岡本安晴　2002.8, 2019.3

判別分析

Fisher's Linear Discriminant Function

1. モデル

各個体から p 変量のデータが得られているとする。これらの個体が g 個のグループに分けられているとき、第 k グループの i 番目の個体のデータを

$$\mathbf{X}_{ki} = \begin{bmatrix} X_{ki1} \\ \vdots \\ X_{kip} \end{bmatrix}$$

と表す。X_{kij} は第 k グループの i 番目の個体の j 番目の変量の値を表す。

データ全体の平均を

$$\mathbf{m} = \frac{1}{N}\sum_{k=1}^{g}\sum_{i=1}^{N_k}\mathbf{X}_{ki}$$

とおく。ここで、$N = \sum_{k=1}^{g} N_k$ であり、N_k は第 k グループのデータ（個体）の総数を表す。

第 k グループの平均を以下のようにおく。

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{i}\mathbf{X}_{ki}$$

データを中心化した（平均を 0 とした）ものを

$$\mathbf{x}_{ki} = \mathbf{X}_{ki} - \mathbf{m}$$

とおく。このとき、$\mathbf{x}_{ki}$ のデータ全体にわたる平均は

$$\bar{\mathbf{x}}_{\bullet\bullet} = \frac{1}{N}\sum_{k,i}\mathbf{x}_{ki} = \mathbf{m} - \mathbf{m} = \mathbf{0}$$

であり、グループ k における平均は

$$\bar{\mathbf{x}}_{k\bullet} = \frac{1}{N_k}\sum_{i}\mathbf{x}_{ki} = \mathbf{m}_k - \mathbf{m}$$

となる。

判別関数を $\mathbf{x}_{ki}$ の変量の線形結合

$$y_{ki} = \mathbf{v}'\mathbf{x}_{ki}$$

として構成し、y_{ki} のグループ間の変動とグループ内の変動との比が最大になるように $\mathbf{v}$ を決める。グループ間の変動は以下のように算出される。

$$SS_{between} = \sum_{k}\sum_{i}\left(\mathbf{v}'\bar{\mathbf{x}}_{k\bullet} - \mathbf{v}'\bar{\mathbf{x}}_{\bullet\bullet}\right)^2$$

$$
\begin{aligned}
&= \sum_k \sum_i \left(\mathbf{v}'\bar{\mathbf{x}}_{k\bullet}\right)^2 \\
&= \sum_k \sum_i \left\{\mathbf{v}'\left(\mathbf{m}_k - \mathbf{m}\right)\right\}\left\{\mathbf{v}'\left(\mathbf{m}_k - \mathbf{m}\right)\right\}' \\
&= \mathbf{v}'\left\{\sum_k \sum_i \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)'\right\}\mathbf{v} \\
&= \mathbf{v}'\mathbf{A}\mathbf{v}
\end{aligned}
$$

ここで、

$$
\begin{aligned}
\mathbf{A} &= \sum_{k=1}^{g} \sum_{i=1}^{N_k} \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)' \\
&= \sum_k N_k \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)'
\end{aligned}
$$

である。

また、

$$\mathbf{A} = \left(N_1(\boldsymbol{m}_1 - \boldsymbol{m}) \quad \cdots \quad N_g(\boldsymbol{m}_g - \boldsymbol{m})\right)\left((\boldsymbol{m}_1 - \boldsymbol{m}) \quad \cdots \quad (\boldsymbol{m}_g - \boldsymbol{m})\right)'$$

であるが、

$$N_1(\boldsymbol{m}_1 - \boldsymbol{m}) + \cdots + N_g(\boldsymbol{m}_g - \boldsymbol{m}) = \mathbf{0}$$

から

$$\mathrm{rank}\left(\left(N_1(\boldsymbol{m}_1 - \boldsymbol{m}) \quad \cdots \quad N_g(\boldsymbol{m}_g - \boldsymbol{m})\right)\right) \leq g - 1$$

となり、

$$\mathrm{rank}\mathbf{A} \leq \mathrm{g} - 1$$

となる。

グループ内変動は以下のように算出される。

$$
\begin{aligned}
SS_{within} &= \sum_{k=1}^{g} \left\{ \sum_{i=1}^{N_k} \left(\mathbf{v}'\mathbf{x}_{ki} - \mathbf{v}'\bar{\mathbf{x}}_{k\bullet}\right)^2 \right\} \\
&= \sum_{k=1}^{g} \left[\sum_{k=1}^{N_k} \left\{\mathbf{v}'\left(\mathbf{X}_{ki} - \mathbf{m}\right) - \mathbf{v}'\left(\mathbf{m}_k - \mathbf{m}\right)\right\}^2 \right] \\
&= \sum_{k=1}^{g} \left[\sum_{k=1}^{N_k} \left\{\mathbf{v}'\left(\mathbf{X}_{ki} - \mathbf{m}_k\right)\right\}^2 \right] \\
&= \sum_{k=1}^{g} \left[\sum_{k=1}^{N_k} \left\{\mathbf{v}'\left(\mathbf{X}_{ki} - \mathbf{m}_k\right)\right\}\left\{\mathbf{v}'\left(\mathbf{X}_{ki} - \mathbf{m}_k\right)\right\}' \right]
\end{aligned}
$$

