ISSN　0385-2148

研究所報

No.54

統計的モデリング

２０２１年１１月

法政大学
日　本　統　計　研　究　所

ISSN　0385-2148

研究所報

No.54

統計的モデリング

２０２１年１１月

法政大学
日 本 統 計 研 究 所

はじめに

第 1 回目は統計的モデリングについての研究集会を行った。統計的モデリング、と言っても関連する分野は幅広くあるので、本研究集会では特に角度統計やコピュラを用いたモデリングに関連する研究者らに最近の研究内容の紹介をしていただくことにした。

最初に、小方浩明先生（東京都立大学）に直線上の多次元高次マルコフデータに対するヴァインコピュラモデリングを紹介していただき、それの円周上のデータに対する方法についての提案やその課題について講演していただいた。次に、シリンダーモデリングの分野で世界的に活躍している気鋭の研究者である井本智明先生（静岡県立大学）にシリンダーモデルやトーラス上のモデリングについて最新の統計モデルの提案、サッカーなどのスポーツのデータに対するモデルの適用例を講演していただいた。塩濱敬之先生（南山大学）には高次の角度マルコフ過程の紹介や mixture transition distribution modeling について講演していただいた。宮田庸一先生（高崎経済大学）には非凸な罰則項によるスパース推定、モデルの拡張型確率的コンプレキシティに対するラプラス近似に関する研究について講演していただいた。最後に江村剛志先生（久留米大学）にはコピュラとフレイルティ混合効果から導かれる多変量故障時間分布について講演していただいた。

本研究集会では講演中にも活発にディスカッションがされ、また、発表後も研究情報の意見交換ができ、大変有意義な研究集会となった。

2021 年 11月　日本統計研究所

目次

講演
円周上の高次マルコフ過程におけるペアサーキュラモデリング
1
小方浩明

講演
シリンダー分布構成法の提案について
6
井本智明

講演
Mixture transition distribution modeling for higher order circular Markov processes
8
塩濱敬之

講演
非凸な罰則項によるスパース推定における拡張型確率的コンプレキシティについて
10
宮田庸一

講演
コピュラと混合効果から生成される多変量故障時間分布
19
江村剛志

円周上の高次マルコフ過程におけるペアサーキュラモデリング*

東京都立大学　小方浩明

1　はじめに

方向統計学は円周上の点によって表される周期的なデータを扱う統計学の一分野である．風向などの角度を表すデータや，病院に運び込まれる患者の到着時刻などの，24 時間表記で表される時刻のデータなどが代表的な例である．方向統計学に関するテキストとしては，例えば Jammalamadaka and SenGupta (2001) や Mardia and Jupp (1999) 等が挙げられる．

先に挙げた風向データなどは，ある一定期間に集められた時系列データであることが多く，ゆえに系列相関を持つと考えるのが自然である．本報告では，円周上の多次元時系列データのモデリングを考察する．まずは同時分布を，各周辺分布と，コピュラの円周上版に相当する多次元サーキュラスの積に分解して表現する．多次元のサーキュラスを更にペアサーキュラスの積に分解することにより，各確率変数の従属性が解釈しやすい格好となる．これは直線上の出時系列データに対する，ヴァインコピュラによるモデリングに類似している．実際には，時系列データに強定常性と高次のマルコフ性を課すことにより，モデルの大きさを制限し，現実的に推定可能なレベルまで落とし込む．

本報告の構成は以下である．2 章で，Smith (2015) による直線上の多次元高次マルコフデータに対するヴァインコピュラモデリングを紹介し，3 章で円周上のデータに対する方法を提案する．4 章で簡単なシミュレーション結果を提示し，5 章にまとめと課題を述べる．

2　ヴァインコピュラによる多次元マルチオーダーマルコフ過程

本章では，Smith (2015) による，多次元時系列データにおける従属性をヴァインコピュラによって表現するモデルを紹介する．以下のような $\mathbb{R}$ 上の m 次元時系列データ $\boldsymbol{Y}_t = (Y_{1,t}, \ldots, Y_{m,t})'$, $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_T')'$ を考える．$\boldsymbol{Y}$ は $N = Tm$ 次元の確率ベクトルであり, 添え字の簡略化のため以降では $\boldsymbol{Y} = (Y_1, \ldots, Y_N)'$ というような表現も用いる．つまり，添え字がカンマで分けられていれば要素と時刻のインデックスを明示した形であり，カンマがなければ全ての要素を一列に並べてまとめて番号を振っていると理解されたい．このとき，$\boldsymbol{Y}$ の同時密度関数は以下のように与えられる．

$$f(\boldsymbol{y}) = c(\boldsymbol{u}) \prod_{t=1}^{T} \prod_{j=1}^{m} f(y_{j,t}).$$

ここに，$\boldsymbol{u}' = (u_1, \ldots, u_N) = (F(x_1), \ldots, F(x_N))$, $f(y_{j,t})$ は $y_{j,t}$ の周辺密度関数，$F(y_i) = F(y_{j,t})$ は対応する分布関数，そして $c(\boldsymbol{u})$ は $\boldsymbol{U} = (U_1, \ldots, U_N)' = (F(Y_1), \ldots, F(Y_N))'$ の同時密度関数であり，一般にコピュラ密度関数と呼ばれるものである．コピュラ密度関数は $[0,1]^N$ 上の関数であるが，以下のように $N(N-1)/2$ 個のペアコピュラ（D-ヴァインコピュラと呼ばれる）に分解することを考える．

$$c(\boldsymbol{u}) = \prod_{i=2}^{N} \prod_{j=1}^{i-1} c_{i,j}(u_{i|j+1}, u_{j|i-1}). \tag{1}$$

*本研究は JSPS 科研費 18K11193 の助成を受けたものです。

ここに，$u_{i|i} \equiv u_i$ であり，$i > j$ に対しては $u_{i|j+1} \equiv F(u_i|u_{i-1}, \ldots, u_{j+1})$, $u_{j|i-1} \equiv F(u_j|u_{i-1}, \ldots, u_{j+1})$ である．$c(\boldsymbol{u})$ に比べ $c_{i,j}(u_{i|j+1}, u_{j|i-1})$ はペアごとであるので，相関構造がより解釈しやすくなる．

また，ペアコピュラをブロックごとにグルーピングすることによって，要素間での相関と系列相関を分けて表現することができる．$\boldsymbol{u}_t = (u_{1,t}, \ldots, u_{m,t})'$ とし，異なる時刻間 $t_1 > t_2$ に対して

$$K_{t_1,t_2}(\boldsymbol{u}_{t_2}, \ldots, \boldsymbol{u}_{t_1}) = \prod_{i=a(t_1)}^{b(t_1)} \prod_{j=a(t_2)}^{b(t_2)} c_{i,j}(u_{i|j+1}, u_{j|i-1})$$

と定義し，同時刻 t に対しては

$$K_{t,t}(\boldsymbol{u}_t) = \prod_{i=a(t)+1}^{b(t)} \prod_{j=a(t)}^{i-1} c_{i,j}(u_{i|j+1}, u_{j|i-1})$$

と定義する．ここに $a(t) = (t-1)m + 1$, $b(t) = tm$ である．すると

$$c(\boldsymbol{u}) = c(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T) = \left\{ \prod_{t=1}^{T} K_{t,t}(\boldsymbol{u}_t) \right\} \left\{ \prod_{t=2}^{T} \prod_{i=1}^{t-1} K_{t,t-i}(\boldsymbol{u}_{t-i}, \ldots, \boldsymbol{u}_t) \right\} \tag{2}$$

と表現できる．一番目の中括弧内が同時刻における要素間の従属性を表現する部分であり，二番目の中括弧内が異なる時刻間での従属性，すなわち系列相関を表現する部分である．

ここで時系列に強定常過程の構造を入れると，(2) 式中の関数 $K_{t,t}$ 並びに $K_{t,t-i}$ は t に依存せず，加えて p 次マルコフ性（過去全てのデータを所与としたときの現時点での条件付分布が，過去 p 時点までを所与としたときの現時点での条件付分布に等しい）の構造を入れると，$i > p$ において $K_{k,k-i} = 1$ となるため，(2) 式は

$$c(\boldsymbol{u}) = c(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T) = \left\{ \prod_{t=1}^{T} K_0(\boldsymbol{u}_t) \right\} \left\{ \prod_{t=2}^{T} \prod_{i=1}^{\min(t-1,p)} K_i(\boldsymbol{u}_{t-i}, \ldots, \boldsymbol{u}_t) \right\}$$

と書き換えられ，モデリングする際に使用するペアコピュラ関数の数が抑えられる．

3　円周上のマルチオーダーマルコフ過程

ヴァインコピュラの概念を用いて，円周上のマルチオーダーマルコフ過程を規定することを考える．円周上の分布におけるコピュラ密度関数に相当するものついては，Wehrly and Johnson (1980) などさまざまな論文で提示されており，Jones et al. (2015) がサーキュラス (circulas) という形で統一的にまとめている．

円周上 (Π で表すこととする) の m 次元時系列データを考え，2 章に倣い $\boldsymbol{\Theta}_t = (\Theta_{1,t}, \ldots, \Theta_{m,t})'$, $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1', \ldots, \boldsymbol{\Theta}_T')' = (\Theta_1, \ldots, \Theta_N)'$ のように表記する．つまり $\boldsymbol{\Theta}$ は Π^N 上での確率ベクトルとなる．Π 上の各周辺密度関数，周辺分布関数をそれぞれ $f_i(\theta_i)$, $F_i(\theta_i)$ $(i = 1, \ldots, N)$ で表す．このとき，$\boldsymbol{\Theta}$ の同時密度関数は以下のように書ける．

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = (2\pi)^N c(2\pi F_1(\theta_1), \ldots, 2\pi F_N(\theta_N)) \prod_{i=1}^{N} f_i(\theta_i). \tag{3}$$

ここに $c(\theta_1, \ldots, \theta_N)$ は，サーキュラスとよばれる，Π^N 上で定義された同時密度関数であり，次の二つの性質を持つ．

(i) 各周辺分布がΠ上での一様分布，すなわち各周辺密度関数が$c_i(\theta_i) = (2\pi)^{-1} \quad (i = 1, \ldots, N)$となる．

(ii) 各変数において周期性

$$c(\theta_1 + 2k_1\pi, \ldots, \theta_N + 2k_N\pi) = c(\theta_1, \ldots, \theta_N),\ (\theta_1, \ldots, \theta_N) \in \Pi^N,\ k_1, \ldots, k_N = 0, \pm 1, \pm 2, \ldots$$

を満たす．

(ii) の周期性が要請されていることにより，サーキュラスは，通常のコピュラを単にリスケールしたものとは異なる．

(1) に倣い，N 次元サーキュラスをペアサーキュラスに分解すると

$$c(\theta_1, \ldots, \theta_N) = (2\pi)^{-N} \prod_{i=2}^{N} \prod_{j=1}^{i-1} (2\pi)^2 c_{i,j}(2\pi\theta_{i|j+1}, 2\pi\theta_{j|i-1}) \tag{4}$$

となる．ここに $\theta_{i|i} \equiv (2\pi)^{-1}\theta_i$ であり，$i > j$ に対しては $\theta_{i|j+1} \equiv F(\theta_i|\theta_{i-1}, \ldots, \theta_{j+1})$, $\theta_{j|i-1} \equiv F(\theta_j|\theta_{i-1}, \ldots, u_{j+1})$ である．また $c_{i,j}(\theta_i, \theta_j)$ は Π^2 上のサーキュラスであり，具体的に

$$c_{i,j}(\theta_i, \theta_j) = (2\pi)^{-1} g_{i,j}(\theta_j - q_{i,j}\theta_i) \tag{5}$$

とすることによりサーキュラスの条件を満足する．ここに $g_{i,j}$ は Π 上の任意の密度関数であり，$q_{i,j} \in \{-1, 1\}$ は，θ_i の符号を決める非確率的な要素である．全ての i, j において $g_{i,j}(\theta) = (2\pi)^{-1}$($\Pi$ 上の一様分布) と設定すると，(3)–(5) より

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \prod_{i=1}^{N} f_i(\theta_i)$$

となり，独立モデルとなる．また Jones et al. (2015) の 2.3 章において，$g_{i,j}$ の集中度が高ければ Θ_i と Θ_j の相関が高くなることが言及されており，$g_{i,j}$ の集中度パラメータをそのまま Θ_i と Θ_j の相関度ととらえることができる．

また，ブロッキングについても 2 章と同様に，異なる時刻間 $t_1 > t_2$ に対して

$$K_{t_1,t_2}(\boldsymbol{\theta}_{t_2}, \ldots, \boldsymbol{\theta}_{t_1}) = \prod_{i=a(t_1)}^{b(t_1)} \prod_{j=a(t_2)}^{b(t_2)} (2\pi)^2 c_{i,j}(2\pi\theta_{i|j+1}, 2\pi\theta_{j|i-1}),$$

また同時刻 t に対しては

$$K_{t,t}(\boldsymbol{\theta}_t) = \prod_{i=a(t)+1}^{b(t)} \prod_{j=a(t)}^{i-1} (2\pi)^2 c_{i,j}(2\pi\theta_{i|j+1}, 2\pi\theta_{j|i-1})$$

と定義し，

$$c(\boldsymbol{\theta}) = c(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T) = (2\pi)^{-N} \left\{ \prod_{t=1}^{T} K_{t,t}(\boldsymbol{\theta}_t) \right\} \left\{ \prod_{t=2}^{T} \prod_{i=1}^{t-1} K_{t,t-i}(\boldsymbol{\theta}_{t-i}, \ldots, \boldsymbol{\theta}_t) \right\}$$

と表現できる．また，強定常性と p 次マルコフ性を仮定したときに，2 章と同様に

$$c(\boldsymbol{\theta}) = c(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T) = (2\pi)^{-N} \left\{ \prod_{t=1}^{T} K_0(\boldsymbol{\theta}_t) \right\} \left\{ \prod_{t=2}^{T} \prod_{i=1}^{\min(t-1,p)} K_i(\boldsymbol{\theta}_{t-i}, \ldots, \boldsymbol{\theta}_t) \right\}$$

と表現できる．

4　シミュレーション

シミュレーションによって，$m = 2$ の角度時系列データを $T = 50$ 個 $((\theta_{1,1}, \theta_{2,1})', \ldots, (\theta_{1,50}, \theta_{2,50})')$ 発生させる．要素間においても時点間においても独立かつ同一の分布に従っており，各周辺分布は位置パラメータ $\mu = 0$, 集中度パラメータ $\kappa = 3$ のフォン・ミーゼス分布（$\mathrm{vM}(\mu = 0, \kappa = 3)$ と表記する）という設定である．フォン・ミーゼス分布の確率密度関数は以下で与えられる．

$$f_{\mathrm{vM}}(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\}.$$

ここに，$I_r(\kappa)$ は第 1 種の r 次修正ベッセル関数であり，

$$I_r(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos(r\theta) \exp(\kappa \cos\theta)\, d\theta, \quad r = 0, \pm 1, \ldots$$

で定義される．集中度パラメータが $\kappa = 0$ のとき，円周上の一様分布に一致する．

上記のようにして発生させられたシミュレーションデータに対して，3 章で考察したモデルをあてはめる．具体的には，まず $p = 1$ 次のマルコフ強定常過程をあてはめ，各周辺分布に $f_i \sim \mathrm{vM}(\mu = 0, \kappa_i)$ $(i = 1, 2)$ をあてはめる．また，ペアサーキュラスを (5) 式で構築し，同時刻における部分においては $g_{2,1}^{(0)} \sim \mathrm{vM}(\mu = 0, \kappa_{2,1}^{(0)})$，ラグ 1 の時刻間における部分においては $g_{i,j}^{(1)} \sim \mathrm{vM}(\mu = 0, \kappa_{i,j}^{(1)})$ $(i, j = 1, 2)$ をあてはめる．(5) 式内の符号を決める非確率的な部分については，すべて $q = 1$ と設定する．データ生成の仕方により，真値は $\kappa_1 = \kappa_2 = 3$, $\kappa_{2,1}^{(0)} = \kappa_{1,1}^{(1)} = \kappa_{1,2}^{(1)} = \kappa_{2,1}^{(1)} = \kappa_{2,2}^{(1)} = 0$ となる．

実際のパラメータ推定は Rstan によるマルコフ連鎖モンテカルロ法 (MCMC) で行う．MCMC の設定は chain = 2, iter = 1500, warmup = 150, iter = 1 であり，全てのパラメータにおいて無情報事前分布とする．各パラメータにおける事後分布の要約を表 1 に示す．おおむね良好な推定結果が得られた．

表 1: MCMC による各パラメーターの事後分布要約

	平均	標準偏差	2.5% 点	50% 点	97.5% 点
κ_1	3.28	0.58	2.26	3.26	4.48
κ_2	3.25	0.58	2.24	3.21	4.46
$\kappa_{21}^{(0)}$	0.28	0.18	0.02	0.25	0.70
$\kappa_{11}^{(1)}$	0.16	0.12	0.01	0.14	0.44
$\kappa_{12}^{(1)}$	0.16	0.12	0.01	0.13	0.46
$\kappa_{21}^{(1)}$	0.22	0.15	0.01	0.19	0.57
$\kappa_{22}^{(1)}$	0.15	0.12	0.01	0.13	0.45

5　まとめと課題

本研究では，円周上のマルチオーダーマルコフ過程に対してサーキュラスを用いたモデリングを提案した．サーキュラスは円周上版のコピュラに対応する概念である．多次元のサーキュラスはペアサーキュラスに分解することにより，確率変数の従属性がより解釈しやすいモデリングとなる．ペアサーキュラスは任意の円周上確率密度関数によって構築でき，その確率密度の集中度がそのまま従属性を表す指標となる．

また，シミュレーションによって発生させられた独立同一分布に従う 2 次元の角度時系列データに対して，円周上の 1 次マルコフ強定常過程をあてはめ，MCMC によってパラメーターの事後分布を見た．おおむね良好な推定結果が得られた．

現時点ではごく簡単なシミュレーションしか実行できておらず，今後は実データなどに対してより高次の円周上マルコフ過程をあてはめ，パラメーターを推定する予定である．そのためには計算負荷の軽減が必須である．具体的には，(4) 式の引数である条件付分布関数の値を求める部分に計算負荷がかかると思われ，その部分の計算の工夫が必要となる．

参考文献

Jammalamadaka, S. A., SenGupta, A. S. (2001). *Topics in circular statistics*. World Scientific Publishing Co., New York.

