

統計數理

Vol. 73, No.2

(通巻 142 号)

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

目 次

特集「生存時間解析と信頼性解析」

「特集 生存時間解析と信頼性解析」について	
江村 剛志・太田 修平	143
非一様マルコフ過程から見たソフトウェア信頼性モデリングの諸相 [原著論文]	
土肥 正・岡村 寛之	147
相関した競合リスクを伴う左側切断・右側打ち切りデータのベイズ流解析 [研究詳解]	
道前 洋史	167
従属打ち切りの下での治療効果推定のための要因計画・要因分析	
—コピュラに基づく方法— [研究詳解]	
江村 剛志・室谷 健太	189
天文学における生存時間解析：銀河の光度関数推定 [研究詳解]	
竹内 努	213
COM-Poisson 分布について—過小分散も過大分散も表現するポアソン分布— [総合報告]	
長塚 豪己	239
共役ガンマ事前分布を用いたワイブル分布のベイズ推定 [原著論文]	
作村 建紀・柳本 武美	253
信頼性解析とツールの活用 [研究詳解]	
岡村 寛之・鄭 俊俊・土肥 正	267
極値統計論に基づくモデリング [総合報告]	
吉田 拓真・北野 利一	287
スポーツ外傷予測への信頼性解析の応用：大相撲力士に対する実証分析 [研究詳解]	
太田 修平・木村 光宏	317
生存解析における樹木法と最近の発展 [総合報告]	
杉本 知之・丸尾 和司・下川 敏雄	335
System signature の計算方法とその応用	
—Consecutive- k -out-of- n :F システムを例として— [研究詳解]	
中村 太信・周 蕎	353

多様体学習を用いた銀河進化の新しい定量化 [研究詳解]

竹内 努・クレスチエータ・山形 大青・曹 愛奈・内田 舜也・池田 思朗・福水 健次・ 加納 龍生・大森 清顕クリストファ・馬 海霞・施文・松井 瀬奈	375
---	-----

2025年12月

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

〒190-8562 東京都立川市緑町10-3 電話 050-5533-8500(代)

本号の内容はすべて <https://www.ism.ac.jp/editsec/toukei/> からダウンロードできます

統
計
數
理

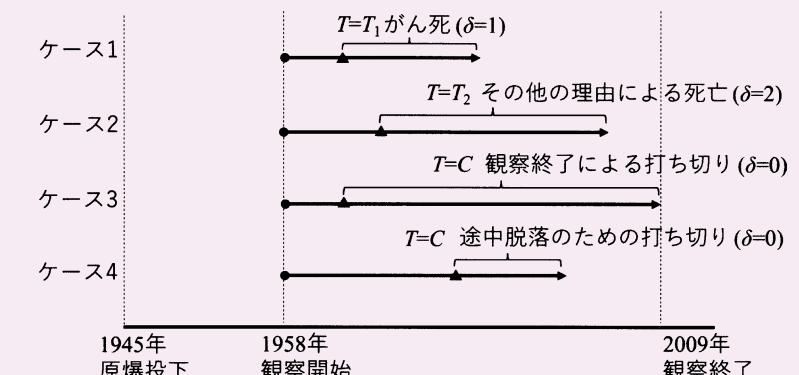
統計数理

Vol. 73, No.2

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

● : 登録日
▲ : がんの発症日
 T_1 : がんの発症からがん死までの時間
 T_2 : がんの発症からがん死以外の理由による死亡までの時間
 C : がんの発症から打ち切りまでの時間



第73巻 第2号

2025

統計数理研究所

ISSN 2760-2125

統計數理

(年2回発行)

編集委員長 小山 慎介

編集委員 庄 建倉

野間 久史

林 慶浩

Figueira Lourenço, Bruno

村上 隆夫

特集担当編集委員 江村 剛志 (広島大学)

太田 修平 (神奈川大学)

編集室

池田 広樹 川合 純華 長嶋 昭子

「統計數理」は、統計數理研究所の研究成果を掲載する機関誌「彙報」として1953年に創刊され、1985年に誌名を現在の「統計數理」に改めました。2025年からはオンラインジャーナルとして新たな形での発行を開始しています。本ジャーナルは、統計科学全般に関する論文を広く受け付け、統計科学の深化と発展、さらに統計科学を通じた社会への貢献を目指しています。

投稿を受け付けるのは、次の6種です。

- a. 原著論文
- b. 総合報告
- c. 研究ノート
- d. 研究詳解
- e. 統計ソフトウェア
- f. 研究資料

投稿された原稿は、編集委員会が選定・依頼した査読者の審査を経て、掲載の可否を決定します。投稿規程、執筆要項は、本誌最終頁をご参照ください。

また、上記以外にも統計科学に関して編集委員会が重要と認める内容について、編集委員会が原稿作成を依頼することがあります。

その他、「統計數理」に関するお問い合わせは、各編集委員にお願いします。

All communications relating to this publication should be addressed to associate editors of the Proceedings.

大学共同利用機関法人 情報・システム研究機構
統計數理研究所

〒190-8562 東京都立川市緑町10-3 電話 050-5533-8500(代)

<https://www.ism.ac.jp/>

© The Institute of Statistical Mathematics 2025

印刷：笠氣出版印刷株式会社

表紙の図は本誌168ページを参照

PROCEEDINGS OF THE INSTITUTE OF STATISTICAL MATHEMATICS

Vol. 73, No. 2

Contents

Special Topic: Survival and Reliability Analyses

On the Special Topic "Survival and Reliability Analyses"

Takeshi EMURA and Shuhei OTA 143

Some Aspects in Software Reliability Modeling from the Viewpoint of Non-homogeneous Markov Processes

Tadashi DOHI and Hiroyuki OKAMURA 147

Bayesian Analysis of Left-truncated Dependent Competing Risks Data

Hirofumi MICHIMAE 167

A Review of Factorial Survival Analysis for Treatment Effects under Dependent Censoring
—A Copula-based Approach—

Takeshi EMURA and Kenta MUROTANI 189

Survival Analysis in Astrophysics: Estimation of Galaxy Luminosity Function

Tsutomu T. TAKEUCHI 213

On COM-Poisson Distribution—An Extended Poisson Model for Under- and Over-dispersed Count Data—

Hideki NAGATSUKA 239

Bayesian Estimation under the Conjugate-gamma Prior for the Weibull Distribution

Takenori SAKUMURA and Takemi YANAGIMOTO 253

Reliability Analysis and Utilization of Tools

Hiroyuki OKAMURA, Junjun ZHENG and Tadashi DOHI 267

Statistical Modeling Using Extreme Value Theory

Takuma YOSHIDA and Toshikazu KITANO 287

Application of Reliability Analysis to Sports Injury Prediction: A Case Study for Sumo Wrestlers

Shuhei OTA and Mitsuhiro KIMURA 317

Tree-structured Approaches and Recent Advances in Survival Analysis

Tomoyuki SUGIMOTO, Kazushi MARUO and Toshio SHIMOKAWA 335

Computation of the System Signature and Their Application—Case of Consecutive- k -out-of- n :F Systems—

Taishin NAKAMURA and Lei ZHOU 353

Research Review

New Quantification of Galaxy Evolution with Manifold Learning

Tsutomu T. TAKEUCHI, Sucheta COORAY, Taisei D. YAMAGATA, Aina May So, Shun-Ya S. UCHIDA, Shiro IKEDA, Kenji FUKUMIZU, Ryusei R. KANO, Kiyoshi Christopher OMORI, Hai-Xia MA, Wen SHI and Sena A. MATSUI 375

