsymbol{c_2} \\ \vdots \\ \boldsymbol{c_p} \end{pmatrix} (\boldsymbol{c_1} \quad \boldsymbol{c_2} \quad \cdots \quad \boldsymbol{c_p}) = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pn} \end{pmatrix} \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{p1} \\ c_{12} & c_{22} & \cdots & c_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1n} & c_{2n} & \cdots & c_{pn} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{k=1}^{n} c_{1k}c_{1k} & \sum_{k=1}^{n} c_{1k}c_{2k} & \cdots & \sum_{k=1}^{n} c_{1k}c_{pk} \\ \sum_{k=1}^{n} c_{2k}c_{1k} & \sum_{k=1}^{n} c_{2k}c_{2k} & \cdots & \sum_{k=1}^{n} c_{2k}c_{pk} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} c_{pk}c_{1k} & \sum_{k=1}^{n} c_{pk}c_{2k} & \cdots & \sum_{k=1}^{n} c_{pk}c_{pk} \end{pmatrix}$$

分散・共分散行列という名前ですが、行列の因子は、平方和あるいは積和です。実際には、これらの値を自由度で割ったものが分散や共分散ですが。もし、求めるものが母集団の分散や共分散であれば、自由度は$n-1$あるいは$n-2$ですが、ここで問題にしているのは標本手段の分布ですから、自由度はnで良いでしょう。

$$\boldsymbol{E}(\boldsymbol{cc}^T) = \frac{1}{n} \begin{pmatrix} \sum_{k=1}^{n} c_{1k}c_{1k} & \sum_{k=1}^{n} c_{1k}c_{2k} & \cdots & \sum_{k=1}^{n} c_{1k}c_{pk} \\ \sum_{k=1}^{n} c_{2k}c_{1k} & \sum_{k=1}^{n} c_{2k}c_{2k} & \cdots & \sum_{k=1}^{n} c_{2k}c_{pk} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} c_{pk}c_{1k} & \sum_{k=1}^{n} c_{pk}c_{2k} & \cdots & \sum_{k=1}^{n} c_{pk}c_{pk} \end{pmatrix}$$

$\boldsymbol{E}(\boldsymbol{cc}^T)$ は平均値ですから、私たちが母集団からｎ個のサンプルを取り出した時の期待値です。一般にはこの行列を$\boldsymbol{\Sigma}$と表します。したがって、次のように、行列の各因子は分散・共分散を使って σ_{ij}と表せます。.

$$\boldsymbol{\Sigma} = \boldsymbol{E}(\boldsymbol{cc}^T) = \frac{1}{n}\begin{pmatrix} \sum_{k=1}^{n} c_{1k}c_{1k} & \sum_{k=1}^{n} c_{1k}c_{2k} & \cdots & \sum_{k=1}^{n} c_{1k}c_{pk} \\ \sum_{k=1}^{n} c_{2k}c_{1k} & \sum_{k=1}^{n} c_{2k}c_{2k} & \cdots & \sum_{k=1}^{n} c_{2k}c_{pk} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} c_{pk}c_{1k} & \sum_{k=1}^{n} c_{pk}c_{2k} & \cdots & \sum_{k=1}^{n} c_{pk}c_{pk} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{p2} \end{pmatrix}$$

式 67

この行列は$\sigma_{ij} = \sigma_{ji}$ですから対称行列で。対角因子が分散で、$\sigma_{ii}$を一般に $\sigma_i{}^2$のように表しますが、σ_{ii} のように書いておいた方がわかりやすいかもしれません。この分散・共分散行列から、次のように計算すれば、相関行列$\boldsymbol{\rho}$が得られます。

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$$

$$\boldsymbol{\rho} = \begin{pmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{11}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$$

式 68

この式からわかるように、nは関係なくなるので、$\boldsymbol{\rho}$は$\boldsymbol{\Sigma}$から直接計算できます。
また、分散行列 $\boldsymbol{V}$は、次のような対角行列です。

$$\boldsymbol{V} = \begin{pmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{pmatrix}$$

対角行列ですから、その平方根は次のようになります。

$$\boldsymbol{V}^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{pmatrix}$$

以上より$\boldsymbol{\Sigma}$, $\boldsymbol{\rho}$, $\boldsymbol{V}$ の関係は次のように表せます。

$$\boldsymbol{V}^{\frac{1}{2}}\boldsymbol{\rho}\boldsymbol{V}^{\frac{1}{2}} = \boldsymbol{\Sigma}$$

$$\boldsymbol{V^{-\frac{1}{2}}\Sigma V^{-\frac{1}{2}} = \rho}$$

式 69

確かめてみます

$$\begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{pmatrix}^{-1}$$

$$\begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}} \end{pmatrix}$$

$$\begin{pmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\sigma_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\sigma_{pp}}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{11}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix} = \boldsymbol{\rho}$$

$$\boldsymbol{V^{-\frac{1}{2}}\Sigma V^{-\frac{1}{2}} = \rho}$$