距離行列や推移確率行列に関する補足

前川眞一

20200604,05,06,2020. 6. 11

0.1 距離行列

q 次元空間内の二つの点 i, j の間のユークリッド距離の２乗は、各点の座標を $q \times 1$ のベクトル $\boldsymbol{x}_j,\ , j = 1, 2, \ldots, N$ とすれば、

$$d_{ij}^2 = \sum_{a=1}^{q} (x_{ia} - x_{ja})^2 \ = \ (\boldsymbol{x_i} - \boldsymbol{x}_j)'(\boldsymbol{x_i} - \boldsymbol{x}_j) \tag{1}$$

と書ける。これをまとめて、d_{ij}^2 を要素として持つ $N \times N$ の行列を $\boldsymbol{D}^2$ とすれば,

$$\boldsymbol{D}^2 = \mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{1}\boldsymbol{1}' - 2\boldsymbol{X}\boldsymbol{X}' + \boldsymbol{1}\boldsymbol{1}'\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right) \tag{2}$$

と書くことが出来る。

なぜなら、$\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)$ の第 i 番目の要素は、$\boldsymbol{x}_i'\boldsymbol{x}_i$ であり、$\boldsymbol{1}\boldsymbol{1}'$ は全ての要素が 1 である $N \times \ \ N$ の行列であるから、$\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{1}\boldsymbol{1}'$ はその第 i 行がすべて $\boldsymbol{x}_i'\boldsymbol{x}_i$ である行列となるため、その第 i, j 要素は $\boldsymbol{x}_i'\boldsymbol{x}_i$ である。同様に、 $\boldsymbol{1}\boldsymbol{1}$'$\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)$ 行列は、その第 j 列が全て $\boldsymbol{x}_j'\boldsymbol{x}_j$ である行列であるからその第 i, j 要素は $\boldsymbol{x}_j'\boldsymbol{x}_j$ である。また、$\boldsymbol{X}'\boldsymbol{X}$ 行列の第 i, j 要素は $\boldsymbol{x}_i'\boldsymbol{x}_j$ であるため $\boldsymbol{D}^2$ 行列の第 i, j 要素は

$$d_{ij}^2 = \boldsymbol{x}_i'\boldsymbol{x}_i \ - \ 2\,\boldsymbol{x}_i'\boldsymbol{x}_j \ + \ \boldsymbol{x}_j'\boldsymbol{x}_j \tag{3}$$

$$= (\boldsymbol{x}_i - \boldsymbol{x}_j)'(\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{4}$$

となり、先に示した２点間のユークリッド距離の２乗となっていることが分かる。

このことを利用して、個体間距離行列が与えられたときに、以下のようにして、個体の座標行列を求めることが出来る。いま、中心化行列を $\boldsymbol{J} = \boldsymbol{I} - (1/N)\boldsymbol{1}\boldsymbol{1}'$ とすると、$\boldsymbol{1}'\boldsymbol{J} = \boldsymbol{0}'$ ならびに $\boldsymbol{J}\boldsymbol{1} = \boldsymbol{0}$ であるため、Eq.(2) の両辺の左右から $\boldsymbol{J}$ を掛けそれを -2 倍すると

$$-2\,\boldsymbol{J}\,\boldsymbol{D}^2\,\boldsymbol{J} = -2\,\boldsymbol{J}\ \left(\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{1}\boldsymbol{1}' - 2\boldsymbol{X}\boldsymbol{X}' + \boldsymbol{1}\boldsymbol{1}'\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)\right)\boldsymbol{J} \tag{5}$$

$$= -2\left(\boldsymbol{J}\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{1}\boldsymbol{1}'\boldsymbol{J} \ - \ \boldsymbol{J}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{J} \ + \ \boldsymbol{J}\boldsymbol{1}\boldsymbol{1}'\mathrm{diag}\left(\boldsymbol{X}\boldsymbol{X}'\right)\boldsymbol{J}\right) \tag{6}$$

$$= \boldsymbol{J}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{J} \ = \ \boldsymbol{J}\boldsymbol{X}(\boldsymbol{J}\boldsymbol{X})' \tag{7}$$

