周期時系列の統計解析

（17）コサイン関数の変数クラスター分析

nino

2021 年 2 月 1 日

複数のコサイン関数に変数クラスター分析を適用した場合，どのようなクラスターが形成されるのかについて検討した．その結果，コサイン関数は周期別にクラスターが形成され，また，それらクラスター内では位相差により分類されると考えられた．

周期時系列

周期時系列として，次のコサイン関数を用いた．

$$x_{T,\alpha}= \cos[2\pi(t+\alpha)\ /T] \qquad (1)$$

ここで，T：周期，α：位相，そして，t：時間（t＝0,1,・・,m）である．

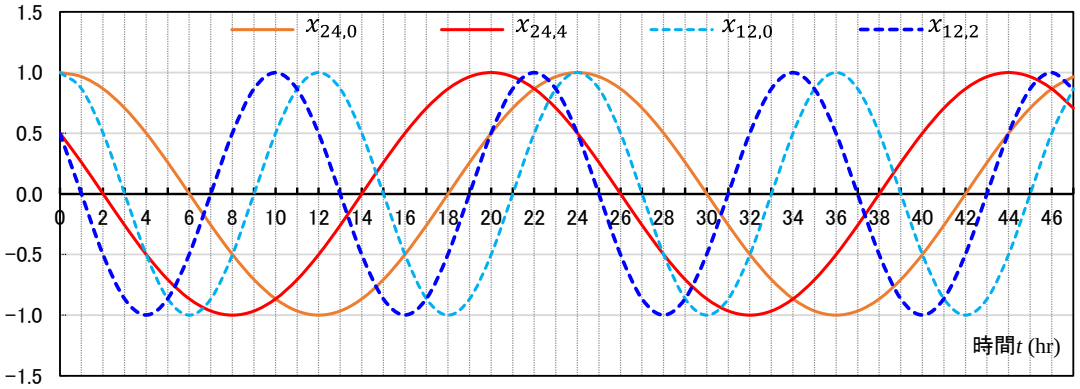
具体例として，1 hr間隔で，期間が 2 日間（n=24×2＝48hr，t =0,1,・・,47hr）の時系列について考える．時間と角度の関係は，1hr＝360°／24hr＝15°に相当する．
例えば，周期T＝24hrで位相α＝0hrの時系列$x_{24,0}$は，

$$x_{24,0}= \cos[2\pi(t+0)\ /24] \qquad (2)$$

で表される．同様に，時系列$x_{24,4}$，$x_{12,0}$，$x_{12,4}$はそれぞれ次式で表される．

$$x_{24,4}= \cos[2\pi(t+4)\ /24] \qquad (3)$$
$$x_{12,0}= \cos[2\pi(t+0)\ /12] \qquad (4)$$
$$x_{12,2}= \cos[2\pi(t+2)\ /12] \qquad (5)$$

$x_{24,4}$は周期が$x_{24,0}$と同じ24hrだが位相はそれより4hr（360°×4hr／24hr＝60°）進み，$x_{12,0}$は周期が12hrで位相が0hr，そして，$x_{12,2}$は周期が$x_{12,0}$と同じだが位相はそれより2hr（360°×2hr／12hr＝60°）進みである．
これら４つの時系列$x_{24,0}$，$x_{24,4}$，$x_{12,0}$，$x_{12,2}$（「Case１」とする）を図１に示した．
図１（Case１）



Case１の変数クラスター分析

最初に，Case1について変数クラスター分析を適用し検討した．距離関数は相関係数rを含む$\sqrt{2(1-r)}$を，また，クラスタリング手法としてウォード法を用いた．距離関数$\sqrt{2(1-r)}$の意味については後述する．

表１にCase１の相関係数rを示した．表中では，$x_{24,0}$と$x_{24,4}$ および$x_{12,0}$と$x_{12,1}$ についてそれぞれ□で囲んである．また，図２にクラスター数を２とした場合における樹形図を示した．

表１

相関係数	$x_{24,0}$	$x_{24,4}$	$x_{12,0}$	$x_{12,2}$
$x_{24,0}$	1			
$x_{24.4}$	0.500	1		
$x_{12.0}$	0.000	0.000	1	
$x_{12.2}$	0.000	0.000	0.500	1

