

連結ベクトルパターンによる多次元順位相関の尺度について

千葉大学 理学部 田栗 正章
岡山大学 教養部 脇本 和昌

1. 図式化の方法

(1) 表 1 のような n 組の順位データ $(p_i, r_{i1}, \dots, r_{ki})$ ($i=1, \dots, n$) が与えられたとする。ここで一般性を失うことなく

$$p_i = i, \quad (i=1, \dots, n); \quad r_{ji} \in \{1, 2, \dots, n\}, \quad (i=1, \dots, n; j=1, \dots, k), \quad (1-1)$$

$$p_i \neq p_{i'}, \quad r_{ji} \neq r_{ji'}, \quad (i \neq i'; i, i'=1, \dots, n; j=1, \dots, k)$$

とする。目的変数 p および k 個の説明変数 r_j の n 個の値を、次のような変換により $0 \sim \pi$ ラジアン の角度に変換する。

$$\eta_i = \frac{i-1}{n-1} \pi, \quad \xi_{ji} = \frac{r_{ji}-1}{n-1} \pi, \quad (i=1, \dots, n; j=1, \dots, k). \quad (1-2)$$

p	r_1	r_2	\dots	r_j	\dots	r_k
1	r_{11}	r_{12}	\dots	r_{j1}	\dots	r_{k1}
2	r_{21}	r_{22}	\dots	r_{j2}	\dots	r_{k2}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	r_{i1}	r_{i2}	\dots	r_{ji}	\dots	r_{ki}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	r_{n1}	r_{n2}	\dots	r_{jn}	\dots	r_{kn}

表1 与えられた順位データ

η	ξ_1	ξ_2	\dots	ξ_j	\dots	ξ_k
η_1	ξ_{11}	ξ_{21}	\dots	ξ_{j1}	\dots	ξ_{k1}
η_2	ξ_{12}	ξ_{22}	\dots	ξ_{j2}	\dots	ξ_{k2}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
η_i	ξ_{1i}	ξ_{2i}	\dots	ξ_{ji}	\dots	ξ_{ki}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
η_n	ξ_{1n}	ξ_{2n}	\dots	ξ_{jn}	\dots	ξ_{kn}

表2 変換されたデータ

(2) 表2のデータに対してそれぞれ次のようなベクトル $\vec{\eta}_i$, $\vec{\xi}_{jk}$ を対応させ、さらにベクトル $\vec{\zeta}_i$ を次式で定義する(図1参照)。

$$\begin{cases} |\vec{\eta}_i| = 1, \\ \arg(\vec{\eta}_i) = \eta_i, \end{cases} \quad \begin{cases} |\vec{\xi}_{jk}| = 1, \\ \arg(\vec{\xi}_{jk}) = \xi_{jk}, \end{cases} \quad (i=1, \dots, n; j=1, \dots, k), \quad (1-3)$$

$$\vec{\zeta}_i = \sum_{j=1}^k w_j \vec{\xi}_{jk}, \quad (i=1, \dots, n). \quad (1-4)$$

ただし $w_j (j=1, \dots, k)$ は各説明変数の重みで、

$$\begin{cases} \sum_{j=1}^k w_j = 1, \\ w_j \geq 0, (j=1, \dots, k), \end{cases} \quad (1-5)$$

を満足する。($|\vec{\zeta}_i| \leq 1$, $\arg(\vec{\zeta}_i) \leq \pi$, $i=1, \dots, n$)

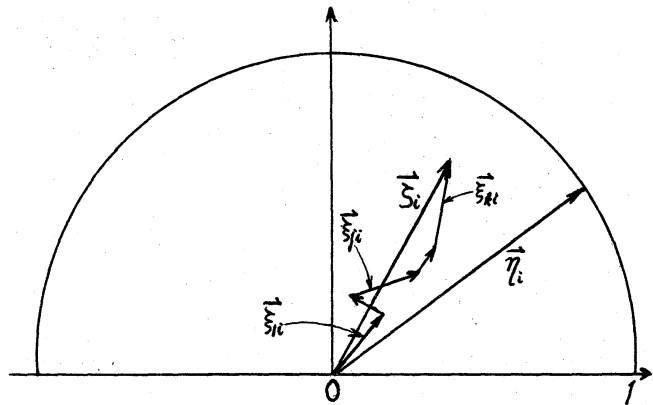


図1 i番目のデータに対応するベクトル $\vec{\eta}_i$, $\vec{\zeta}_i$

(3) 次に原点 θ から出発して $\vec{\eta}_1, \dots, \vec{\eta}_n$ および $\vec{\zeta}_1, \dots, \vec{\zeta}_n$ をそれぞれこの順につなぎ合わせ、二次元平面上に図示する(図2参照)。

