

クロス表の SVD および内積による表現と座標変換

グラフ図示による例示

○岡本安晴
日本女子大学

1 はじめに

クロス表の分析は、数量化法、対応分析、双対尺度法などよく知られている方法があり、適当なデータの前処理を行った後、固有分解などが行われる。クロス表は、2次元の配列データであるが、2次元配列データは、データサイエンスでは特異値分解（SVD）により行情報と列情報を取り出すということが行われている（Brunton & Kutz, 2019）。岡本（2022）は、クロス表を直接 SVD により分析して、行情報と列情報の関係を見るために同じ座標軸に重ねて表示し、行情報と列情報の関係の視覚的表現を得ている。クロス表の SVD と内積に関するコメント（岡太, 2022）に基づき、内積モデルとして表現すると、クラスター分析や直交回転が可能になる。以下に、論じる。

	C1	C2	C3	C4	C5
R1	10	5	0	0	0
R2	5	10	5	0	0
R3	0	5	10	3	0
R4	0	0	3	10	5
R5	0	0	0	5	10

2. クロス表の SVD

表 1 のクロス表 X の SVD を

$$X = \sum s_i u_i v_i'$$

とおく。特異値は大きさの順に

$$s_1 \geq \dots \geq s_r > 0$$

と添え字が付けられているものとする。行情報 $\sqrt{s_i} u_i$ と列情報 $\sqrt{s_i} v_i$ を第 i 軸上に重ねて表示する。第 1 番目と第 2 番目の特異値に注目して、第 1 番目を横軸、第 2 番目を縦軸にとり表示すると図 1 のようになる。関係の強い行項目と列項目が近接して配置され、表 1 に見られる 2 つのグループに纏まっていることが分かる。

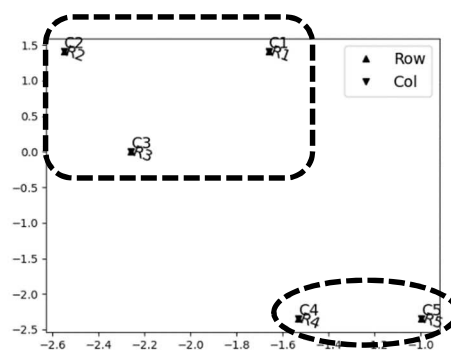


図 1 BiPlot

3. 内積モデル

SVD は以下のように変形することができる。

$$X = (u_1 \ \dots \ u_r) \begin{pmatrix} s_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_r \end{pmatrix} \begin{pmatrix} v_1' \\ \vdots \\ v_r' \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_M \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}'$$

ここで、 X は (M, N) 型とし、

$$\begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_M \end{pmatrix} = (s_1^{\alpha_1} \mathbf{u}_1 \quad \cdots \quad s_r^{\alpha_r} \mathbf{u}_r), \quad \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_N \end{pmatrix} = (s_1^{\beta_1} \mathbf{v}_1 \quad \cdots \quad s_r^{\beta_r} \mathbf{v}_r), \quad \alpha_t + \beta_t = 1$$

である。このとき、

$$X = (x_{ij}) = (\mathbf{a}_i \mathbf{b}_j') \tag{1}$$

と書ける。すなわち、クロス表 X の各セルの値を内積で表すことができる。ベクトル \mathbf{a}_i は、 i 番目の行項目の座標であり、ベクトル \mathbf{b}_j は j 番目の列項目の座標である。これらのベクトルは内積によってクロス表のデータを表しているのだから、同じ空間内のベクトルと考えられる。セルの値が大きいとき、ベクトル \mathbf{a}_i とベクトル \mathbf{b}_j のなす角は 0 に近く、またベクトルの長さも長い。ベクトルの長さは、その表す項目の強さと考えられる。セルの値が負であれば、ベクトルのなす角は 90 度より大きい。すなわち、ベクトルは反対方向である。クロス表の内積モデルは、セルの値が負の場合も分析可能である。

ここで、 α_t と β_t は、第 t 軸における行項目と列項目のそれぞれの座標値の2乗和が項目数に比例するように決める。第 t 軸における行項目の座標値の2乗和 $SS_{t,u}$ と列項目の座標値の2乗和 $SS_{t,v}$ は

$$SS_{t,u} = \|s_t^{\alpha_t} \mathbf{u}_t\|^2 = s_t^{2\alpha_t}, \quad SS_{t,v} = \|s_t^{\beta_t} \mathbf{v}_t\|^2 = s_t^{2\beta_t}$$