図２

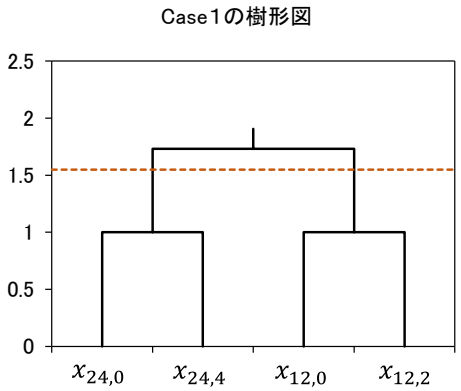


表１をみると，$x_{24,0}$と$x_{24,4}$ および$x_{12,0}$と$x_{12,2}$ はそれぞれ同じグループを形成し，相関係数はともに0.５（＝cos60°）を示した．相関係数は周期時系列の位相差のコサインで表されるからである（参考文献１）．また，$x_{24,\alpha}$と$x_{12,4\alpha}$ は周期が異なるため，両者の相関係数はすべてゼロを示した（参考文献２）．

樹形図（図２）では，$x_{24,0}$と$x_{24,4}$ および$x_{12,0}$と$x_{12,1}$ はそれぞれクラスターを形成し，両クラスター内での距離はともに１（＝$\sqrt{2(1-0.5)}$）で同じ値を示した．また，両クラスター間における合併後の距離は1.73となった．

距離関数の意味について

今回用いた距離関数$\sqrt{2(1-r)}$はどのように解釈すれば良いのだろうか？距離関数の意味について検討してみた．

クラスター分析では，ベクトル間の類似性を表す指標として，コサイン類似度を用いる場合がある．そして，このコサイン類似度を用いてユークリッド距離を表すことができる（詳細は参考文献３，４を参照）．ここでは，２次元ベクトルの場合について概要を記した．

２点P(p_1,p_2)，Q(q_1,q_2)に対してベクトル$\vec{p}(p_1,p_2)$，$\vec{q}(q_1,q_2)$とすると，この２点のユークリッド距離d(P,Q)は，

$$
\begin{aligned}
d(\mathrm{P},\mathrm{Q}) &= |\vec{p}-\vec{q}| \\
&= \sqrt{|\vec{p}-\vec{q}|^2} \\
&= \sqrt{|\vec{p}|^2 - 2\vec{p}\cdot\vec{q} + |\vec{q}|^2}
\end{aligned}
\qquad (6)
$$

となる．

ここで，両ベクトルの内積 $\vec{p},\vec{q} = |\vec{p}||\vec{q}|\cos(\vec{p},\vec{q})$ を式(6)に代入すると，

$$d(\mathrm{P},\mathrm{Q}) = \sqrt{|\vec{p}|^2 - 2|\vec{p}||\vec{q}|\cos(\vec{p},\vec{q}) + |\vec{q}|^2} \quad (7)$$

となり，さらに，$\vec{p}$，$\vec{q}$を正規化すると，$|\vec{p}| = 1$，$|\vec{q}| = 1$であるから，

$$d(\mathrm{P},\mathrm{Q}) = \sqrt{2(1 - \cos(\vec{p},\vec{q}))} \quad (8)$$

が得られた．式(8)は，正規化した場合におけるユークリッド距離d(P,Q)とコサイン類似度$\cos(\vec{p},\vec{q})$の関係を示している．
一方，コサイン類似度$\cos(\vec{p},\vec{q})$は相関係数rに等しいから（参考文献５），

$$d(\mathrm{P},\mathrm{Q}) = \sqrt{2(1-r)} \quad (9)$$

が成り立つ．式(9)は，今回用いた距離関数に等しい．このように，ベクトルを用いてクラスター分析におけるユークリッド距離が導出された．では，周期時系列を用いた場合は，どうだろうか？

同一周期Tの２つの時系列$x_{T,\alpha i}$と$x_{T,\alpha j}$の距離$d(x_{T,\alpha i}, x_{T,\alpha j})$について考える．両者の相関係数$r$と位相差$\alpha_i - \alpha_j = \varphi$の関係は$r = \cos\varphi$であるから(参考文献１)，式(9)より次式が得られる．

$$d(x_{T,\alpha i}, x_{T,\alpha j}) = \sqrt{2(1 - \cos\varphi)} \quad (10)$$

式(10)を式(8)と対比すると，$\cos(\vec{p},\vec{q}) = \cos\varphi$となる．すなわち，２つのベクトルの角度と２つの周期時系列の位相差は等しい（参考文献６）．