$$= \mathbf{v}'\left\{\sum_{k,i}(\mathbf{X}_{ki}-\mathbf{m}_k)(\mathbf{X}_{ki}-\mathbf{m}_k)'\right\}\mathbf{v}$$

$$= \mathbf{v}'\mathbf{W}\mathbf{v}$$

ここで、

$$\mathbf{W} = \sum_{k=1}^{g}\sum_{i=1}^{N_k}(\mathbf{X}_{ki}-\mathbf{m}_k)(\mathbf{X}_{ki}-\mathbf{m}_k)'$$

である。

級間変動と級内変動の比を

$$Q_0 = \frac{SS_{between}}{SS_{within}} = \frac{\mathbf{v}'\mathbf{A}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}}$$

とおく。Q_0 の最大化を

$$\mathbf{v}'\mathbf{W}\mathbf{v} = 1$$

の条件の下で行う。このために次の目的関数をおく。

$$Q = \mathbf{v}'\mathbf{A}\mathbf{v} - \lambda(\mathbf{v}'\mathbf{W}\mathbf{v} - 1)$$

Q の偏導関数は次式で与えられる。

$$\frac{\partial Q}{\partial \mathbf{v}} = 2\mathbf{A}\mathbf{v} - 2\lambda\mathbf{W}\mathbf{v}$$

上式の値を０とおいて次式を得る。

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{W}\mathbf{v} \qquad (1)$$

（１）式を満たす $\mathbf{v}$ を求める。まず、$\mathbf{W}$ の固有値分解を考える。

$$\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{U}'$$

$$= \mathbf{U}\mathbf{L}^{1/2}\left(\mathbf{U}\mathbf{L}^{1/2}\right)' \qquad (2)$$

（２）式を（１）式に代入すると以下のようになる。

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{U}\mathbf{L}^{1/2}\left(\mathbf{U}\mathbf{L}^{1/2}\right)'\mathbf{v}$$

$$\left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\mathbf{A}\left\{\left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\right\}'\cdot\left(\mathbf{U}\mathbf{L}^{1/2}\right)'\mathbf{v} = \lambda\left(\mathbf{U}\mathbf{L}^{1/2}\right)'\mathbf{v} \qquad (3)$$

ここで、

$$\left[\left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\mathbf{A}\left\{\left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\right\}'\right]' = \left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\mathbf{A}\left\{\left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\right\}'$$

が成り立っている。すなわち、λ と $\left(\mathbf{U}\mathbf{L}^{1/2}\right)'\mathbf{v}$ は実対称行列 $\left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\mathbf{A}\left\{\left(\mathbf{U}\mathbf{L}^{1/2}\right)^{-1}\right\}'$ の固有値

と固有ベクトルになっている。

（3）式を満たすλと$\mathbf{v}$は（1）式より次式を満たす。

$$\mathbf{v}'\mathbf{A}\mathbf{v} = \lambda\mathbf{v}'\mathbf{W}\mathbf{v}$$

したがって、

$$Q_0 = \frac{SS_{between}}{SS_{within}} = \frac{\mathbf{v}'\mathbf{A}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}} = \lambda$$

が成り立つ。すなわち、固有値λは群間変動量と群内変動量の比の値になっている。

0 でない固有値の数をnとおき、n個の固有値は

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

と並べられているとする。

j番目の固有値に対する固有ベクトル$\mathbf{v}_j$によって構成される判別関数の分散θ_jは以下のように計算される。

$$y_{ki}^{(j)} = \mathbf{v}_j'\mathbf{x}_{ki} = \mathbf{v}_{ki}'\left(\mathbf{X}_{ki} - \mathbf{m}\right)$$

とおけば、

$$\begin{aligned}
\theta_j &= \frac{1}{N}\sum_{k=1}^{g}\sum_{i=1}^{N_k}\left(y_{ki}^{(j)}\right)^2 \\
&= \frac{1}{N}\sum_{k,i}\left\{\mathbf{v}_j'\left(\mathbf{X}_{ki} - \mathbf{m}\right)\right\}\left\{\mathbf{v}_j'\left(\mathbf{X}_{ki} - \mathbf{m}\right)\right\}' \\
&= \mathbf{v}_j'\left\{\frac{1}{N}\sum_{k,i}\left(\mathbf{X}_{ki} - \mathbf{m}\right)\left(\mathbf{X}_{ki} - \mathbf{m}\right)'\right\}\mathbf{v}_j \\
&= \mathbf{v}_j'\left(\frac{1}{N}\mathbf{T}\right)\mathbf{v}_j
\end{aligned}$$