Jones, M.C., Pewsey, A., Kato, S. (2015). On a class of circulas: copulas for circular distributions. *Annals of the Institute of Statistical Mathematics*, 67, 843–862.

Mardia, K. V., Jupp, P. E. (1999). *Directional statistics*. Wiley, Chichester.

Smith, M. S. (2015). Copula modelling of dependence in multivariate time series. *International Journal of Forecasting*, 31, 815–833.

Wehrly, T. E., Johnson, R. A. (1980). Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, 67, 255–256.

シリンダー分布構成法の提案について

井本智明
静岡県立大学　経営情報学部

ある地点での風向や樹木の生長方向のような角度に関するデータは様々な社会科学分野で現れる。こうしたデータを適切に分析するためには，正規分布やポアソン分布のように直線上で定義される確率分布ではなく，角度の持つ周期性を考慮に入れた円周上の確率分布を用いたモデリングが必要となる。また，ある地点の温度と風向の組合せからなるデータや樹木の配置地点と生長方向の組合せからなるデータのように，直線上変量と円周上変量の組合せからなるデータを分析する際にはシリンダー上分布が必要となる。

本報告では，平均 0，分散 1 の実数値全体を台とする直線上対称確率密度関数 $g_L(\cdot)$ と平均方向 0 の $(-\pi, \pi]$ を台とする円周上確率密度関数 $g_C(\cdot)$ から構成される確率密度関数

$$\begin{aligned} f(x,\theta) &= \left[1+\lambda\sin\left\{2\mathrm{Arctan}\left(\frac{x-\mu_L}{\tau}\right)\right\}\sin(\theta-\mu_C)\right]\frac{1}{\tau}g_L\left(\frac{x-\mu_L}{\tau}\right)g_C(\theta-\mu_C), \\ &= \left[1+\frac{2\lambda\left(\frac{x-\mu_L}{\tau}\right)\sin(\theta-\mu_C)}{1+\left(\frac{x-\mu_L}{\tau}\right)^2}\right]\frac{1}{\tau}g_L\left(\frac{x-\mu_L}{\tau}\right)g_C(\theta-\mu_C), \qquad (1) \\ &\qquad\qquad -\infty < x < \infty,\ -\pi < \theta \le \pi \end{aligned}$$

を持つシリンダー上確率変数 (X, Θ) について考える。この関数に含まれる $-1 \le \lambda \le 1$ は直線上変量 X と円周上変量 Θ の相関を操るパラメータ，$-\infty < \mu_L < \infty$ と $\tau > 0$ はそれぞれ X の平均と標準偏差を表すパラメータ，$-\pi < \mu_C \le \pi$ は Θ の平均方向を表すパラメータである。$G_L(\cdot)$ と $G_C(\cdot)$ をそれぞれ $g_L(\cdot)$ と $g_C(\cdot)$ の分布関数とすると，分布 (1) は接合関数を

$$c(x,\theta) = 1 + \frac{2\lambda G_L^{-1}(x)\sin G_C^{-1}(\theta)}{1+\{G_L^{-1}(x)\}^2}$$

としたときのコピュラ分布としてみることもできる。

この分布の X と Θ の周辺確率密度関数 $f_X(x)$ と $f_\Theta(\theta)$ は

$$f_X(x) = \frac{1}{\tau}g_L\left(\frac{x-\mu_L}{\tau}\right), \qquad f_\Theta(\theta) = g_C(\theta-\mu_C),$$

であることがわかる。つまり，直線上変量が実数値全体で定義されており，その分布が対称であることを仮定するならば，提案法 (1) を用いることで周辺分布を任意に設定してシリンダー上分布の構成が行える。しかも，この方法によって構成された分布には複雑な関数が追加的に含まれることがないので統計推測のための計算コストを抑えることができる。また，Θ を所与としたときの X の条件付き確率密度関数 $f_{X|\Theta}(x \mid \theta)$ や X を所与としたときの Θ の条件付き確率密度関数 $f_{\Theta|X}(\theta \mid x)$ も容易に

$$\begin{aligned} f_{X|\Theta}(x \mid \theta) &= \left[1+\frac{2\lambda\left(\frac{x-\mu_L}{\tau}\right)\sin(\theta-\mu_C)}{1+\left(\frac{x-\mu_L}{\tau}\right)^2}\right]\frac{1}{\tau}g_L\left(\frac{x-\mu_L}{\tau}\right), \\ f_{\Theta|X}(\theta \mid x) &= \left[1+\frac{2\lambda\left(\frac{x-\mu_L}{\tau}\right)\sin(\theta-\mu_C)}{1+\left(\frac{x-\mu_L}{\tau}\right)^2}\right]g_C(\theta-\mu_C) \end{aligned}$$

となることがわかる。特に後者は，[1] による sine-skewed 分布族に属する分布となっていることから回帰分析を行う際にも役立つ性質を有している。

確率変数 X と Θ についてのモーメントを，$i=\sqrt{-1}$ を虚数単位として，

$$\mathrm{E}\left[\frac{X-\mu_L}{\tau}\right]=\mu_p, \qquad \mathrm{E}[e^{ip(\Theta-\mu_C)}]=\alpha_p+i\beta_p$$

とおくと，分布 (1) の同時モーメントは

$$\mathrm{E}\left[\frac{X-\mu_L}{\tau}e^{iq(\Theta-\mu_C)}\right]=\begin{cases}\mu_p(\alpha_q+i\beta_q), & p\text{が偶数のとき},\\ \lambda A_p\{(\beta_{q+1}-\beta_{q-1})-i(\alpha_{q+1}-i\alpha_{q-1})\}, & p\text{が奇数のとき},\end{cases}$$

と表されることがわかる。ここで，$A_p=\displaystyle\int_{-\infty}^{\infty}\frac{x^p}{1+x^2}f_X(x)dx$ である。これより，X と $(\cos\Theta,\sin\Theta)$ の正準相関係数として定義される X と Θ に関するシリンダー上相関係数 [2, 3] は，

$$r_{X\Theta}^2=2\lambda^2A_1^2(1-\alpha_2)$$

と表されることがわかる。このようなシンプルさから，モーメントによって定義される特徴量に対する各パラメータの役割が確認しやすくなる利点もある。

こうした分布 (1) のシンプルな特徴から，スポーツにおけるのパス地点とパス方向からなるデータやある地域での樹木配置と生長方向からなるデータのようなシリンダー上データの分析を容易に実行できる可能性を秘めている。今後は，提案手法の多変量拡張や変量間の相関を強める拡張を考えることで，シリンダー上データの統計分析に役立たせていく予定である。

参考文献

[1] T. Abe and A. Pewsey. Sine-skewed circular distributions. *Statistical Papers*, 52:683–707, 2011.

[2] K. V. Mardia. Linear–circular correlation coefficients and rhythmometry. *Biometrika*, 63:403–405, 1976.

[3] T. E. Wehrly and R. A. Johnson. Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, 66:255–256, 1979.

Mixture Transition Distribution Modeling for Higher-Order Circular Markov Processes

Takayuki Shiohama
Department of Data Science, Nanzan University
Hiroaki Ogata
Faculty of Economics and Business Administration
Tokyo Metropolitan University

Abstract

The stationary higher-order Markov process for circular data is considered. We employ the mixture transition distribution (MTD) model to incorporate higher-order dependency in the circular time series. The underlying circular transition distribution is based on the Wehrly and Johnson's bivariate circular models. The structure of the circular autocorrelation function is found to be similar to the autocorrelation function of the autoregressive process on the line. The validity of the model is assessed by applying it to a series of real directional data.
keywords : circular statistics, Markov process, mixture transition distribution.

1 Introduction

Statistical analysis related to the data taking values on unit vectors is called directional statistics, where the direction is more important than the magnitude. Directional data occurs in many areas, namely geostatistics, natural sciences, environmental sciences, biology, information science, genetics, among others. The basic manifold of the high dimensional sphere is the unit circle which is embedded in the Euclidean space $\mathbb{R}^2$, and the statistical methods for the unit circle are known as circular statistics.

While many data taking values on the unit circle or sphere are characterized with time series structure, the statistical analysis for circular time series are not fully developed so far. We can observe many time series data for the circle, such as the wind or wave directions, time records of a certain event, and animal movement trajectories, which show typical periodic patterns. As for the correlation coefficient of two consecutive events in circular data, Fisher and Lee (1983) proposed circular correlation coefficients, and considered basic circular time series models as linked autoregressive and moving average models, and wrapping process for real valued time series (Fisher and Lee (1994)). The comprehensive analysis of circular time series modeling can be found in Breckling (1989).

One of the major approaches for modeling circular time series is to apply the Markov models for the circular transition densities. Wehrly and Johnson (1980) proposed joint distribution for the circular random variables. Holzmann et al. (2006) applied it for time series modeling using Hidden Markov models. Abe et al. (2017) showed the circular autocorrelation structure of the circular Markov process proposed by Wehrly and Johnson (1980). Some other Markov based circular time series are considered by Kato (2010) where he considered the Möbius transformation of the circular random variables to produce the transition densities. His models are extended

by Jones et al. (2015), where they considered the circular joint distributions called 'circulas' for analyzing correlated structures for circular random variables.

Whilst all these approaches are based on the first order Markov modeling for circular time series, investigating higher-order Markov models on the circle is primary importance. It is worthwhile to mention that, as far as the authors knowledge, there exist no approaches for higher-order circular Markov models in literature. Modeling discrete valued Higher-order Markov chains are very useful for the analysis of complex temporal relationships. Raftery (1985) considered the mixture transition distribution (MTD) model. Since then, the MTD model has been developed and improved in various ways, see for details, Raftery and Tavaré (1994), Le et al. (1996), and Berchtold et al. (2002).

In this study, we develop the higher-order Markov process on the circle using the transition density of Wehrly and Johnson (1980)'s model together with the MTD approaches. The autocorrelation structure of the proposed models are investigated. Since the probabilistic structure of the proposed models are quite complex and is not included in the class of regular parametric models, special attentions must be paid for parameter estimation. For this, we proposed the methods of maximum likelihood estimation, and provide some information criteria for model selections. Since the MTD model is analogous to the standard autoregressive model in that the autocorrelations satisfy a system of linear equations similar to the Yule-Walker equations, the proposed higher-order circular Markov models have similar autocorrelation structures. The asymptotic properties for the proposed models are considered.

References

Abe, T., H. Ogata, T. Shiohama, and H. Taniai (2017). Circular autocorrelation of stationary circular markov processes. *Statistical Inference for Stochastic Processes 20*(3), 275–290.

Berchtold, A., A. Raftery, et al. (2002). The mixture transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science 17*(3), 328–356.

Breckling, J. (1989). *The analysis of directional time series: applications to wind speed and direction*, Volume 61 of *Lecture Notes in Statistics*. Springer Science & Business Media.

Fisher, N. and A. Lee (1994). Time series analysis of circular data. *Journal of the Royal Statistical Society: Series B (Methodological) 56*(2), 327–339.

Fisher, N. I. and A. Lee (1983). A correlation coefficient for circular data. *Biometrika 70*(2), 327–332.

Holzmann, H., A. Munk, M. Suster, and W. Zucchini (2006). Hidden markov models for circular and linear-circular time series. *Environmental and Ecological Statistics 13*(3), 325–347.

Jones, M., A. Pewsey, and S. Kato (2015). On a class of circulas: copulas for circular distributions. *Annals of the Institute of Statistical Mathematics 67*(5), 843–862.

Kato, S. (2010). A markov process for circular data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(5), 655–672.

Le, N. D., R. D. Martin, and A. E. Raftery (1996). Modeling flat stretches, bursts outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association 91*(436), 1504–1515.

Raftery, A. and S. Tavaré (1994). Estimation and modelling repeated patterns in high order markov chains with the mixture transition distribution model. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 43*(1), 179–199.

Raftery, A. E. (1985). A model for high-order markov chains. *Journal of the Royal Statistical Society: Series B (Methodological) 47*(3), 528–539.

Wehrly, T. E. and R. A. Johnson (1980). Bivariate models for dependence of angular observations and a related markov process. *Biometrika 67*(1), 255–256.

非凸な罰則項によるスパース推定における拡張型確率的コンプレキシティについて

宮田 庸一 *

第一回 日本統計研究所　研究集会
統計的モデリングと統計推測理論

2021 年 8 月 12 日

概要

非凸な罰則関数を用いたスパース推定の一つであるブリッジ推定におけるチューニングパラメーターの選択に関して考える. Wang et al. (2009) によりベイズ型情報量規準によるチューニングパラメーターの選択方法が提案されているが, モデルの拡張型確率的コンプレキシティに対するラプラス近似の主要項として正当化できることについて報告を行った.

1　はじめに

記号 $'$ は, 行列の転置を表すものとする. 観測ベクトル $(Y_i, \boldsymbol{x}_i')'$ は, 以下のスパースな線形モデルに従うとする.

$$\begin{aligned} Y_i &= \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_{k_n}^* x_{k_n,i} + u_i, \quad (i = 1, ..., n) \\ &= \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_{k_n}^* x_{k_n,i} + 0 \cdot x_{k_n+1,i} + \cdots + 0 \cdot x_{p_n i} + u_i, \end{aligned}$$

ただし, $\boldsymbol{x}_i = (x_{1i}, \cdots, x_{p_n i})'$ は p_n 次元の説明変数を表すベクトルとし, Y_i は被説明変数とする. また $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, .., \beta_{k_n}^*, 0, \cdots, 0)' \in \mathbb{R}^{p_n+1}$ は真のパラメーターのベクトルとし, $\beta_j^* \neq 0$ $(j = 1, ..., k_n)$, および $\beta_j^* = 0$ $(j = k_n + 1, ..., p_n)$ であるとする. u_i は期待値 $E(u_i) = 0$, 分散 $Var(u_i) = \sigma_0^2$ となる撹乱項とする. ここで, 切片以外の真のパラメーターベクトルにおける非ゼロの要素の添え字の集合を $S_{0,n} = \{j \in \{1, ..., p_n\} | \beta_{0,j} \neq 0\}$ とし, その要素の個数を $k_n := |S_{0,n}|$ で表すとことにする. ここでは k_n と説明変数の個数 p_n は, 標本の大きさ n に依存してもよいことにする. このモデルにより生成された標本 $(Y_i, \boldsymbol{x}_i')'$ $(i = 1, ..., n)$ に対して, 以下線形モデルを当てはめることを考える.

$$Y_i = \beta_0 + \beta_{j_1} x_{j_1 i} + \cdots + \beta_{j_k} x_{j_k i} + \epsilon_i, \qquad (i = 1, ...n),$$

* 高崎経済大学・経済学部, 〒370-0801 高崎市上並榎町 1300

ただし, $j_1, ..., j_k \in \{1, ..., p_n\}$, $k = 1, ..., p_n$ とし, ϵ_i $(i = 1, ..., n)$ は独立に正規分布 $N(0, \sigma^2)$ に従うとする. このとき, どのようにして適切な説明変数を選ぶのかが問題となるが, p_n の次元がそれほど大きくない場合には, AIC(Akaike, 1973), BIC(Schwarz, 1978) などの情報量規準を用いて総当たり法を行うのが標準的なアプローチでとなる. しかしその一方で, p_n の次元が高くなるにつれ, 評価すべきモデルの数が指数的に増加するため, 総当たり法を行うのが困難になる. このような問題に対する一つの標準的なアプローチは, Tibshirani (1996) の LASSO(Least absolute shrinkage and selection operator) に代表される罰則付き最小二乗推定量を用いることである. 具体的には, $\boldsymbol{y} = (Y_1, ..., Y_n)'$, $\boldsymbol{X}_1 = (x_{ji})$:$n \times p_n$ 行列, $\mathbf{1}_n = (1, ..., 1)'$, $\boldsymbol{X} = \begin{pmatrix} \mathbf{1}_n & \boldsymbol{X}_1 \end{pmatrix}$ とおき, 損失関数

$$L_{0,n}(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \|\boldsymbol{\beta}\|_1 \tag{1}$$