で与えられる。

$$SS_{t,u}/SS_{t,v} = M/N$$

とおくと、

$$\alpha_t = \frac{1 \log(s_t^2 M/N)}{4 \log s_t}, \quad \beta_t = \frac{1 \log(s_t^2 N/M)}{4 \log s_t}$$

を得る。

4. クラスター分析

行項目と列項目に与えられた座標ベクトルが同一空間にあると考えると、これらの座標値に基づいてクラスター分析を行うことができる。表1のデータの内積モデル(1)に基づいてクラスター分析(Ward法)を行った結果を図2に示す。まず、行項目 C_i と列項目 R_i がクラスターとして纏められ、続いて表1のデータに示されている関係の強さに対応してクラスターが構成されていることが分かる。内積の角に基づいてクラスター分析を行うこともできる。

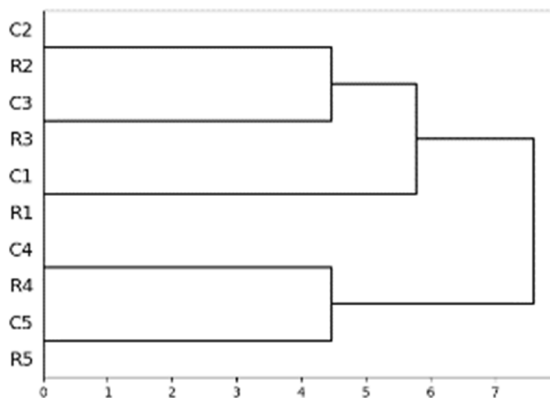


図2 クラスター分析 (Ward 法)

5. 座標変換 (正射影)

内積モデル(1)は、次のように書くこともできる。

$$X = AB', \quad A = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_M \end{pmatrix}, \quad B = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_N \end{pmatrix}$$

したがって、

$$X = (AT)(BT)', \text{ただし、} TT' = I \tag{2}$$

が成り立ち、座標行列 A と B は直交行列 T により座標変換を行うことが出来る。最初に、SVD により与えられた布置 A と B (図 1 に対応するもの) が分かりやすい布置でないときは、座標変換を行うことにより解り易いものになる可能性がある。その例を以下に報告する。

6. 座標変換 (主成分分析、正射影)

表 2 に示されているクロス表を分析してみる。行項目 R_i と列項目 C_i との関係は強く、組 R_i と C_i は他の組と循環的に緩く繋がっている。このクロス表の SVD を求め、布置を描くと図 3 のようになる。表 2 に示されている循環的關係から予想される円環構造が直接には認められない。これは、円環構造が表示されるように座標系が設定されていないことによる。図 3 の布置は、第 1 次元と第 2 次元の座標によるものである。全ての座標次元に基づくクラスター分析の結果は図 4 のようになる。表 2 のデータの示す円環情報が反映されているように見える。

	C1	C2	C3	C4
R1	10	1	0	0
R2	0	10	1	0
R3	0	0	10	1
R4	1	0	0	10

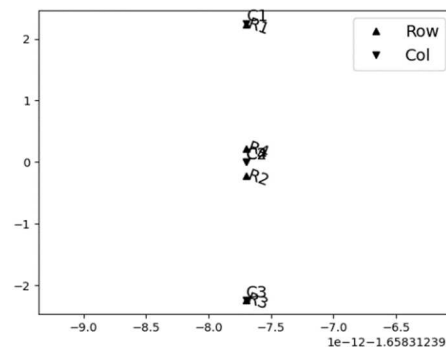


図 3 表 2 のデータの SVD による布置

式 (2) に示されているように、座標系は直交変換により任意に設定することができる。座標系の変換方法を探るために、図 3 の布置の 3 次元図を描いてみる (図 5)。図 5 の左側の図は、ベクトルの先端が第 1 軸 (X 軸) に垂直な 1 つの平面上に並んでいることを示し、右側の図はそれらのベクトルの先端が円環上に並んでいることを表している。この円環構造をよく反映する座標系を得るための式 (2) における変換を以下のように求める。

