2021 年 2 月 28 日

方向ベクトルの 2 次元分布を用いた統計モデルに関する研究
A Study on Statistical Models Using Two-Dimensional Distribution of Directional Vectors

19N7100014C
経営システム工学専攻田寺 凌太

概要

本稿では実風データについて, 角度データを扱うモデル, 風速データを扱うモデル, 両方を同時に表すモデルを用い解析する.
さらに観測風データについて風向風速を同時に表し、それを視覚的に認知出来るようなモデルの作成を目標とした。また, 風速観測における左側打ち切りデータに考慮したモデルについても考察を行うこととした.

1 角度データとは

風向データや渡り鳥の移動方向のような, 個々の観測が $[0, 2\ \pi)$ の角度として表されるデータを指す [1]. このデータにおいては実数値データのための統計的手法をそのまま用いることができない場合がある. 例えば 2 つの角度 $\pi/4$, $7\ \pi/4$ の実数値における標本平均はπであると計算できるが, この値を ” 角度の平均 “ とするのは自然な定義とは言えない. 同様に分散, 歪度, 尖度などの要約統計量も, このままでは不自然な定義となってしまう.

1.1 Von-mises 分布

先のような特異性を持つデータを扱うために本稿ではガウス分布の周期変数への一般化である von-mises 分布がある.[1]von-mises 分布の確率密度関数は

$$\boldsymbol{x} = (\cos\theta, \sin\theta)' \tag{1}$$

$$\boldsymbol{\mu} = (\cos\eta, \sin\eta)' \tag{2}$$

に対して次のように与えられる.

$$VM(\theta|\boldsymbol{\mu}, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^{'} \boldsymbol{x}\right) \quad 0 \leq \theta < 2\pi \tag{3}$$

ただし式中のパラメータ $\boldsymbol{\mu}$ は分布のモード,κ は集中度, 関数 $I_0(\kappa)$ は 0 次の第一種変形 Bessel 関数

$$I_\nu(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} \cos(\nu\theta)\exp[\kappa\cos\theta]d\theta \tag{4}$$

$$= \sum_{r=0}^{\infty} \frac{1}{\Gamma(\nu + r + 1)r!}(\frac{\kappa}{2})^{2r+\nu} \tag{5}$$

を表す. パラメータ κ は集中度,μ は平均方向を表し, 平均合成ベクトル長は $A(\kappa) = I_1(\kappa)/I_0(\kappa)$ で与えられる.

1.1.1 混合分布

Von-mises 分布は単峰の分布であるため多峰性のデータをうまく表現することは出来ない. そこで 2 つ以上の Von-mises 分布の混合モデル (von-mises mixture) を扱う. その確率密度関数は K コの Von-mises 分布 $VM(\mu_1, \kappa_1)...VM(\mu_k, \kappa_k)$ の確率密度関数をそれぞれの混合比率 $p_k(0 \leq p_k \leq 1)$ で混合して,

$$f(X) = \sum_{k=1}^{K} p_k VM_k(X|\boldsymbol{\Theta_k}) \tag{6}$$

と表現される.

1.2 混合分布の解析手法

1.2.1 最尤法

混合分布の形状はパラメータ $\boldsymbol{p} = \{p_1, ...p_k\}, \boldsymbol{\mu} = \{\mu_1, ...\mu_k\}, \boldsymbol{\kappa} = \{\kappa_1, ..., \kappa_k\}$ によって決まる, これらのパラメータの値を求める方法の 1 つに最尤推定法がある. 解析データ $\boldsymbol{X} = \{x_1, ..., x_n\}$ についての対数尤度関数は,(6) の尤度関数が

$$L(\boldsymbol{X}|\boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \prod_{n=1}^{N}\sum_{k=1}^{K} p_k VM_k(x_n|\boldsymbol{\Theta_k}) \tag{7}$$

と表されるため, 対数をとり

$$lnL(\boldsymbol{X}|\boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \sum_{n=1}^{N} ln\{\sum_{k=1}^{K} p_k VM_k(x_n|\boldsymbol{\Theta_k})\} \tag{8}$$

になる. 対数の内部に k についての和があるため, パラメータについての最尤解は閉形式の解析解では得られない. そのため本稿では尤度関数を最大にするアプローチとして,EM アルゴリズムを用いて推定する [2].