を最小にする $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p_n})'$ の値 $\hat{\boldsymbol{\beta}}_L$ を求めればよい. ただし, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p_n} |\beta_j|$ とする. この推定量は LASSO 推定量と呼ばれ, $\lambda_n \to \infty$ $(n \to \infty)$ とその他の適切な条件のもとで漸近正規性 (Knight and Fu, 2000) を持つことが知られている. 一方で, LASSO 推定量をモデル選択手法の観点から見たときには, いくつかの問題がある. それは収束レートが $\sqrt{n}$ である漸近正規性が成り立つために必要となるチューニングパラメーター λ_n のオーダーの仮定の下では, 真のモデルを選ぶ確率は漸近的に 1 にならないことが知られている (Zou, 2006, p.1419). この問題を改善するために, 式 (1) における罰則項を, 非凸の形のものに置き換えた以下の損失関数を考え, これを最小にする推定量 $\hat{\boldsymbol{\beta}}$ を求める:

$$L_n(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p_n} |\beta_j|^\gamma, \tag{2}$$

ただし $0 < \gamma < 1$ とする. この推定量はブリッジ推定量と呼ばれている. 当然のことながら, ブリッジ推定においても適切な λ_n の選び方が重要になる.Wang et al. (2009) では, 線形モデルの下で, モデル選択に関して一致性を持つ BIC 型の情報量規準を与えている. Huang et al. (2008) では, オラクル性と呼ばれる, 推定量 $\hat{\boldsymbol{\beta}}$ が漸近正規性, 漸近有効性を持ち, なおかつ説明変数に関しても一致性を持つような条件を与えている. 例えば $\gamma = 1/2$ として, $k_n = k$(定数), $p_n = \log n$ とおいた場合, λ_n のオーダーに関する条件は, $\lambda_n/(n^{1/4}(\log n)^{3/4}) \to \infty$ $(n \to \infty)$, $\lambda_n = o(n^{1/2})$ となるが, そのような λ_n の選択肢は, $\lambda_n = 3n^{3/8}$ でも良いし, $\lambda_n = 10n^{3/8}$ でもよいことになる. すなわち Huang et al. (2008) の結果は λ_n の選択に関しては, 一意に定まらないことになる. 一方で, Huang et al. (2008) で与えられた条件の下では, ブリッジ推定量の漸近的な良さが保証されていることもわかる. このため, 本報告では, Huang et al. (2008) の条件の下で λ_n を選ぶための規準を, 拡張型確率的コンプレキシティの近似として与えられ, それが Wang et al. (2009) の規準と同様の形になることを示す. なお, 今回の話は Miyata (2021) の紹介となる.

2 ベイジアンアプローチ

ここでは, 前章の (2) 式で与えた, ブリッジ推定量をベイズ統計学の観点からの解釈するとともに, チューニングパラメーター λ_n のベイズ型の選択方法を提案する.

まずパラメーター $\beta_1, ..., \beta_{p_n}$ は, 以下の指数ベキ (exponential-power) 分布 $EP(0, \lambda_{0,n}, \sigma^2)$ の確率密度関数に独立に従うとする (Box and Tiao, 1973, p.157).

$$\pi(\beta_j|\lambda_{0,n}) = \frac{\gamma\lambda_{0,n}^{1/\gamma}}{2\Gamma(1/\gamma)(2\sigma^2)^{1/\gamma}} \exp\left\{-\frac{\lambda_{0,n}}{2\sigma^2}|\beta_j|^\gamma\right\}.$$

$\lambda_{0,n}$ は定数でもよいが, サンプルサイズ n に依存してもよい形にする. このため, $\boldsymbol{\beta}$ の事前密度関数は

$$\pi(\boldsymbol{\beta}|\lambda_{0,n}) = \pi(\beta_0)\prod_{j=1}^{p_n}\frac{\gamma(\lambda_{0,n})^{1/\gamma}}{2\Gamma(1/\gamma)(2\sigma^2)^{1/\gamma}} \exp\left\{-\frac{\lambda_{0,n}}{2\sigma^2}|\beta_j|^\gamma\right\}.$$

で与えられる. $\pi(\beta_0)$ は β_0 に対する事前分布で, 十分に滑らかなものとする. また記号を簡略化するため, 切片に対する事前分布を除いたものを $\pi_1(\boldsymbol{\beta}|\lambda_{0,n}) = \prod_{j=1}^{p_n}\pi(\beta_j|\lambda_{0,n})$ とする. そして, $\boldsymbol{X}$ および $\boldsymbol{\beta}$ を与えたときの $\boldsymbol{Y}$ の条件付き確率密度関数は多変量正規分布 $N_n(\boldsymbol{X\beta}, \sigma^2\mathbf{I}_n)$,

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{X\beta}||^2\right\}.$$

に従うものとする. 通常, $\boldsymbol{\beta}$ の事後密度関数は $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})\pi(\boldsymbol{\beta}|\lambda_{0,n})$ に比例する形になるが, ここではこの事後密度関数をさらに一般化して, $\boldsymbol{\beta}$ の事後密度関数は $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})^{\epsilon_n}\pi(\boldsymbol{\beta}|\lambda_{0,n})$ に比例するものを考える. この疑似事後密度関数を $p_Q(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y})$ とする. ただし ϵ_n は正の定数, もしくは $n \to \infty$ としたときに減少する列とする. なお ϵ_n が減少するということは, $\boldsymbol{\beta}$ に関する尤度 $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})$ の情報を弱めることを意味している.

ここで, $\boldsymbol{\beta}$ の損失関数を $\mathrm{H}_n(\boldsymbol{\beta}) = -\log\{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})^{\epsilon_n}\pi_1(\boldsymbol{\beta}|\lambda_{0,n})\}$ とおくと, この疑似事後密度関数は $p_Q(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})^{\epsilon_n}\pi(\boldsymbol{\beta}|\lambda_{0,n}) = \exp\{-\mathrm{H}_n(\boldsymbol{\beta})\}\pi(\beta_0)$ となる. これより, 切片の事前分布 $\pi(\beta_0)$ を無視して $\exp\{-\mathrm{H}_n(\boldsymbol{\beta})\}$ を最大化する疑似的な MAP(maximum a posteriori) 推定量 $\hat{\boldsymbol{\beta}}_{\lambda_n}$ を考えると,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\lambda_n} &= \underset{\boldsymbol{\beta}}{\mathrm{argmax}}\{-\mathrm{H}_n(\boldsymbol{\beta})\} \\
&= \underset{\boldsymbol{\beta}}{\mathrm{argmax}}\{\epsilon_n \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) + \log\pi_1(\boldsymbol{\beta}|\lambda_{0,n})\} \\
&= \underset{\boldsymbol{\beta}}{\mathrm{argmin}}\Big\{||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda_n\sum_{j=1}^{p_n}|\beta_j|^\gamma\Big\}.
\end{aligned}$$

となる. ただし $\lambda_n = \lambda_{0,n}/\epsilon_n$ と置いた. つまり $\hat{\boldsymbol{\beta}}_{\lambda_n}$ は, 非凸な罰則項 $\lambda_n\sum_{j=1}^{p_n}|\beta_j|^\gamma$ を持つブリッジ推定量は同等なことがわかる.

次に, ブリッジ推定量により選ばれたモデルの良さを評価する. $\hat{S} := \{j \in \{1, ..., p_n\}|\hat{\beta}_{\lambda_n,j} \neq 0\}$ は, $\hat{\boldsymbol{\beta}}_{\lambda_n} = (\hat{\beta}_{\lambda_n,0}, \hat{\beta}_{\lambda_n,1}, \cdots, \hat{\beta}_{\lambda_n,p_n})'$ において非ゼロな推定量の添え字の集合とする. $|\hat{S}|$ は集合 $\hat{S}$ における要素の個数を表す記号とする.

$\mathbf{X}(\hat{S})$ は, 集合 $\hat{S}$ の添え字に対応する説明変数の行列 $\boldsymbol{X}$ の部分行列とする. 例えば, $p_n = 3$, $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,0}, \hat{\beta}_{\lambda,1}, 0, \hat{\beta}_{\lambda,3})'$ とすると,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{31} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{pmatrix}, \quad \mathbf{X}(\hat{S}) = \begin{pmatrix} 1 & x_{11} & x_{31} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{3n} \end{pmatrix} \quad \mathbf{X}_1(\hat{S}) = \begin{pmatrix} x_{11} & x_{31} \\ \vdots & \vdots \\ x_{1n} & x_{3n} \end{pmatrix}$$
$$\hat{\boldsymbol{\beta}}_\lambda(\hat{S}) = (\hat{\beta}_{\lambda,0}, \hat{\beta}_{\lambda,1}, \hat{\beta}_{\lambda,3})', \quad \boldsymbol{\beta}(\hat{S}) = (\beta_0, \beta_1, \beta_3)'$$

となる.

3 拡張型確率的コンプレキシティ

ここでは, ブリッジ推定により選ばれたモデルの良さを評価するために, $L(Y_i, \boldsymbol{x}_i(\hat{S})'\boldsymbol{\beta}(\hat{S})) := -\log p(Y_i|\boldsymbol{x}_i(\hat{S}), \boldsymbol{\beta}(\hat{S})) = (1/2)\log(2\pi\sigma^2) + \{Y_i - \boldsymbol{x}_i(\hat{S})'\boldsymbol{\beta}(\hat{S})\}^2/(2\sigma^2)$ とおき, Yamanishi (1998) により提案された拡張型確率的コンプレキシティ (Extended Stochastic Complexity, ESC) を考える:

$$\mathrm{ESC}(\hat{S}, \lambda_{0,n}, \epsilon_n) := -\frac{1}{\epsilon_n}\log \int \exp\left\{-\epsilon_n \sum_{i=1}^{n} L(Y_i, \boldsymbol{x}_i'(\hat{S})\boldsymbol{\beta}(\hat{S}))\right\} \pi(\boldsymbol{\beta}(\hat{S})|\lambda_{0,n}) d\boldsymbol{\beta}(\hat{S}). \quad (3)$$

ただし $\pi(\boldsymbol{\beta}(\hat{S})|\lambda_{0,n})$ は, 事前密度関数 $\pi(\boldsymbol{\beta}|\lambda_{0,n})$ における $\boldsymbol{\beta}(\hat{S})$ の周辺確率密度関数とする. ESC を用いるときには, 関数 $L(Y_i, \boldsymbol{x}_i(\hat{S})'\boldsymbol{\beta}(\hat{S}))$ が, 観測される確率変数 Y_i と, その予測値 $\hat{Y}_i(\hat{S}) := \boldsymbol{x}_i(\hat{S})'\boldsymbol{\beta}(\hat{S})$ の損失関数の形になっている必要があるが, 我々は正規線形モデルを考えているためこの条件は見たされる. ここで (3) 式に対する解釈を述べる. 積分内の $\exp\left\{-\epsilon_n \sum_{i=1}^{n} L(Y_i, \hat{Y}_i(\hat{S}))\right\}$ であるが, これは大きな値であるほど, モデル $\hat{S}$ により与えられた予測量 $\hat{Y}_i(\hat{S})$ で, Y_i をうまく予測できることを表している. すなわち, (3) 式の積分は, 予測の良さを表す指標 $\exp\left\{-\epsilon_n \sum_{i=1}^{n} L(Y_i, \hat{Y}_i(\hat{S}))\right\}$ を, 事前分布の下で平均を取ったものと解釈することができる. ここでは (3) 式に対する直接的な解釈を与えたが, 逐次型確率予測問題からの解釈が山西健司 (2010) の補題 16(p.114) で与えらえている. こちらの方が, ESC が事後平均の関数で与えられるという意味で, ベイズ的な解釈ができる.

なお (3) 式は

$$-\frac{1}{\epsilon_n}\log \int_{\mathbb{R}^{|\hat{S}|+1}} p(\boldsymbol{y}|\boldsymbol{X}(\hat{S}), \boldsymbol{\beta}(\hat{S}))^{\epsilon_n} \pi(\boldsymbol{\beta}(\hat{S})|\lambda_{0,n}) d\boldsymbol{\beta}(\hat{S}) \quad (4)$$

と表すことができるため, 対数周辺尤度を拡張した形と関連があることもわかる. もし $\epsilon_n = 1$ のときには,

$$\mathrm{ESC}(\hat{S}, \lambda_{0,n}, 1) := -\log \int_{\mathbb{R}^{|\hat{S}|+1}} p(\boldsymbol{y}|\boldsymbol{X}(\hat{S}), \boldsymbol{\beta}(\hat{S})) \pi(\boldsymbol{\beta}(\hat{S})|\lambda_{0,n}) d\boldsymbol{\beta}(\hat{S}), \quad (5)$$

となるため, $\hat{S}$ に対応するモデルの周辺尤度に対数を取り, (-1) を乗じたものになることがわかる[*1].

4 拡張型確率的コンプレキシティに対するラプラス近似

記号を簡略化するために, $\boldsymbol{\Sigma}_n = n^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1$, $\boldsymbol{\Sigma}_{1n} = n^{-1}\boldsymbol{X}_1(S_0)'\boldsymbol{X}_1(S_0)$ とする. また $\rho_{1n} = \lambda_{\min}(\boldsymbol{\Sigma}_n)$, $\rho_{2n} = \lambda_{\max}(\boldsymbol{\Sigma}_n)$ とし, $\tau_{1n} = \lambda_{\min}(\boldsymbol{\Sigma}_{1n})$, $\tau_{2n} = \lambda_{\max}(\boldsymbol{\Sigma}_{1n})$ とおく. ただし,$\lambda_{\min}(\boldsymbol{A})$ は行列 $\boldsymbol{A}$ の最小固有値を表す記号とし, $\lambda_{\max}(\boldsymbol{A})$ は $\boldsymbol{A}$ の最大固有値を表す記号とする. このとき, 以下の条件を仮定する:

(A1) $\sum_{i=1}^n x_{ji} = 0$, $\dfrac{1}{n}\sum_{i=1}^n x_{ji}^2 = 1$ $(j = 1, ..., p_n)$

(A2) ϵ_i $(i = 1, 2, ...)$ は独立で同一分布に従い, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $(0 < \sigma^2 < \infty)$ とする.

(A3) (a) $\rho_{1n} > 0$ for all n, (b) $\dfrac{p_n + \lambda_n k_n}{n\rho_{1n}} \to 0$ $(n \to \infty)$.

(A4) (a) $\dfrac{\lambda_n k_n^{1/2}}{n^{1/2}} \to 0$ $(n \to \infty)$ (b) $\dfrac{\lambda_n \rho_{1n}^{2-\gamma}}{n^{\gamma/2} p_n^{(2-\gamma)/2}} \to \infty$ $(n \to \infty)$

(A5) ある定数 $0 < b_0 < b_1 < \infty$ が存在して,

$$b_0 \leq \min\{|\beta_{0,j}| \,|\, 1 \leq j \leq k_n\} \leq \max\{|\beta_{0,j}| \,|\, 1 \leq j \leq k_n\} \leq b_1.$$

(A6) $\boldsymbol{X}_1(S_0) = (\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$ とおく. このとき, ある正の定数 $0 < \tau_1 < \tau_2 < \infty$ が存在して,$\tau_1 \leq \tau_{1n} \leq \tau_{2n} \leq \tau_2$ for $\forall\, n$. (b)

$$n^{-1/2} \max_{1 \leq i \leq n} \boldsymbol{w}_i'\boldsymbol{w}_i \to 0, \quad (n \to \infty)$$

が成り立つ.

(B1) 以下のことが成り立つ.

$$\frac{k_n \log n}{\epsilon_n n} \to 0, \quad \text{and} \quad \epsilon_n n^{1/2} \to \infty \qquad (n \to \infty) \tag{6}$$

条件 (A1)–(A6) は Huang et al. (2008) とまったく同じものである. 条件 (A2) は, 線形回帰モデルでは標準的な仮定である. 条件 (A3) (a) は $\rho_{1n} \to 0$ $(n \to \infty)$ であることは許していることから, フルモデルにおいては, $\boldsymbol{\Sigma}_n$ が漸近的に退化する状況, 例えば漸近的に多重共線性が起こるような状況は許していることになる. (A3) (b) は, ブリッジ推定量の一致性を示すときに必要になる. 条件 (A4) は, ブリッジ推定量がオラクル性, および漸近正規性を示すために必要になる. 条件 (A6)(a) は, $\boldsymbol{\Sigma}_{1n}$ が, n に関して一様に正値定符号であることを仮定している. また (A5)(b) は非ゼロの回帰係数に対するブリッジ推定量の漸近正規性を示す際に必要になる. 条件 (B1) は, 一般化周辺尤度において漸近展開するために必要になるが, 真のモデルにおける説明変数の数 k_n の増加のスピードはかなり遅いものと考えるのが一般的であるため, これはかなり弱い条件となる.

