角度データの統計

——wrapped normal 分布モデル——

統計数理研究所　馬場康維

(1980年1月　受付)

Statistics of Angular Data

—Wrapped Normal Distribution Model—

Yasumasa Baba
(The Institute of Statistical Mathematics)

There are situations where it is required to obtain measures describing structure of multivariate angular data.

In this paper we discuss the measures and propose statistical methods based on a multivariate wrapped normal distribution model. Using simulated bivariate angular data it is investigated to estimate correlation coefficient and regression line.

1. はじめに

観測値が"方向"あるいは角度の場合はそれが2次元の方向であれば単位円周上の点で，3次元の方向であれば単位球面上の点で表現できる．このようなデータは医学，生物学，気象学その他の種々の広範な分野で見出せる一般的なものであり，"directional data"として知られている．directional data の統計に関する最近までの種々の報告は Mardia [1], [2] によって総括されている．本稿では2次元の方向即ち円周上の点の統計について述べる．

円周上の点は適当な方向から測った角度によって表わせる．我々が必要とする統計量は原点となる方向の選び方や角度の範囲をどう選ぶかに依存しないものでなければならない．ところが通常の平均，分散等の統計量はこの要請を満たしていない．例えば平均について考えてみる．x 軸から反時計まわりに角度を測りその範囲は $[0, 2\pi]$ とする．2つの角度 $\pi/4$ と $7\pi/4$ とを考えてみる．我々が直観的に捉える平均的な角度は0又は 2π であろう．ところが算術平均は π となり直観とは一致しない．一方角度を測る原点を y 軸とした場合2つの角度は $7\pi/4, 5\pi/4$ となりその算術平均 $6\pi/4$ は直観的な方向と一致する．さらにもう一つの場合を考えてみる．角度の原点を x 軸とし角度は区間 $[-\pi, \pi]$ で表わすことにする．この場合上記の2つの角度は $\pm\pi/4$ と表現され算術平均は0となる．これは直観的な平均方向と一致する．このように算術平均は原点や角度の範囲の選び方に依存し角度の統計量には適していない．分散，相関係数等の他の統計量も同様である．従って我々は通常の統計量とは異った統計量を構成する必要がある．

角度変量による統計的解析の手法の現状はどうか．一変量に関しては記述統計を始めとして推定論，検定論等はかなり研究されている [1]．2変量の場合 [3] は，相関の尺度についてはいくつかの統計量が提案されており [2], [4], [5], [6], [7]，回帰モデルが Gould によって提案されている [8]．多変量についてはまだ不十分である．

最近，結晶蛋白質分子やある種の tRNA の3次構造の解析に"torsion angle"といわれる

角度が用い得ることが Kitamura によって指摘され，角度変量によるクラスター分析 [9]，相関分析，回帰分析 [10], [11] が行われている．

このような現状を踏まえた上で，本稿では一変量から多変量まで統一的にデータ構造を記述する方法として wrapped normal 分布を基礎にした統計量を提案する（第3節）．

第2節ではこのための準備として角度の原点や定義域の選び方に依存しない統計量として平均方向，円分散，相関係数等を通常の統計量との類推から導出する．

第4節では wrapped normal 分布に対していくつかの相関係数がどういう性質を持っているかを比較検討する．

2. 記 述 測 度

角度データに対しては通常の統計量とは異るものが必要である．この節では矛盾のない形でデータの構造を表現する統計量について述べる．なおこの節の第1項，第2項は主として Mardia [1] の文献によっている．第3項では Baba, Tanemura, Kitamura によって提案された相関係数について述べる．

2.1 平均方向と円分散

単位円周上の位置を角度そのもので表わすと不都合が起きることは本稿の始めに述べた．そこで単位円周上の点を角度 θ のかわりに単位ベクトル $e^{i\theta}$ によって表現することにする．そうすれば，円の回転，角度の範囲の選び方等による不都合は起きない．

n 個の観測値 $\theta_1, \theta_2, \cdots, \theta_n$ から作られる合成ベクトルの平均を

$$\frac{1}{n}\sum_{k=1}^{n} e^{i\theta_k} = \bar{R}\, e^{i\bar{\theta}} \tag{2.1}$$

と表わすことにすると，$\bar{\theta}$ は平均ベクトルの方向，$\bar{R}$ は平均ベクトルの長さである．この $\bar{\theta}$ を平均方向 (mean direction) と呼びこれを角度の平均と見做すことにすれば，第1節で述べたような平均に関する問題は起きない．ただし $\bar{R}=0$ のときは $\bar{\theta}$ は不定になる．以下ではこういう場合は扱わない．

次に $\bar{R}$ の役割を考える．n 個の円周上の位置が同一であるときは $\bar{R}=1$ である．n 個の点が円周を n 等分する点に1個ずつ位置する場合は $\bar{R}=0$ になる．従って $\bar{R}$ は観測値のバラツキに依存する量である．$\bar{R}$ のかわりに

$$V_c = 1-\bar{R} \tag{2.2}$$

を用いる．これは観測値の分布の仕方により0と1の間の値をとりバラツキの測度に適している．この V_c を円分散 (circular variance) と呼ぶ．

観測値から実際に $\bar{\theta}$, $\bar{R}$ を求めるには次のようにすればよい．

$$\bar{c} = \frac{1}{n}\sum_{i=1}^{n} \cos\theta_i, \qquad \bar{s} = \frac{1}{n}\sum_{i=1}^{n} \sin\theta_i \tag{2.3}$$