さらに，式(10)を三角関数の半角公式を用いて変形すると，

$$\begin{aligned} d(x_{T,\alpha i}, x_{T,\alpha j}) &= \sqrt{2(1 - cos\varphi)} \\ &= \sqrt{2(1 - (1 - 2\sin^2(\varphi/2)))} \\ &= 2\sin(\varphi/2) \end{aligned} \quad (11)$$

となり，同一周期の時系列間の距離は位相差φを含むサイン関数で表された．コサイン類似度がコサイン関数で表されているのと対比している．

このサイン関数で表された距離とコサイン類似度はどのような関係にあるのかを調べるため，２次元ベクトル図を用いて検討した．

例として，図３に半径１の円弧を含むベクトル図を示した．
点Pと点Qは半径１の円弧上にあり，それぞれを原点からのベクトル$\vec{p}$, $\vec{q}$とすると，$|\vec{p}| = |\vec{q}| = 1$である．これは先述の正規化したベクトルの場合に等しい．また，$\vec{p}$と$\vec{q}$のなす角度を∠P0Q＝φとする．点Sは原点0から線分$\overline{\mathrm{PQ}}$に垂線を下した交点，点Rは点Qから線分$\overline{0\mathrm{P}}$に垂線を下した交点である．したがって，∠P0S＝∠Q0S＝$\varphi/2$であり，⊿0PQは線分$\overline{\mathrm{PQ}}$を底辺とする二等辺三角形となる．

図３

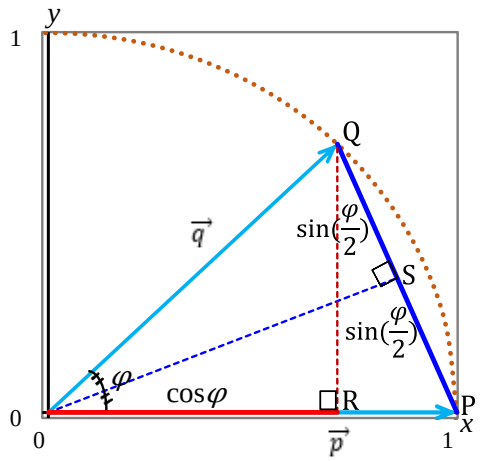


図３において，$\overline{0R}$がコサイン類似度$\cos\varphi$に相当し，$\vec{q}$のx軸($\overline{0P}$)への射影で表される．コサイン類似度は$\vec{q}$の$\vec{p}$との重なり程度と解釈できる．両ベクトルが完全に重なった時，すなわち，$\varphi=0^\circ$ の時にコサイン類似度は最大値１を示す．また，$\varphi=90^\circ$ の時にゼロを，$\varphi=180^\circ$ の時に最小値－１をとる．

一方，$\overline{PS}=\overline{QS}=\sin(\varphi/2)$であるから，$\overline{PQ}=2\sin(\varphi/2)$が成り立つ．これが距離に相当し，角度$\varphi$と連動している．距離は$\vec{q}$の$\vec{p}$ との乖離の程度を表しているといえる．距離は$\varphi=0^\circ$ の時にゼロを，$\varphi=90^\circ$ の時に$2\times\sin45^\circ \fallingdotseq 1.414$を示す．また，$\varphi=180^\circ$ の時に最大値２を，$\varphi=-180^\circ$ の時に最小値－２を示す．距離は「コサイン類似度」と対比して「サイン乖離度」と命名できる．
なお，図３では，便宜上 $\vec{p}$をx軸上に置いて説明したが，$\vec{p}$と$\vec{q}$の関係はあくまで相対的なものである．

参考までに，表２，表３にCase１の位相差φと距離$2\sin(\varphi/2)$を示した．なお，表２の位相差φは式(2)～式(4)より明らかだが，$\arccos(r)$からも求めることができる．