[*1] 小さな値を取るほど良いモデル.

ここで, 記号を簡略化するため, 標準化された疑似 log posterior を

$$h(\boldsymbol{\beta}(\hat{S})) := -\frac{1}{n}\log p(\boldsymbol{y}|\boldsymbol{X}(\hat{S}),\boldsymbol{\beta}(\hat{S}))^{\epsilon_n}\pi(\boldsymbol{\beta}(\hat{S})|\lambda_{0,n})$$

とする. またその $\boldsymbol{\beta}(\hat{S})$ に関する 1 次の導関数を $\boldsymbol{\beta}(\hat{S})=\hat{\boldsymbol{\beta}}_{\lambda_n}(\hat{S})$ で評価したものを

$$\begin{aligned} Dh :&= \frac{\partial}{\partial\boldsymbol{\beta}(\hat{S})}h(\boldsymbol{\beta}(\hat{S}))\Big|_{\boldsymbol{\beta}(\hat{S})=\hat{\boldsymbol{\beta}}_{\lambda_n}(\hat{S})} \\ &= -\frac{\epsilon_n}{\sigma^2}\boldsymbol{X}(\hat{S})'(\boldsymbol{y}-\boldsymbol{X}(\hat{S})\boldsymbol{\beta}(\hat{S}))+\frac{\lambda_{0,n}\gamma}{2\sigma^2 n}\boldsymbol{\psi}_{1n}^{+}, \end{aligned} \tag{7}$$

で定義する. ただし, $\hat{\boldsymbol{\Sigma}}_{1n}^{+}=n^{-1}\boldsymbol{X}(\hat{S})'\boldsymbol{X}(\hat{S})$, $\hat{k}_n:=|\hat{S}|$,
$\boldsymbol{\psi}_{1n}^{+} := \left(0,|\hat{\beta}_{\lambda_n,1}|^{\gamma-1}\mathrm{sgn}(\hat{\beta}_{\lambda_n,1}),...,|\hat{\beta}_{\lambda_n,\hat{k}_n}|^{\gamma-1}\mathrm{sgn}(\hat{\beta}_{\lambda_n,\hat{k}_n})\right)'$ とおいた. 本来, $\hat{S}$ の集合は $\hat{j}_1,\hat{j}_2,...,\hat{j}_{\hat{k}_n}$ と書くべきであるが, 記号を簡略化するために, $\hat{S}=\{1,2,...,\hat{k}_n\}$ とした. 実際, このような簡略化は一般性を失わずにできる. また

$$\boldsymbol{\Psi}_{2n}^{+} := \gamma(\gamma-1)\begin{pmatrix} 0 & & & \\ & |\hat{\beta}_{\lambda_n,1}|^{\gamma-2} & & 0 \\ & & \ddots & \\ & 0 & & \ddots \\ & & & & |\hat{\beta}_{\lambda_n,\hat{k}_n}|^{\gamma-2} \end{pmatrix} \tag{8}$$

とし, その $\boldsymbol{\beta}(\hat{S})$ に関する 2 次の導関数を $\boldsymbol{\beta}(\hat{S})=\hat{\boldsymbol{\beta}}_{\lambda_n}(\hat{S})$ で評価したものを

$$\begin{aligned} D^2h &:= \frac{\partial^2}{\partial\boldsymbol{\beta}(\hat{S})\partial\boldsymbol{\beta}(\hat{S})'}h(\boldsymbol{\beta}(\hat{S}))\Big|_{\boldsymbol{\beta}(\hat{S})=\hat{\boldsymbol{\beta}}_{\lambda_n}(\hat{S})} \\ &= \frac{\epsilon_n}{\sigma^2}\hat{\boldsymbol{\Sigma}}_{1n}^{+}+\frac{\lambda_{0,n}}{2\sigma^2 n}\boldsymbol{\Psi}_{2n}^{+} \end{aligned}$$

とする. このとき, 以下が成り立つ.

定理 1 条件 (A1)–(A6), (B1) の下で,

$$\exp\left\{-\epsilon_n\mathrm{ESC}(\hat{S},\lambda_{0,n},\epsilon_n)\right\}=\exp\left\{-\epsilon_n\widehat{\mathrm{ESC}}(\hat{S},\lambda_{0,n},\epsilon_n)\right\}\{1+o_p(1)\}, \tag{9}$$

が成り立つ. ただし $|\boldsymbol{M}|$ は行列 $\boldsymbol{M}$ の行列式を表すものとし, $\hat{k}_n^{+}=\hat{k}_n+1$,

$$\exp\left\{-\epsilon_n\widehat{\mathrm{ESC}}(\hat{S},\lambda_{0,n},\epsilon_n)\right\}=(2\pi)^{\hat{k}_n^{+}/2}\left|nD^2h\right|^{-1/2}p(\boldsymbol{y}|\boldsymbol{X}(\hat{S}),\hat{\boldsymbol{\beta}}_{\lambda_n}(\hat{S}))^{\epsilon_n}\pi(\hat{\boldsymbol{\beta}}_{\lambda_n}(\hat{S})|\lambda_{0,n})$$

とする.

5　ベイズ型情報量規準

これまでは, σ^2 を既知として扱ってきたが, 実際にデータ解析を行うためには何らかの方法で推定する必要がある. ここでは, Wang et al. (2009) と同様にして

$$\hat{\sigma}^2_{\lambda_n} = \frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\lambda_{0,n}/\epsilon_n}||^2.$$

を用いて推定を行う. ここで定理 1 で得られた近似式 (9) に対して, 対数をとり, $(-2)/\epsilon_n$ をかけると

$$2\mathrm{ESC}(\hat{S}, \lambda_{0,n}, \epsilon_n) = n\log\left(\hat{\sigma}^2_{\lambda_n}\right) + \frac{\lambda_{0,n}}{\epsilon_n\hat{\sigma}^2_{\lambda_n}}\sum_{q\in\hat{S}}|\hat{\beta}_{\lambda_n,q}|^\gamma + \frac{\hat{k}_n+1}{\epsilon_n}\log(n\epsilon_n) + O_p\left(\frac{\hat{k}_n}{\epsilon_n}\right) + o_p(1).$$

となる. $\lambda_{0,n}$ と ϵ_n に対しては様々な取り方が考えられるが, ここでは $\lambda_{0,n} = \lambda_0$(定数), $\epsilon_n \downarrow 0$ $(n\to\infty)$, $|\hat{S}| = o_p(\log n)$ とおき, $O_p\left(\frac{\hat{k}_n\log n}{\epsilon_n}\right)$ より低いオーダーの項を取り除くと, 以下の形のベイズ型情報量規準を得ることができる.

$$\begin{aligned}\mathrm{BtIC}(\hat{S}) &= n\log(\hat{\sigma}^2_{\lambda_n}) + \hat{k}^+_n\log(n\epsilon_n)/\epsilon_n \\ &= n\log(\hat{\sigma}^2_{\lambda_n}) + C^*_n\hat{k}^+_n\log(n),\end{aligned} \tag{10}$$

ただし

$$C^*_n = \frac{1+(\log\epsilon_n)/\log n}{\epsilon_n} \tag{11}$$

となる. C^*_n は, 条件 (A1)–(A6) の下では多項式オーダーで増加する項, 即ち $C^*_n = O(n^{\alpha_n})$ $(0 < \alpha_n < 1/2)$ となる. これに関しては, 次の章で述べる. 通常の BIC であれば, $n\log(\hat{\sigma}^2_{\lambda_n}) + \hat{k}^+_n\log(n)$ となるので, 式 (10) は異なる形になっているが, これは Wang et al. (2009) で与えらえた情報量規準と同じ形になっている. Wang et al. (2009) の結果は, $n\log(\hat{\sigma}^2_{\lambda_n}) + C^*_n|\hat{S}|\log(n)$ の形の情報量規準を形式的に作っておいて, それがモデル選択に関して一致性を持つような C^*_n に対する条件を求めたものであるが, 今回, Huang et al. (2008) で与えられた条件の下で, 拡張型確率的コンプレキシティの近似として, Wang et al. (2009) の情報量規準を導出できたことにより, Wang et al. (2009) の結果に対するある種の妥当性を与えたといえる.

6　Wang et al. (2009) の情報量規準との相違点

拡張型確率的コンプレキシティの近似から情報量規準 (10) が導出されたわけであるが, 厳密に考えた場合, Wang et al. (2009) の情報量規準との相違点がある. Wang et al. (2009) では, C^*_n は多項式オーダーよりも低い任意のオーダーで増加することも認めている. またその論文のシミュレーションにおいては, $C^*_n = \log(\log p_n)$ が採用されている. 一方で, 我々の導出した結果においては, C^*_n がどのようなオーダーで増加するかを考えてみる.

最初に $\epsilon_n = \epsilon_0 n^{-\alpha_n}$ とおき, α_n の取りうる範囲を求めよう. ただし, 話を簡単にするため, $k_n = O(1)$, $\rho_{1n} = O(1)$ とする. このとき, 条件 (A4)(a), (b) は

$$\frac{\lambda_n}{n^{1/2}} \to 0, \qquad \frac{\lambda_n}{n^{\gamma/2} p_n^{(2-\gamma)/2}}$$

となる. これより,

$$\frac{\gamma}{2} + \frac{2-\gamma}{2} \log_n p_n < \alpha_n < \frac{1}{2}$$

となる. もし $\gamma = 1/2$ とおくとき,

$$\frac{1}{4} + \frac{3}{4} \log_n p_n < \alpha_n < \frac{1}{2} \tag{12}$$

となる. この結果と (11) 式の形から, 我々の仮定の下では, C_n^* は多項式オーダーで増加することがわかる. これは Wang et al. (2009) で提案されたものより速いオーダーで無限大にいくため, 実際には定数 $\epsilon_0 > 0$ を調整する必要がある.

また Wang et al. (2009) においては, 線形モデルの説明変数は切片を持たない確率変数を考えているのに対して, Miyata (2021) においては, 説明変数は切片を持ち, 固定されたもの (すなわち計画行列) を考えている. この点においても相違点がある.

参考文献

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281.

Box, G. E. P. and G. C. Tiao (1973). *Bayesian inference in statistical analysis*. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist. 36*(2), 587–613.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Ann. Statist. 28*(5), 1356–1378.

Miyata, Y. (2021). A Bayesian approach for selecting a tuning parameter in sparse linear models (for submission).

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*(2), 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B Stat. Methodol. 71*(3), 671–683.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. Inform. Theory 44*(4), 1424–1439.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*(476), 1418–1429.

山西健司 (2010). **情報論的学習理論**. 共立出版.

コピュラとフレイルティ混合効果から導かれる多変量故障時間分布

王尹辰（Yin-Chen Wang）
国立中央大学・統計研究所; peacebubu@gmail.com

江村剛志（Takeshi Emura）
久留米大学・バイオ統計センター; takeshiemura@gmail.com

要旨: コピュラモデルとフレイルティ（混合効果）モデルは，多変量故障時間分布における相関構造をモデリングするための一般的なモデル族である．本稿は，コピュラモデルとフレイルティモデルを含むさらに一般的な多変量故障時間モデル族を調査することを目的とする．この目的のために，先行研究である Marshall and Olkin(1988)の多変量故障時間モデルの族を再訪する．彼らの研究は，特殊なコピュラのみ，また二変量モデルに限定されていた．これに対して，我々の研究は，応用上よく用いられるコピュラに焦点をあて，また3次元以上の多変量モデルも考察する．部分族として，共有フレイルティモデルにより生成されるモデル族「フレイルティ-コピュラ」を定義・提案する．また，この族に属する個々の分布の名称を系統的に与える方法を提案する．また，先行研究では考慮されてこなかった複雑なフレイルティ分布(対数正規，切断正規，および折り畳み正規)の有用性も示す．本稿は同著者による Wang and Emura(2021)を邦訳し，やや簡潔にまとめたものである．

キーワード: 二変量分布，コピュラ, FGM コピュラ，フレイルティ，信頼性，生存時間解析

1. はじめに

寿命試験や臨床試験などでは，観測対象から複数の「故障」や「死亡」に関する値が得られる．観測値間の相関構造をモデリングには，多変量分布が必要となる(Duchateau and Janssen 2008; Balakrishnan and Lai 2009; Crowder 2012; Emura et al. 2019; 江村･道前 2020; 塚原 2021)．「コピュラ」と「フレイルティ」は，多変量故障時間モデルの構築のための良く知られたツールである．

フレイルティモデルでは，複数の故障時間に影響を与える未観測要因を考慮することで相関構造をモデリングする(Vaupel et al. 1979; Aalen 1994; Hougaard 1995)．フレイルティモデルは，様々なタイプの多変量故障時間の相関構造をモデリングするための一般的なツールとして重要な役割を担ってきた(Hougaard 1995; Duchateau et al. 2002; Duchateau and Janssen 2008; Rondeau et al. 2015; Ha et al. 2017; Ha and Lee 2021)．フレイルティモデルでは，未観測因子の分散は故障リスクの「不均一性」の程度を表すとともに，故障時間の間での相関構造の強さも表す．

コピュラは，不均一性の概念なしに純粋な相関構造を特徴づけるモデルを与える．コピュラ関

数(Nelsen 2006)は多数の周辺分布を結合し，同時分布を生成する関数のことを指す．コピュラ関数の選択は，フレイルティモデルとは異なり，周辺分布には影響しない．この性質により，多変量またはクラスター化された故障時間をモデリングするための明解な技法となっている（Shih 2014; Prenen et al. 2018; Emura et al. 2019; Campos et al. 2021; Sofeu et al. 2021; kwon et al. 2021; 塚原 2021）．スクラーの定理によって（Sklar 1959; Nelsen 2006），任意の多変量故障時間分布はコピュラとして表現されるが，一部の分布はフレイルティモデルとして表現されない場合がある．

コピュラモデルとフレイルティモデルは仕組みが概念的に異なる[see Section 3.3.4 of Duchateau and Janssen (2008)]．例えば，フレイルティモデルで定義された区分的定数ハザードモデル(e.g. Lo et al. 2017; Schneider al. 2020)は，コピュラモデルで定義された区分的定数ハザードモデルと異なる(e.g. Emura and Michimae 2017; Lipowski et al. 2021)ため，2 つのモデル異なる解釈を持つ．

上述の議論は，相関構造と不均一性という 2 つの異なる側面を統一する多変量故障時間モデルを構築する動機付けとなる．このようなモデルの研究は，コピュラとフレイルティが別々に議論されてきたため，極めて限られている．Marshall and Olkin(1988)がコピュラとフレイルティの性質を同時に持つモデルを提案したが，彼らの研究は特殊な二変量コピュラ(独立コピュラ，フレシェ−ヘフディング限界コピュラ，FGM コピュラ)に焦点を当てた．一方本稿では，クレイトンとグンベルコピュラなどの応用上重要なものや，多変量 FGM コピュラに焦点を当てる．部分族として，共有フレイルティモデルにより生成されるモデル族「フレイルティ−コピュラ」を定義・提案する．また，この族に属する個々の分布の名称を系統的に与える方法を提案する．また，先行研究では考慮されてこなかった複雑なフレイルティ分布(対数正規，切断正規，および折り畳み正規)の有用性も示す．

本稿は以下のように構成されている．2節で我々が提案するモデルを与える．3節〜5節では，クレイトンコピュラ，グンベルコピュラ，および多変量 FGM コピュラによってそれぞれ生成されたモデルを紹介する．6節で結論と今後の課題を述べる．

2. 提案するモデル

本節では，不均一性と相関構造という2つの異なる側面を持つ多変量故障時間モデルの基本的な枠組みを提案する． 2.1 節では，フレイルティでモデリングされる不均一性の概念を紹介する．2.2 節では，フレイルティ項を所与とした条件付きコピュラモデルを紹介し，「フレイルティ−コピュラモデル」という我々の提案する多変量故障時間分布族を定義する．

2.1 フレイルティ分布

故障のリスクに影響を与える未観測な因子を記述する「フレイルティ」の概念(Vaupel et al. 1979)を復習しよう．部品の母集団を考えたとき，ある部品は壊れやすい一方で，他の部品は頑健であろう．この不均一性をモデリングするために，リスクに影響を与える正の確率変数Zを考え，これをフ