となる。（なおこの操作のことを２重中心化 double centering と呼ぶ。）ここで、

$$\boldsymbol{Y} = \boldsymbol{J}\boldsymbol{X} \tag{8}$$

は、もとの座標行列の列平均を $\mathbf{0}$ とした座標行列（中心化された座標行列）であるから、Eq.(7) は $-2\,\boldsymbol{J}\,\boldsymbol{D}^2\,\boldsymbol{J}$ をある行列 $\boldsymbol{Y}$ とその転置の積に分解できれば、それが中心化された座標行列であること示している。そして、このような分解は、固有値と固有ベクトルを用いて行うことが出来る。

なお、ユークリッド距離はその原点と座標軸の向きに関して不定性を持つため、座標の原点と軸の向きは一意には定まらない。すなわち、$\boldsymbol{c}$ を任意のベクトル、$\boldsymbol{T}$ を任意の直交行列とすると

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{T} \;+\; \mathbf{1}\boldsymbol{c}' \tag{9}$$

から計算された距離行列は $\boldsymbol{X}$ から計算された距離行列に等しくなる。

0.2 推移確率行列

離散的な確率変数 X を考え、その値は p 個の値 $a_1, a_2, \ldots, a_p$ のどれかであるとする。そして、確率変数の列 $X^{(t)}, t = 1, 2, \ldots$ を考え、第 t 回目に X が取る値、すなわち $X^{(t)}$ は、X の一つ前の時点 $t-1$ での X の値 $X^{(t-1)}$ にのみ依存し、時点 $t-1$ に $X = a_i$ であったものが次の時点 t で $X = a_j$ となる推移確率、すなわち $\Pr\left(X^{(t)} = a_j | X^{(t-1)} = a_i\right)$ は時点に依存しないものとする。このような確率過程のことを時間的に均一なマルコフ過程と呼ぶ。

いま、推移確率行列（推移核）を

$$\boldsymbol{K} = \{k_{ij}\},\; k_{ij} = \Pr\left(X^{(t)} = a_j | X^{(t-1)} = a_i\right), \quad , i, j = 1, 2, \ldots, p \tag{10}$$

とすれば、$X^{(t)}$ は必ず p 個の a_j のどれかの値になるから

$$\sum_{j=1}^{p} k_{ij} \;=\; \sum_{j=1}^{p} \Pr\left(X^{(t)} = a_j | X^{(t-1)} = a_i\right) \;=\; 1 \tag{11}$$

である。次に、時点 0 における X の値（状態）を $\boldsymbol{x}^{(0)}$ という $p \times 1$ のベクトルで表す。ただし、このベクトルはどれか一つの要素が 1 で残りの要素は 0 であるようなダミー変数ベクトルであるとする。すると、時点 1 に於いて確率変数 X の値が a_j になっている確率は

$$\Pr\left(X^{(1)} = a_j\right) = \sum_{i=1}^{p} \Pr\left(X^{(t)} = a_j | X^{(t-1)} = a_i\right) \times x_i^{(0)} \tag{12}$$

$$= \Pr\left(X^{(1)} = a_j\right) \;=\; \sum_{i=1}^{p} k_{ij} \times x_i^{(0)} \tag{13}$$

で与えられる。例えば、時点 0 に於いて X の値が a_i であったとすれば $\boldsymbol{x}^{(0)}$ はその第 i 番目の要素のみが 1 であるため、$\Pr\left(X^{(1)} = a_j\right) = k_{ij}$ である。また、この p 個の確率を縦に並べたものを

$$\boldsymbol{x}^{(1)} = \{x_j^{(1)}\}, \quad x_j^{(1)} = \Pr\left(X^{(1)} = a_j\right), \quad j = 1, 2, \ldots, p \tag{14}$$

と置くと、

$$\boldsymbol{x}^{(1)'} = \boldsymbol{x}^{(0)'}\boldsymbol{K} \tag{15}$$

と書けるが、先ほどの例のように、時点 0 に於いて X の値が a_i であったとすれば $\boldsymbol{x}^{(0)}$ はその第 i 番目の要素のみが 1 であるため、 $\boldsymbol{x}^{(1)}$ は $\boldsymbol{K}$ 行列の第 i 行を転置したものとなる。そして、これは、時点 0 で X の値が $\boldsymbol{x}_{(0)}$ であった場合の時点 1 に於ける X が取るべき値の生起確率となっている。