として

$$\bar{\theta} = \arctan(\bar{s}/\bar{c}) \tag{2.4}$$

$$\bar{R} = \sqrt{\bar{c}^2+\bar{s}^2} \tag{2.5}$$

ただし $\bar{\theta}$ の範囲は $\bar{c}, \bar{s}$ の符号によって決定されるものとする．ところで，原点を平均方向 $\bar{\theta}$ にすると，

$$\sum_{i=1}^{n} \sin(\theta_i-\bar{\theta}) = 0 \tag{2.6}$$

(2.7) $$\sum_{i=1}^{n}\cos(\theta_i-\bar{\theta})=n\bar{R}$$

が成り立つ．従って

(2.8) $$V_c=1-\frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i-\bar{\theta})$$

である．

2.2 円周上の2点の距離

円周上の2点 θ_1, θ_2 を結ぶ弦の長さ

$$1-\cos(\theta_1-\theta_2)$$

は2点が一致していれば0，最も離れている場合（円周上の反対側にある場合）には最大値2をとる．従ってある種の距離の尺度となる [9]．そうすると円周上の任意の点 α に関する n 個の点の分散の尺度として

(2.9) $$V(\alpha)=\frac{1}{n}\sum_{i=1}^{n}\{1-\cos(\theta_i-\alpha)\}$$

が定義できる [1]．この式を最小にする α の値を求めると

(2.10) $$\frac{\partial V(\alpha)}{\partial\alpha}=0$$

より

(2.11) $$\sum_{i=1}^{n}\sin(\theta_i-\alpha)=0$$

でなければならない．(2.6) 式と (2.11) 式の比較により $\alpha=\bar{\theta}$ を得る．即ち平均方向は円分散を最小にする方向である．これは通常の直線上の位置の観測値 $x_1, x_2, \cdots, x_n$ の任意の点 α に関する分散を

(2.12) $$D=\frac{1}{n}\sum_{i=1}^{n}(x_i-\alpha)^2$$

としたとき D を最小にする α は

(2.13) $$\sum_{i=1}^{n}(x_i-\alpha)=0$$

の解であること，即ち平均であることに対応している．従って $\sin(\theta_i-\bar{\theta})$ は観測値 θ_i の平均からの偏差に対応している．

もう一つの変量 ϕ_i を導入し

$$\alpha=a+b\,\phi_i$$

とする．この α を (2.9) 式に代入すると

(2.14) $$V(a,b)=\frac{1}{n}\sum_{i=1}^{n}\{1-\cos(\theta_i-a-b\phi_i)\}$$

を得る．$V(a,b)$ を最小にする a, b を求めれば，ϕ を与えたときの θ の回帰直線を求めることができる．これは最小2乗法に対応している．Gould は円周上の誤差分布が von Mises 分布であるという仮定から ϕ, θ の回帰直線の推定は (2.14) 式で定義される $V(a,b)$ を最小にする a, b を求めることに帰着されることを示している [8]．

2.3 相関係数

2つの角度 θ, ϕ 間の相関を表わす尺度について述べる.

(2.11) 式と (2.13) 式との対応関係から $\sin(\theta-\bar{\theta})$ が平均からの偏差に対応していることは前に述べた通りである. このことから θ, ϕ の共分散を次式で定義する.

$$V(\theta,\phi)=\frac{1}{n}\sum_{i=1}^{n}\sin\theta_i^*\sin\phi_i^* \tag{2.15}$$

ここで

$$\begin{aligned}\theta_i^* &= \theta_i-\bar{\theta},\\ \phi_i^* &= \phi_i-\bar{\phi}\end{aligned} \tag{2.16}$$

であり $\bar{\theta}, \bar{\phi}$ はそれぞれ θ, ϕ の平均方向である. ϕ_i^*, θ_i^* に線型の関係があるとき $V(\theta,\phi)$ は c の正負に応じた符号をとる. このことから, (2.15) 式で θ を ϕ と置きかえたものを $V(\phi,\phi)$, ϕ を θ と置きかえたものを $V(\theta,\theta)$ と書いて

$$r_{ss}=\frac{V(\theta,\phi)}{\sqrt{V(\theta,\theta)\,V(\phi,\phi)}}=\frac{\sum\sin\theta_i^*\sin\phi_i^*}{\sqrt{(\sum\sin^2\theta_i^*)\,(\sum\sin^2\phi_i^*)}} \tag{2.17}$$

を2つの角度間の相関係数と定義する (Baba, Tanemura and Kitamura [5], [6], [7]).

一般に

$$|r_{ss}|\leqq 1$$

であり, $r_{ss}=\pm 1$ となるのは $\phi^*=\pm\theta^*$ のときである.