1.2.2 EM アルゴリズム

このアルゴリズムは E-step と M-step と呼ばれる 2 つの step を収束するまで繰り返すことで最大化を行う。初めに p が隠れ変数 z によってカテゴリー分布で表されているとして、z の事後分布 γ (z) を求める。(E-step) その後、対数尤度 L を最大化するそれぞれのパラメータを求め、対数尤度がどのくらい更新されるか計算する。(M-step)

Algorithm 1 von-mises 分布における EM アルゴリズム

入力：

X={$x_1,\cdots,x_n$}(角度データ)

出力：

p_k, $\boldsymbol{\mu}_k$, κ_k, k =1,⋯K (K 個のクラスター毎の混合確率密度関数のパラメータ)

repeat:

{The E (Expectation) step of EM}

for i = 1 to n **do**

for k = 1 to K **do**

$f_k(x_i|,\Theta_k) \leftarrow \exp\left(\kappa_k \boldsymbol{\mu}'_k \boldsymbol{x}_i\right)/I_0(\kappa)$

End for

for k = 1 to k do

$\gamma\,(\mathrm{k}|x_i,\theta) \leftarrow \pi_k f_k(x_i|,\Theta_i)\Big/\sum_{i=1}^{K}\pi_i f_i(x_i|,\Theta_i)$

end for

end for

{The M (Maximization) step of EM}

for k = 1 to K **do**

$\pi_k \leftarrow \frac{1}{n}\sum_{i=1}^{n}\gamma\,(\mathrm{k}|x_i,\theta)$

$\mu_k \leftarrow \sum_{i=1}^{n} x_i\,\gamma\,(\mathrm{k}|x_i,\theta)$

$\bar{r} \leftarrow ||\mu_k||/(\mathrm{n}\pi_k)$

$\mu_k \leftarrow \mu_k/||\mu_k||$

$\kappa_k \leftarrow \frac{\bar{r}d-\bar{r}^3}{1-\bar{r}^2}$

End for

Until convergence

図 1: 混合 Von-mises 分布における EM アルゴリズム
論文:[8] 引用

2 風速データ

風速データを扱うモデルについては正規分布やワイブル分布 [4] 等が挙げられるが, 本稿では 0 を多く含むデータにおいて 0 を値として持たないワイブル分布を当てはめるかについて述べる. ここでは,0 の値を 1 点退化分布とし, それ以外の範囲を通常のワイブル分布と同様とした”Zero-Infrated-Weibull distribution”観測できるデータの最小値以下を 0 として扱っていると仮定している”Left-censored-Weibull distribution”の 2 つについて扱う.

2.1 Weibull 分布

ワイブル分布は, 右方向に歪んだデータ, 左方向に歪んだデータ, 対称のデータをモデル化できる. したがって, 真空管, コンデンサ, ボールベアリング, リレー, 材料強度など幅広い応用分野の信頼性を評価するために使用される. また, 減少, 増加, または安定するハザード関数をモデル化することも可能で, アイテムの寿命のあらゆる段階を表せる.[5]

その確率密度関数 $p(X;m,\eta)$ は

$$p(X;m,\eta) = \frac{m}{\eta} * (\frac{x}{\eta} * \exp\{-(\frac{x}{\eta})^m\}) \tag{9}$$

累積分布関数は

$$P(X;m,\eta) = 1 - exp\{-(\frac{x}{\eta})^m\} \tag{10}$$

と表される. ただしパラメータ m はワイブル係数と呼ばれる分布の性質を決定する形状パラメータであり,η は尺度パラメータである. この分布では $x = 0$ の時値をとらない特徴がある. そのため, 本稿では $x = 0$ の時を扱うため, 以下の分布を考えることとした.

2.2 Zero-Infrated-Weibull 分布

0 の値を 1 点退化分布とし, それ以外の範囲を通常のワイブル分布と同様としたもので, その確率密度関数は

$$ZIweibull(X;m,\eta) = \begin{cases} \frac{x_{|x=0}}{N} & x_i = 0 \\ \frac{m}{\eta} * (\frac{x}{\eta})^{m-1} * \exp\{-(\frac{t}{\eta})^m\} & otherwise \end{cases} \tag{11}$$

となり, その尤度はデータから $x = 0$ の場合のデータを除いた際のワイブル分布と同様である.