この連結された図式パターンをそれぞれ P パターンおよび R パターンと呼ぶ。このようにして得られたものが、目的変数と説明変数との相関度を表現する連結ベクトルパターンである。

[註] P パターンと R パターンの終点は必ず一致し、ON 軸上に

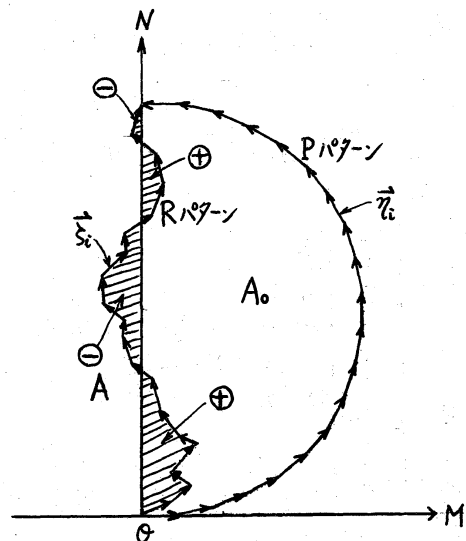


図2 PパターンとRパターンの一例

存在する(図2参照)。なぜならば(1-3)~(1-5)より

$$\begin{aligned}\sum_{i=1}^n \cos \delta_i &= \sum_{i=1}^n \sum_{j=1}^k w_j \cos \xi_{ji} = \sum_{j=1}^k w_j \sum_{i=1}^n \cos \xi_{ji} = \sum_{i=1}^n \cos \eta_i = 0, \\ \sum_{i=1}^n \sin \delta_i &= \sum_{i=1}^n \sum_{j=1}^k w_j \sin \xi_{ji} = \sum_{j=1}^k w_j \sum_{i=1}^n \sin \xi_{ji} = \sum_{j=1}^k w_j \sum_{i=1}^n \sin \eta_i = \sum_{i=1}^n \sin \eta_i.\end{aligned}\tag{1-6}$$

2. 相関度の読みとり方

連結ベクトルパターンから目的変数と説明変数との相関度を読みとる方法は、PパターンとRパターンとの接近度、なめらかさおよび形状の比較等により得られる。ここでPパターンおよびRパターンのなめらかさとは、 $|\arg(\bar{\eta}_i) - \arg(\bar{\eta}_{i-1})|$ および $|\arg(\bar{\xi}_{ji}) - \arg(\bar{\xi}_{j,i-1})|$ ($i=2, \dots, n$)のことであり、この値が小なるときそれぞれのパターンはなめらかであるという。

一方目的変数と説明変数との相関は、おおよそ次のように分類してよいであろう。

(a) 相関のない場合

(b) 部分的に相関のある場合 $\left\{ \begin{array}{l} (b-1) \text{ 正の相関の場合} \\ (b-2) \text{ 負の相関の場合} \end{array} \right.$

(c) 全体的に相関のある場合 $\left\{ \begin{array}{l} (c-1) \text{ 正の相関のみの場合} \\ (c-2) \text{ 負の相関のみの場合} \\ (c-3) \text{ 正負の相関が混在する場合} \end{array} \right.$

まずPパターンとRパターンとの接近度について考える。

(1-2)~(1-5)を考慮すれば、(c-1)の場合はPパターンとRパタ

ーンは似たような形状を示すことが判り、従って2つのパターンは接近していると考えられる。又(c-2)の場合はPパターンとRパターンはON軸に関して対称な形状を示すことになり、2つのパターンは離れていると考えられる。さらに(a)の場合にはRパターンはほぼON軸に沿った形となることが判り、従って2つのパターンの接近度は(c-1)と(c-2)の場合の中間の値となることが予想される。

次になめらかさについては、やはり(1-2)~(1-5)を考慮すれば次のようなことが判る。(c-1)および(c-2)の場合にはRパターンはかなりなめらかであることが予想され、又(a)の場合にはRパターンは極端になめらかではなくなるであろう。

3. 相関度の2つの尺度

3.1 面積比相関係数

次のように記号を定める。

A_0 : PパターンとON軸とで囲まれる部分の面積。

A : RパターンとON軸とで囲まれる部分の面積。ただしON軸より左側にある場合はその面積に負の符号をつける。RパターンがON軸の両側に渡って存在するときはその和をとる(図2参照)。

このとき目的変数と説明変数との相関度 a を次式で定義し、

面積比相関係数と呼ぶことにする。

$$a = A / A_0. \quad (3-1)$$

ここで A は重み w_1, \dots, w_k の関数と考えられるが、 $\max_{\{w_j\}} |A| \leq A_0$ が成立するから、

$$|a| \leq 1 \quad (3-2)$$

となる。さらに第2節で調べたように ρ と (r_1, \dots, r_k) の間に正の相関関係が強ければ A が A_0 に近いような重みが存在するはずであり、従って a は 1 に近いと考えられる。又負の相関関係が存在すれば A が $-A_0$ に近いような重みが存在するはずであり、従って a は -1 に近いと考えられる。無相関の場合には重みをどのように変化させても $|a|$ は 0 に近いと考えられる。