December, 2025

Research Organization of Information and Systems

The Institute of Statistical Mathematics

10-3 Midori-cho, Tachikawa, Tokyo 190-8562, JAPAN

「特集 生存時間解析と信頼性解析」について

江村 剛志¹・太田 修平² (オーガナイザー)

生存時間解析と信頼性解析では、被験者の生存時間や製品の寿命などの「イベント発生時間」に関するデータ解析法やモデル構築法を扱う(武富・山本, 2023)。イベント発生時間を観測するには、被験者や製品の一定期間の追跡が必要であるから、観測の途中で、打ち切り・切断・競合リスクなどの現象が生じ、データは「不完全」となることが多い(Klein and Moeschberger, 2003; 宮村, 1982; 道前, 2023)。加えて、経時的なイベント発生機序をモデル化するために、点過程やハザード関数を用いることがある。このように、生存時間解析・信頼性解析では、イベント発生時間の構造に適合した固有の統計モデルや解析法が必要となる。

医学研究や製品開発で用いられる生存時間解析・信頼性解析の統計手法は古典的なものが殆どであったが、近年では最新の研究成果を用いた高度な統計手法を用いる実証研究も増えている。例えば、がんの治療効果を調べるために被験者から測定する「代替エンドポイント」の検証(江村・大庭, 2024)では、illness-death モデルや joint frailty-copula モデルなどの最新の生存時間解析のモデルが用いられている(Okui et al., 2024; 2025; Simon et al., 2025)。これら新手法の多くは複雑であり、ユーザーが手法を正しく使うには、その理論と方法の理解が必須である。本特集号で紹介する統計手法も新しいものが多く、それらは、理論と方法を正しく理解した上で実践されるべきである。

生存時間解析の研究は、統計学の中でも極めて活発な学術領域であり、現在も世界中の研究者が発展に貢献している。2021 年に統計関連学会連合が発行する国際誌 Japanese Journal of Statistics and Data Science(JJSD)では、諸外国から研究成果 17 報を集めた生存時間解析の特集が組まれた(Emura and Ha, 2021)。2025 年に発表された JJSD の Biostatistics 特集号(Ueki and Araki, 2025)では、18 報のうち 8 報が生存時間解析の研究であった(Cook and Lawless, 2025; Rouzbahani et al., 2025; Chiou et al. 2025, など)。邦文誌でも、2023 年には日本統計学会誌の「特集：生存時間解析・イベントヒストリー分析」で 11 報が出版、翌年 2024 年には計量生物学誌の「特集：生存時間データ解析法」で 13 報が出版された。しかしながら、「信頼性解析」に関する纏まった結果を報じた特集号はほとんど見あたらない。

そこで本特集では、生存時間解析と信頼性解析を併せ、統計学的なデータ解析法、統計推測法、統計的モデル化の第一線で活躍する研究者に最新の研究結果の紹介・総説(レビュー)・ソフトウェア解説を提供頂いた。専門家による厳密な査読の結果、以下で解説する合計 11 報が掲載となった。

統計モデルのパラメータ推定方法の研究は、統計学において重要なテーマであり、最尤法やベイズ法をはじめ、様々な方法が開発されている。長塚氏の論文では、生存時間解析・信頼性解析の両分野で注目されている「COM-Poisson 分布」を紹介し、パラメータ推定法(最小二乗法、最尤法、条件付き最尤法(Tomio and Nagatsuka, 2022)を詳解している。また、治癒モデルや

¹ 広島大学 情報科学部：〒739-8511 広島県東広島市鏡山一丁目 3 番 2 号; takeshiemura@gmail.com

² 神奈川大学 工学部：〒221-8686 神奈川県横浜市神奈川区六角橋三丁目 27 番 1 号; ota@kanagawa-u.ac.jp

確率過程モデルへの応用も解説している。作村・柳本氏の論文では、ワイブル分布の2つのパラメータに共役事前分布を導入した上で、適切なベイズ推定量を提案し、シミュレーションを用いて、提案する推定量の性能が優れていることを示している。竹内氏の論文では、天文学の観測データが観測装置の検出限界による「切断」を伴うことを指摘した上で、2変数光度関数の新しいノンパラメトリックな推定法を提案している。

複雑な打ち切り・切断・競合リスクを伴うデータは、適切なモデルや解析法を構築しなければバイアスのある結論を導く。道前氏の論文では、Michimae et al. (2024)の内容をベースに、競合リスクを伴う左側切・右側打ち切りのある生存時間データに基づくベイズ流解析法を紹介している。特に、競合リスク間の従属性を適切にモデル化することが、バイアスの無い結論を導くと報告している。江村・室谷氏の論文では、Emura et al. (2024)を基に、従属打ち切り下でのバイアスの無い治療効果の推定法を解説している。尚、これら論文で用いられる主要な統計モデルはコピュラであり、その理論体系は成書に纏められている (Nelsen, 2006; 江村, 2025)。尚、コピュラの信頼性解析における最近の話題は、太田 (2023)を参照、生存時間解析における話題は、杉本・田中 (2023), 江村 (2023), 道前 (2024), 中蘭 他 (2025)を参照されたい。

生存時間解析法を用いた予測モデルの構築は、今後の発展が期待される機械学習・データサイエンス分野で重要な役割を果たすと思われる。杉本・丸尾・下川氏の論文では、統計的機械学習ツールとして、広く応用されるようになっている決定木やフォレスト法などを生存データへ適用する手法に加え、最近の因果フォレストへの展開を報告している。太田・木村氏の論文では、スポーツ選手の外傷リスクを経時的に評価・予測するための点過程モデル (Ota and Kimura, 2023)を詳説し、大相撲力士に対する外傷予測の実証分析を行っている。

信頼性解析では、指数分布、対数正規分布、ガンマ分布、ワイブル分布、パレート分布などのパラメトリック分布を用いてきた (武富・山本, 2023)。極値統計学は、確率分布の裾に焦点を当て、稀な事象が生起する確率メカニズムを理解するための枠組みを提供する。吉田・北野氏の論文では、極値統計学の考え方や統計ソフト R のパッケージを紹介しながら、降雨量データへ極値統計モデルの適用例を示している。中村・周氏の論文では、Nakamura and Yamamoto (2022)を基に、システム全体の信頼性を個々のコンポーネントの信頼性から導出する際に必要となる system signature の算出方法について、石油パイプラインシステムなどに代表される Consecutive-k-out-of-n:F システムを例に詳説している。

ソフトウェアの信頼性評価は、機械装置などのハードウェアを対象とした信頼性評価と同様に重要なテーマである。また、信頼性解析では適切な統計解析ツールなどのソフトウェアの利用が必要となる。土肥・岡村氏の論文では、非一様マルコフ過程に着目し、ソフトウェア信頼性予測および関連問題を扱うための統一的アプローチを紹介し、実際のソフトウェアフォールトデータを用いてモデルの適合性と予測性能を明らかにしている。岡村・鄭・土肥氏の論文では、信頼性分野で用いられる代表的な信頼性評価手法と統計的手法を概説している。さらに、信頼性解析ツールについて、商用ソフトウェア、オープンソースソフトウェア、学術分野で開発されたツールなどを例に挙げ、それぞれの特徴や機能について解説している。

本特集号が、生存時間解析と信頼性解析の研究や実務に役立てば幸いである。

謝 辞

本特集号では、多数の専門家から建設的かつ重要なアドバイスを頂きながら編集にあたった。特に、道前洋史氏、古川恭治氏、杉本知之氏、井上真二氏、松本晃太郎氏、斎藤哲雄氏、西郷達彦氏、田中智大氏、木村光宏氏、その他多くの方々に感謝を申し上げたい。