となる。ここで、

$$\mathbf{T} = \sum_{k=1}^{g}\sum_{i=1}^{N_k}\left(\mathbf{X}_{ki} - \mathbf{m}\right)\left(\mathbf{X}_{ki} - \mathbf{m}\right)'$$

はデータの全変動量を表す。

したがって、j番目の判別関数を平均が０、分散を１になるように標準化したもの（因子、factor）は以下のように与えられる。

$$f_{ki}^{(j)} = \frac{1}{\sqrt{\theta_j}}y_{ki}^{(j)}$$

$$= \frac{1}{\sqrt{\theta_j}} \mathbf{v}'_j \mathbf{x}_{ki}$$

$$= \frac{1}{\sqrt{\theta_j}} \mathbf{v}'_j \left(\mathbf{X}_{ki} - \mathbf{m} \right)$$

$$= \mathbf{b}'_j \left(\mathbf{X}_{ki} - \mathbf{m} \right)$$

ここで、

$$\mathbf{b}_j = \theta_j^{-1/2} \mathbf{v}_j$$

とおいた。

$\mathbf{X}_{ki}$ あるいは $\mathbf{x}_{ki}$ を標準化した値は次式で与えられる。

$$\mathbf{z}_{ki} = \mathbf{D}_{diag}^{-1/2} \mathbf{x}_{ki} = \mathbf{D}_{diag}^{-1/2} \left(\mathbf{X}_{ki} - \mathbf{m} \right)$$

ここで、

$$\mathbf{D}_{diag} = diag\left(\frac{1}{N} \mathbf{T} \right)$$

である。

以上より、$\mathbf{z}_{ki}$ と $f_{ki}^{(j)}$ との相関係数ベクトル $\mathbf{s}^{(j)}$ は以下のように算出される。

$$\begin{aligned}
\mathbf{s}^{(j)} &= \frac{1}{N} \sum_{k,i} \mathbf{z}_{ki} f_{ki}^{(j)} \\
&= \frac{1}{N} \sum_{k,i} \mathbf{z}_{ki} \left(\mathbf{b}'_j \mathbf{x}_{ki} \right)' \\
&= \frac{1}{N} \sum_{k,i} \mathbf{z}_{ki} \left(\mathbf{D}_{diag}^{1/2} \mathbf{z}_{ki} \right)' \mathbf{b}_j \\
&= \left(\frac{1}{N} \sum_{k,i} \mathbf{z}_{ki} \mathbf{z}'_{ki} \right) \mathbf{D}_{diag}^{1/2} \mathbf{b}_j \\
&= \mathbf{R} \mathbf{c}_j
\end{aligned}$$

ここで、

$$\mathbf{R} = \frac{1}{N} \sum_{k,i} \mathbf{z}_{ki} \mathbf{z}'_{ki} \text{、}$$

$$\mathbf{c}_j = \mathbf{D}_{diag}^{1/2}\mathbf{b}_j$$

である。

したがって、因子構造行列は次式で与えられる。

$$\begin{aligned}\mathbf{S} &= \left[\mathbf{s}^{(1)} \quad \cdots \quad \mathbf{s}^{(n)}\right] \\ &= \mathbf{R}\left[\mathbf{c}_1 \quad \cdots \quad \mathbf{c}_n\right] \\ &= \mathbf{RC}\end{aligned}$$

ここで、

$$\mathbf{C} = \left[\mathbf{c}_1 \quad \cdots \quad \mathbf{c}_n\right]$$

とおいた。

k 番目までの判別関数を採用したとき、残りの $n-k$ 個の判別関数による判別が有意であることの χ^2 検定は次式により行うことができる。

$$\chi^2 = -\left(N - \frac{p+g}{2} - 1\right)\log\Lambda' \text{、} \quad df = (p-k)(g-k-1) \qquad (4)$$

ここで、

$$\Lambda' = \prod_{j=k+1}^{n}\frac{1}{1+\lambda_j}$$

であり、p は $\mathbf{X}_{ki}$ における変数（変量）の数、g はグループの数、N はデータの総数 $\sum_k N_k$ である。

$k=0$ のときの（4）式は、グループ間の差を１元配置の多変量分散分析で検定するものとなっている。

参考文献

Cooley,W.W. and Lohnes,P.R. (1971) *Multivariate data analysis.* Wiley.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate analysis*. Academic Press.

Overall,J.E. and Klett,C.J. (1972) *Applied multivariate analysis*. McGraw-Hill Book Company.