3.2 なめらかさ相関係数

次にもう1つの相関係数について考えてみよう。隣り合うデータの各変数についての順位の差の加重和を用いて、なめらかさ相関係数 Δ を次式で定義する。

$$\Delta = 1 - \frac{2}{n(n-1)} \sum_{i=2}^n \left(\sum_{j=1}^k w_j |r_{ji} - r_{j,i-1}| \right). \quad (3-3)$$

ここで

$$0 \leq \sum_{i=2}^n |r_{ji} - r_{j,i-1}| \leq \frac{1}{2} n(n-1), \quad (j=1, \dots, k)$$

が成立するから、

$$0 \leq \Delta \leq 1 \quad (3-4)$$

となる。さらに第2節で調べたように p と (r_1, \dots, r_k) の間に正又は負の相関関係が強ければ $\sum_{i=2}^n |r_{pi} - r_{pi-1}|$ の値は小さくなるはずであり、従って Δ が1に近くなるような重みが存在するはずである。又相関のない場合には、重みをどのように変化させても Δ は0に近いであろうと考えられる。

4. 最適な重みの計算

4.1 面積比相関係数 a のベクトル表現

3.1 で定義した A_0 はデータ数 n が与えられれば決まる定数であるが、これを計算すると次のようになる。

$$\begin{aligned} A_0 &= \frac{1}{2} \sum_{i=2}^n \sin \eta_i \left(\sum_{k=1}^{i-1} \cos \eta_k + \sum_{k=1}^i \cos \eta_k \right) \\ &= \sum_{i=1}^n \sin \eta_i \sum_{k=1}^i \cos \eta_k - \frac{1}{2} \sum_{i=1}^n \sin \eta_i \cos \eta_i = \text{constant}. \end{aligned} \quad (4-1)$$

次に A について考える。まず記号を定義する。

$$\begin{cases} \gamma_{ji} = \cos \xi_{ji}, \\ \theta_{ji} = \sin \xi_{ji}, \end{cases} \quad (i=1, \dots, n; j=1, \dots, k), \quad (4-2)$$

$$\begin{cases} \gamma'_i = (\gamma_{1i}, \gamma_{2i}, \dots, \gamma_{ki}), \\ \theta'_i = (\theta_{1i}, \theta_{2i}, \dots, \theta_{ki}), \end{cases} \quad (i=1, \dots, n),$$

$$w' = (w_1, w_2, \dots, w_k).$$

ただし w は (1-5) を満足する重みベクトルとする。これらを用いて A を計算すると次のようになる。

$$\begin{aligned}
A &= \frac{1}{2} \left(\sum_{j=1}^k w_j \sin \xi_{j\ell} \right) \left(\sum_{j=1}^k w_j \cos \xi_{j\ell} \right) + \frac{1}{2} \sum_{i=2}^n \left(\sum_{j=1}^k w_j \sin \xi_{ji} \right) \left\{ \sum_{j=1}^{i-1} \left(\sum_{j=1}^k w_j \cos \xi_{j\ell} \right) + \sum_{j=i}^k \left(\sum_{j=1}^k w_j \cos \xi_{j\ell} \right) \right\} \\
&= \sum_{i=1}^n \left(\sum_{j=1}^k w_j \theta_{ji} \right) \left\{ \sum_{j=1}^i \left(\sum_{j=1}^k w_j r_{j\ell} \right) \right\} - \frac{1}{2} \sum_{i=1}^n \left(\sum_{j=1}^k w_j \theta_{ji} \right) \left(\sum_{j=1}^k w_j r_{ji} \right) \\
&= \sum_{i=1}^n (w' \theta_i) \left(\sum_{j=1}^i r_{j\ell}' w \right) - \frac{1}{2} \sum_{i=1}^n (w' \theta_i) (r_i' w) \\
&= w' \left\{ \sum_{i=1}^n \theta_i \left(\sum_{j=1}^i r_{j\ell}' \right) \right\} w - \frac{1}{2} w' \left(\sum_{i=1}^n \theta_i r_i' \right) w \\
&= w' \left\{ \sum_{i=1}^n \theta_i \left(\sum_{j=1}^i r_{j\ell}' - \frac{1}{2} r_i' \right) \right\} w.
\end{aligned}$$