2.3 Left-Censored-Weibull 分布

データの観測において観測できるデータの最小値以下を 0 として扱っていると仮定しているモデルで, その確率密度関数は

$$LCweibull(X;m,\eta) = Weibull(x_i;m,\eta)^{\delta} + Weibull(c;m,\eta)^{1-\delta_i} \tag{12}$$

ただし式中の δ_i は

$$\delta_i = \begin{cases} 1 & x_i > c \\ 0 & otherwise \end{cases} \tag{13}$$

である. 尤度関数 L は

$$L(m,\eta,c) = \{F(c)\}^z \prod_{i=1}^{n} f(x_i;m,\eta) \tag{14}$$

ただし,z は c 以下の観測データの数である.

3 風速風向を同時に表すモデル

ここでは風向データ風速データを同時に表現することができるモデルを扱う.

3.1 Johnson-Wehrly モデル

Johnson-Wehrly モデルは Johnson-Wehrly が表した [6] 角度モデルと直線上モデルを同時に表すことのできるモデルで, その確率密度関数 $f(\theta,x)$ は

$$f(\theta,x) = 2\pi g[2\pi(F_1(\theta) - F_2(x))]f_1(\theta)f_2(x) \tag{15}$$

で表される. ただし式中の $g(\cdot)$ は円型分布 $F_1(\theta)$ は $f_1(\theta)$ の累積分布関数 $F_2(x)$ は $f_2(x)$ の累積分布関数である.

3.1.1 パラメータ推定

Johnson-Wehrly モデルのパラメータ推定アルゴリズムは次のように表される.[7]

Step.1. 密度関数 $\hat{f}_1(\theta)$,$\hat{f}_2(x)$ に対応する累積分布関数 $\hat{F}_1(\theta)$, $\hat{F}_2(x)$ の値を求める.

Step.2. 密度関数Step.1で求めた値から $\{2\pi(\hat{F}_1(\theta) - \hat{F}_2(x))\}_{i=1}^{n}$ を計算し円型分布 $g(\cdot)$ の値を求める.

Step.3. 確率密度関数 $f(\theta,x)$ を

$$f(\theta,x) = 2\pi\hat{g}[2\pi(\hat{F}_1(\theta) - \hat{F}_2(x))]\hat{f}_1(\theta)\hat{f}_2(x) \tag{16}$$

として表す.

4　数値実験

4.1　混合 Von-mises 分布

4.1.1　実データによる実験

用いるデータは気象庁より公開されている各地点の１時間ごとの風向を 16 方位にて表したデータである. 本稿では地点を東京, 期間は 2007 年を対象とした.

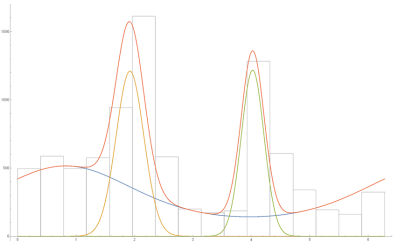

図 2: 2007 年東京風向データのヒストグラムと混合分布

図 2 は 2007 年東京の 1 時間ごとの風向きデータのヒストグラムとその確率密度表したものである. 先ほどのヒストグラムと同様横軸が角度, 縦軸が頻度となっている, 赤色のラインが混合された分布, オレンジ, 青, 緑が混合前の単峰 von-mises 分布を表している. この図からヒストグラムと密度関数の峰の位置, すなわちモードパラメータ μ と峰の形すなわち形状パラメータ κ は良い値をとれていることが確認できた. また,1 年間のデータを四季毎に分け解析を行うこととした. 春と秋の峰は 2 つで位置は 1 年間で行ったそれとほとんど変わらない. 夏と春はそれぞれ, 南からの風, 北からの風が多く吹いていることがわかる.

4.2　ワイブル分布

2020 年 9 月 3 日から 9 月 9 日の 1 分ごとの風速データ [m/s] に対し 3 つのワイブル分布に対しあてはめ, 解析を行った. データ数は 8000, 値 0 をとるデータは 139 存在した.