参考文献

- Chiou, S. H., Su, C. L. and Lin, F. C. (2025). Fitting cure models using pseudo-observations approaches with the R package pseudoCure, *Japanese Journal of Statistics and Data Science*, **8**(1), 603–639.
- Cook, R. J. and Lawless, J. F. (2025). Methodological challenges in studying disease processes using observational cohort data, *Japanese Journal of Statistics and Data Science*, **8**(1), 323–345.
- 江村剛志 (2023). 生存時間の2標本問題—コピュラに基づく従属打ち切り問題への対処—, 日本統計学会誌, **52**(2), 295–317.
- 江村剛志 (2025). 『コピュラ理論の基礎』, コロナ社, 東京.
- Emura, T. and Ha, I. D. (2021). Special feature: Recent statistical methods for survival analysis, *Japanese Journal of Statistics and Data Science*, **4**(2), 889–894.
- 江村剛志, 大庭幸治 (2024). 生存時間変数に対する代替性評価—メタアナリシスアプローチ—, 計量生物学, **45**(1), 67–85.
- Emura, T., Ditzhaus, M., Dobler, D. and Murotani, K. (2024). Factorial survival analysis for treatment effects under dependent censoring, *Statistical Methods in Medical Research*, **33**(1), 61–79.
- Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York.
- 道前洋史 (2023). 競合リスクを伴う左側切断・右側打ち切りデータの解析—現状と今後の展望について—, 日本統計学会誌, **52**(2), 203–220.
- 道前洋史 (2024). コピュラで構成した同時事前分布に基づくベイズリッジ推定量—Cox回帰モデルを例に—, 計量生物学, **45**(1), 15–35.
- Michimae, H., Emura, T., Miyamoto, A. and Kishi, K. (2024). Bayesian parametric estimation based on left-truncated competing risks data under bivariate Clayton copula models, *Journal of Applied Statistics*, **51**(13), 2690–2708.
- 宮村鍛夫 (1982). 競合モデルに関する統計的手法, オペレーションズ・リサーチ, **27**, 871–876.
- Nakamura, T. and Yamamoto, H. (2022). Algorithm for computing the system signatures of linear and circular connected-(1, 2)-or-(2, 1)-out-of-(m, n):F lattice systems, *International Journal of Reliability, Quality and Safety Engineering*, **29**(04), <https://doi.org/10.1142/S0218539321410072>.
- 中薗孝輔, 魚住龍史, 江村剛志 (2025). パラメトリックコピュラモデルを用いた相関のある生存時間データにおけるMann-Whitney効果の感度分析, 日本統計学会誌, **55**(1), 85–113.
- Nelsen, R. B. (2006). *An introduction to Copulas: Second Edition*, Springer, New York.
- Okui, J., Nagashima, K., Matsuda, S., et al. (2024). Recurrence-free survival as a surrogate endpoint for overall survival after neoadjuvant chemotherapy and surgery for oesophageal squamous cell carcinoma, *British Journal of Surgery*, **111**(2), <https://doi.org/10.1093/bjs/znae038>.
- Okui, J., Matsuda, S., Nagashima, K., et al. (2025). Recurrence-free survival as a surrogate endpoint for overall survival in resectable esophageal cancer: Integrated analysis of individual patient data from phase III trials, *Annals of Surgery*, <https://doi.org/10.1097/SLA.0000000000006919>.
- 太田修平 (2023). 信頼性データ解析への多変量 Farlie-Gumbel-Morgenstern コピュラの応用, 日本統計学会誌, **52**(2), 177–201.
- Ota, S. and Kimura, M. (2023). Statistical injury prediction for professional sumo wrestlers: Modeling and perspectives, *PLoS one*, **18**(3), <https://doi.org/10.1371/journal.pone.0283242>.
- Rouzbahani, M., Akhoond, M. R. and Chinipardaz, R. (2025). A new bivariate survival model with a cured fraction: A mixed Poisson frailty–copula approach, *Japanese Journal of Statistics and Data Science*, **8**(1), 367–391.
- Simon, F., Ligtvoet, R., Robrecht, S., et al. (2025). Endpoint surrogacy in first-line chronic lymphocytic leukemia, *Journal of Clinical Oncology*, **43**(4), 381–391.
- 杉本知之, 田中健太 (2023). 2変量生存時間モデルにおけるコピュラとその利用, 日本統計学会誌, **52**(2),

153–176.

武富奈菜美, 山本和嬉 (2023). 生存時間解析・信頼性解析のための統計モデル, 日本統計学会誌, **52**(2), 69–112.

Tomio, Y. and Nagatsuka, H. (2022). A conditional maximum likelihood estimation of the COM-Poisson distribution and its uniqueness and existence, *Total Quality Science*, **7**(3), 137–148.

Ueki, M. and Araki, Y. (2025). Special feature: Recent advances in biostatistics, *Japanese Journal of Statistics and Data Science*, **8**(1), 319–322.

非一様マルコフ過程から見たソフトウェア信頼性 モデリングの諸相

土肥 正[†]・岡村 寛之[†]

(受付 2024 年 11 月 30 日；改訂 2025 年 2 月 21 日；採択 2 月 27 日)

要　　旨

本稿では、ソフトウェア信頼性モデリングの進展を体系的に整理し、非一様マルコフ過程(NHMP)を基盤とした統一的なアプローチについて外観する。従来の一様マルコフ過程(Homogeneous Markov Process: HMP)および非一様ポアソン過程(Non-homogeneous Poisson Process: NHPP)モデルを一般化し、一般化 2 項過程(Generalized Binomial Process: GBP)や一般化ポリア過程(Generalized Polya Process: GPP)といった柔軟性の高い新しいモデルを紹介する。さらに、ソフトウェアのプロダクトおよびプロセスメトリクスを活用した信頼性予測モデルについても論じ、その有用性を検証する。これらの実証結果は、実際のソフトウェア開発における信頼性評価に寄与するものであり、将来のより精緻なモデル構築やソフトウェア信頼性評価へ適用可能性を広げる。

キーワード：ソフトウェア信頼性モデル、非一様マルコフ過程、一般化 2 項過程、一般化ポリア過程、ソフトウェアメトリクス、信頼性予測。

1. はじめに

過去 50 年間の長きにわたり、多くのソフトウェア信頼性モデルがソフトウェア工学の分野において開発されてきた (Lyu, 1996; Musa et al., 1987; Xie, 1991)。特に、一様マルコフ過程(Homogeneous Markov Process: HMP)に基づくソフトウェア信頼性モデルと非一様ポアソン過程(Non-homogeneous Poisson Process: NHPP)に基づくソフトウェア信頼性モデルは、テスト段階で検出されるソフトウェアフォールトの累積数の確率的挙動を記述する扱いやすさとフォールトデータへの適合性の良さから、これまで広く利用されてきた。ソフトウェアフォールトデータは、フォールト検出時間(時間領域)データと時間間隔(グループ)データに分類される。時間領域データでは各イベントの発生時刻が記録されており、統計的完全データとみなされる。一方、グループデータは観測点ごとに発生したイベントの累積数を記録したもので、正確なイベント発生時刻が不明であるため、統計的不完全データとされる。実際、ソフトウェア開発のシステムテスト段階で観測されるフォールトデータの多くはグループデータであり、分散開発環境下でソフトウェアをテスト・デバッグすることが一般的である。

