因子分析の基本

因子分析とは
今nについて 3 種類のテスト成績x_1,x_2,x_3が得られているとする。各々の成績には高い相関が見られる時、因子分析（factor analysis）ではその相関が少数個の潜在的因子を想定することで説明する。因子が 1 つである時、

$$\left.\begin{aligned} x_{1i} &= \mu_1 + a_1 f_i + e_{1i} \\ x_{2i} &= \mu_2 + a_2 f_i + e_{2i} \\ x_{3i} &= \mu_3 + a_3 f_i + e_{3i} \end{aligned}\right\}(i = 1,2,...n)$$

ここで、f_iはi番目の個体の潜在的な共通因子（common facto）の値 = 因子得点（factor score）であり、この値は直接測定できず、平均 0、分散 1 とおく。係数a_1,a_2,a_3を因子負荷行列（factor loading）、e_{1i},e_{2i},e_{3i}を独自因子（unique factor）あるいは特殊因子（specific factor）と呼び、各テスト固有の変動を表す。e_1,e_2,e_3は平均 0、分散d_1^2,d_2^2,d_3^2を持つ確率変数で互いに無相関。またe_1,e_2,e_3とfも無相関とする
この時、分散共分散行列は、

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{bmatrix} = \begin{bmatrix} a_1^2 + d_1^2 & a_1 a_2 & a_1 a_3 \\ & a_2^2 + d_2^2 & a_2 a_3 \\ & & a_3^2 + d_3^2 \end{bmatrix}$$

である。

観測値から得られた分散共分散行列に対して、上の関係を満たす$a_1,a_2,a_3,d_1^2,d_2^2,d_3^2$が存在すればこの現象は 1 因子モデルで良く説明できていると考える。
今、p変量、n個体のデータが得られているとする。p個の変量間の相関をm個の因子で説明する。

$$\left.\begin{aligned} x_{1i} &= \mu_1 + a_{11} f_{1i} + ... + a_{1m} f_{mi} + e_{1i} \\ x_{pi} &= \mu_p + a_{p1} f_{1i} + ... + a_{pm} f_{mi} + e_{pi} \end{aligned}\right\}(i = 1,2,...n)$$

あるいは、
$\mathbf{x}_i = \boldsymbol{\mu} + \mathbf{A}\mathbf{f}_i + \mathbf{e}_i, (i = 1,2,...n)$
$\mathbf{x} : p \times 1, \boldsymbol{\mu} : 1 \times 1, \mathbf{A} : p \times m, \mathbf{f} : m \times 1, \mathbf{e} : p \times 1$
$\boldsymbol{\mu}$は平均ベクトル、$\mathbf{A}$は因子負荷行列（factor loading matrix）、$\mathbf{f}_i$は因子得点ベクトル（factor score vector）、$\mathbf{e}_i$は独自因子の得点ベクトルである。$f_1,...,f_m$は平均 0、分散 1、

$e_1,...,e_p$ は平均 0、分散 $d_1^2,...,d_p^2$ 、独自因子間および独自因子と共通因子の間は無相関。共通因子 $f_1,...,f_m$ 間の間を無相関と仮定する場合を直交因子（orthogonal factor）、そうでない場合を斜交因子（oblique factor）と呼ぶ。
直交因子の場合、 p 変量ベクトル $\mathbf{x}$ の分散共分散行列 $\mathbf{\Sigma}$ は、

$$\mathbf{\Sigma} = \mathbf{AA}' + \mathbf{D}$$

$\mathbf{D}$ は $p \times p$ 行列で、 $\mathbf{D} = diag\left(d_1^2,...,d_p^2\right)$ である。（分散共分散行列 $\mathbf{\Sigma}$ の共通因子分解）

観測値 $\mathbf{x}$ に基づいてこの関係を満たす因子負荷量 $\mathbf{A}$ と独自因子の分散 $\mathbf{D}$ を求める。
また、 $\mathbf{\Sigma}$ の対角要素は、

$$\sigma_{jj} = h_j^2 + d_j^2, h_j^2 = a_{j1}^2 + ... + a_{jm}^2$$

と表される。 h_j^2 は共通因子 f による変動を表し、共通性（communality）と呼ばれる。

ところで、

$$\mathbf{x}_i = \boldsymbol{\mu} + \mathbf{Af}_i + \mathbf{e}_i, (i = 1,2,...n)$$