レイルティ項と呼ぶ(Hougaard 1995; Duchateau and Janssen 2008; Ha et al. 2017). フレイルティ項Zはハザード関数のスケーリング係数として扱い，大きなZの値は高いハザード(脆弱な部品の場合)に関連付けられる．一方，小さなZの値は低いハザード（頑健な部品の場合)に関連付けられる．いま$f_Z(z)$をZの確率密度関数とし，パラメタベクトル$\boldsymbol{\eta}$によって特徴づけられるとする．一般的によく用いられるZのモデルは，ガンマ分布，対数正規分布，安定分布，逆ガウス分布である(Whitmore and Lee 1991; Joe 1993; Hougaard 1995; Duchateau and Janssen 2008; Ha et al. 2017; Piancastelli et al. 2020)．一方，数学的な利便性から我々が提案するのは，次の密度関数である．

1. 2パラメタガンマ分布； $Z\sim Gamma(\alpha,\beta)$

$$f_Z(z)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}z^{\alpha-1}\exp\left(-\frac{z}{\beta}\right),\quad z>0,\quad \alpha>0,\quad \beta>0$$

2. 対数正規分布； $Z\sim LN(\mu,\sigma^2)$

$$f_Z(z)=\frac{1}{z\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\log(z)-\mu)^2}{2\sigma^2}\right\},\quad z>0,\quad -\infty<\mu<\infty,\quad \sigma>0$$

3. 切断正規分布； $Z\sim TN(\mu,\sigma^2)$

$$f_Z(z)=\frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\},\quad z>\mu,\quad -\infty<\mu<\infty,\quad \sigma>0$$

4. 折り畳み正規分布(Leone et al. 1961)； $Z\sim FN(\mu,\sigma^2)$

$$f_Z(z)=\frac{1}{\sigma\sqrt{2\pi}}\left[\exp\left\{-\frac{(z+\mu)^2}{2\sigma^2}\right\}+\exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}\right],\quad z>0,\ \mu\geq 0,\ \sigma>0$$

上記4つのモデルは次の分布特性を持つ．期待値は$E^{Gamma}(Z)=\alpha\beta$，$E^{LN}(Z)=\exp(\mu+\sigma^2/2)$，$E^{TN}(Z)=\mu+(2/\pi)^{1/2}\sigma$，$E^{FN}(Z)=(2/\pi)^{1/2}\sigma\times\exp\{-\mu^2/(2\sigma^2)\}+\mu\{1-2\Phi(-\mu/\sigma)\}$となる．切断正規分布と折り畳み正規分布は，$\mu=0$場合にのみ同等である．従って，この2つのモデルが一致するのは，ゼロ切断モデルとゼロ折り畳みモデルの場合のみである．折り畳み正規分布で$\mu>3\sigma$の場合，正規分布とほぼ同等であるが，切断正規分布と大きく異なる．しかし，正規分布自体は負の値をとるため，フレイルティ分布とは考えられない．切断正規分布と折り畳み正規分布は多くの文献で考慮されている(Cohen 1951; 1961; 1991; Leone et al. 1961; Sundberg 1974; Lodhi, et al. 2021; Zeng and Gui 2021)にもかかわらず，フレイルティモデルの文献においてほとんど考慮されていない．また4つのモデルは Marshall and Olkin(1988)では考慮されていない．彼らは 1 パラメタガンマ分布，安定分布，負の二項分布，および切断二項分布を考慮した．

2.2 多変量フレイルティ-コピュラモデル

本節では多変量故障時間$(T_1, \ldots, T_k)$に対して「フレイルティ-コピュラモデル」を定義・提案する.

Oakes (1989)に従い，すべての故障時間はフレイルティ値$Z = z$の影響を受けるとする．すなわち，T_jの$Z = z$を所与としたときの条件生存関数を次のように定義する

$$S_{T_j|Z}(t_j|z) \equiv \Pr(T_j > t_j|z) = S_{0j}(t_j)^z,$$

ここで$S_{0j}(\cdot)$は任意の連続型生存関数である．フレイルティ項が正の確率変数であるから，任意の$z > 0$に対して$S_{T_j|Z}(t_j|z)$は生存関数となる.

Marshall and Olkin(1988)に従い，条件付き生存関数をコピュラ$C_\theta: [0, 1]^k \to [0, 1]$を用いて

$$S_{T_1,\ldots,T_k|Z}(t_1, \ldots, t_k|z) \equiv \Pr(T_1 > t_1, \ldots, T_k > t_k|z) = C_\theta\{S_{T_1|Z}(t_1|z), \ldots, S_{T_k|Z}(t_k|z)\}, \quad (1)$$

と定義する．ここでθはコピュラのパラメタである．式 (1) は，任意のコピュラに対して，多変量生存関数となることは明らかである. Marshall and Olkin(1988)は，次のより一般的なモデル

$$C_\theta\{S_{T_1|Z}(t_1|z_1), \ldots, S_{T_k|Z}(t_k|z_k)\},$$

を考えた．ここで$(z_1, \ldots, z_k)$多変量フレイルティ項である．我々はこの一般的なモデルではなく，式(1)の「共有」フレイルティ$z_1 = \cdots = z_k = z$の場合に焦点を当てる．独立コピュラ$C_\theta(u_1, \ldots, u_k) = \prod_{j=1}^k u_j$の場合，モデル(1)は通常のフレイルティモデルに帰着する (Oakes 1989).

Emura et al.(2017, 2019)は$Z = z$をクラスター内の全ての部品（対象者）で共有される値とし，2次元コピュラC_θに対してモデル(1)を考え，これを結合フレイルティ-コピュラ（joint frailty-copula）モデルと呼んでいる．彼らは$-\mathrm{d}\{\log S_{0j}(t)\}/dt$をスプライン関数によってモデリングしている．また$S_{0j}(.)$をワイブル分布でモデリングする場合も Wu et al.(2020)および Shinohara et al.(2020)で提案されている．このようなクラスター化されたデータの解析においては，$Z = z$の値は未観測であるものの，クラスターを観測することにより，その値を類推することが出来る．従って，興味のある相関構造は$Z = z$を所与としたときの条件付きケンドール順位相関係数

$$\tau_\theta^{|Z} = 4\iint S_{T_1T_2|Z}(t_1, t_2|z)dS_{T_1T_2|Z}(t_1, t_2|z) - 1 = 4\iint_{[0,1]\times[0,1]} C_\theta(v, w)dC_\theta(v, w) - 1$$

である.

しかしながら，我々の本稿における目的はクラスター化されたデータの解析ではない．つまり我々は，Zは未観測であるとし，これをその分布に関する積分で除外し

$$P(T_1 > t_1, \ldots, T_k > t_k) \equiv S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = \int_0^\infty S_{T_1 \ldots T_k|Z}(t_1, \ldots, t_k|z) f_Z(z)\, dz$$

を得る．上式の積分内の条件付き生存関数を式(1)で指定し次の定義を与える．

定義(多変量フレイルティ-コピュラモデル): 次式の多変量生存関数の族を定義する．

$$P(T_1 > t_1, \ldots, T_k > t_k) \equiv S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = \int_0^\infty C_\theta\{ S_{01}(t_1)^z, \ldots, S_{0k}(t_k)^z \} f_Z(z)\, dz \qquad (2)$$

ここで$f_Z(.)$は正値確率変数の密度関数，$C_\theta(.)$はk変量コピュラ，および$S_{0j}(.),\ j \in \{1, \ldots, k\}$は任意の生存関数とする．

提案するモデル族(2)は，コピュラによって与えられる相関構造を共有フレイルティの効果でより一般化した族と見做すことができる．式(2)は Marshall and Olkin(1988)が与えたモデル族の部分族であるが，彼らは共有フレイルティ分布とコピュラの選択について十分に議論していない．Charpentier et al.(2014)は，極値コピュラ族の下で式(2)と類似の族を定義したが，クレイトン型コピュラや FGM 型コピュラなど，極値コピュラ族に属さない多くの有用なコピュラが存在する．

式(2)のフレイルティ-コピュラモデルは広い分布族であり，様々なモデルを含む．この名称「フレイルティ-コピュラ」に従い，この族に属する個々のモデルの名称を系統的に与えることが出来る．たとえば，フレイルティがガンマ分布，コピュラがクレイトン型コピュラで決まるモデルは「ガンマ-クレイトンモデル」と呼ぶことにする．

式(2)の積分をどのように計算するかは最も重要な問題であるが，Marshall and Olkin(1988)では議論が不十分であった．以下の我々の議論では，その問題の解答のいくつかを与える．

2.3 提案モデルの一般的性質

数学的利便性の観点から，フレイルティ-コピュラモデルの式(2)を次のように書き換える

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = H_{\theta,\boldsymbol{\eta}}\{\Lambda_{01}(t_1), \ldots, \Lambda_{0k}(t_k)\}. \qquad (3)$$

ここで$\Lambda_{0j}(t_j) = -\log S_{0j}(t_j)$は累積ハザード関数，関数$H_{\theta,\boldsymbol{\eta}}$は後に例を通して議論する．式(2)はモデルの構築法を明示する一方，式(3)は統計的モデリングに便利である．たとえば，累積ハザード関数にワイブル回帰モデルを設定する場合，$\Lambda_{0j}(t_j) = \lambda_j t_j^\nu \exp(\beta_j x_j)$とし，ここで尺度パラメタ$\lambda_j > 0$，形状パラメタ$\nu > 0$，共変量$x_j$をおいた．共変量が無い場合，$\Lambda_{0j}(t_j) = \lambda_j t_j^\nu$とおく．

我々の興味は，式(2)の積分を計算する方法や，$H_{\theta,\boldsymbol{\eta}}$の式を求める方法である．コピュラやフレイルティ分布を上手く選び，式(2)または式(3)が簡潔な形で表されれば，そのモデルは実用的であ

ろう．この問題に対し Charpentier et al.(2014)のアイデアを利用し，式(2)を

$$S_{T_1 \dots T_k}(t_1, \dots, t_k) = \int_0^\infty \exp[-\ell_\theta\{z\Lambda_{01}(t_1), \dots, z\Lambda_{0k}(t_k)\}] f_Z(z)\, dz,$$

としてみる．ここで$\ell_\theta = -\log(C_\theta)$は依存関数（dependence function）と呼ばれる．この依存関数が条件$\ell_\theta(zt_1, \dots, zt_k) = z\ell_\theta(t_1, \dots, t_k)$を満足すると，次のように積分が計算できる

$$\begin{aligned} S_{T_1 \dots T_k}(t_1, \dots, t_k) &= L_{\boldsymbol{\eta}}[\ell_\theta\{\Lambda_{01}(t_1), \dots, \Lambda_{0k}(t_k)\}] \\ &= L_{\boldsymbol{\eta}}[-\log C_\theta\{\exp(-\Lambda_{01}(t_1)), \dots, \exp(-\Lambda_{0k}(t_k))\}], \end{aligned}$$

ここで

$$L_{\boldsymbol{\eta}}(s) = \int_0^\infty \exp(-zs) f_Z(z|\boldsymbol{\eta})\, dz$$

はラプラス変換である．残念ながら，この条件は常に満足されるものでなく，また積分が計算出来るための必要条件でもない．この積分の計算問題は，3節～5節でさらに個別に議論していく．

提案するモデル（2）の周辺分布を見てみよう．周辺生存関数は

$$S_{T_j}(t_j) = \int_0^\infty S_{0j}(t_j)^z f_Z(z|\boldsymbol{\eta})\, dz = L_{\boldsymbol{\eta}}[-\log S_{0j}(t_j)], \quad j = 1, 2, \dots, k, \tag{4}$$

となるので，これはフレイルティ分布に依存する．式(4)を利用して，式(2)を次のように書き換える

$$S_{T_1 \dots T_k}(t_1, \dots, t_k) = C_{\theta,\boldsymbol{\eta}}\{S_{T_1}(t_1), \dots, S_{T_k}(t_k)\}, \tag{5}$$

ここで$C_{\theta,\boldsymbol{\eta}}$はコピュラである．明らかに$C_{\theta,\boldsymbol{\eta}} \neq C_\theta$であり，実際

$$C_{\theta,\boldsymbol{\eta}}(u_1, \dots, u_k) = \int_0^\infty C_\theta\left[\exp\{-zL_{\boldsymbol{\eta}}^{-1}(u_1)\}, \dots, \exp\{-zL_{\boldsymbol{\eta}}^{-1}(u_k)\}\right] f_Z(z|\boldsymbol{\eta})\, dz$$

である．上式から次の基本的性質が導かれる．

(i) $C_{\theta,\boldsymbol{\eta}}$は$S_{0j}(\cdot)$または$\Lambda_{0j}(\cdot)$に依存しない．

(ii) $C_\theta(u_1, \dots, u_k) = \prod_{j=1}^k u_j$の場合，$C_{\theta,\boldsymbol{\eta}} = L_{\boldsymbol{\eta}}[L_{\boldsymbol{\eta}}^{-1}(u_1) + \cdots + L_{\boldsymbol{\eta}}^{-1}(u_k)]$である．

(iii) $C_\theta = \min_i(u_i)$の場合，$C_{\theta,\boldsymbol{\eta}} = \min_i(u_i)$である．

(iv) $f_Z(z|\boldsymbol{\eta})$が点$z = 1$に退化するとき，$C_{\theta,\boldsymbol{\eta}} = C_\theta$である．

性質(i)は，$C_{\theta,\boldsymbol{\eta}}$と$S_{0j}(\cdot)$が別々にモデリングできることを示す．性質(ii)は，$C_{\theta,\boldsymbol{\eta}}$が純粋なフレイルティモデルに帰着することを示す．性質(iii)は，コピュラが上限に達した場合，フレイルティ分布は不定（任意である）ことを示す．性質(iv) は，$C_{\theta,\boldsymbol{\eta}}$が純粋なコピュラ$C_\theta$に帰着することを示す．$C_{\theta,\boldsymbol{\eta}}$のより具体的な性質は，コピュラとフレイルティ分布を実際に与えた3節～5節で調べる．

2.4 ケンドール順位相関係数

故障時間$(T_1, \ldots, T_k)$の間の相関の強さを表す指標であるケンドール順位相関係数を考える．それには，フレイルティ-コピュラモデル(2)の下で適当なペア(T_i, T_j)を選び，2変量コピュラ$C_{\theta,\boldsymbol{\eta}}(1, \ldots, u_i, \ldots, u_j, \ldots, 1)$に関する順位相関係数を考えれば良い．従って，$k = 2$の場合を議論すれば十分である. Nelsen(2006)に習い，$k = 2$の場合のケンドール順位相関係数は

$$\tau_{\theta,\boldsymbol{\eta}} = 4\iint S_{T_1T_2}(t_1, t_2)dS_{T_1T_2}(t_1, t_2) - 1 = 4\iint_{[0,1]^2} C_{\theta,\boldsymbol{\eta}}(v, w)dC_{\theta,\boldsymbol{\eta}}(v, w) - 1, \qquad (6)$$

である．ここで$C_{\theta,\boldsymbol{\eta}}(v,w)$は$0 \le v, w \le 1$で定義される二変量コピュラである．$C_{\theta,\boldsymbol{\eta}}$が特定のコピュラになる場合，積分を計算し，$\tau_{\theta,\boldsymbol{\eta}}$が簡潔な式で表示できる．しかしながら，$\tau_{\theta,\boldsymbol{\eta}}$の式は一般には求めることが難しく，数値計算を要する．Marshall and Olkin(1988)に従い，我々も次のシミュレーションアルゴリズムを採用して，$\tau_{\theta,\boldsymbol{\eta}}$の値を近似することを推奨する:

ケンドール順位相関$\tau_{\theta,\boldsymbol{\eta}}$を求めるアルゴリズム:

ステップ 1: $Z \sim f_Z(z|\boldsymbol{\eta})$を生成.

ステップ 2: $(U_1, U_2) \sim C_\theta(v,w)$を生成.

ステップ 3: $T_1 = S_{01}^{-1}\left(U_1^{1/Z}\right)$; $T_2 = S_{02}^{-1}(U_2^{1/Z})$とおく.

ステップ 4 : ステップ 1～3 をR回繰り返し，次を得る

$$\hat{\tau}_{\theta,\boldsymbol{\eta}} = \binom{R}{2}^{-1} \sum_{1 \le i < j \le R} \mathrm{sgn}\{(T_1^i - T_1^j)(T_2^i - T_2^j)\}.$$

上記$\hat{\tau}_{\theta,\boldsymbol{\eta}}$は$\tau_{\theta,\boldsymbol{\eta}}$の不偏推定量である.

3. ガンマ-クレイトンモデル

クレイトン型コピュラ(Clayton 1978)は，次のように定義される

$$C_\theta(u_1, \ldots, u_k) = (u_1^{-\theta} + \cdots + u_k^{-\theta} - k + 1)^{-1/\theta}, \qquad \theta > 0.$$

クレイトン型コピュラは，生存時間解析において最も広く用いられているコピュラである(Rivest and Wells 2001; Rotolo et al. 2013; Emura et al. 2014; Emura and Chen 2016; Moradian et al. 2019; Kwon et al. 2021; Emura et al. 2020; Emura, Sofeu and Rondeau 2021). 統計的工程管理にも用いられる（Long et al. 2014; Sun et al. 2020; Huang and Emura 2021; Emura, Lai and Sun 2021; Kim

et al. 2021; 江村 2021).