時間領域データを取り扱う際、基礎となるフォールト検出過程が広いクラスの単調非減少確率点過程に従う場合であっても、尤度関数を閉形式で求めることが比較的容易であり、最尤推

[†] 広島大学 大学院先進理工系科学研究科：〒739-8527 広島県東広島市鏡山 1-4-1; dohi@hiroshima-u.ac.jp, okamu@hiroshima-u.ac.jp

定やベイズ推定を含む統計的推論を行うことが可能となる。一方、グループデータを扱う際には、イベント累積数の確率法則を把握し、累積フォールト数に関する状態空間上で尤度関数を構築する必要がある。これにより、多くの推定可能なソフトウェア信頼性モデルは、いわゆるマルコフ性を持つモデルとして記述されてきた。例えば、Jelinski and Moranda (1972), Moranda (1979), Xie (1989) による古典的なソフトウェア信頼性モデルは、マルコフ純逆死滅過程およびマルコフ純出生過程として記述され、異なる状態依存遷移率をもつ HMP として分類される。一方、NHPP に基づくソフトウェア信頼性モデル (Abdel-Ghaly et al., 1986; Achcar et al., 1998; Erto et al., 2020; Goel and Okumoto, 1979; Littlewood, 1984; Goel, 1985; Gokhale and Trivedi, 1998; Musa and Okumoto, 1984; Ohba, 1984; Ohishi et al., 2009; Okamura et al., 2013; Xiao, 2015; Yamada et al., 1983; Zhao and Xie, 1996) では、状態とは独立で時間依存遷移率をもつマルコフ過程として分類される。

一方、ソフトウェア工学の分野では、上記のようなフォールトデータだけを扱うのではなく、ソフトウェアの開発中に観測可能な特徴量(共変量)に着目した研究も数多く行われている。ソフトウェアの特徴量としては、ソフトウェア規模、複雑性、品質、プロセス、環境などが挙げられ、それらはソフトウェアメトリクスと呼ばれる。これらのメトリクスは、ソフトウェアの信頼性に大きく影響を与える要因であり、ソフトウェアの信頼性予測において重要な役割を果たすと考えられている。例えば、ソフトウェア規模はソフトウェア信頼性に影響を与える主要な要因であり、ソフトウェアの規模が大きくなるとソフトウェアの信頼性が低下する傾向にあることは直感的に明らかである。このような外的要因を考慮したソフトウェア信頼性モデルは、ソフトウェアの信頼性予測においてより現実的な解を提供することが期待され、実証ソフトウェア工学における中心的な役割を演じると言っても過言ではない。Rinsaka et al. (2006) や Li et al. (2022) は Lawless (1987) の比例強度モデルを適用したメトリクスに基づくソフトウェア信頼性モデルを提案しており、Shibata et al. (2006), Kuwa and Dohi (2013a) は Cox 比例ハザードモデルの枠組みを NHPP モデルに適用している。また、Okamura et al. (2010), Okamura and Dohi (2014), Ikemoto et al. (2013), Kuwa and Dohi (2013b), Okamura and Dohi (2015) はロジスティック回帰やポアソン回帰などの一般化線形モデルへの拡張を提案している。

本稿は二つの話題で構成される。一つ目は、HMP および NHPP に基づくソフトウェア信頼性モデルの一般化である非一様マルコフ過程(Non-homogeneous Markov Process: NHMP)に着目し、ソフトウェア信頼性予測および関連問題を扱うための統一的アプローチについて議論する。最近、Li et al. (2023) は NHMP の 2 つのサブクラスである一般化 2 項過程(Generalized Binomial Process: GBP)および一般化ポリア過程(Generalized Polya Process: GPP)に注目し、GBP と GPP がそれぞれ状態と時間に依存した遷移率を持つマルコフ逆死滅過程(Shanthikumar, 1981)とマルコフ出生過程(Kendall, 1948; Shaked et al., 2002)となり、既存のソフトウェア信頼性モデルの多くを含む汎用的モデルであることを示した。過去の研究において、Shanthikumar (1981) は Goel and Okumoto (1979) による最も古典的な指数形 NHPP モデルと Jelinski and Moranda (1972) による HMP モデルによって記述されたソフトウェア信頼性モデルが GBP モデルに含まれる事實を示しているが、NHMP モデルの汎用性や柔軟性について理論的にも実証的にも十分な議論がなされていなかった。本稿の前半では、GBP および GPP に基づく統一的なモデリングの枠組みを紹介し、ソフトウェア信頼性モデルにおけるそれらの特性を概観するとともに、実際のソフトウェアフォールトデータを用いて適合性と予測性能を調査する。具体的には、NHPP, GBP, GPP によって記述されるソフトウェア信頼性モデルを比較し、それらの適合性および予測性能を評価する。特に、Okamura and Dohi (2008) で扱われた NHPP を一般化した非一様な混合ポアソン過程(Non-homogeneous Mixed Poisson Process: NHMPP)に関する新たな知見として、单一プロジェクトデータのみを用いたソフト

ウェア信頼性評価において、NHMPP は NHPP よりも常に適合性の面で劣ることが確認された。本稿の後半では、ソフトウェアメトリクスに着目し、ソフトウェアの信頼性予測におけるメトリクス活用の重要性について議論する。特に、Okamura and Dohi (2015) によって導入された一般化線形モデル(GLM)に基づくソフトウェア信頼性モデルを紹介し、文献 Okamura and Dohi (2015) で取り扱ったデータを含むより多くのデータを用いた検証を行い、ソフトウェア信頼性予測におけるメトリクスモデルの有用性を示す。

2. ソフトウェア信頼性の基本モデル

2.1 HMP によるモデリング

ソフトウェアテストの最終段階であるシステムテストにおいて、時刻 t までに検出されたフォールト数を $\{N(t), t \geq 0\}$ とする。いま、システムテストにおけるフォールト検出過程に対して、次の仮定を設ける。

仮定 A. システムテスト実施前にプログラム中に含まれる総フォールト数 N は一定かつ有限である。

仮定 B. 各々のフォールトは互いに独立に検出され、フォールト検出時間は同一の連続形確率分布関数 $F(t)$ をもつ非負の確率変数である。検出されたフォールトは直ちに修正・除去される。

仮定 A と仮定 B の下で、時刻 t までに検出される累積フォールト数は、以下の 2 項分布

$$(2.1) \quad P(N(t) = k \mid N) = \binom{N}{k} F(t)^k (1 - F(t))^{N-k}$$

によって記述される。このとき、確率過程 $N(t)$ は連続時間離散状態マルコフ過程となり、特に時刻 t においてソフトウェアに内在する残存フォールト数 $N - N(t)$ は状態数 0 で吸収状態となる死滅過程 ($N(t)$ は状態数 N で吸収状態となる出生過程) と等価となる。