次に、時点 2 に於ける X の値の生起確率は、マルコフ連鎖の仮定から、

$$\Pr\left(X^{(2)}=a_j\right)=\sum_{i=1}^{p}\Pr\left(X^{(2)}=a_j|X^{(1)}=a_i\right)\times\Pr\left(X^{(1)}=a_i\right) \tag{16}$$

$$=\sum_{i=1}^{p}k_{ij}\times x_i^{(1)} \tag{17}$$

であるが、これまとめて書くと

$$\boldsymbol{x}^{(2)'}=\boldsymbol{x}^{(1)'}\boldsymbol{K}\ =\ \boldsymbol{x}^{(0)'}\boldsymbol{K}\ \boldsymbol{K}\ =\ \boldsymbol{x}^{(0)'}\boldsymbol{K}^2 \tag{18}$$

と書き表される。そして、この一般形は

$$\boldsymbol{x}^{(t)'}=\boldsymbol{x}^{(0)'}\boldsymbol{K}^t \tag{19}$$

である。

マルコフ連鎖がいくつかの条件を満たせば、核行列をべき乗していくとそれがやがて

$$\lim_{t\to\infty}\boldsymbol{K}^t=\mathbf{1}\boldsymbol{\pi}' \tag{20}$$

すなわち、$\boldsymbol{K}^t$ の各行が定常分布と呼ばれる $\boldsymbol{\pi}$ の転置と等しくなることが知られている。

したがって、どのような初期値 $\boldsymbol{x}_{(0)}$ から出発しても、最終的には

$$\lim_{t\to\infty}\boldsymbol{x}^{(t)'}=\lim_{t\to\infty}\boldsymbol{x}^{(0)'}\boldsymbol{K}^t \tag{21}$$

$$=\boldsymbol{x}^{(0)'}\lim_{t\to\infty}\boldsymbol{K}^t\ =\ \boldsymbol{x}^{(0)'}\mathbf{1}\boldsymbol{\pi}'\ =\ (\boldsymbol{x}^{(0)'}\mathbf{1})\boldsymbol{\pi}' \tag{22}$$

$$=\boldsymbol{\pi}' \tag{23}$$

となり、定常分布が得られることを示している。

また、 $\lim_{t\to\infty}\boldsymbol{K}^t$ を $\boldsymbol{K}^\infty$ と書けば

$$\boldsymbol{\pi}'\boldsymbol{K}^\infty=\boldsymbol{\pi}' \tag{24}$$

であるが、これは定常分布が $\boldsymbol{K}^\infty$ 行列の 1 という値を持つ固有値に対応する左固有ベクトルであることを示している。すなわち、$\boldsymbol{\pi}$ は $\boldsymbol{K}^{\infty'}$ 行列の 1 という値を持つ固有値に対応する（右）固有ベクトルである

例えば

$$\boldsymbol{K}=\begin{bmatrix}0.5 & 0.3 & 0.2\\ 0.1 & 0.6 & 0.3\\ 0.1 & 0.1 & 0.8\end{bmatrix} \tag{25}$$

の場合

$$\boldsymbol{K}^\infty=\begin{bmatrix}0.16667 & 0.26667 & 0.56667\\ 0.16667 & 0.26667 & 0.56667\\ 0.16667 & 0.26667 & 0.56667\end{bmatrix} \tag{26}$$

となり、定常分布は

$$\boldsymbol{\pi} = (0.16667, 0.26667, 0.56667)' \tag{27}$$

である。

なお、核行列を

$$k^*_{ij} = \Pr\left(X^{(t)} = a_i | X^{(t-1)} = a_j\right) = k_{ji} \tag{28}$$

と定義した場合には

$$\boldsymbol{x}^{(t)} = \boldsymbol{K}^{*t}\, \boldsymbol{x}^{(0)} \;=\; (\boldsymbol{K}')^t\, \boldsymbol{x}^{(0)} \tag{29}$$

となる。(mat12.pdf ではこの方式を用いているので注意されたい。)