4.2.1　Zero-Infrated-Weibull 分布

式 (11) から

$$m = 1.58568482234895, \eta = 1.40220331686096 \tag{17}$$

と推定値が得られ, その対数尤度は

$$L = -8481.611 \tag{18}$$

であった図 3.

4.2.2　Left-Censored-Weibull 分布

観測されたデータの最小値である 0.04[m/s] を閾値 c として, 式 (LCW) から

$$m = 1.366955, \eta = 1.509025 \tag{19}$$

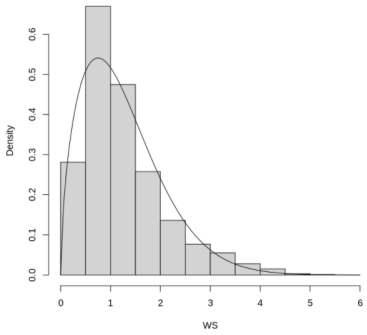

図 3: Zero-Infrated-Weibull

と推定値が得られ, その対数尤度は

$$L = -8736.246740 \tag{20}$$

であった 2 つのモデルを AIC(赤池情報量基準) で比較すると [9]
Zero-Infrated-Weibull は $AIC = 16989.222$、、Left-Censored-Weibull は $AIC = 17498.49348$ であったため,Zero-Infrated-Weibull の方が当てはまりの良いモデルであるといえる.

4.3　Johnson-Wehrly モデル

Johnson-Wehrly モデルの風データへの当てはめを行う. 先述のワイブル分布の際に使用したデータを使い解析を行った. 円型分布を Von-mises 分布, 直線上分布を正規分布, ワイブル分布として 2 つの場合の比較を行う.

4.3.1　直線上分布を正規分布とした場合

直線上分布を正規分布とした場合モデルの密度関数は $f_1(\theta)$ を $\frac{1}{2\pi}$,
$f_2(x)$ を正規分布 $\phi(x;\mu_2,\sigma)$,$g(\cdot)$ を Von-mises 分布 $(\theta;\mu_1,\kappa)$ とするので,
$F_1(\theta) = \frac{\theta}{2\pi}$,$F_2(x) = \Phi(x;\mu_2,\sigma)$ とすると

$$p_1(\theta,x) = \frac{1}{2\pi I_0(\kappa)} * \exp\{\kappa\cos(\theta - 2\pi\Phi(x;\mu_2,\sigma) - \mu_1)\}\phi(x;\mu_2,\sigma) \tag{21}$$

となる.
ここで先述のアルゴリズムを適用すると,

$$\begin{cases} \mu_1 = 2.582 \\ \kappa = 0.553 \\ \mu_2 = 1.253 \\ \sigma = 0.8352 \end{cases} \tag{22}$$

という結果が得られた

4.3.2　直線上分布をワイブル分布とした場合

直線上分布をワイブル分布とした場合モデルの密度関数は $f_1(\theta)$ を $\frac{1}{2\pi}$,
$f_2(x)$ をワイブル分布 $\zeta(x;m,\eta)$,$g(\cdot)$ を Von-mises 分布 $(\theta;\mu,\kappa)$ とするので,

$$F_1(\theta) = \frac{\theta}{2\pi}, F_2(x) = Z(x; m, \eta) \tag{23}$$

とすると

$$p_2(\theta, x) = \frac{1}{2\pi I_0(\kappa)} * \exp\{\kappa \cos(\theta - 2\pi Z(x; m, \eta) - \mu)\} \tag{24}$$

となる.
ここで先述のアルゴリズムを適用すると,

$$\begin{cases} \mu = 0.4542 \\ \kappa = 1.957 \\ m = 1.5856 \\ \eta = 1.4022 \end{cases} \tag{25}$$

という結果が得られた. ここで 2 つを俯瞰して見るとそれぞれの峰の位置から角度は 0 を東に反時計回りに 2π となっているため, 東から北になるにつれて, 風力の強い風が吹きやすくなっていることが確認できた. これはデータの観測点ボストンが東に海が見える港町であるため海からの東風が多いこと, データの観測時期が 10 月という現地での秋にあたることから, 北風が強くなったと考えられる.