に対して、任意の $p \times p$ 直交行列を $\mathbf{T}$ とすると（ $\mathbf{TT}' = \mathbf{I}$ ）、

$$\begin{aligned}\mathbf{x}_i &= \mathbf{Af}_i + \mathbf{e}_i = \mathbf{A}\left(\mathbf{TT}'\right)\mathbf{f}_i + \mathbf{e}_i \\ &= \mathbf{AT} \times \mathbf{T}'\mathbf{f}_i + \mathbf{e}_i \\ &= \mathbf{A}^*\mathbf{f}_i^* + \mathbf{e}_i\end{aligned}$$

が成り立つ。ここで、 $\mathbf{Af}_i = \mathbf{A}^*\mathbf{f}_i^*$ なので、もし m 個の因子 f が x の共分散を良く説明しているのであれば、因子 f^* も同様に説明しており、 $\mathbf{A}^*$ と $\mathbf{f}_i^*$ もまた解である。 $\mathbf{T}$ が直交行列でなくても、 $\mathbf{TT}' = \mathbf{I}$ であったので、 $\mathbf{A}^* = \mathbf{AT}$ 、 $\mathbf{f}_i^* = \mathbf{T}^{-1}\mathbf{f}_i$ とおけば同様である。

このように因子には座標軸の変換＝回転の任意性があり、 $\mathbf{A}$ や $\mathbf{f}_i$ は不定である。 $\mathbf{T}$ が直交行列である場合を直交回転、そうでない場合を斜交回転という。

主因子法
$\mathbf{\Sigma} = \mathbf{AA}' + \mathbf{D}$ の関係から分散共分散行列の対角要素を共通性で置き換えた $p \times p$ 行列、

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} h_1^2 & \rho_{12} & .... & \rho_{1p} \\ \rho_{21} & h_2^2 & .... & \rho_{2p} \\ .... & .... & .... & .... \\ \rho_{p1} & \rho_{p2} & .... & h_p^2 \end{bmatrix} = \boldsymbol{\Sigma} - \mathbf{D}$$

は共通因子によって説明される部分の分散共分散を表し、もしモデルがよく当てはまっているのであれば、この$\boldsymbol{\Sigma}^*$は階数mの行列$\mathbf{A}$と$\mathbf{A}'$の積に分解できる。その時$\boldsymbol{\Sigma}^*$の階数もmになり、そのp個の固有値は、$\lambda_1 \geq \lambda_2 \geq ...\lambda_m \geq \lambda_{m+1} = ... = \lambda_p = 0$のように$m$個が正、$p-m$個が0になる。

このうち、固有値$\left(\boldsymbol{\Sigma}^* - \lambda\mathbf{I}\right)c = 0$の固有値$\lambda_1,...,\lambda_m$に対応する固有ベクトルを、

$c_1,...,c_m\left(\left\|c_j^2\right\| = 0\right)$とすると、$\boldsymbol{\Sigma}^*$は以下のように展開できる。

$$\boldsymbol{\Sigma}^* = \lambda_1 c_1 c_1' + ... + \lambda_m c_m c_m' = \left[\sqrt{\lambda_1}c_1,....,\sqrt{\lambda_1}c_m\right]\left[\sqrt{\lambda_m}c_1,....,\sqrt{\lambda_m}c_m\right]$$

したがって、

$$\mathbf{A} = \left[\sqrt{\lambda_1}c_1,....,\sqrt{\lambda_m}c_m\right]$$

とおけば因子負荷行列$\mathbf{A}$が得られる。
しかしこの方法で得られる$\mathbf{A}$は各変量の単位の取り方に依存する。実際には各変量の分散を 1 に標準化し分散共分散の変わりに相関行列を用いる方法が主流である。主因子法ではこの性質を利用して因子負荷量を反復アルゴリズムにより推定する。
（以下、現時点では略。また文中すべての添え字は要チェック）