いま、マルコフ過程の定常推移確率を

$$(2.2) \quad P_k(t) = \Pr\{N(t) = k \mid N(0) = 0\}, \quad k = 0, 1, 2, \dots$$

とし、各状態への推移率を λ_k ($k = 0, 1, 2, \dots$) のように表す。ソフトウェア信頼性で最も初期のモデルである Jelinski and Moranda モデル (Jelinski and Moranda, 1972) では、フォールト検出時間分布にパラメータ b (> 0) の指数分布 $F(t) = 1 - \exp(-bt)$ を仮定しており、確率過程 $N(t)$ は時間に関して一定の推移率をもつ連続時間マルコフ連鎖に帰着される。この連続時間マルコフ連鎖の推移率は $\lambda_k = (N - k)b$ ($k = 0, 1, \dots, N - 1$) によって与えられる。この事実は、定常推移確率に関するコルモゴロフの前進方程式(微分差分方程式)

$$(2.3) \quad \frac{d}{dt} P_0(t) = -\lambda_0 P_0(t),$$

$$(2.4) \quad \frac{d}{dt} P_k(t) = \lambda_{k-1} P_{k-1}(t) - \lambda_k P_k(t), \quad k = 1, 2, \dots, N - 1,$$

$$(2.5) \quad \frac{d}{dt} P_N(t) = \lambda_{N-1} P_{N-1}(t)$$

を初期条件 $P_0(0) = 1, P_k(0) = 0$ ($k = 1, \dots, N$) の下で解くことにより容易に確認できる。すなわち、

$$(2.6) \quad P_k(t) = P(N(t) = k \mid N) = \binom{N}{k} \{1 - e^{-bt}\}^k e^{-b(N-k)t}, \quad k = 0, 1, \dots, N$$

となる。

仮定 A と仮定 B とは離れて、オープンソースソフトウェアのようにシステムの運用期間中においてもなお、フォールトの作りこみが発生するような状況を考える。一般の推移率 λ_k が $\sum_{k=0}^{\infty} \lambda_k^{-1} = \infty$ を満たすとき、定常推移確率は

$$(2.7) \quad P_k(t) = \lambda_{k-1} e^{-\lambda_k t} \int_0^t e^{\lambda_k x} P_{k-1}(x) dx$$

を満足する純出生過程となる。Moranda (1979) は推移率を $\lambda_k = ab^k$ ($k = 0, 1, \dots$) のように仮定し、geometric de-eutrophication モデルと呼ばれる異なる HMP モデルを提案しており、Boland and Singh (2003) はその定常推移確率 $P_k(t)$ の無限級数表現を与えている。

2.2 NHPP によるモデリング

システムテストの長さ如何にかかわらず、テスト前にソフトウェア内に残存する初期フォールト数 N は未知であるので、 N を非負の整数值確率変数とみなすことが可能である。すなわち、 N の(混合比分布と呼ばれる)確率関数を $P(N = n) = \pi_N(n|\cdot)$ とすれば、 $N(t)$ の確率関数は

$$(2.8) \quad P(N(t) = k) = \sum_{n=k}^{\infty} P(N(t) = k|n) \pi_N(n|\cdot)$$

のような混合分布によって与えられる。特別な場合として、

$$(2.9) \quad \pi_N(n|\omega) = \frac{\omega^n e^{-\omega}}{n!}$$

のようなパラメータ ω (> 0) のポアソン分布を仮定すれば、確率過程 $N(t)$ の確率関数は

$$(2.10) \quad P(N(t) = k) = \frac{(\omega F(t))^k}{k!} e^{-\omega F(t)}$$

となり、平均値関数 $E[N(t)] = \Lambda(t) = \omega F(t)$ の NHPP となる。ここで、確率分布 $F(t)$ は t に関して非減少関数であるので、 $\lim_{t \rightarrow \infty} \Lambda(t) = \omega < \infty$ となり、平均値関数 $\Lambda(t)$ は上に有界となる。

Moranda (1979) による geometric de-eutrophication モデルと同様に、 $\lim_{t \rightarrow \infty} \Lambda(t) \rightarrow \infty$ を満たす非有界な平均値関数をもつ NHPP モデルはいくつか知られており、Cretois and Gaudoin (1998)、Duane (1964)、Littlewood (1984) による幂法則モデルや Musa and Okumoto (1984)、Musa et al. (1987) による対数ポアソン実行時間モデルがソフトウェア信頼性評価においてもよく使われている。平均値関数の有界性に関するこれら二つのモデルの違いは、無限長のテスト期間や運用期間を想定したとき、前者では累積フォールト数の確率関数がパラメータ ω のポアソン分布 $\lim_{t \rightarrow \infty} P_n(t) = \omega^n \exp(-\omega)/n!$ となるのに対し、後者は確率 1 で発散する ($\lim_{t \rightarrow \infty} N(t) \rightarrow \infty$) ことである。実際、 $P(N(t) \leq k) = P(T_k \geq t)$ を満たす各イベントの発生時間を表す確率変数 T_k ($k = 0, 1, 2, \dots; T_0 = 0$) を k 番目のフォールトが検出される時間(到着時間)とすれば、有界な平均値関数をもつ NHPP の到着時間分布は defective となり、任意の k に対して確率変数 T_k の有限次元モーメントは発散する。これより、平均値関数 $\Lambda(t) = \omega F(t)$ をもつ NHPP では、平均フォールト検出時間 $E[T_k]$ を評価することは意味をなさないと言える。

表 1において、代表的なフォールト検出時間分布に対応した NHPP モデルを整理する。Okamura and Dohi (2013) では、表 1 の 11 種類の NHPP モデルを実装したソフトウェア信頼性評価ツール SRATS (Software Reliability Assessment Tool on the Spreadsheet) を開発している。このように、NHPP モデルの表現能力はフォールト検出時間分布 $F(t)$ の表現能力に依存

表 1. 代表的な NHPP モデル。

Models	$\Lambda(t)$ ($\omega > 0, a > 0, b > 0, c > 0$)
Exponential dist. (Exp) Goel and Okumoto (1979)	$\Lambda(t) = \omega F(t)$ $F(t) = 1 - e^{-bt}$
Gamma dist. (Gamma) Yamada et al. (1983), Zhao and Xie (1996)	$\Lambda(t) = \omega F(t)$ $F(t) = \int_0^t \frac{c^b s^{b-1} e^{-cs}}{\Gamma(b)} ds$
Pareto dist. (Pareto) Abdel-Ghaly et al. (1986)	$\Lambda(t) = \omega F(t)$ $F(t) = 1 - \left(\frac{b}{t+b}\right)^c$
Truncated normal dist. (Tnorm) Okamura et al. (2013)	$\Lambda(t) = \omega \frac{F(t) - F(0)}{1 - F(0)}$ $F(t) = \frac{1}{\sqrt{2\pi}b} \int_{-\infty}^t e^{-\frac{(s-c)^2}{2b^2}} ds$
Log-normal dist. (Lnorm) Achcar et al. (1998), Okamura et al. (2013)	$\Lambda(t) = \omega F(\ln t)$ $F(t) = \frac{1}{\sqrt{2\pi}b} \int_{-\infty}^t e^{-\frac{(s-c)^2}{2b^2}} ds$
Truncated logistic dist. (Tlogist) Ohba (1984)	$\Lambda(t) = \omega F(t)$ $F(t) = \frac{1 - e^{-bt}}{1 + ce^{-bt}}$
Log-logistic dist. (Llogist) Gokhale and Trivedi (1998)	$\Lambda(t) = \omega F(\ln t)$ $F(t) = \frac{(bt)^c}{1 + (bt)^c}$
Truncated extreme-value max dist. (Txvmax) Ohishi et al. (2009)	$\Lambda(t) = \omega \frac{F(t) - F(0)}{1 - F(0)}$ $F(t) = e^{-e^{-\frac{t-c}{b}}}$
Log-extreme-value max dist. (Lxvmax) Ohishi et al. (2009)	$\Lambda(t) = \omega F(\ln t)$ $F(t) = e^{-(\frac{t}{b})^{-c}}$
Truncated extreme-value min dist. (Txvmin) Ohishi et al. (2009)	$\Lambda(t) = \omega \frac{F(0) - F(t)}{F(0)}$ $F(t) = e^{-e^{-\frac{t-c}{b}}}$
Log-extreme-value min dist. (Lxvmin) Goel (1985)	$\Lambda(t) = \omega (1 - F(-\ln t))$ $F(t) = e^{-e^{-\frac{t-c}{b}}}$
Logarithmic Poisson (Log) Musa and Okumoto (1984), Musa et al. (1987)	$\Lambda(t) = a \ln(1 + bt)$
Power-law (Plaw) Cretois and Gaudoin (1998), Duane (1964), Littlewood (1984)	$\Lambda(t) = at^b$
Generalized truncated logist. dist. (Gtlogist) Erto et al. (2020), Xiao (2015)	$\Lambda(t) = \omega \frac{1 - e^{-at^c}}{1 + be^{-at^c}}$ $F(t) = \frac{1 - e^{-at^c}}{1 + be^{-at^c}}$