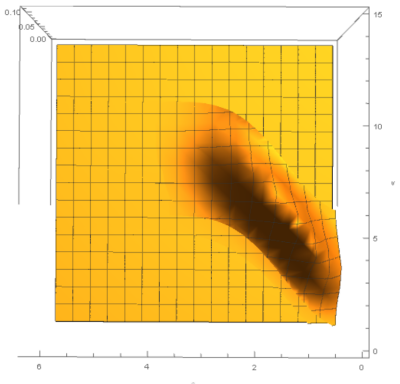

図 4: Weibull-vonmises 分布俯瞰図

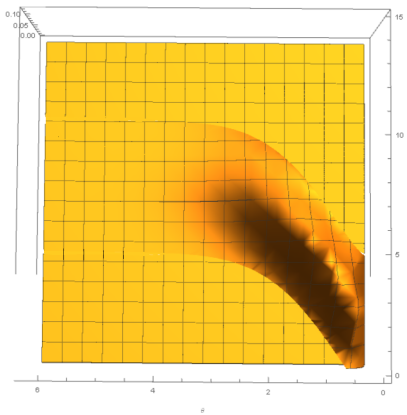

図 5: Weibull-vonmises 分布俯瞰図

また正規分布とワイブル分布を比較した場合, ワイブル分布の方が峰の並び方の形としてなだらかに湾曲しており, 線形的に峰が並んでいる正規分布の場合に比べてより柔軟に風力の特徴を表せていることがわかる.

結論

本稿では, 観測風データについて, 風向風速 2 つのデータからデータの解析, モデルの検討を行った. 風向データにおいては,2 つ以上の峰を持つ風向データにおいて単峰である Von-mises 分布の混合モデルを用いることで 2 つ以上の峰を表現することが出来た, 風速データにおいては, は 0 を多く含むデータにおいて 0 を値として持たないワイブル分布を当てはめるかについて,2 つのモデルを考え比較した. またその 2 つを同時に表すモデルとして, 直線上分布と英分布 2 つを扱う Johnson-Werhly 分布のワイブル分布と正規分布の 2 つの場合を比較し, 観測値における風の特徴を確認することができた.

今後の課題

今回使用したデータは風向風速データ以外にも雨量や気温なども含まれており, それらが風向風速と相関している新たなモデルの作成を目標としていたのだが, 本稿では風向風速の 2 つでしか, 本稿ではできなかったため, 他のデータに関しても行いたいと考える. また,Johnson-wehrly モデルにおいて, 本稿では直線上分布をワイブル分布と正規分布で行ったが, 他の一般的な分布や, 本稿で取り上げた 0 を多く含むデータについて顧慮されたワイブル分布についての比較も行いたいと考える.

参考文献

[1] 清水邦夫 (2018) 『角度データのモデリング』 近代科学社

[2] C. M. Bishop (2011) 『Pattern Recognition and Machine Learning』 Springer 社

[3] Arindam Banerjee, et al. ” Clustering on the Unit Hypersphere using von Mises-Fisher Distributions ” Journal of Machine Learning Research 6 (2005) 1345–1382

[4] J.V. Seguro, T.W. Lambert. ” Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis ” Journal of Wind Engineering and Industrial Aerodynamics 85 (2000) 75 84

[5] 信頼性分析におけるワイブル分布
”https://support.minitab.com/ja-jp/minitab/18/help-and-how-to/modeling-statistics/reliability/supporting-topics/distribution-models/weibull-distribution/”
Minitab 18 サポート

[6] Richard A. Johnson and Thomas E. Wehrly ”Some Angular-Linear Distributions and Related Regression Models” Journal of the American Statistical Association, 73, 602-606.

[7] Eduardo Garc´ıa–Portugu´es1 ”An algorithm for estimating circular–linear densities” X Congreso Galego de Estat´ıstica e Investigaci´on de Operaci´ons Pontevedra, 3–4–5 de novembro de 2011

[8] Arindam Banerjee ”Clustering on the Unit Hypersphere using von Mises-Fisher Distributions” Journal of Machine Learning Research 6 (2005) 1345–1382

[9] Akaike, H., ”Information theory and an extension of the maximum likelihood principle”, Proceedings of the 2nd International Symposium on Information Theory, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest: 267-281 (1973).