θ_i^*, ϕ_i^* が原点のまわりに集中しているとき,

$$\sin\theta_i^*\simeq\theta_i^*,\qquad \sin\phi_i^*\simeq\phi_i^* \tag{2.18}$$

という近似をすれば r_{ss} は通常の相関係数になる.

r_{ss} と類似した相関係数が Thompson によって提案されている [12]. これを r_T とすると

$$r_T=\frac{V\left(\frac{1}{2}\theta,\ \frac{1}{2}\phi\right)}{\sqrt{V\left(\frac{1}{2}\theta,\ \frac{1}{2}\theta\right)\,V\left(\frac{1}{2}\phi,\ \frac{1}{2}\phi\right)}} \tag{2.19}$$

である. ここで $V(\theta,\phi)$ は (2.15) 式で定義されるものである. これは円分散が通常の分散に対応すること, 円分散は

$$2V\left(\frac{1}{2}\theta,\ \frac{1}{2}\theta\right)=\frac{1}{n}\sum(1-\cos\theta_i^*)=\frac{2}{n}\sum\sin^2\frac{\theta_i^*}{2}$$

で与えられることから共分散を

$$2V\left(\frac{1}{2}\theta,\ \frac{1}{2}\phi\right)=\frac{2}{n}\sum\sin\frac{\theta_i^*}{2}\sin\frac{\phi_i^*}{2}$$

と定義して得られるものである.

r_T は平均方向と反対側のデータに敏感である. $\theta_i^*=\pm\pi$ の点は $\theta^*\to\sin\frac{\theta_i^*}{2}$ という変換によって ± 1 に変換される. したがって平均方向の僅かな変動によって平均方向から最も離れたデータが共分散に大きな変動を及ぼすことになる. したがって r_T は円周上で拡がった分布を

持つデータの相関係数としては不適当である．

一方 r_{ss} は上記のような不都合はおきない．むしろ分布のすそ（$\pm\pi$ の点）の影響をあまり受けない．二つの相関係数のこの相違は r_{ss} は θ,ϕ の周期 2π の関数を用いた統計量であるのに対して r_T は周期 π の関数を用いたものであることによる．

3. wrapped normal 分布モデル

前節で導出した統計量は角度の原点や定義域の選び方に依存しないものという利点を持っているが，データの構造を見るにはこれを変換した方が解り易い．例えば円分散よりは標準偏差を用いて角度のバラツキが $\pm\alpha$ というような表現の方が解り易い．ここでは分布のモデルとして wrapped normal 分布を用いて前節で求めた統計量よりも通常のものに近い統計量への変換式を求める．

我々が必要とする分布は1変量の場合は円周上で定義されるものであり，2変量の場合はトーラス上の分布，多変量の場合は多次元トーラス上の分布である．

1変量の場合についていえば直線上の正規分布に対応する円周上の分布は von Mises 分布である．したがって分布モデルとして von Mises 分布が考えられる．しかしながら von Mises 分布はトーラス上の分布への拡張がむずかしい．

1変量の wrapped normal 分布は正規分布を円周に巻きつけたもので von Mises 分布に近い性質を持っている．これは2変量の正規分布をトーラスに巻きつけることによってトーラス上の分布に簡単に拡張できる．さらに多変量正規分布を多次元トーラスに巻きつければ多変量の wrapped normal 分布が定義できる．

wrapped normal 分布をモデルとする利点は2変量以上の多変量に拡張した場合，ユークリッド空間で定義される多変量正規分布と同様に相関係数あるいは分散共分散行列が具体的な意味を持つことにある．さらに周辺分布はやはり wrapped normal 分布になるという扱いやすさも持っている．

3.1 wrapped normal 分布

wrapped normal 分布と von Mises 分布

円周上で定義される確率分布は周期性を持つ必要がある．即ち p.d.f. を $f(\theta)$ としたとき

$$f(\theta+2\pi)=f(\theta) \tag{3.1}$$

$$\int_0^{2\pi} f(\theta)\,d\theta=1. \tag{3.2}$$

この要請を満たす分布の代表的なものは von Mises 分布と wrapped normal 分布である．

von Mises 分布は直線上の正規分布に対応する分布で "circular normal 分布" ともいわれる．

正規分布は位置の母数 μ の最尤推定量が標本平均である分布として特徴づけられる．これに対応して von Mises 分布は円周上の位置の母数 μ の最尤推定量が平均方向である分布として特徴づけられる [13].

von Mises 分布の p.d.f. は

$$f(\theta)=\frac{1}{2\pi I_0(\kappa)}\exp\{\kappa\cos(\theta-\mu)\} \tag{3.3}$$

で与えられる．ここで $I_m(\kappa)$ は m 次の第 I 種変形ベッセル関数である．$\kappa=0$ のとき分布は円周上の一様分布となる．κ が大きいとき分布は正規分布 $N(\mu,\kappa^{-1})$ に近づく．

wrapped normal 分布の p.d.f. は次式で与えられる．

$$f(\theta)=\frac{1}{\sqrt{2\pi}\,\sigma}\sum_{k=-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}(\theta-2\pi k-\mu)^2\right\} \tag{3.4}$$

この分布は σ が大きいとき円周上の一様分布に近づき，σ が小さいとき正規分布 $N(\mu,\sigma^2)$ に近づく．

両極限では wrapped normal 分布と von Mises 分布は一致する．中間領域については Stephens により二つの分布が近いことが数値的に確かめられている [14]．二つの分布を描いたものが第1図である．ここでは von Mises 分布および wrapped normal 分布の平均を0と置いている．κ と σ は二つの分布の円分散を等しいとしたときの関係式