表２

位相差(°)	$x_{24,0}$	$x_{24,4}$	$x_{12,0}$	$x_{12,2}$
$x_{24,0}$	0			
$x_{24,4}$	60.0	0		
$x_{12,0}$	90.0	90.0	0	
$x_{12,2}$	90.0	90.0	60.0	0

表３

距離	$x_{24,0}$	$x_{24,4}$	$x_{12,0}$	$x_{12,2}$
$x_{24,0}$	0			
$x_{24,4}$	1.000	0		
$x_{12,0}$	1.414	1.414	0	
$x_{12,2}$	1.414	1.414	1.000	0

表２で留意すべき点は，24hr周期と12hr周期の時系列間における位相差が$\varphi=90^\circ$ で示されていることである．$\varphi=90^\circ$ は，同一周期の時系列間における位相差であるが，この場合は周期が異なるため，相関係数がゼロになる（表１）ことに起因している（参考文献６）．そのため，表３に示すように，異周期の時系列間の距離は，同一周期で位相差$\varphi=90^\circ$ の時系列間の距離（$2\sin45^\circ \fallingdotseq 1.414$）と同じ値になった．また，$x_{24,0}$と$x_{24,4}$ および$x_{12,0}$と$x_{12,2}$ の距離はともに$2\sin30^\circ = 1$ を示した．

Case１に周期時系列$x_{24,2}$を加えた５つの周期時系列の変数クラスター分析

Case１に時系列$x_{24,2}$を加えた５つの時系列（$x_{24,0}$，$x_{24,2}$，$x_{24,4}$，$x_{12,0}$，$x_{12,2}$：Case２とする）の変数クラスター分析について検討した．

$$x_{24,2}= \cos[2\pi(t+2)\ /24] \qquad (12)$$

$x_{24,2}$は周期が24hrで位相が2hr（360°×2hr／24hr＝30°）の時系列である．

図４にCase２の時系列を示した．

図４（Case2）

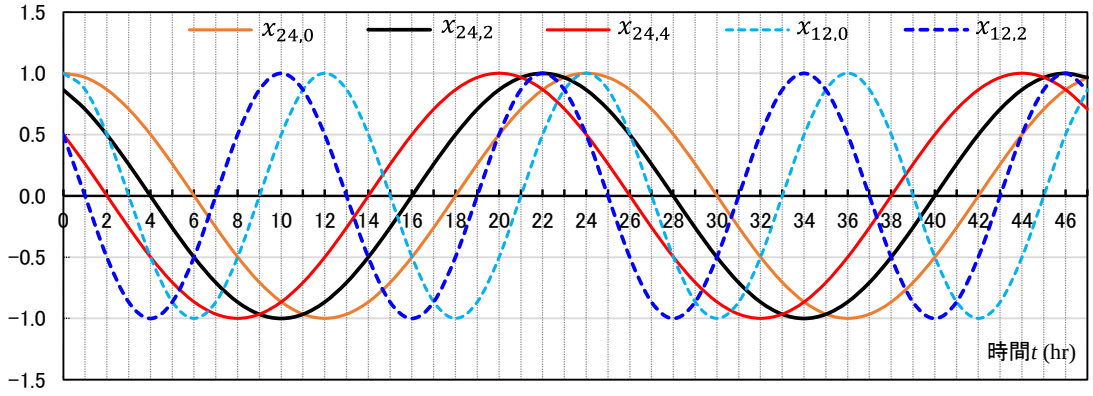


表４，表５，表６にCase２の相関係数，位相差，距離を示した．また，図５にクラスター数を２とした場合の樹形図を示した．

表４

相関係数	$x_{24,0}$	$x_{24,2}$	$x_{24,4}$	$x_{12,0}$	$x_{12,2}$
$x_{24,0}$	1				
$x_{24,2}$	0.866	1			
$x_{24,4}$	0.500	0.866	1		
$x_{12,0}$	0.000	0.000	0.000	1	
$x_{12,2}$	0.000	0.000	0.000	0.500	1