0.3 ソシオメトリの行列

n 人の人のうち、 i さんが j さんに話しかける場合に $s_{ij} = 1$ そうでない場合には $s_{ij} = 0$ と定義した $n \times n$ の行列を考える。ただし $s_{ii} = 1$ とする。この場合、もしも $s_{ij} = 1$ であれば、i さんから j さんへの長さが 1 の情報の伝達ルートがあるという。この s_{ij} を要素として持つ $n \times n$ の行列を $\boldsymbol{S}$ とし、その第 i 行を $\boldsymbol{s}'_i$、 第 j 列を $\boldsymbol{s}_{(j)}$ と置く。すると、 $\boldsymbol{s}_i$ は i さんが他のどの人と話しているかを示すベクトルとなり、$\boldsymbol{s}_{(j)}$ は j さんがどの人から話を聞くかを示すベクトルとなっている。

なお、この行列は対称行列であるとは限らない。話を聞くだけで自分からは話しかけない人も居るからである。

例えば以下の図には $n = 4$ の場合が示されているが、矢印の向きが情報の移動を示している。また、これを表現した $\boldsymbol{S}$ 行列は以下のようになるがその対角要素には 0 を入れている。この n 人の間には長さが 1 のルートが 6 つ存在する。

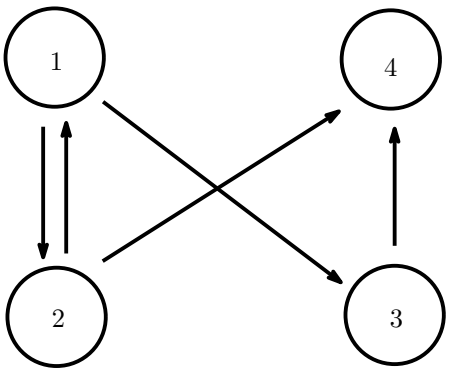

Figure1　4人の関係

$$\boldsymbol{S} \;=\; \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 1 & 0 \\ 2 & 1 & 0 & 0 & 1 \\ 3 & 0 & 0 & 0 & 1 \\ 4 & 0 & 0 & 0 & 0 \end{array} \tag{30}$$

次に、この行列の積 $\boldsymbol{S}^2 = \boldsymbol{S}\boldsymbol{S}$ を考え、その要素を

$$\boldsymbol{S}^2 = \{s^{(2)}_{ij}\} \tag{31}$$

とすると、それは i さんと j さんとの間の長さが 2 のルートの数を示している。この $s_{ij}^{(2)}$ は、行列の積の定義から

$$s_{ij}^{(2)} = \sum_{k=1}^{n} s_{ik}s_{kj} \;=\; \boldsymbol{s}_i'\boldsymbol{s}_{(j)} \tag{32}$$

であるが、i さんが、少なくとも一人の人、たとえば a さんに話しかければ $s_{ia} = 1$ であり、また、j さんがその a さんから話しかけられるとすれば $s_{aj} = 1$ であるから、上記の $s_{ia}s_{aj}$ は 1 になる。また、同様に、i さんが b さんに話しかけ、b さんは j さんに話しかけるとすると、$s_{ib}s_{bj} = 1$ である。したがって、$\boldsymbol{S}^2$ の i, j 要素は i さんから j さんに情報が伝わる長さが 2 のルートの数を示しており、その値が 0 の場合は i さんから j さんに情報は伝わらない。また、この数には、i さんから i さんへ a さんを介して戻ってくるルートも含まれることになる。例えば上記の例では $\boldsymbol{S}^2$ 行列は以下のようになっているが、1 から 1 へのルートで長さが 2 のものの数が 1 となっているが、これは 1 から 2 へ行ってまた 1 へもどって来るというルートのことである。

$\boldsymbol{S}^2 =$

	1	2	3	4
1	1	0	0	2
2	0	1	1	0
3	0	0	0	0
4	0	0	0	0

, $\boldsymbol{S}^3 =$

	1	2	3	4
1	0	1	1	0
2	1	0	0	2
3	0	0	0	0
4	0	0	0	0

, $\boldsymbol{V}^{(3)} =$

	1	2	3	4
1	0	0	1	0
2	0	0	0	2
3	0	0	0	0
4	0	0	0	0

次に $\boldsymbol{S}^3 = \boldsymbol{S}\boldsymbol{S}^2$ を考えてみる。いま、$\boldsymbol{S}^2$ 行列の第 j 列を $\boldsymbol{s}_{(j)}^{(2)}$ と置くと

$$s_{ij}^{(3)} = \sum_{k=1}^{n} s_{ik}s_{kj}^{(2)} \;=\; \boldsymbol{s}_i'\boldsymbol{s}_{(j)}^{(2)} \tag{33}$$