するため、より広いクラスの確率分布を考えることは有効である。Okamura and Dohi (2016) は $F(t)$ に連続形相型分布を導入することで、汎用的な NHPP モデルの適用可能性について言及している。Dohi et al. (2024) は多項式近似を導入することで、異なる観点から NHPP モデルを一般化している。

式(2.8)の混合分布表現において、初期フォールト数 N の分布関数 $\pi_N(n|\cdot)$ にポアソン分布以外の確率分布を仮定することで、NHPP モデルとは異なる時間依存遷移率をもつマルコフ過程を構成することが可能である。例えば、混合比分布の確率関数が

$$(2.11) \quad \pi_N(n|\gamma_1, \gamma_2) = \binom{n + \gamma_1 - 1}{n} \left(\frac{\gamma_2}{\gamma_2 + 1} \right)^{\gamma_1} \left(\frac{1}{\gamma_2 + 1} \right)^n$$

のような負の 2 項分布に従うものとすれば、 $N(t)$ の確率関数は

$$(2.12) \quad P(N(t) = k) = \binom{k + \gamma_1 - 1}{k} \left(\frac{\gamma_2}{\gamma_2 + F(t)} \right)^{\gamma_1} \left(\frac{F(t)}{\gamma_2 + F(t)} \right)^k$$

となり、ポリア-ランドバーグ過程の一般化でもある非一様な混合ポアソン過程(Non-homogeneous Mixed Poisson Process: NHMPP)となる。このように、異なる混合比分布を仮定することで様々な非一様点過程を導出することができるが、次節で明らかになるように、結果として統計的に重要な意味をなさないモデルになることが理論的に示される。上述の結果は、式(2.10)で与えられるポアソン分布のパラメータ ω が確率密度関数 $dG(t | \theta_G)/dt = \gamma_2^{\gamma_1} \omega^{\gamma_1-1} \exp(-\gamma_2 \omega) / (\gamma_1 - 1)!$ ($\theta_G = (\gamma_1, \gamma_2)$) のガンマ分布に従う場合に相当する。

2.3 NHMP によるモデリング

次に、一般化されたソフトウェア信頼性モデルについて紹介する。マルコフ過程の推移率が状態 k と時刻 t の両方に依存する場合、すなわち、 $\lambda_k(t)$ ($k = 0, 1, 2, \dots$) の場合について考える。このような計数過程は非一様マルコフ過程(Non-homogeneous Markov process: NHMP)と呼ばれ、HMP モデルと NHPP モデルを特殊な場合として包含することに注意する。先述のモデルと同様に、Li et al. (2023) によって示されたように、NHMP モデルにおいても吸収状態をもつ場合ともたない場合を考え、以下のような 2 種類の推移率を考える。

$$(2.13) \quad \lambda_k(t) = (\beta - \alpha k) \kappa(t), \quad k = 0, 1, 2, \dots, \gamma_{\alpha, \beta},$$

$$(2.14) \quad \lambda_k(t) = (\alpha k + \beta) \kappa(t), \quad k = 0, 1, 2, \dots$$

ここで、 $\alpha (\geq 0)$ と $\beta (\geq 0)$ は非負の実数パラメータであり、 $\kappa(t)$ は t に関する任意の連続関数、 $\gamma_{\alpha, \beta} = \lceil \beta/\alpha \rceil$ であり、 $\lceil x \rceil$ は x を超えない整数値を示す。本稿では、 $\kappa(t)$ と $\Lambda(t) = \int_0^t \kappa(x) dx$ をそれぞれベースライン強度関数 (baseline intensity function)、累積ベースライン強度関数 (cumulative baseline intensity function) と呼ぶことにする。式 (2.13) と式 (2.14) で与えられる線形推移率は、それぞれ Kendall (1948) と Konno (2010) によって導入され、対応する NHMP は一般化 2 項過程 (generalized binomial process: GBP) 並びに一般化ポリア過程 (generalized Polya process: GPP) と呼ばれる。Shanthikumar (1981, 1983) は $\lambda_k(t) = (\beta - \alpha k) \kappa(t)$ ($n = 0, 1, 2, \dots, \gamma_{\alpha, \beta}$) および $\kappa(t) = \omega b \exp(-bt)$ となる逆死滅過程タイプの NHMP のコルモゴロフ前進方程式

$$(2.15) \quad \frac{d}{dt} P_0(t) = -\beta \kappa(t) P_0(t),$$

$$(2.16) \quad \frac{d}{dt} P_k(t) = \{\beta - \alpha(k-1)\} \kappa(t) P_{k-1}(t) - \{\beta - \alpha k\} \kappa(t) P_k(t), \\ k = 0, 1, \dots, \gamma_{\alpha, \beta}$$

$$(2.17) \quad \frac{d}{dt} P_{\gamma_{\alpha, \beta}}(t) = \{\beta - \alpha(\gamma_{\alpha, \beta} - 1)\} \kappa(t) P_{\gamma_{\alpha, \beta}-1}(t)$$

を初期条件 $P_0(0) = 1, P_k(0) = 0$ ($k = 1, 2, \dots, \gamma_{\alpha, \beta}$) の下で解き、定常推移確率が

$$(2.18) \quad P_k(t) = \binom{\gamma_{\alpha, \beta}}{k} A(t)^{\gamma_{\alpha, \beta}-k} (1 - A(t))^k, \quad k = 0, 1, \dots, \gamma_{\alpha, \beta}$$

によって与えられることを示した。ここで、

$$(2.19) \quad A(t) = e^{-\alpha \Lambda(t)},$$

$$(2.20) \quad \Lambda(t) = \int_0^t \kappa(x) dx$$

であり、式(2.18)の確率分布が平均と分散

$$(2.21) \quad E[N(t)] = \gamma_{\alpha,\beta} \{1 - e^{-\alpha\Lambda(t)}\},$$

$$(2.22) \quad \text{Var}[N(t)] = \gamma_{\alpha,\beta} \{1 - e^{-\alpha\Lambda(t)}\} e^{-\alpha\Lambda(t)}$$

をもつ2項分布となることを示した。式(2.21)と式(2.22)から、 $E[N(t)] > \text{Var}[N(t)]$ となることがわかり、GBPモデルは常にunder-dispersionとなる。式(2.22)の分散は時間 t の減少関数となるため、テスト時間が経過するとともに不確実性が減少することが理解できる。特に、 $\alpha = 0$ および $\kappa(t) = 1$ のとき、式(2.18)の確率関数は強度関数 $\kappa(t)$ をもつNHPPモデルとJelinski and Moranda(1972)によるHMPモデルに帰着される。表1の $\Lambda(t)$ を式(2.20)に代入することで、NHPPモデルに対応したGBPモデルを構築することが可能となる。