ここで行列 T を次のようにおく。

$$\begin{cases} T = [t_{\alpha\beta}], \\ t_{\alpha\beta} = \sum_{i=1}^n \theta_{\alpha i} \left(\sum_{j=1}^i r_{j\ell} - \frac{1}{2} r_{\alpha i} \right), \end{cases} \quad (\alpha=1, \dots, k; \beta=1, \dots, k). \quad (4-3)$$

これを用いれば、

$$A = w' T w \quad (4-4)$$

となる。従って面積比相関係数 a は、(4-1) によって定まる定数 A_0 を用いれば次のような w の 2 次形式として表現できる。

$$a = \frac{1}{A_0} w' T w. \quad (4-5)$$

4.2 なめらかさ相関係数 Δ のベクトル表現

4.1 と同様な方法で Δ を計算する。ベクトル b を次のようにおく。

$$\begin{cases} b' = (b_1, \dots, b_{\alpha}, \dots, b_k), \\ b_{\alpha} = \frac{2}{n(n-1)} \sum_{i=2}^n \sum_{j=1}^k w_j |r_{ji} - r_{j,i-1}|, \end{cases} \quad (\alpha=1, \dots, k). \quad (4-6)$$

このとき Δ は次のようになる。

$$\begin{aligned}
\Delta &= 1 - \frac{2}{n(n-1)} \sum_{i=2}^n \sum_{j=1}^k w_j |r_{ji} - r_{j,i-1}| \\
&= 1 - \sum_{j=1}^k w_j \sum_{i=2}^n \frac{2}{n(n-1)} |r_{ji} - r_{j,i-1}|
\end{aligned}$$

$$= 1 - b'w \quad (4-7)$$

すなわちなめらかな相関係数は、 w の1次式として表現できる。

4.3 最適な重みの計算

1組のデータが与えられたとき、2つの尺度 a と b を用いて説明変数と目的変数との相関度を適切に表現する方法について考えてみよう。第2節で述べたように、代表的な相関関係は(a)~(c-3)と考えてよいであろう。そこでここでは、これらの相関関係の内のどれか1つが最も強くなるように、各説明変数に加えるべき重みを求める。

まず (c-1), (c-2) の場合には最適な重みを与えたとすれば λ は 1 に近いと考えられ、 $|\lambda|$ も 1 に近くなるはずである。従って λ の値が大きな範囲内で $|\lambda|$ を最大化するような重みを求めればよい (又はその逆でもよい)。次に (c-3) の場合であるが、最適な重みを与えられたとき $|\lambda|$ はそれ程大きくないであろうが、 λ はかなり 1 に近いことが予想できる。従って λ を最大にするような重みを求めればよい。(b-1), (b-2) の場合も同様に考えれば、 $|\lambda|$ 又は λ を最大にするような重みを求めればよいことが判る。ここで $a < 0$ となる場合には順位のつけ方を逆にして、常に $a > 0$ となるようにしておけば、最適な重みを計算する問題は制約条件を持つ次の 2 つのタイプの非線型最適化問題に帰着することができる。

$$\begin{aligned}
 1^\circ \quad f(w) &= \frac{1}{A_0} w' T w \rightarrow \max, \\
 &\begin{cases} \mathbb{1}' w = 1, & w \geq 0, \\ \Delta_l \leq 1 - b' w \leq \Delta_u, \end{cases} \quad (4-8)
 \end{aligned}$$

ただし $f(w)$: 最適化すべき目的関数,

$$\mathbb{1}' = (1, 1, \dots, 1),$$

Δ_l : λ のとりうる値の下限,

Δ_u : λ のとりうる値の上限。

[註] 制約条件 (4-8) の上の 2 式は (1-5) によるものであり、下の式は λ の値をある範囲内に制限することからでてくる。

$$2^\circ \quad f(w) = 1 - b'w \rightarrow \max ,$$

$$\begin{cases} 1'w = 1, & w \geq 0 \\ a_l \leq \frac{1}{A_0} w' T w \leq a_u, \end{cases} \quad (4-9)$$

ただし a_l : a のとりうる値の下限,

a_u : a のとりうる値の上限.

上述のように、最適な重みは上記 1°, 2° の 2 つの非線型最適化問題の解として得られる。1° は形式的には 2 次計画の問題であるが、行列 T は一般には正值ではないので、有限回の iteration での解の収束性や、最適解を得る可能性については保証されない。又 2° は 2 次の制約条件を持つ線形計画問題である。

参 考 文 献

- [1] Kendall, M.G. and Stuart, A., *The Advanced Theory of Statistics*, Vol.2, London : Charles W. Griffin and Co., Ltd., 1961.
- [2] Wakimoto, K. and Taguri, M., "On the Representation Method of Multiple Correlation by Pattern of Connected Vectors" (in Japanese), *Journal of the Japan Statistical Society*, 5 (October 1974), 9-24.