行項目と列項目の座標を 1 つにまとめる。

$$Y = \begin{pmatrix} A \\ B \end{pmatrix}$$

Y の各行が空間内の点を表すとみたとき、これらの点の分布が最大になる方向に座標系を設定する。行列 Y を中心化する。

$$Z = Y - \mathbf{1}_{M+N} \bar{Y}$$

ここで、

$$\bar{Y} = \frac{1}{M+N} \mathbf{1}'_{M+N} Y$$

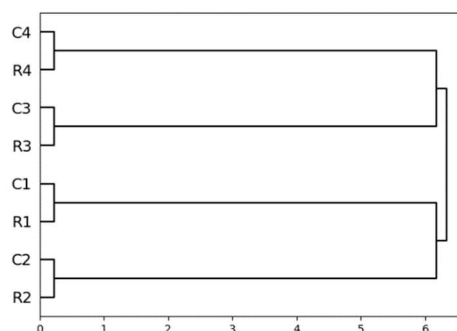


図 4 表 2 のデータの SVD による布置に基づくクラスター

である。内積モデルにおける座標変換なので、中心化のみを行い標準化は行わない。 Z の表す点の布置の分散の最大(極大)方向 w を、次式を最大化(極大化)するものとして求めることができる。

$$(Zw)'(Zw), \quad w'w = 1$$

得られたベクトルを w_1, \dots, w_r とおくとき、変換行列を次式

$$P = (w_1 \ \dots \ w_r)$$

で設定する。この行列は式(2)の条件、 $PP' = I$ を満たす。座標変換を

$$YP \tag{3}$$

と行う。中心化した Z ではなく、 Y を変換しているのは、内積モデルであるので原点を移動できないからである。

ベクトル a_i および b_j の変換後は

$$a_i P = (a_i w_1, \dots, a_i w_r), \quad b_j P =$$

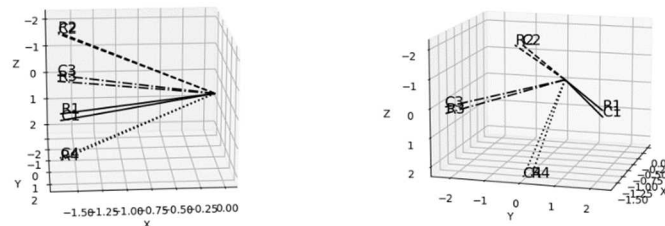


図5 布置の3D図

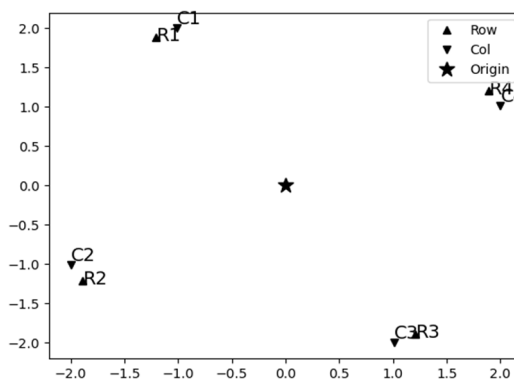


図6 座標変換後の布置

$(b_j w_1, \dots, b_j w_r)$ となる。これは、直交座標系 $\{w_1 \ \dots \ w_r\}$ への座標変換である。表2のデータのSVDに座標変換(3)を適用すると、図6の結果を得る。行項目と列項目の組が円環上に並んでいる。

7. 結語

対応分析では、内積とデータとの関係は、 $(f_{kl} - \pi_{kl})/\pi_{kl} \approx r'_k c_l$ 、 $\pi_{kl} = f_{k+} f_{+l}/N$ で表され、交互作用の相対値が内積で表されている(足立、2003)。これに対して本稿の内積モデル(1)では、元のデータ値が内積で表されている。 $\alpha_t = 1, \beta_t = 0$ と置くと、 $XX' = AA'$ となるので、このとき A の行ベクトルの長さの2乗が X の行の2乗和に等しくなる。 $\alpha_t = 0, \beta_t = 1$ と置くと、 $X'X = BB'$ となるので、このとき B の行ベクトルの長さの2乗が X の列の2乗和に等しくなる。行と列の交互作用は、 A と B の行ベクトルのなす角に反映される。すなわち、主効果がベクトルの長さに反映され、交互作用がベクトルのなす角に反映されると解釈できる。なお、本稿のモデル(1)が対象とするクロス表の値は度数以外でもよく、実数値であればよい。例として、SD法における対象の評定値のクロス表が挙げられるが、分析例を発表当日に報告する。

8. 引用文献

- (1) 足立浩平(2003) 対応分析と多重対応分析と同時対応分析. 心理学評論、45、547-563.
- (2) Brunton, S. L., & Kutz, J. N. (2019). *Data-driven science and engineering*. Cambridge Univ. Press.
- (3) 岡太彬訓(2022) 私信.
- (4) 岡本安晴(2022) 官能評価に使える統計解析ツール: Python 編. 日本官能評価学会誌 26、2、102-106