であるが、$\boldsymbol{s}_{(j)}^{(2)}$ の第 k 列には、k さんから j さんへの長さが 2 のルートの数が入っている。したがって、上式は、i さんと k さんが長さが 1 のルートで繋がっており、その k さんと j さんの間の長さが 2 のルートが $s_{kj}^{(2)}$ 個あれば、結局 i さんと j さんとの間の k さんを介する長さが 3 のルートが $s_{kj}^{(2)}$ 個あることを示している。これを全ての k さんに関して加えたものが長さ 3 のルートの総数となる。

一般的には

$$\boldsymbol{S}^{\ell} = \{s_{ik}s_{ij}^{(\ell)}\} \tag{34}$$

が i さんと j さんの間の長さが ℓ のルートの数を表している。また、この行列の要素はは i さんと j さんの間に $\ell - 1$ 人を介したルートの数と解釈することも出来るが、その際、自分自身も仲介する人の数に数えられることに注意しなければならない。例えば上記の例では、$s_{12}^{(3)} = 1$ であるが、二人を介して 1 から 2 へと情報を伝えるルートは $1- > 2- > 1- > 2$ しかなく 1 と 2 がそれぞれ仲介者として数えられている。

このような冗長なルートの数え方を無くすことは難しいが、任意の二人の間を仲介者無しに行き来するルート数えないためには

$$\boldsymbol{V}^{(1)} = \boldsymbol{S}, \quad v_{ij}^{(\ell)} = \sum_{k \neq i,j} s_{ik}v_{kj}^{(\ell-1)} \tag{35}$$

という行列を用いれば良い。しかし、上記の $\boldsymbol{V}^{(3)}$ 行列に示す様に、これでも 1 から 3 への長さ 3 のルート（間に二人が介在するルート）が一つあることになるが、それは $1- > 2- > 1- > 3$ というルートが数えられることになる。

なお、この行列は以下のようにして計算することが出来る。

$$\boldsymbol{V}^{(\ell)} = \boldsymbol{S}\boldsymbol{V}^{(\ell-1)} - \mathrm{diag}\,(\boldsymbol{S})\boldsymbol{V}^{(\ell-1)} - \boldsymbol{S}\mathrm{diag}\left(\boldsymbol{V}^{(\ell-1)}\right) + \mathrm{diag}\,(\boldsymbol{S})\mathrm{diag}\left(\boldsymbol{V}^{(\ell-1)}\right) \quad (36)$$

最後に、i さんから j さんに最低何人を介して情報が伝わるかを表す行列を $\boldsymbol{D}$ とすれば、それは S^{ℓ} 行列を作っていく過程で最初に i, j 要素が 0 で無くなった時点の ℓ を $d_{ij} = \ell$ として記録していけば作成することが出来る。ただし、$d_{ii} = 0$ であり、最後まで 0 で残った要素は、例えば、n を代入すれば良い。なお、この行列は対称であるとは限らないので、厳密には距離行列と呼ぶことは出来ない。

$$\boldsymbol{D} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 1 & 2 \\ 2 & 1 & 0 & 2 & 1 \\ 3 & 1 & 2 & 0 & 1 \\ 4 & 2 & 1 & 1 & 0 \end{array} \quad (37)$$

この $\boldsymbol{D}$ 行列を元に例えば多次元尺度法、クラスター分析等を用いて個人のグルーピングをする事が出来るが、一番単純な方法は、$\boldsymbol{D}$ 行列の行と列を並べ替えて、対角要素付近に小さい値が、対角要素から離れるにしたがって大きな値が入るようにする事である。このような性質を持つ行列を Robinson Matrix と呼ぶが、その並べ替えの方法に関しては例えば以下を参考のこと。

References

Laurent, M. & Seminaroti, M. *Combinatorial and algorithmic properties of Robinson matrices.* https://lipn.univ-paris13.fr/Lagos2017/talks/Laurent.pdf