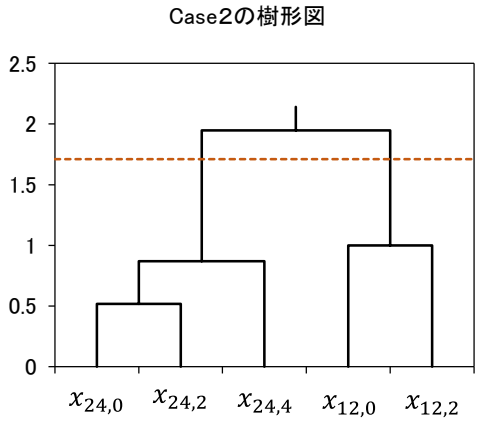
表５

位相差(°)	$x_{24,0}$	$x_{24,2}$	$x_{24,4}$	$x_{12,0}$	$x_{12,2}$
$x_{24,0}$	0				
$x_{24,2}$	30.0	0			
$x_{24,4}$	60.0	30.0	0		
$x_{12,0}$	90.0	90.0	90.0	0	
$x_{12,2}$	90.0	90.0	90.0	60.0	0

表６

距離	$x_{24,0}$	$x_{24,2}$	$x_{24,4}$	$x_{12,0}$	$x_{12,2}$
$x_{24,0}$	0				
$x_{24,2}$	0.518	0			
$x_{24,4}$	1.000	0.518	0		
$x_{12,0}$	1.414	1.414	1.414	0	
$x_{12,2}$	1.414	1.414	1.414	1.000	0

図５



相関係数（表４）によると，Case２は，Case１の場合と同様に，24hr周期と12hr周期の２つのグループ間で無相関を示した．位相差（表５）や距離（表６）も同様であった．表６についてみると，$x_{24,0}$と$x_{24,2}$の距離および$x_{24,2}$と$x_{24,4}$の距離は，いずれの位相差も30°であ

るので，0.518（＝2sin15°）を示した．
樹形図（図５）では，Case１の場合と同様に，周期別にクラスターが形成された．12hr周期の$x_{12,0}$と$x_{12,2}$の距離はCase１と同様に１を示した．しかし，24hr周期のクラスターでは，まず，$x_{24,0}$と$x_{24,2}$のクラスターを形成し，そのクラスターと$x_{24,4}$とで新たにクラスターを形成している．$x_{24,0}$と$x_{24,2}$の距離は0.518，そのクラスターと$x_{24,4}$の合併後の距離は0.869を示した．したがって，24周期のクラスター内では３つの周期時系列は相互の位相差によって分類されると考えられた．さらに，周期別のクラスター間距離は1.95となった．

なお，このウォード法による合併後の距離を試算したところ（参考文献７），当然のことだが，一致した．試算した理由は，複数の市販ソフトやフリーソフトを使用して，距離を求めたところ異なる結果が得られることもあったためである．一般に統計ソフトによって得られた計算結果をそのまま使用する場合が多いと思うが，モデルデータなどを使用して，結果を確認しておくことが望ましい．

これまでの検討結果から，複数の周期時系列に変数クラスター分析を適用すると，時系列は周期別にクラスターを形成するとともに，それら同一周期の各クラスター内では位相差によりさらに分類されると考えられた．主成分分析もクラスター分析と同様な傾向を示しており（参考文献６），両者には共通するものがある．クラスター分析と主成分分析はともに相関係数に基づいた統計手法であるためと考えられる．

参考文献

１．物理のかぎしっぽ，周期時系列の統計解析(1)相関係数と位相差
http://hooktail.sub.jp/contributions/shuki01.pdf
２．物理のかぎしっぽ，周期時系列の統計解析(13)周期時系列における主成分分析の意味
http://hooktail.sub.jp/contributions/shuki13.pdf
３．ユークリッド距離さえあればコサイン類似度が計算できる．
https://qiita.com/obake_kaiware/items/36104a479582063308f0
４．Cosine Distance as Similarity Measure in KMeans [duplicate]
https://stats.stackexchange.com/questions/299013/cosine-distance-as-similarity-measure-in-kmeans
５．相関係数とコサイン類似度
https://qiita.com/wsuzume/items/59ef9db4b3fb9d750dc1
６．物理のかぎしっぽ，周期時系列の統計解析(12) 横断面データの主成分分析
http://hooktail.sub.jp/contributions/shuki12.pdf
７．階層型クラスタリング(ウォード法，群平均法など)の計算過程
https://analysis-navi.com/?p=1805