クレイトン型コピュラの下で，提案するモデル(2)は次のようになる

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = \int_0^\infty \left\{S_{01}(t_1)^{-\theta z} + \cdots + S_{0k}(t_k)^{-\theta z} - k + 1\right\}^{-\frac{1}{\theta}} f_Z(z)\, dz.$$

上記積分はどのようなフレイルティ分布でも陽に書けない．ガンマフレイルティ分布の場合を下で解説する．対数正規フレイルティ分布の場合の説明は Wang and Emura(2021)に譲る．

$Z \sim Gamma(\alpha, \beta)$の下で生成されるガンマ-クレイトンモデルは次式となる

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = \int_0^\infty \left\{S_{01}(t_1)^{-\theta z} + \cdots + S_{0k}(t_k)^{-\theta z} - k + 1\right\}^{-\frac{1}{\theta}} \frac{1}{\Gamma(\alpha)\beta^\alpha} z^{\alpha-1} \exp\left(-\frac{z}{\beta}\right) dz.$$

ここで$\theta > 0,\ \alpha > 0,\ \beta > 0$である．上記の同時生存関数を次のように書き換える

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = H_{\theta,\alpha,\beta}\{\Lambda_{01}(t_1),\ \ldots,\ \Lambda_{0k}(t_k)\}, \tag{7}$$

ここで

$$H_{\theta,\alpha,\beta}(a_1, \ldots, a_k) = \int_0^1 \left(w^{-a_1\theta\beta} + \cdots + w^{-a_k\theta\beta} - k + 1\right)^{-\frac{1}{\theta}} \frac{(-\log w)^{\alpha-1}}{\Gamma(\alpha)} dw, \quad a_j > 0$$

である．この積分は陽に書けないが，$H_{\theta,\alpha,\beta}(.)$を数値計算することは容易である．モデル(7)は，ガンマフレイルティモデルの一般化と見なすことができる．すなわち，$\theta \downarrow 0$で退化したモデルは

$$H_{0,\alpha,\beta}(a_1, \ldots, a_k) = \left(1 + \beta \sum_{j=1}^k a_j\right)^{-\alpha}$$

となる．式(7)の周辺生存関数は$S_{T_j}(t_j) = \{1 + \beta\Lambda_{0j}(t_j)\}^{-\alpha},\ j = 1, \ldots, k,$ となる．したがって

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = C_{\theta,\alpha}\{S_{T_1}(t_1), \ldots, S_{T_k}(t_k)\},$$

と書け，ここで

$$C_{\theta,\alpha}(u_1, \ldots, u_k) = \int_0^1 \{w^{(1-u_1^{-1/\alpha})\theta} + \cdots + w^{(1-u_k^{-1/\alpha})\theta} - k + 1\}^{-1/\theta} \frac{(-\log w)^{\alpha-1}}{\Gamma(\alpha)} dw$$

はβに依存しないコピュラである．

二変量の場合，同時生存関数は

$$S_{T_1T_2}(t_1, t_2) = H_{\theta,\alpha,\beta}\{\Lambda_{01}(t_1), \Lambda_{02}(t_2)\}; \qquad \theta > 0,\ \alpha > 0,\ \beta > 0$$

であり，ここで

$$H_{\theta,\alpha,\beta}(a_1, a_2) = \int_0^1 \left(w^{-a_1\theta\beta} + w^{-a_2\theta\beta} - 1\right)^{-\frac{1}{\theta}} \frac{(-\log w)^{\alpha-1}}{\Gamma(\alpha)} dw, \quad a_1 > 0, \quad a_2 > 0$$

である．この関数は，次の R コードを使用して簡単に計算できる．

```
H=function(w){
    (w^(-a1*theta*beta)+w^(-a2*theta*beta)-2+1)^(-1/theta)*(-log(w))^(alpha-1)/gamma(alpha)
}
integrate(H,0,1)$value
```

ガンマ分布のパラメタを$\alpha = 1/\eta,\ \beta = \eta > 0$とおくと，平均値$E(Z) = 1$，分散$Var(Z) = \eta$となる．このとき，フレイルティ分布は分散パラメタ$\eta$によって特徴付けられる．

図1は$S_{T_1T_2}(t_1, t_2)$の等高線$S_{T_1T_2}(t_1, t_2) = p,\ p \in \{0.1, 0.2, \ldots, 0.9\}$を表す．$\theta$が 1 から 8 に増えると等高線が垂直および水平になるのがわかる．また，ηが 0.5 から 1 に増えると，低い等高線が縦軸・横軸に近づくことがわかる．よって，コピュラのパラメタとフレイルティ分布のパラメタが異なる働きをしている．θとηが増えると，完全正相関の等高線$\min\{S_{T_1}(t_1), S_{T_2}(t_2)\} = p$に収束することも理解できる．$\theta$と$\eta$が増えるとケンドール順位相関係数の値も増加する（図1）．

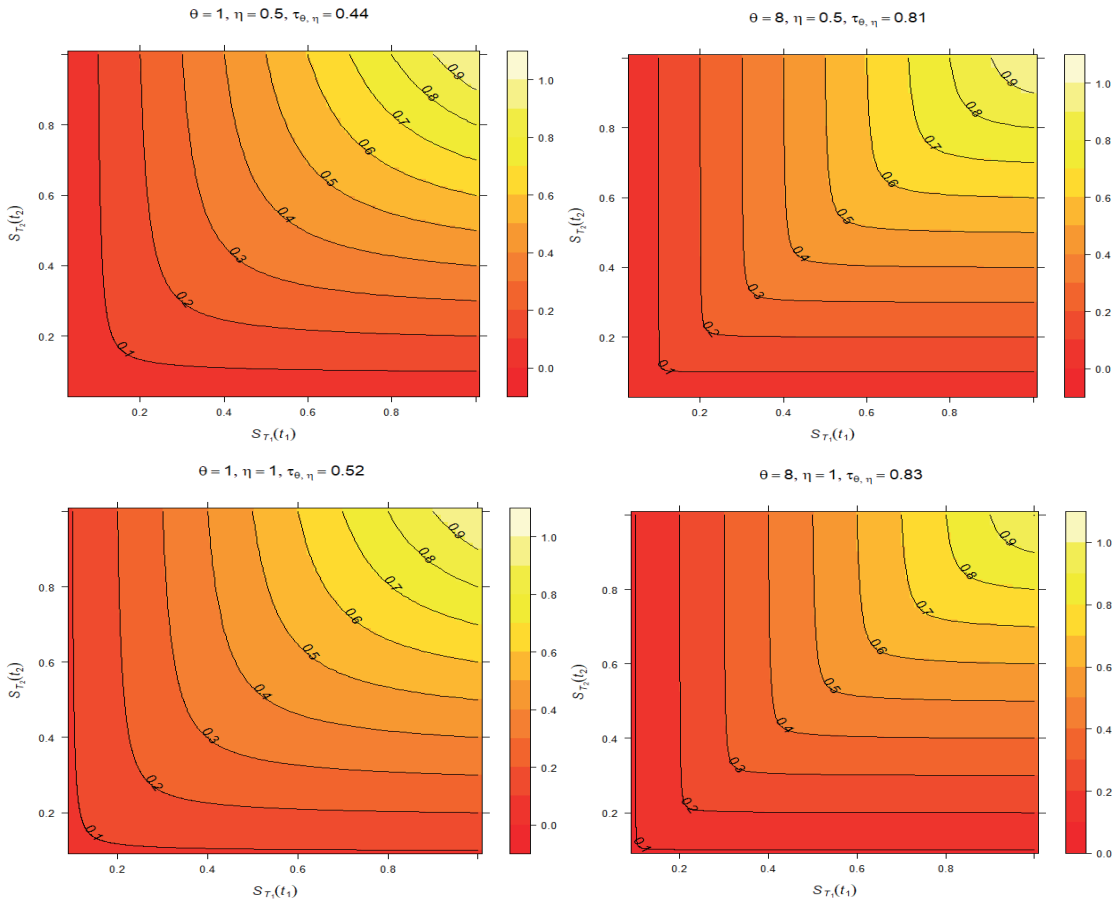

図 1:ガンマ-クレイトンモデルの生存関数の等高線．

4. グンベルコピュラ

グンベル型コピュラ(Gumbel 1960)は，次のように定義される

$$C_\theta(u_1, \ldots, u_k) = \exp[-\{(-\log u_1)^{\theta+1} + \cdots + (-\log u_k)^{\theta+1}\}^{1/(\theta+1)}], \qquad \theta \geq 0.$$

グンベル型コピュラも生存時間解析で多用される(Peng et al. 2018; Wang et al. 2020; Kawakami et al. 2021; Sofeu et al. 2021)．グンベル型コピュラの下で，提案するモデル(2)は

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = \int_0^\infty \exp(-zA_\theta) f_Z(z)\, dz$$

となり，ここで

$$A_\theta = A(t_1, \ldots, t_k; \theta) = \left[\sum_{j=1}^{k} \{-\log S_{0j}(t_j)\}^{\theta+1} \right]^{\frac{1}{\theta+1}} = \left[\sum_{j=1}^{k} \Lambda_{0j}^{\theta+1}(t_j) \right]^{\frac{1}{\theta+1}}.$$

上記の積分は，多くのフレイルティ分布の場合で陽に計算することができる．積分の計算を以下のフレイルティ分布(ガンマ，切断正規，折り畳み正規)の場合で詳説する．

4.1 ガンマ-グンベルモデル:

$Z \sim Gamma(\alpha, \beta)$，$\alpha > 0$，$\beta > 0$とすると，ガンマ-グンベルモデルは次のように計算できる

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = \int_0^\infty \exp(-zA_\theta) \frac{1}{\Gamma(\alpha)\beta^\alpha} z^{\alpha-1} \exp\left(-\frac{z}{\beta}\right) dz = (1 + \beta A_\theta)^{-\alpha}.$$

周辺生存関数は$S_{T_j}(t_j) = \{1 + \beta\Lambda_{0j}(t_j)\}^{-\alpha}$であるから，上式は次のように表せる

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = \left(1 + \left[\{S_{T_1}(t_1)^{-1/\alpha} - 1\}^{\theta+1} + \cdots + \{S_{T_k}(t_k)^{-1/\alpha} - 1\}^{\theta+1}\right]^{1/(\theta+1)}\right)^{-\alpha}.$$

したがって，このモデルのコピュラは次のようになる

$$C_{\theta,\alpha}(u_1, \ldots, u_k) = \left[1 + \left\{\left(u_1^{-1/\alpha} - 1\right)^{\theta+1} + \cdots + \left(u_k^{-1/\alpha} - 1\right)^{\theta+1}\right\}^{\frac{1}{\theta+1}}\right]^{-\alpha}.$$

このコピュラはβに依存しない．上記コピュラはアルキメデス型$C_{\theta,\alpha}(u_1, \ldots, u_k) = \phi_{\theta,\alpha}^{-1}[\phi_{\theta,\alpha}(u_1) + \cdots + \phi_{\theta,\alpha}(u_k)]$であり，生成素$\phi_{\theta,\alpha}(t) = (t^{-1/\alpha} - 1)^{\theta+1}$を持つ(Nelsen 2006)．

二変量の場合，同時生存関数は

$$S_{T_1T_2}(t_1, t_2) = \left(1 + \beta\left[\Lambda_{01}^{\theta+1}(t_1) + \Lambda_{02}^{\theta+1}(t_2)\right]^{\frac{1}{\theta+1}}\right)^{-\alpha}; \quad \theta \geq 0,\ \alpha > 0,\ \beta > 0.$$

であり，二変量コピュラ は

$$C_{\theta,\alpha}(u_1, u_2) = \left[1 + \left\{\left(u_1^{-1/\alpha} - 1\right)^{\theta+1} + \left(u_2^{-1/\alpha} - 1\right)^{\theta+1}\right\}^{\frac{1}{\theta+1}}\right]^{-\alpha}.$$

Genest and MacKay (1986)の公式により，ケンドール順位相関係数は次のように計算される

$$\tau_{\theta,\,\alpha} = 1 + 4\int_0^1 \frac{\phi_{\theta,\alpha}(t)}{d\phi_{\theta,\alpha}(t)/dt}dt = 1 - \frac{2}{(\theta+1)\left(\frac{1}{\alpha}+2\right)}.$$

フレイルティ分布を1パラメタに制約した，$Gamma(\alpha = 1, \beta) = Exp(\beta)$や$Gamma(\alpha = \gamma/2, \beta = 2) = \chi^2_{df=\gamma}$も考慮できる．制約$\alpha = 1/\eta$と$\beta = \eta$の下では，二変量コピュラは

$$C_{\theta,\eta}(u_1, u_2) = \left[1 + \{(u_1^{-\eta} - 1)^{\theta+1} + (u_2^{-\eta} - 1)^{\theta+1}\}^{\frac{1}{\theta+1}}\right]^{-1/\eta}$$

である．この$C_{\theta,\eta}$は，BB1 コピュラ(Joe(1997)の P.150)である．Nelsen(2006)は，2 パラメタのコピュラの例として挙げている(彼の書籍の例 4.22 と 4.26 を参照)．コピュラ$C_{\theta,\eta}$は$\theta = 0$のときクレイトンコピュラに，$\eta \to 0$のときグンベルコピュラに帰着する．ηがクレイトンコピュラからの乖離を示すため，頑健性の研究にも有用である(Emure, Sofeu, and Rondeau(2021)の Supplementary Material)．

BB1 コピュラは，Lu と Bhattacharyya(1990)によって与えられた 2 変量ワイブルモデルからも導かれる．Wang et al.(2020)はガンマ-グンベル分布を競合リスクデータ分析に用いている．

4.2 切断正規-グンベルモデル:

驚くべきことに，$Z \sim TN(\mu, \sigma^2)$の場合でも，同時生存関数が次のように陽に書くことが出来る

$$S_{T_1 \ldots T_k}(t_1, \ldots, t_k) = 2\exp\left(\frac{\sigma^2}{2}A_\theta^2 - \mu A_\theta\right)\Phi(-\sigma A_\theta) \tag{8}$$

ここで，Φは$N(0, 1)$の累積分布関数である．導出は付録を参照されたい．周辺生存関数は

$$S_{T_j}(t_j) = 2\exp\left(\frac{\sigma^2}{2}\,\Lambda_{0j}^2(t_j) - \mu\,\Lambda_{0j}(t_j)\right)\Phi(-\sigma\,\Lambda_{0j}(t_j))$$

となる．式(8)は陽に書くことが出来たが，対応するコピュラとケンドール順位相関係数は陽に書くことが出来ない(下記)．単純な指数分布のモデル$\Lambda_{01}(t) = t$を考えると，周辺生存関数は$S_{T_1}(t) = 2\exp(\sigma^2 t^2/2 - \mu t)\Phi(-\sigma t)$となる．

制約$E(Z) = \mu + \sqrt{2/\pi}\sigma = 1$を与え，基準化された1パラメタのフレイルティ分布を考える．このフレイルティ分布のラプラス変換を

$$L_{\sigma^2}(s) = \int_{1-\sqrt{2/\pi}\sigma}^{\infty} \exp(-zs)\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\exp\left[-\frac{\{z-(1-\sqrt{2/\pi}\sigma)\}^2}{2\sigma^2}\right]dz$$

とおき，その逆関数を$L_{\sigma^2}^{-1}(.)$とする．このときコピュラは次のようになる

$$C_{\theta,\,\sigma^2}(u_1,\,u_2) = 2\exp\left[\frac{\sigma^2}{2}\left\{L_{\sigma^2}^{-1}(u_1)^{\theta+1} + L_{\sigma^2}^{-1}(u_2)^{\theta+1}\right\}^{\frac{2}{\theta+1}} - (1-\sqrt{2/\pi}\sigma)\left\{L_{\sigma^2}^{-1}(u_1)^{\theta+1} + L_{\sigma^2}^{-1}(u_2)^{\theta+1}\right\}^{\frac{1}{\theta+1}}\right]$$
$$\times\,\Phi\left[-\sigma\left\{L_{\sigma^2}^{-1}(u_1)^{\theta+1} + L_{\sigma^2}^{-1}(u_2)^{\theta+1}\right\}^{\frac{1}{\theta+1}}\right].$$