$$\exp(-\sigma^2/2)=I_1(\kappa)/I_0(\kappa) \tag{3.5}$$

によって対応させたものである．

2変量 wrapped normal 分布

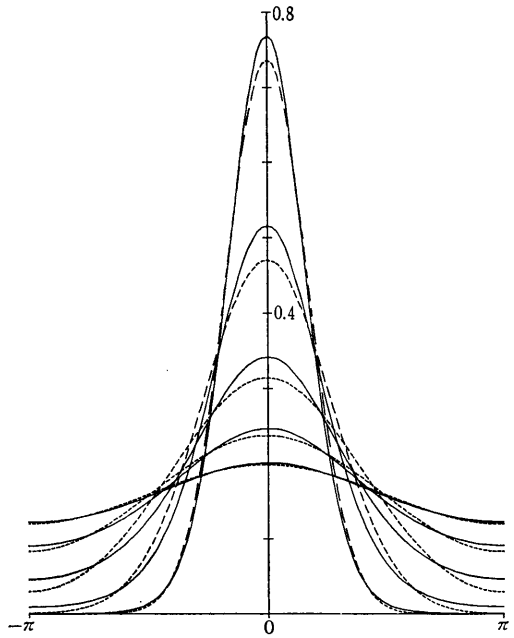
2変量の wrapped normal 分布は次式で定義される．

$$f(\theta,\phi)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\sum_k\sum_l\exp\{-Q/2\}$$

$$Q=\left\{\frac{(\theta-2\pi k-\mu_1)^2}{{\sigma_1}^2}-\frac{2\rho(\theta-2\pi k-\mu_1)(\phi-2\pi l-\mu_2)}{\sigma_1\sigma_2}+\frac{(\phi-2\pi l-\mu_2)^2}{{\sigma_2}^2}\right\}\Big/(1-\rho^2) \tag{3.6}$$

周辺分布は

$$\int_0^{2\pi}f(\theta,\phi)\,d\phi=\frac{1}{\sqrt{2\pi}\,\sigma_1}\sum_k\exp\left\{-\frac{1}{2{\sigma_1}^2}(\theta-2\pi k-\mu_1)^2\right\} \tag{3.7}$$



第1図 von Mises 分布と wrapped normal 分布．
実線は von Mises 分布で頂点の高い順に $\kappa=4, 2, 1, 1/2, 1/4$．破線は von Mises 分布を近似する wrapped normal 分布で，頂点の高い順に $\sigma=0.542, 0.848, 1.270, 1.683, 2.043$

となり1変量の wrapped normal 分布である．特性関数は

$$(3.8)\qquad c(p,q)=\mathrm{E}(e^{i\theta p+i\phi q})$$
$$=\exp\left\{i\mu_1 p+i\mu_2 q-\frac{1}{2}\sigma_1{}^2p^2-\frac{1}{2}\sigma_2{}^2q^2-\rho\sigma_1\sigma_2 pq\right\}$$

となり通常の正規分布の特性関数と同じ形をしている．ただし $f(\theta,\phi)$ の周期性から，p, q は整数値以外は取り得ない．

σ_1, σ_2 が小さい極限では (2.6) 式の主要な項は $k=0$, $l=0$ の項である．従って σ_1, σ_2 が小さい極限では wrapped normal 分布は正規分布 $N(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)$ である．σ_1, σ_2 が大きい極限はフーリエ級数展開を用いると簡単に求まる．(3.8) を用いて

$$(3.9)\quad f(\theta,\phi)=\frac{1}{(2\pi)^2}\left[1+2\sum_{p=1}^{\infty}\cos p\theta_1{}^*\,e^{-\sigma_1{}^2p^2/2}+2\sum_{q=1}^{\infty}\cos q\theta_2{}^*\,e^{-\sigma_2{}^2q^2/2}\right.$$
$$\left.+4\sum_{p=1}^{\infty}\sum_{q=1}^{\infty}\{\cos(p\theta_1{}^*+q\theta_2{}^*)e^{\rho\sigma_1\sigma_2pq}+\cos(p\theta_1{}^*-q\theta_2{}^*)e^{-\rho\sigma_1\sigma_2pq}\}e^{-(\sigma_1{}^2p^2+\sigma_2{}^2q^2)/2)}\right]$$

と表わされる．ここで

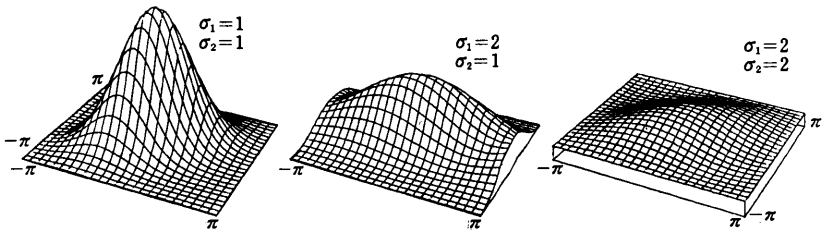
$$\theta_1{}^*=\theta-\mu_1,\qquad \theta_2{}^*=\phi-\mu_2$$

である．$|\rho|<1$ で σ_1, σ_2 が大きい極限では

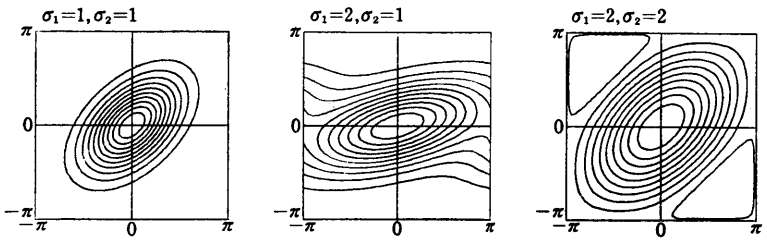
$$(3.10)\qquad f(\theta,\phi)\simeq\frac{1}{(2\pi)^2}$$