NHPPモデルと同様に、累積ベースライン強度関数 $\Lambda(t)$ が有界、すなわち $\lim_{t \rightarrow \infty} \Lambda(t) = \omega$ のとき、 $\lim_{t \rightarrow \infty} E[N(t)] = \gamma_{\alpha,\beta} \{1 - \exp(-\alpha\omega)\}$ および $\lim_{t \rightarrow \infty} \text{Var}[N(t)] = \gamma_{\alpha,\beta} \exp(-\alpha\omega) \{1 - \exp(-\alpha\omega)\}$ を確認することができる。逆に、 $\lim_{t \rightarrow \infty} \Lambda(t) \rightarrow \infty$ ならば、 $\lim_{t \rightarrow \infty} E[N(t)] = \gamma_{\alpha,\beta}$ 、 $\lim_{t \rightarrow \infty} \text{Var}[N(t)] = 0$ となる。GBPモデルがサンプルパスの意味で $N(t) = \gamma_{\alpha,\beta}$ において吸収状態をもつことに注意すれば、有界な累積ベースライン強度関数 $\Lambda(t)$ において $\lim_{t \rightarrow \infty} P_{\gamma_{\alpha,\beta}}(t) = \{1 - \exp(-\alpha\omega)\}^{\gamma_{\alpha,\beta}}$ が成立し、無限長のシステムテスト期間においてソフトウェアに残存するすべてのフォールトが検出され得ない正の確率が存在することになる。さらに、GBPモデルをNHMPとして眺めたとき、任意の状態 k から $k+1$ ($k = 0, 1, \dots, \gamma_{\alpha,\beta} - 1$)に推移するまでの時間分布はdefectiveとなり、確率過程 $N(t)$ は漸近的に $\gamma_{\alpha,\beta}$ に収束する。よって、フォールト検出の観点から $\gamma_{\alpha,\beta} \{1 - \exp(-\alpha\omega)\}$ は検出可能な正味累積フォールト数として解釈することができる。一方、累積ベースライン強度関数 $\Lambda(t)$ が非有界のとき、大数の法則により、 $N(t)$ は $\gamma_{\alpha,\beta}$ に収束することがわかる。

GBPによるモデル化の枠組を理解すると、これまでに詳細に調べられることのなかった既存のソフトウェア信頼性モデルの構造を容易に理解することが可能となる。例えば、表1の幂法則タイプの平均値関数 $\Lambda(t) = at^b$ を考えたとき、 $\kappa(t) = (a/(1+b))t^{b+1}$ および $\lambda_k(t) = (\beta - \alpha k)(a/(1+b))t^{b+1}$ となるので、 $b' = 1 + b$ 、 $\beta' = (a/b')\beta$ 、 $\alpha' = (a/b')\alpha$ のような変換を行うと $\lambda_k(t) = (\beta' - \alpha' k)t^{b'}$ を得る。特に $b' = 2$ ($b = 1$)、 $b' > 2$ ($b > 1$)のとき、このNHMPモデルは、それぞれSchick and Wolverton(1978)とWagoner(1973)によるソフトウェア信頼性モデルと一致する。また、Littlewood(1980, 1981)はベイズ推論の枠組において、パレートタイプのフォールト検出時間分布を扱い、推移率が $\lambda_k(t) = (\beta - \alpha k)/(t + \gamma)$ ($k = 0, 1, 2, \dots, \gamma_{\alpha,\beta}$)になるモデルを提案しており、 $\kappa(t) = (t + \gamma)^{-1}$ および $\Lambda(t) = \ln(1 + t/\gamma)$ のGBPとなる。このモデルはMusa and Okumoto(1984)による対数ポアソン実行時間モデルの拡張になっており、期待累積フォールト数は $E[N(t)] = \gamma_{\alpha,\beta} \{1 - \gamma(\gamma + t)^{\alpha}\}$ によって与えられる。

次に、 $\lambda_k(t) = (\alpha k + \beta)\kappa(t)$ ($k = 0, 1, 2, \dots$)のGPPについて考えよう。Konno(2010)は出生過程タイプのNHMPとして以下のようなコルモゴロフの前進方程式

$$(2.23) \quad \frac{d}{dt} P_0(t) = -\beta\kappa(t)P_0(t),$$

$$(2.24) \quad \frac{d}{dt} P_k(t) = \{\alpha(k-1) + \beta\}\kappa(t)P_{k-1}(t) - \{\alpha k + \beta\}\kappa(t)P_k(t), \quad k = 1, 2, \dots$$

を初期条件 $P_0(0) = 1, P_n(0) = 0$ の下で解き、定常推移確率

$$(2.25) \quad P_n(t) = \binom{n + \frac{\beta}{\alpha} - 1}{n} A(t)^{\frac{\beta}{\alpha}} (1 - A(t))^n, \quad n = 0, 1, \dots$$

を導出し、平均と分散がそれぞれ

$$(2.26) \quad E[N(t)] = \gamma_{\alpha,\beta} \{e^{\alpha\Lambda(t)} - 1\},$$

$$(2.27) \quad \text{Var}[N(t)] = \gamma_{\alpha,\beta} e^{\alpha\Lambda(t)} \{e^{\alpha\Lambda(t)} - 1\}$$

となる負の 2 項分布に従うことを示した。Konno (2010) と Gat (2016) はこの確率過程をそれぞれ一般化ポリア過程、一般化ユール過程と呼んでいるが、本稿では統一的に GPP と略記する。式 (2.26) と式 (2.27) から、 $E[N(t)] < \text{Var}[N(t)]$ となり、GBP は over-dispersion となる。累積ベースライン強度関数 $\Lambda(t)$ について $\lim_{t \rightarrow \infty} \Lambda(t) = \omega$ が成立すれば、 $\lim_{t \rightarrow \infty} E[N(t)] = \gamma_{\alpha,\beta} \{\exp(\alpha\omega) - 1\}$ および $\lim_{t \rightarrow \infty} \text{Var}[N(t)] = \gamma_{\alpha,\beta} \exp(\alpha\omega) \{\exp(\alpha\omega) - 1\}$ となり、累積ベースライン強度関数が時刻無限大で発散すれば $\lim_{t \rightarrow \infty} E[N(t)] = \lim_{t \rightarrow \infty} \text{Var}[N(t)] \rightarrow \infty$ となる。GPP モデルが吸収状態をもたない NHMP であるのに対し、有界な累積ベースライン強度関数であっても累積期待フォールト数の期待値は $E[N(t)] \rightarrow \gamma_{\alpha,\beta} \{\exp(\alpha\omega) - 1\} (t \rightarrow \infty)$ となる。GPP モデルは Okamura and Dohi (2008) による非一様混合ポアソンモデルの一般化としてもみなすことができ、式 (2.18) と式 (2.25) の確率関数が $\alpha = 0$ で連続であるので、 $\kappa(t) = 1$ と $\alpha = 0$ のとき HMP モデルと NHPP モデルは、それぞれ GBP モデルと GPP モデルの特別な場合に帰着される (Gat, 2016)。

3. 最尤推定法

3.1 HMP モデルと NHPP モデル

HMP および NHPP に基づくソフトウェア信頼性モデルでは、推移率(強度関数)が決定された後、モデルパラメータを推定するための一般的な手法として、対数尤度関数を最大化する最尤推定法が用いられる。先にも述べたように、ソフトウェアフォールトデータは時間領域データとグループデータの 2 種類に分類される。 $t_0 = 0$ として、 m 個のフォールト検出時刻 $t_i (i = 1, 2, \dots, m)$ が観測され、このとき $t_e (\geq t_m)$ は観測(または打ち切り)時刻を示す。一方、グループデータは観測時刻とソフトウェアの累積フォールト検出数 $(\tau_i, n_i) (i = 1, 2, \dots, m)$ のペアで構成される。この場合、 m はテストの実施日数、週数、または月数を示し、初期状態として $(\tau_0, n_0) = (0, 0)$ が設定される。