図2は$S_{T_1T_2}(t_1,\,t_2)$の等高線$S_{T_1T_2}(t_1,\,t_2) = p,\ p \in \{0.1, 0.2, \ldots, 0.9\}$を表す．$\theta$が 1 から 8 に増えると等高線が垂直および水平になるのがわかる．また，σが 0.1 から 1 に増えると，低い等高線が縦軸・横軸に近づくことがわかる．よって，コピュラのパラメタとフレイルティ分布のパラメタが異なる働きをしている．θとσが増えると，完全正相関の等高線$\min\{S_{T_1}(t_1),\,S_{T_2}(t_2)\} = p$に収束することも理解できる．$\theta$と$\sigma$が増えるとケンドール順位相関係数の値も増加する（図 2）．

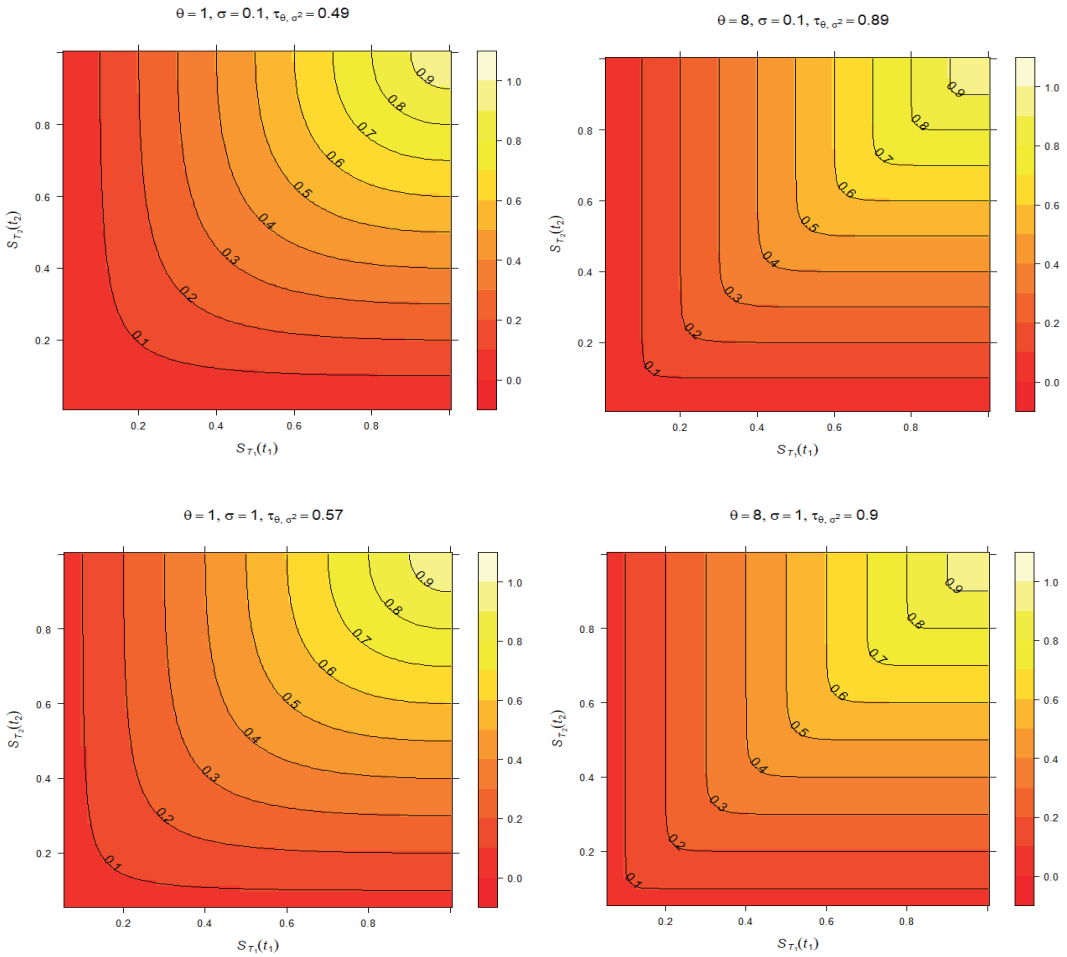

図 2. 切断正規−グンベルモデルの生存関数の等高線．

4.3 折り畳み正規−グンベルモデル

さらに驚くべきことに，折り畳み正規分布$Z\sim FN(\mu,\,\sigma^2)$のような複雑な密度関数形式であっても，同時生存関数が次のように陽に書くことが出来る

$$S_{T_1\cdots T_k}(t_1,\ldots,t_k)=\exp\left(\frac{\sigma^2}{2}A_\theta^2+\mu A_\theta\right)\Phi\left(-\frac{\mu}{\sigma}-\sigma A_\theta\right)+\exp\left(\frac{\sigma^2}{2}A_\theta^2-\mu A_\theta\right)\Phi\left(\frac{\mu}{\sigma}-\sigma A_\theta\right).$$

導出の詳細については，付録を参照されたい．ゼロで折り畳んだ($\mu=0$)場合，

$$S_{T_1\cdots T_k}(t_1,\ldots,t_k)=2\exp\left(\frac{\sigma^2}{2}A_\theta^2\right)\Phi(-\sigma A_\theta)$$

となり，ゼロ切断正規−グンベルモデルの生存関数に等しい(4.2 節)．モデルの性質は，切断正規−グンベルモデルと同じように議論できる．詳細は Wang and Emura(2021)を参照されたい．

5. ガンマ−FGM モデル

多変量 FGM コピュラ(Kotz et al. 2000)は，次のように定義される

$$C_\theta(u_1,\ldots,u_k)=\prod_{i=1}^{k}u_i+\theta\prod_{i=1}^{k}u_i(1-u_i),\qquad \theta\in[-1,1] \tag{9}$$

優れた数学的性質から，FGM コピュラは理論・応用上重要である(Domma and Giordano 2013; Shih and Emura 2018; 2020; 2021; Mohtashami−Borzadaran et al. 2019; Shih et al. 2019; de Oliveira et al. 2021; Kawakami et al. 2021; Ota and Kimura 2021; Šeliga et al. 2021)．提案するモデル(2)は次のようになる:

$$S_{T_1\cdots T_k}(t_1,\ldots,t_k)\;=\int_0^\infty\left[\;\prod_{i=1}^{k}S_{0i}(t_i)^z+\theta\prod_{i=1}^{k}S_{0i}(t_i)^z\{1-S_{0i}(t_i)^z\}\;\right]f_Z(z)\,dz.$$

Marshall and Olkin(1988)は$k=2$かつ$Z\sim Gamma(\alpha,\,\beta)$の場合，積分が陽に書けることを示した．

以下で，$k\geq 2$の場合への一般化を考える．$Z\sim Gamma(\alpha,\,\beta)$のとき，

$$\begin{aligned}S_{T_1\cdots T_k}(t_1,\ldots,t_k)&=\int_0^\infty\left[\;\prod_{i=1}^{k}S_{0i}(t_i)^z+\theta\prod_{i=1}^{k}S_{0i}(t_i)^z\{1-S_{0i}(t_i)^z\}\;\right]\frac{z^{\alpha-1}\exp(-z/\beta)}{\Gamma(\alpha)\beta^\alpha}dz\\&=\left[\;1+\beta\left\{\sum_{i=1}^{k}\Lambda_{0i}(t_i)\right\}\;\right]^{-\alpha}+\theta\left[\;1+\beta\left\{\sum_{i=1}^{k}\Lambda_{0i}(t_i)\right\}\;\right]^{-\alpha}\\&\quad-\theta\sum_{i_1=1}^{k}\left[\;1+\beta\left\{\;\Lambda_{0i_1}(t_{i_1})+\sum_{i=1}^{k}\Lambda_{0i}(t_i)\;\right\}\;\right]^{-\alpha}\\&\quad+\theta\sum_{j=2}^{k}(-1)^j\sum_{1\leq i_1<\cdots<i_j\leq k}\left[\;1+\beta\left\{\;\sum_{i=1}^{k}\Lambda_{0i}(t_i)+\sum_{m=1}^{j}\Lambda_{0i_m}(t_{i_m})\;\right\}\;\right]^{-\alpha}\end{aligned}$$

である．従って，ガンマ-FGM モデルは，複雑な形式ではあるが陽に書くことが出来る．周辺生存関数は$S_{T_j}(t_j)=\{1+\beta\Lambda_{0j}(t_j)\}^{-\alpha}$であるから，コピュラは次のようである

$$
\begin{aligned}
C_{\alpha,\beta,\theta}(u_1,...,u_k)=&\left[1-\left\{\sum_{i=1}^{k}(1-u_i^{-1/\alpha})\right\}\right]^{-\alpha}+\theta\left[1-\left\{\sum_{i=1}^{k}(1-u_i^{-1/\alpha})\right\}\right]^{-\alpha}\\
&-\theta\sum_{m=1}^{k}\left\{u_m^{-1/\alpha}-\sum_{i=1}^{k}(1-u_i^{-1/\alpha})\right\}^{-\alpha}\\
&+\theta\sum_{j=2}^{k}(-1)^j\sum_{1\le i_1<\cdots<i_j\le k}\left[1-\left\{\sum_{i=1}^{k}(1-u_i^{-1/\alpha})+\sum_{m=1}^{j}(1-u_{i_m}^{-1/\alpha})\right\}\right]^{-\alpha}.
\end{aligned}
$$

このコピュラは文献には存在していないようである．コピュラがβに依存しないことに注意する．

$k=2$かつ$\alpha=1/\eta$の場合，コピュラは

$$
\begin{aligned}
C_{\theta,\eta}(u_1,u_2)=&(u_1^{-\eta}+u_2^{-\eta}-1)^{-1/\eta}+\theta(u_1^{-\eta}+u_2^{-\eta}-1)^{-1/\eta}\\
&-\theta(2u_1^{-\eta}+u_2^{-\eta}-2)^{-1/\eta}-\theta(u_1^{-\eta}+2u_2^{-\eta}-2)^{-1/\eta}\\
&+\theta(2u_1^{-\eta}+2u_2^{-\eta}-3)^{-1/\eta}.
\end{aligned}
$$

上記のコピュラは Cook and Johnson (1986)が提案したものと同じであり，上記のコピュラの導出は Marshall and Olkin(1988)と同じである．$\theta=0$でクレイトンコピュラに帰着し，$\eta\to 0$のとき FGM コピュラに帰着する．このコピュラは，他の一般化 FGM コピュラ(Bairamov and Bayramoglu 2013; Huang et al. 2013; Domma and Giordano 2013; Hürlimann 2017; Saminger-Platz et al. 2020)と同様に，FGM コピュラの狭い相関の範囲を拡張することが出来る．

$k=2$のとき，ガンマ-FGM 分布の同時生存関数は次のようである

$$
\begin{aligned}
S_{T_1T_2}(t_1,t_2)=&B_\beta(t_1,t_2)^{-\alpha}+\theta B_\beta(t_1,t_2)^{-\alpha}-\theta\{\beta\Lambda_{01}(t_1)+B_\beta(t_1,t_2)\}^{-\alpha}\\
&-\theta\{\beta\Lambda_{02}(t_2)+B_\beta(t_1,t_2)\}^{-\alpha}+\theta\{2B_\beta(t_1,t_2)-1\}^{-\alpha}.
\end{aligned}
$$

ここで$B_\beta(t_1,t_2)=1+\beta\{\Lambda_{01}(t_1)+\Lambda_{02}(t_2)\}$とおいた．生存関数$S_{T_1T_2}(t_1,t_2)$の等高線は Wang and Emura(2021)を参照されたい．さらに$\theta=0$，$\alpha=1/\eta$，$\beta=\eta$とおくと，

$$
S_{T_1T_2}(t_1,t_2)=\left[1+\eta\{\Lambda_{01}(t_1)+\Lambda_{02}(t_2)\}\right]^{-1/\eta}
$$

となり，これは共有ガンマフレイルティモデルに帰着する．

6.結論

本稿では，多変量故障時間モデルの構築に関する2つの主要なモデル族である，コピュラモデルとフレイルティモデルの統合を目指した．そこで，コピュラモデルとフレイルティモデルの両方を含む，さらに広い多変量故障時間モデルの族である「多変量フレイルティ-コピュラモデル」を提案した．この提案したモデル族は，Marshall and Olkin(1988)によって提案されたものの部分族である．本研究は，Marshall and Olkin(1988)が行った不完全な研究を補完・拡張するものと見做せる．実際，コピュラとフレイルティ分布の組み合わせにより，様々な多変量分布を導き出すことに成功し，そのような新しい分布を生み出す手法を示した．さらに，提案したモデル族に属する分布の名称を系統的に与える手法も提案した．とりわけ魅力的な多変量分布は，ガンマ-クレイトンモデル，ガンマ-グンベルモデル，切断正規-グンベルモデル，折り畳み正規-グンベルモデル，ガンマ-FGM モデルと命名されたものであった．

本稿で提案したいくつかの分布の実データへの当てはめや，これら分布下での統計的推測理論の構築は，今後の重要な課題である．生存時間データを扱うためにこれらの分布を用いることが想定されるが，生存時間データは打ち切り・切断・競合リスクなどにより完全データを得ることが困難である（Klein and Moeschberger 2003; Dörre and Emura 2019; Dörre 2021; Emura and Ha 2021; Li et al. 2021; Wu et al. 2021）．とりわけ，従属打ち切り（dependent censoring）の相関構造は同定が困難であり，コピュラを既知と扱う場合がある（Rivest and Wells 2001; Emura and Chen 2018; Emura and Hsu 2020）．近年この従属打ち切りの相関構造が推定可能なモデルが提案され始めている（Deresa and Van Keilegom 2020）．同様に，従属切断（dependent truncation）の相関構造モデリング，とりわけパラメトリックモデリング（Emura and Konno 2012a,b; Emura and Pan 2020）にも提案したモデルが応用できる可能性がある．

謝辞

阿部俊弘先生には，日本統計研究所，研究集会「統計的モデリングと統計的推測理論」において本研究の口頭発表の機会および所報へ掲載の機会を頂いたことに心より感謝申し上げます．本研究は，台湾科学技術省(MOST 107-2118-M-008-003-MY3)からの助成金によって行われた．

付録: 数学的な導出

切断正規−グンベルモデルの導出:

切断正規分布 $Z\sim TN(\mu,\ \sigma^2)$の下で，以下の計算を行うと，

$$\begin{aligned}
S_{T_1,\ldots,T_k}(t_1,\ldots,t_k) &= \int_{-\infty}^{\infty} S_{T_1,\ldots,T_k|Z}(t_1,\ldots,t_k \mid Z) f_Z(z)dz \\
&= \int_{\mu}^{\infty} \exp(-zA_\theta)\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}dz \\
&= \int_{\mu}^{\infty} \frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\exp\left[-\frac{1}{2\sigma^2}\{2\sigma^2 A_\theta z+(z-\mu)^2\}\right]dz \\
&= 2\exp\left\{\frac{(\mu-\sigma^2A_\theta)^2-\mu^2}{2\sigma^2}\right\}\int_{\mu}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{\{z-(\mu-\sigma^2A_\theta)\}^2}{2\sigma^2}\right]dz,
\end{aligned}$$

となる．これに変数変換$w=\{z-(\mu-\sigma^2A_\theta)\}/\sigma$を適用すると，求める結果を得る

$$S_{T_1\ldots T_k}(t_1,\ldots,t_k)=2\exp\left(\frac{\sigma^2}{2}A_\theta^2-\mu A_\theta\right)\Phi(-\sigma A_\theta).$$

折り畳み正規−グンベルモデルの導出:

折り畳み正規分布$Z\sim FN(\mu,\ \sigma^2)$の下で，次式を計算する必要がある

$$S_{T_1\ldots T_k}(t_1,\ldots,t_k)=\int_0^{\infty}\exp(-zA_\theta)\frac{1}{\sigma\sqrt{2\pi}}\left[\exp\left\{-\frac{(z+\mu)^2}{2\sigma^2}\right\}+\exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}\right]dz.$$

先の切断正規分布の場合と同様の計算を行い，次の積分を得る

$$\int_0^{\infty}\exp(-zA_\theta)\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(z+\mu)^2}{2\sigma^2}\right\}dz=\exp\left(\frac{\sigma^2}{2}A_\theta^2+\mu A_\theta\right)\Phi\left(-\frac{\mu}{\sigma}-\sigma A_\theta\right).$$

μを$-\mu$で置き換えると，もう一つの必要な積分を得る．よって

$$S_{T_1\ldots T_k}(t_1,\ldots,t_k)=\exp\left(\frac{\sigma^2}{2}A_\theta^2+\mu A_\theta\right)\Phi\left(-\frac{\mu}{\sigma}-\sigma A_\theta\right)+\exp\left(\frac{\sigma^2}{2}A_\theta^2-\mu A_\theta\right)\Phi\left(\frac{\mu}{\sigma}-\sigma A_\theta\right).$$

参考文献

Aalen OO (1994). Effects of frailty in survival analysis. Statistical Methods in Medical Research 3(3): 227-243.

Bairamov I, Bayramoglu K (2013). From the Huang-Kotz FGM distribution to Baker's bivariate distribution, Journal of Multivariate Analysis, 113: 106-115.