となる．即ちトーラス上の一様分布である．

2変量 wrapped normal 分布の例を第2図，第3図に示した．



第2図　$\rho=0.5$ の wrapped normal 分布
この立体図は文献[17]所載のプログラムを用いて描いたものである．



第3図　$\rho=0.5$ の wrapped normal 分布の等高線
この等高線は文献[17]所載のプログラムを用いて描いたものである．

多変量 wrapped normal 分布

2変量の wrapped normal 分布を拡張して p 変量 wrapped normal 分布の p.d.f. を次式で定義する.

$$f(\theta_1, \theta_2, \cdots, \theta_p) = \frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|} \sum_{\boldsymbol{k}} \exp\left(-\frac{1}{2} Q\right) \tag{3.11}$$

$$Q = (\boldsymbol{\theta} - 2\pi\boldsymbol{k} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - 2\pi\boldsymbol{k} - \boldsymbol{\mu})$$

ここで

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \qquad \boldsymbol{k} = \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ k_p \end{pmatrix}$$

である. $\boldsymbol{\theta}$ は確率ベクトル, $\boldsymbol{\mu}$ は平均であり $\boldsymbol{k}$ は成分が整数のベクトルである. また

$$\boldsymbol{\Sigma} = [\sigma_{ij}]$$

はトーラスに巻きつける正規分布の分散共分散行列である. 以下 θ_i に関する分散を ${\sigma_i}^2 = \sigma_{ii}$ と書く. また θ_i, θ_j の通常の相関係数を ρ_{ij} とすると $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ である. なお式中の和の記号 $\sum_{\boldsymbol{k}}$ は $\boldsymbol{k}$ の成分 $k_1, k_2, \cdots, k_p$ に関する和を表わす.

正規分布は退化していないものとする. 即ち

$$|\boldsymbol{\Sigma}| \neq 0$$

であるものとしておく. 2変量の場合と同様にこの分布の二つの極限は正規分布およびトーラス上の一様分布である.

3.2 統 計 量

p 変量 wrapped normal 分布に対して2節の統計量に対応するものを求める. 2節の標本に関する期待値は wrapped normal 分布の期待値で置きかえるものとする.

$$\mathrm{E}(\sin(\theta_i - \mu_i)) = 0 \qquad (i = 1, 2, \cdots, p) \tag{3.12}$$

がなりたつ. したがって母平均方向は μ_i である.

平均ベクトルの長さ $\bar{R}$ に対応するものを R_{oi} とすると

$$R_{oi} = \mathrm{E}(\cos(\theta_i - \mu_i)) = \exp\left(-\frac{1}{2}{\sigma_i}^2\right) \qquad (i = 1, \cdots, p) \tag{3.13}$$

したがって

$$\sigma_i = \sqrt{-2\log_e R_{oi}} \tag{3.14}$$

である. ここで $0 \leqq R_{oi} \leqq 1$ であるから σ_i は $[0, \infty)$ で定義される. R_{oi} を θ_i の標本による平均ベクトルの長さ $\bar{R}_i$ で置き換えて得られる量

$$s_i = \sqrt{-2\log_e \bar{R}_i} \tag{3.15}$$

を θ_i の標準偏差と定義する. これは1変量の場合の定義と同じものである [15].

次に (2.15) 式で定義される共分散と (2.17) 式で定義される相関係数に対応するものを求める.

$$V_{ij} = \mathrm{E}(\sin(\theta_i - \mu_i)\sin(\theta_j - \mu_j))$$

とすると

$$V_{ij}=\exp\left\{-\frac{1}{2}(\sigma_i^2+\sigma_j^2)\right\}\sinh\sigma_{ij} \tag{3.16}$$

したがって第2節の r_{ss} に対応する θ_i, θ_j の相関係数を c_{ij} とすると

$$c_{ij}=\frac{\sinh\sigma_{ij}}{\sqrt{\sinh\sigma_i^2\sinh\sigma_j^2}} \tag{3.17}$$

である．ところで(3.13)式と(3.16)式より

$$V_{ij}=R_{oi}R_{oj}\sinh\sigma_{ij} \tag{3.18}$$

である．したがって

$$\sigma_{ij}=\operatorname{arc\,sinh}\{V_{ij}/(R_{oi}R_{oj})\} \tag{3.19}$$

がなりたつ．あるいは(3.17)式より

$$\sigma_{ij}=\operatorname{arc\,sinh}(c_{ij}\sqrt{\sinh\sigma_i^2\sinh\sigma_j^2}) \tag{3.20}$$

である．c_{ij} を r_{ss} で，σ_i, σ_j を(3.15)式で定義される s_i, s_j で置き換えることにより得られる統計量

$$s_{ij}=\operatorname{arc\,sinh}(r_{ss}\sqrt{\sinh s_i^2\sinh s_j^2}) \tag{3.21}$$

を θ_i, θ_j の共分散と定義する．

以上の変換により $\theta_1, \theta_2, \cdots, \theta_p$ に関する分散共分散行列

$$\boldsymbol{s}=[s_{ij}] \tag{3.22}$$

が求まる．

相関行列を

$$\boldsymbol{r}=[r_{ij}] \tag{3.23}$$

とする．r_{ij} は θ_i, θ_j の相関係数で次式によって定義する．

$$r_{ij}=s_{ij}/(s_is_j) \tag{3.24}$$

3.3 議　　論

ここで定義した標準偏差，分散共分散行列，相関係数を与える式(3.15)，(3.21)，(3.24)等は2節で与えられた統計量からの変換式である．この変換によって得られる統計量は2節のものより通常のものに近い．母集団が wrapped normal かどうかにかかわらず上の変換式を用いて統計量を定義すれば角度の原点や定義域には依存しない統計量が得られたことになる．

