

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

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

であり、グループ k における平均は

$$\bar{\mathbf{x}}_{k\cdot} = \frac{1}{N_k} \sum_i \mathbf{x}_{ki} = \mathbf{m}_k - \mathbf{m}$$

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。

$$SS_{between} = \sum_k \sum_i (\mathbf{v}' \bar{\mathbf{x}}_{k\cdot} - \mathbf{v}' \bar{\mathbf{x}}_{..})^2$$

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

ここで、

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

である。

グループ内変動は以下のように算出される。

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

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。

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

(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。

$$\begin{aligned} \mathbf{W} &= \mathbf{U}\mathbf{L}\mathbf{U}' \\ &= \mathbf{U}\mathbf{L}^{1/2}(\mathbf{U}\mathbf{L}^{1/2})' \end{aligned} \quad (2)$$

(2) 式を (1) 式に代入すると以下のようになる。

$$\begin{aligned} \mathbf{A}\mathbf{v} &= \lambda\mathbf{U}\mathbf{L}^{1/2}(\mathbf{U}\mathbf{L}^{1/2})' \mathbf{v} \\ (\mathbf{U}\mathbf{L}^{1/2})^{-1} \mathbf{A} \{(\mathbf{U}\mathbf{L}^{1/2})^{-1}\}' \cdot (\mathbf{U}\mathbf{L}^{1/2})' \mathbf{v} &= \lambda(\mathbf{U}\mathbf{L}^{1/2})' \mathbf{v} \end{aligned} \quad (3)$$

ここで、

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

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

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$$

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。

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

とおけば、

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

となる。ここで、

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

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。

$$\begin{aligned} f_{ki}^{(j)} &= \frac{1}{\sqrt{\theta_j}} y_{ki}^{(j)} \\ &= \frac{1}{\sqrt{\theta_j}} \mathbf{v}_j' \mathbf{x}_{ki} \\ &= \frac{1}{\sqrt{\theta_j}} \mathbf{v}_j' (\mathbf{X}_{ki} - \mathbf{m}) \\ &= \mathbf{b}_j' (\mathbf{X}_{ki} - \mathbf{m}) \end{aligned}$$

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。

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

ここで、

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

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。

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

ここで、

$$\begin{aligned} \mathbf{R} &= \frac{1}{N} \sum_{k,i} \mathbf{z}_{ki} \mathbf{z}'_{ki} \\ \mathbf{c}_j &= \mathbf{D}_{diag}^{1/2} \mathbf{b}_j \end{aligned}$$

である。

したがって、因子構造行列は次式で与えられる。

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

ここで、

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

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。

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

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.