Balakrishnan N, Lai CD. (2009). Continuous Bivariate Distributions. Springer.

Campos E, Braekers R, de Souza DJ, Chaves LM. (2021). Factor copula models for right-censored clustered survival data. Lifetime Data Analysis, 27: 499–535.

Charpentier A, Fougères AL, Genest C, Nešlehová JG. (2014). Multivariate archimax copulas. Journal of Multivariate Analysis, 126, 118-136.

Clayton DG (1978). A model for association in bivariate life tables and its application to epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 65, 141-51

Cohen, A (1951). On estimating the mean and variance of singly truncated normal frequency distributions from the first three sample moments. Annals of the Institute of Statistical Mathematics 3, 37–44.

Cohen, A (1961). Tables for maximum likelihood estimates: Singly truncated and singly censored samples. Technometrics 3, 535–541.

Cohen, A (1991). Truncated and Censored Samples: Theory and Applications, CRC Press.

Cook RD, Johnson ME. (1986). Generalized Burr-Pareto-logistic distributions with applications to a uranium exploration data set. Technometrics, 28(2), 123-131.

Crowder MJ (2012). Multivariate Survival Analysis and Competing Risks, CRC Press

de Oliveira RP, de Oliveira Peres MV, dos Santos MR, Martinez EZ, Achcar JA (2021). A Bayesian inference approach for bivariate Weibull distributions derived from Roy and Morgenstern methods. Statistics, Optimization & Information Computing, 9(3), 529-554.

Deresa NW, Van Keilegom I (2020). A multivariate normal regression model for survival data subject to different types of dependent censoring. Computational Statistics & Data Analysis, 144, 106879.

Domma F, Giordano S (2013). A copula-based approach to account for dependence in stress-strength models. Statistical Papers, 54(3), 807-826.

Dörre, A. (2021). Semiparametric likelihood inference for heterogeneous survival data under double truncation based on a Poisson birth process. Japanese Journal of Statistics and Data Science, doi:10.1007/s42081-021-00128-w.

Dörre A, Emura T (2019), Analysis of Doubly Truncated Data, An Introduction, JSS Research Series in Statistics, Springer.

Duchateau L, Janssen P (2008). The Frailty Model, Springer.

Duchateau L, Janssen P, Lindsey P, Legrand C, Nguti R, Sylvester R (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. Computational Statistics & Data Analysis 40(3), 603-620.

Emura T, Chen YH (2016). Gene selection for survival data under dependent censoring, a copula-based approach, Statistical Methods in Medical Research 25(6): 2840–57.

Emura T, Chen YH (2018). Analysis of Survival Data with Dependent Censoring, Copula-Based Approaches, JSS Research Series in Statistics, Springer.

Emura T, Ha ID (2021). Special feature: Recent statistical methods for survival analysis. Japanese Journal of

Statistics and Data Science, doi:10.1007/s42081-021-00140-0

Emura T, Hsu J H. (2020). Estimation of the Mann–Whitney effect in the two-sample problem under dependent censoring. Computational Statistics & Data Analysis, 150, 106990.

Emura T, Kao FH, Michimae H (2014). An improved nonparametric estimator of sub-distribution function for bivariate competing risk models, Journal of Multivariate Analysis 132: 229-41.

Emura T, Konno Y (2012a). Multivariate normal distribution approaches for dependently truncated data, Statistical Papers 53(1): 133-49

Emura T, Konno Y (2012b). A goodness-of-fit tests for parametric models based on dependently truncated data, Computational Statistics & Data Analysis 56: 2237-50.

Emura T, Lai CC, Sun LH (2021). Change point estimation under a copula-based Markov chain model for binomial time series, Econometrics and Statistics, doi:10.1016/j.ecosta.2021.07.007.

Emura T, Matsui S, Rondeau V (2019). Survival Analysis with Correlated Endpoints, Joint Frailty-Copula Models, JSS Research Series in Statistics, Springer

Emura T, Michimae H (2017). A copula-based inference to piecewise exponential models under dependent censoring, with application to time to metamorphosis of salamander larvae, Environmental and Ecological Statistics 24(1): 151–73.

Emura T, Nakatochi M, Murotani K, Rondeau V (2017). A joint frailty-copula model between tumour progression and death for meta-analysis, Statistical Methods in Medical Research 26: 2649-66.

Emura T, Pan CH (2020) Parametric maximum likelihood inference and goodness-of-fit tests for dependently left-truncated data, a copula-based approach, Statistical Papers 61:479-501.

Emura T, Sofeu C, Rondeau V (2021). Conditional copula models for correlated survival endpoints: individual patient data meta-analysis of randomized controlled trials, Statistical Methods in Medical Research, doi:10.1177/09622802211046390.

Emura T, Shih JH, Ha ID, Wilke RA (2020). Comparison of the marginal hazard model and the sub-distribution hazard model for competing risks under an assumed copula, Statistical Methods in Medical Research 29(8): 2307–27

江村剛志, 道前洋史 (2020). コピュラを用いた生存時間解析-相関のあるエンドポイントとメタ分析の活用-, 統計数理 第 68 巻(第 1 号): 147-174 (in Japanese).

江村剛志 (2021). コピュラ・マルコフ連鎖モデルによる定常時系列解析−パラメトリック推定と統計的工程管理-, 日本統計学会誌 第 51 巻(第 1 号): 41-73 (in Japanese).

Genest C, MacKay RJ (1986). Copules archimédiennes et families de lois bidimensionnelles dont les marges sont données. Canadian Journal of Statistics 14(2): 145-159.

Gumbel EJ (1960) Distributions de valeurs extremes en plusieurs dimensions. PubL Inst Stat. Pari 9: 171-3.

Ha ID, Jeong JH, Lee Y. (2017). Statistical Modelling of Survival Data with Random Effects: H-likelihood Approach, Springer, Singapore.

Ha, ID, Lee Y. (2021), A review of h-likelihood for survival analysis, Japanese Journal of Statistics and Data Science, doi:10.1007/s42081-021-00125-z.

Hougaard P (1995). Frailty models for survival data. Lifetime Data Analysis, 1(3), 255-273.

Huang JS, Dou X, Kuriki S, Lin GD (2013). Dependence structure of bivariate order statistics with applications

to Bayramoglu's distributions, Journal of Multivariate Analysis, 114: 201-208.

Huang X, Emura T (2021). Model diagnostic procedures for copula-based Markov chain models for statistical process control, Communications in Statistics-Simulation and Computation 50(8): 2345-67

Hürlimann W (2017). A comprehensive extension of the FGM copula, Statistical Papers, 58: 373-392.

Joe H (1993). Parametric families of multivariate distributions with given margins. Journal of Multivariate Analysis 46(2), 262-82.

Joe H (1997). Multivariate Models and Dependence Concepts. Chapman & Hall/CRC, New York.

Kawakami R, Michimae H, Lin YH (2021). Assessing the numerical integration of dynamic prediction formulas using the exact expressions under the joint frailty-copula model, Japanese Journal of Statistics and Data Science, doi:10.1007/s42081-021-00133-z.

Kim JM, Baik J, Reller M (2021). Control charts of mean and variance using copula Markov SPC and conditional distribution by copula. Communications in Statistics-Simulation and Computation 50(1): 85-102.

Klein JP, Moeschberger ML (2003). Survival Analysis: Techniques for Censored and Truncated Data, New York: Springer.

Kotz S, Balakrishnan N, Johnson NL (2000). Continuous Multivariate Distributions, Volume 1: Models, New York: Springer.

Kwon S, Ha ID, Shih JH, Emura, T. (2021). Flexible parametric copula modeling approaches for clustered survival data. Pharmaceutical Statistics, doi:10.1002/pst.2153.

Leone FC, Nelson LS, Nottingham RB (1961). The folded normal distribution. Technometrics 3(4),543-550.

Li D, Hu XJ, Wang R (2021). Evaluating association between two event times with observations subject to informative censoring. Journal of the American Statistical Association, doi:10.1080/01621459.2021.1990766.

Lipowski C, Lo SM, Shi S, Ralf A. Wilke RA (2021). Competing risks regression with dependent multiple spells: Monte Carlo evidence and an application to maternity leave, Japanese Journal of Statistics and Data Science, doi.org/10.1007/s42081-021-00110-6.

Lodhi C, Mani Tripathi Y, Kumar Rastogi M (2021). Estimating the parameters of a truncated normal distribution under progressive type II censoring. Communications in Statistics-Simulation and Computation, 50(9): 2757-2781

Lo SM, Stephan G, Wilke RA (2017). Competing risks copula models for unemployment duration: An application to a German Hartz reform. Journal of Econometric Methods 6(1).

Long TH, Emura T (2014). A control chart using copula-based Markov chain models, Journal of the Chinese Statistical Association 52(4): 466-96.

Lu JC, Bhattacharyya GK (1990). Some new constructions of bivariate Weibull models. Annals of the Institute of Statistical Mathematics, 42(3), 543-559.

Marshall AW, Olkin I. (1988). Families of multivariate distributions. Journal of the American Statistical Association, 83(403), 834-841.

Moradian H, Larocque D, Bellavance F (2019). Survival forests for data with dependent censoring. Statistical Methods in Medical Research, 28(2), 445-461.

Mohtashami-Borzadaran V, Amini M, Ahmadi J (2019). A generalized bivariate lifetime distribution based on parallel-series structures. Kybernetika, 55(3), 435-454.

Nelsen RB (2006) An Introduction to Copulas, (2nd ed.). Springer.

Oakes D (1989). Bivariate survival models induced by frailties. Journal of the American Statistical Association 84:487-493.

Ota S, Kimura M (2021). Effective estimation algorithm for parameters of multivariate Farlie–Gumbel–Morgenstern copula. Japanese Journal of Statistics and Data Science, doi:10.1007/s42081-021-00118-y.

Peng M, Xiang L, Wang S (2018) Semiparametric regression analysis of clustered survival data with semi-competing risks, Computational Statistics & Data Analysis 124: 53-70.

Piancastelli LS, Barreto-Souza W, Mayrink VD (2020). Generalized inverse-Gaussian frailty models with application to TARGET neuroblastoma data. Annals of the Institute of Statistical Mathematics doi.org/10.1007/s10463-020-00774-z

Prenen L, Braekers R, Duchateau L (2018). Investigating the correlation structure of quadrivariate udder infection times through hierarchical Archimedean copulas, Lifetime Data Analysis. 24:719-742

Rivest LP, Wells MT (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. Journal of Multivariate Analysis 79(1): 138-155.

Rondeau V, Pignon JP, Michiels S, MACH-NC Collaborative Group. (2015). A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. Statistical Methods in Medical Research 24(6): 711-729.

Rotolo F, Legrand C, Van Keilegom I (2013). A simulation procedure based on copulas to generate clustered multi-state survival data. Computer Methods and Programs in Biomedicine, 109(3), 305-312.

Šeliga A, Kauers M, Saminger-Platz S, Mesiar R, Kolesárová A, Klement EP (2021). Polynomial bivariate copulas of degree five: characterization and some particular inequalities. Dependence Modeling, 9(1), 13-42.

Saminger-Platz S, Kolesárová A, Šeliga A, Mesiar R, Klement EP (2021). The impact on the properties of the EFGM copulas when extending this family. Fuzzy Sets and Systems 415:1-26

Schneider S, Demarqui FN, et al. (2020). An approach to model clustered survival data with dependent censoring. Biometrical Journal 62(1):157-174.

Shih JH (2014). Copula models, in Handbook of Survival Analysis, Chapter 24, pp.489-510, CRC Press.

Shih JH, Chang YT, Konno Y, Emura T (2019). Estimation of a common mean vector in bivariate meta-analysis under the FGM copula, Statistics 53(3): 673-95.

Shih JH, Emura T (2018). Likelihood-based inference for bivariate latent failure time models with competing risks under the generalized FGM copula, Computational Statistics 33(3): 1293-23

Shih JH, Emura T (2019). Bivariate dependence measures and bivariate competing risks models under the generalized FGM copula, Statistical Papers 60(4): 1101-18

Shih JH, Emura T (2021). On the copula correlation ratio and its generalization, Journal of Multivariate Analysis 182: 104708

Shinohara S, Lin YH, Michimae H, Emura T (2020) Dynamic lifetime prediction using a Weibull-based bivariate failure time model: a meta-analysis of individual-patient data, Communications in Statistics-Simulation and Computation, DOI:10.1080/03610918.2020.1855449

Sklar A (1959). Fonctions de répartition à n dimensions et leurs marges", Publ Inst Statist Univ Paris 8: 229-31.

Sofeu C, Emura T, Rondeau V (2021) A joint frailty-copula model for meta-analytic validation of failure time

surrogate endpoints in clinical trials, Biometrical Journal 63(2): 423-46

Sun LH, Huang XW, Alqawba, MS, Kim JM, Emura T (2020). Copula-based Markov Models for Time Series-Parametric Inference and Process Control, JSS Research Series in Statistics, Springer

Sundberg R (1974). On estimation and testing for the folded normal distribution. Communications in Statistics-Theory and Methods, 3(1): 55-72.

塚原英敦(2021). リスク解析における接合関数, 日本統計学会誌 第 51 巻(第 1 号): 101-121 (in Japanese).

Vaupel JW, Manton KG, Stallard E (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography, 16(3), 439-454.

Wang YC, Emura T (2021). Multivariate failure time distributions derived from shared frailty and copulas, Japanese Journal of Statistics and Data Science, doi:0.1007/s42081-021-00123-1

Wang YC, Emura T, Fan TH, Lo SM, Wilke RA (2020). Likelihood-based inference for a frailty-copula model based on competing risks failure time data, Quality and Reliability Engineering International 36(5):1622-38

Whitmore GA, Lee MLT. (1991). A multivariate survival distribution generated by an inverse Gaussian mixture of exponentials. Technometrics 33(1), 39-50.

Wu BH, Michimae H, Emura T (2020). Meta-analysis of individual patient data with semi-competing risks under the Weibull joint frailty-copula model. Computational Statistics 35:1525-52.

Wu K, Wang L, Yan L, Lio Y (2021). Statistical inference of left truncated and right censored data from Marshall–Olkin bivariate Rayleigh distribution. Mathematics, 9(21):2703

Zeng X, Gui W (2021). Statistical inference of truncated normal distribution based on the generalized progressive hybrid censoring. Entropy, 23(2), 186.

研究所報（最近刊行分）

号数	タイトル	刊行年月日
30	国連ミレニアム開発目標と統計	2003. 10. 20
31	Workshops on “the Population Censuses” and “the Use of Census Micro Data”	2003. 12. 20
32	ミクロデータとその利用	2004. 04. 20
33	International Symposia on Population Census and Micro Data Archives	2005. 01. 10
34	政府統計の二次的利用	2005. 04. 20
35	ジェンダー（男女共同参画）統計	2007. 02. 20
36	人口センサスの現状と新展開	2007. 04. 01
37	統計における官学連携	2007. 04. 20
38	ジェンダー（男女共同参画）統計　Ⅱ	2009. 02. 10
39	社会生活基本調査とその利用	2010. 01. 15
40	地方統計の現状と課題	2010. 09. 15
41	Exploring Potential of Individual Statistical Records	2011. 11. 05
42	観光統計	2013. 02. 05
43	国民経済計算関連統計の新たなる展開	2014. 01. 30
44	タウンページデータによる事業所立地分析	2014. 02. 15
45	フィンランドのビジネス・レジスター	2015. 03. 20
46	19 世紀ドイツ営業統計史研究	2015. 07. 20
47	地方統計と統計 GIS	2016. 01. 25
48	首都圏の人口移動	2017. 03. 10
49	宿泊業及び飲食業の実証分析	2018. 08. 01
50	サービス分野の生産物分類	2019. 01. 31
51	全市区町村産業連関表（平成 23 年表）の推計	2019. 10. 15
52	商業統計調査	2021. 01. 31
53	産業連関表から供給・使用表へ	2021. 03. 31

研　究　所　報　No. 54
2021 年 11 月　30 日

発行所　法政大学　日本統計研究所
〒194-0298 東京都町田市相原町 4342
Tel 042-783-2325,6
Fax 042-783-2332
jsri@adm.hosei.ac.jp
発行人　菅　幹雄

B U L L E T I N

OF

JAPAN STATISTICS RESEARCH INSTITUTE

No.54 November 2021

Statistical Modelling

CONTENTS

Foreword

Pair circulas modeling in higher-order Markov process on the circle
Hiroaki OGATA

On a construction of cylinder distribution
Tomoaki IMOTO

Mixture transition distribution modeling for
higher order circular Markov processes
Takayuki SHIOHAMA

Extended stochastic complexity for sparse estimation
via non-convex regularized regression
Yoichi MIYATA

Multivariate failure time distributions generated from a copula and mixing effects
Takeshi EMURA

Edited by
JAPAN STATISTICS RESEARCH INSTITUTE
HOSEI UNIVERSITY
TOKYO, JAPAN