母集団の分布として wrapped normal 分布が仮定できる場合 (3.21)，(3.24) は分散共分散行列および相関行列の推定値を与える．なお平均の推定値は平均方向である．

標本分布に基づく議論は稿を改めることにしたい．ここでは十分 n が大きい場合の点推定量を求めたものとするに留める．推定量の偏り，分散等の性質を考えると十分大きな n とは

$$\exp(\sigma^2)\ll n$$

を満たす n である．

母集団が wrapped normal であるという仮定が少しゆるめられたとしても単峰で平均方向の近傍に集中したデータに対して上の変換式はデータ構造の表現に十分に役立つものと考えられる．これには標準偏差 s_i が小さい場合を考えてみる．そのときは n をデータ数として

$$s_i^2 \simeq \frac{1}{n}\sum_{l=1}^{n}(\theta_{il}-\bar{\theta}_i)^2 \qquad (i=1,\cdots,p)$$

$$s_{ij} \simeq \frac{1}{n}\sum_{l=1}^{n}(\theta_{il}-\bar{\theta}_i)(\theta_{jl}-\bar{\theta}_j) \qquad (i,j=1,\cdots,p)$$

と近似される．これは通常の分散，共分散の定義に他ならない．したがって小さな標準偏差に対しては我々は通常の統計量を求めていることになる．

3.4 例

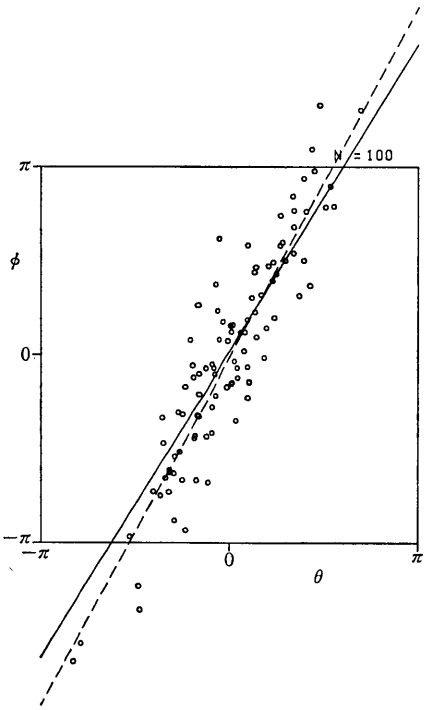
2変量の場合のシミュレーションデータにより解析例を示す．第4図は正規分布 $N(0, 0, \left(\frac{5\pi}{18}\right)^2, \left(\frac{5\pi}{9}\right)^2, 0.9)$ からの 100 個の標本をプロットしたものである．我々がこういうデータを観測した場合は実際には第5図か第6図のような観測値を得る．通常の線型の相関係数を求めてみると，第4図，第5図，第6図それぞれ，0.913, 0.490, 0.353 を得る．一方 r_{ss} は 0.549 で表現のしかたには影響されない．

wrapped normal 分布を用いて得られる分散共分散行列は

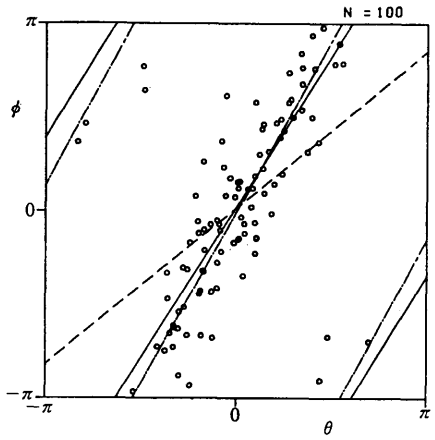
$$\begin{bmatrix} 0.811 & 1.322 \\ 1.322 & 3.000 \end{bmatrix}$$

である．これから相関係数の推定値として 0.848 を得る．平均 μ の推定値は [0.005, 0.075] である．従って θ を与えたときの ϕ の回帰直線を

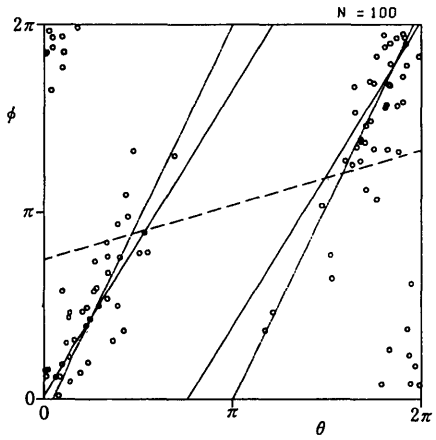
$$\phi = a+b\theta$$



第4図 $N(0, 0, \left(\frac{5\pi}{18}\right)^2, \left(\frac{5\pi}{9}\right)^2, 0.9)$ からの 100 個の標本および回帰直線
実線： wrapped normal 分布モデル，$\phi=1.63\theta+0.07$
破線： 最小2乗法，$\phi=1.85\theta-0.01$



第5図　領域 $[-\pi, \pi]^2$ による表現
実線：　wrapped normal 分布モデル，
$\phi=1.63\theta+0.07$
破線：　最小2乗法，$\phi=0.84\theta+0.04$
一点鎖線：　von Mises 分布モデル，
$\phi=1.86\theta-0.03$



第6図　領域 $[0, 2\pi]^2$ による表現
実線：　wrapped normal 分布モデル，
$\phi=1.63\theta+0.07$
破線：　最小2乗法，$\phi=0.29\theta+2.35$
一点鎖線：　von Mises 分布モデル，
$\phi=2.09\theta-0.30$

とすると

$$\hat{b} = s_{12}/s_1^2 = 1.63$$

$$\hat{a} = \bar{\phi} - b\bar{\theta} = 0.067$$

従って回帰直線は

$$\phi = 0.067 + 1.63\theta$$

で与えられる．第4図，第5図，第6図の実線はこの回帰直線を描いたもので，どの表現でも回帰直線は変わらない．比較のために通常の線型の回帰直線を描いてある．第4図，第5図，第6図の破線がその図のプロットに対応した回帰直線である．表現の仕方によって全く異った直線が得られるのがわかる．

次に Gould の方法によるものを描いたのが第4図，第5図，第6図の一点鎖線である．これは

$$V = 1 - \frac{1}{n}\sum_{i=1}^{n} \cos(\phi_i - a - b\theta_i) \tag{3.25}$$

を最小にする a, b を求めることによって回帰直線を求めたものである．第4図では通常の回帰直線と重なっている．第5図の直線は第4図とほぼ同じものである．第6図だけは他の2つの図の場合と幾分異っている．これは(3.25)式が変換

$$\phi_i \to \phi_i + 2\pi$$

に対して不変であるが，

$$\theta_i \to \theta_i + 2\pi$$

に対しては不変ではないことと関係している．

4. 相関係数

ここではいくつかの相関係数を比較する．なおこの節は Baba, Tanemura の最近の研究に

基づいている [5], [6], [7].

$(\theta_1, \phi_1), (\theta_2, \phi_2), \cdots, (\theta_n, \phi_n)$ を n 個の観測値とする．θ, ϕ は角度である．それぞれの平均方向を $\bar{\theta}, \bar{\phi}$ とし

$$\theta_i^* = \theta_i - \bar{\theta}, \qquad \phi_i^* = \phi_i - \bar{\phi}$$

とする．θ, ϕ のそれぞれの平均ベクトルの長さを $\bar{R}_1, \bar{R}_2$ とする．

以下では θ, ϕ の相関係数として第2節で定義した r_{ss}, Mardia の定義によるもの, Masuyama の symmetrical correlation coefficient を角度の場合に適用したものの3つの相関係数について比較する．

Mardia の相関係数を r_M とすると

$$r_M^2 = \frac{\max(D_+, D_-)}{(1-\bar{R}_1^2)(1-\bar{R}_2^2)} \tag{4.1}$$

$$D_\pm = \left\{\frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i^* \pm \phi_i^*) - \bar{R}_1\bar{R}_2\right\}^2 + \left\{\frac{1}{n}\sum_{i=1}^{n}\sin(\theta_i^* \pm \phi_i^*)\right\}^2 \tag{4.2}$$

である．

ところで単位ベクトル間の相関係数 [16] を r_v とすると，

$$r_v = \left\{\frac{1}{n}\sum\cos(\theta_i^* - \phi_i^*) - \bar{R}_1\bar{R}_2\right\} \Big/ \sqrt{(1-\bar{R}_1^2)(1-\bar{R}_2^2)}$$

である．一方 (4.1) 式の $D_\pm$ の右辺の第2項は wrapped normal 分布のような対称な分布からの標本に対しては小さな値になる．この項を無視すれば，r_M^2 は r_v^2 に等しい．したがって r_M はベクトル相関と考えることができる．

これに対して r_{ss} は単位ベクトルの正弦のみの相関であり，θ, ϕ の関数のスカラー相関である．

Masuyama の symmetrical correlation coefficient [16] を単位ベクトルの場合に適用すると次の相関係数 [7] が得られる．

$$r_{sym} = \frac{|T|}{\sqrt{|A||B|}} \tag{4.3}$$

ここで A, B, T は次式で与えられる行列である．

$$A = \frac{1}{n}\begin{bmatrix} \sum\cos^2\theta_i^* - n\bar{R}_1^2 & \sum\cos\theta_i^*\sin\theta_i^* \\ \sum\sin\theta_i^*\cos\theta_i^* & \sum\sin^2\theta_i^* \end{bmatrix}$$

$$B = \frac{1}{n}\begin{bmatrix} \sum\cos^2\phi_i^* - n\bar{R}_2^2 & \sum\cos\phi_i^*\sin\phi_i^* \\ \sum\sin\phi_i^*\cos\phi_i^* & \sum\sin^2\phi_i^* \end{bmatrix}$$

$$T = \frac{1}{n}\begin{bmatrix} \sum\cos\theta_i^*\cos\phi_i^* - n\bar{R}_1\bar{R}_2 & \sum\cos\theta_i^*\sin\phi_i^* \\ \sum\sin\theta_i^*\cos\phi_i^* & \sum\sin\theta_i^*\sin\phi_i^* \end{bmatrix}$$

A, B, T の成分をそれぞれ A_{ij}, B_{ij}, T_{ij} で表わすと我々の相関係数は

$$r_{ss} = \frac{T_{22}}{\sqrt{A_{22}B_{22}}} \tag{4.4}$$

である．Mardia の相関係数は

$$r_M^2 = \frac{\max(D_+, D_-)}{(A_{11}+A_{22})(B_{11}+B_{22})} \tag{4.1}'$$

(4.2)′　　$D_{\pm}=(T_{11}\mp T_{22})^2(T_{12}\pm T_{21})^2$　　（複号同順）

と表現できる．即ち r_{ss}, r_M, r_{sym} は A, B, T の成分の組合せ方によって異った形になるものである．

2変量 wrapped normal 分布に対して上記の3つの相関係数を求める．$n\to\infty$ として標本による期待値を wrapped normal 分布による期待値に置きかえ，対応する相関係数を ρ_{ss}, ρ_M, ρ_{sym} とすれば

(4.5)　　$$\rho_{ss}=\sinh(\rho\sigma_1\sigma_2)/\sqrt{\sinh\sigma_1^2\sinh\sigma_2^2}$$

(4.6)　　$$\rho_M=\mathrm{sgn}(\rho)\{\exp(|\rho|\sigma_1\sigma_2)-1\}/\sqrt{(\exp\sigma_1^2-1)(\exp\sigma_2^2-1)}$$

(4.7)　　$$\rho_{sym}=\rho_{ss}\sinh^2\frac{\rho\sigma_1\sigma_2}{2}\Big/\sinh^2\frac{\sigma_1^2}{2}\sinh^2\frac{\sigma_2^2}{2}$$

したがって $|\rho|=1$ であっても各相関係数の絶対値は1より小さくなる．

さらに常に

$$|\rho_{sym}|<|\rho_{ss}|$$

がなりたつ．このことから十分大きな n に対して r_{ss} の方が r_{sym} よりも θ,ϕ の線型関係に敏感であるといえる．ρ_M と ρ_{ss} の関係は ρ,σ_1,σ_2 に依存して複雑である．

σ_1,σ_2 が小さいときの各相関係数は

$$\rho_{ss}\simeq\rho\left\{1-\frac{1}{12}(\sigma_1^4-2\rho^2\sigma_1^2\sigma_2^2+\sigma_2^4)\right\}$$

$$\rho_M\simeq\rho\left\{1-\frac{1}{2}(\sigma_1^2-2\rho\sigma_1\sigma_2+\sigma_2^2)\right\}$$

$$\rho_{sym}\simeq\rho^3\left\{1-\frac{1}{8}(\sigma_1^4-2\rho^2\sigma_1^2\sigma_2^2+\sigma_2^4)\right\}$$

となる．従って ρ_{sym} は通常の相関係数の拡張にはなっていない．ρ_{ss}, ρ_M の ρ への近づき方を比較すると ρ_{ss} の方が早い．従って ρ_{ss} の方が通常の線型相関係数の“自然な”拡張になっていると考えられる．

おわりに

この研究を始めるきっかけとなったのは大阪大学薬学部の北村一泰氏との共同研究である．また統計数理研究所の種村正美氏には共同研究を通じ種々の議論をしていただいた．両氏に謝意を表する．

有益な助言をいただいた査読者に感謝する．

参考文献

[1] Mardia, K.V. (1972). *Statistics of Directional Data*, Academic Press, New York.

[2] Mardia, K.V. (1975). Statistics of directional data (with discussion), *J.R. Statist. Soc.*, B, **37**, 349–393.

[3] Puri, M.L. and Rao, J.S. (1977). Problems of association for bivariate circular data and a new test of independence, *Multivariate Analysis* IV (ed. P.R. Krishnaiah), 513–522, North Holland, Amsterdam.

[4] Mardia K.V. and Puri, M.L. (1978). A spherical correlation coefficient robust against scale, *Biometrika*, **65**, 391–395.

[5] 馬場康維，種村正美 (1978)． 円周上の統計，第46回日本統計学会予稿集 93.
[6] 馬場康維，種村正美 (1979)． Directional Data の相関，第47回日本統計学会予稿集 35-36.
[7] Baba, Y., Tanemura, M. and Kitamura, K. Correlation between Angular Variates, 発表予定.
[8] Gould, A.L. (1969). A regression technique for angular variates, *Biomtrics*, **25**, 683–700.
[9] Kitamura, K., Wakahara, A., Hakoshima, T., Mizuno, H. and Tomita, K. (1978). Classification of Nucleotide Comformations Observed in Yeast Phenylalanine tRNA, Application of Cluster Analysis, *Nucleic Acids Res. Special Publication* No. **5**, s373–376.
[10] 北村一泰，箱嶋敏雄，砂田基晴，若原章男，冨田研一，馬場康維 (1979)． ヌクレオサイドの torsion angle 間の相関性について —torsion angle の新しい表現法とその応用—，第6回生体分子の構造に関する討論会講演要旨集 37-38.
[11] 北村一泰，砂田基晴，若原章男，冨田研一，松浦良樹，安岡則武，角戸正夫，馬場康維 (1979)． 結晶蛋白質分子中にみられる Oligopeptide unit の構造特性の抽出 —Protein Data Bank の利用と Comformation の統計的解析—，化学と情報に関する討論会論文集 37-42.
[12] Thompson, J.W. (1975). Discussion of Professor Mardia's paper, *J.R. Statist. Soc.*, B, **37** 379.
[13] Gumbel, E.J., Greenwood, J.A. and Durand, D. (1953). The circular normal distribution: Theory and tables, *J. Amer. Statist. Ass.*, **48**, 131–152.
[14] Stephens, M.A. (1963). Random walk on a circle, *Biometrika*, **50**, 385–390.
[15] 文献 [1] の Chap. 3 参照.
[16] Masuyama, M. (1939). Correlation between tensor quantities, *Proc. Phys-Math. Soc. Japan*, **21**, 638.
[17] 森 正武 (1974)． 曲線と曲面 —計算機による作図と追跡—，教育出版.