まず最初に、打ち切り時刻 t_e までに観測される m 個の時間領域データ $(t_1, t_2, \dots, t_m; t_e)$ が与えられているとする。このとき、推移率 λ_k をもつ HMP モデルと NHPP モデルに対する対数尤度関数は

$$(3.1) \quad \ln L_{HMP}(\boldsymbol{\theta}) = \sum_{i=1}^m \ln \lambda_i(\boldsymbol{\theta}) - \lambda_m(\boldsymbol{\theta}) t_e,$$

$$(3.2) \quad \begin{aligned} \ln L_{NHPP}(\omega, \boldsymbol{\theta}) &= \sum_{i=1}^m \ln \lambda(t_i; \omega, \boldsymbol{\theta}) - \Lambda(t_e; \omega, \boldsymbol{\theta}) \\ &= \sum_{i=1}^m \ln f(t_i; \boldsymbol{\theta}) + \ln \omega^m - \omega F(t_e; \boldsymbol{\theta}) \end{aligned}$$

によって与えられる。ここで、 $\boldsymbol{\theta}$ は各モデルに含まれる未知パラメータベクトルであり、HMP モデルの推移率を $\lambda_k(\boldsymbol{\theta}) (k = 0, 1, 2, \dots)$ 、 $f(t; \boldsymbol{\theta}) = dF(t; \boldsymbol{\theta})/dt$ 、NHPP モデルの平均値関数と強度関数をそれぞれ $\Lambda(t; \omega, \boldsymbol{\theta})$ 、 $\lambda(t; \omega, \boldsymbol{\theta}) = d\Lambda(t; \boldsymbol{\theta})/dt$ のように表記する。

次に、グループデータ $(\tau_i, n_i) (i = 1, 2, \dots, m)$ が与えられているものとする。時間領域データにおいて m が観測された検出フォールト数を表していたのに対し、グループデータでは m は観測データ区間数を表しており、 $(\tau_{i-1}, \tau_i]$ を i 番目の観測区間、 n_i を i 番目の観測区間で検出された累積フォールト数を示す。HMP モデルの対数尤度関数は、マルコフ性から

$$(3.3) \quad \ln L_{HMP}(\boldsymbol{\theta}) = \prod_{i=1}^m \Pr\{N(\tau_i) = n_i \mid N(\tau_{i-1}) = n_{i-1}\}$$

のように与えられるが、任意の推移率に対して常に陽な表現が与えられる訳ではない。例えば、Jelinski and Moranda モデルでは、 $\theta = (N, b)$ に対する対数尤度関数は

$$(3.4) \quad \ln L_{HMP}(\theta) = \sum_{i=1}^m \ln \binom{N - n_{i-1}}{n_i - n_{i-1}} \{1 - e^{-b\tau_i}\}^{n_i - n_{i-1}} e^{-b(N - n_i)\tau_i}$$

となる。一方、NHPP モデルでは、ポアソン分布に従うことから、対数尤度関数は

$$(3.5) \quad \begin{aligned} \ln L_{NHPP}(\omega, \theta) &= \sum_{i=1}^m \{(n_i - n_{i-1}) \ln [\Lambda(\tau_i; \omega, \theta) - \Lambda(\tau_{i-1}; \omega, \theta)] \\ &\quad - \ln[(n_i - n_{i-1})!]\} - \Lambda(\tau_m; \omega, \theta) \\ &= \sum_{i=1}^m \{(n_i - n_{i-1}) \ln [F(\tau_i; \theta) - F(\tau_{i-1}; \theta)] \\ &\quad - \ln[(n_i - n_{i-1})!]\} + \ln \omega^{n_m} - \omega F(\tau_m; \theta) \end{aligned}$$

となる。

より一般に、式 (2.8) で与えられる混合分布表現を考える。NHPP モデルと同様に、NHMPP モデルにおける対数尤度関数は、時間領域データとグループデータに対して、それぞれ

$$(3.6) \quad \ln L_{NHMPP}(\theta_G) = \sum_{i=1}^m \ln f(t_i; \theta) + \ln \int_0^\infty \ln \omega^m e^{-\omega F(t_i; \theta)} dG_\omega(\omega | \theta_G),$$

$$(3.7) \quad \begin{aligned} \ln L_{NHMPP}(\theta_G) &= \sum_{i=1}^m \{(n_i - n_{i-1}) \ln [F(\tau_i; \theta) - F(\tau_{i-1}; \theta)] \\ &\quad - \ln[(n_i - n_{i-1})!]\} + \ln \int_0^\infty (\omega^{n_m} e^{-\omega F(\tau_m; \theta)}) dG_\omega(\omega | \theta_G) \end{aligned}$$

のように求められる。次に示す一連の結果は、NHPP モデルと NHMPP モデルに対する対数尤度関数の大小関係を規定するものである。

性質 1. 任意の $\hat{\omega}$ に対して以下を満たすパラメータ θ_G が存在する。

$$(3.8) \quad G(\omega; \theta_G) = \begin{cases} 0 & (0 < \omega < \hat{\omega}), \\ 1 & (\hat{\omega} \leq \omega < \infty). \end{cases}$$

補題 1. 任意の $n (> 0)$ と $0 \leq F(\tau; \theta) \leq 1$ に対して $\omega^n e^{-\omega F(\tau; \theta)}$ を最大にする $\hat{\omega} (0 < \hat{\omega} < \infty)$ が唯一存在する。

補題 2. 混合比分布 $G(\omega; \theta_G)$ が性質 1 を満たすとき、NHMPP に基づくソフトウェア信頼性モデルの対数尤度関数を最大にする混合比分布は

$$(3.9) \quad G(\omega) = \begin{cases} 0 & (0 < \omega < \hat{\omega}), \\ 1 & (\hat{\omega} \leq \omega < \infty) \end{cases}$$

となる。補題 1 と 補題 2 から、直ちに以下の結果を得る。

定理 1. (1) NHPP モデルの最尤推定値が存在し、NHMPP モデルの混合比分布が性質 1 を満たすものとする。このとき、NHPP モデルと NHMPP モデルの対数尤度の値は等しくなる。

(2) NHMPP モデルの混合比分布が性質 1 を満たさないとき、NHPP モデルの対数尤度関数は NHMPP モデルの対数尤度関数よりも常に大きい値をもつ。

以上の結果から、混合比分布が性質 1 を満たすステップ関数であれば、NHMPP モデルにおける対数尤度関数は NHPP モデルの対数尤度より常に小さくなるため、最尤推定の意味で NHMPP モデルを考える必然性はないことが示される。

3.2 NHMP モデル

次に GBP と GPP に対する対数尤度関数を導出する。NHPP における ω の存在は仮定せず、 θ を累積ベースライン強度関数のパラメータ $\kappa(t) = \kappa(t; \theta)$ および $\Lambda(t) = \Lambda(t; \theta)$ とし、時間領域データが観測されているとする。GBP に対する尤度関数は

$$(3.10) \quad L_{GBP}(\alpha, \beta, \theta) = \prod_{i=1}^m \{\beta - \alpha(i-1)\} \kappa(t_i; \theta) e^{-\int_0^{t_e} \{\beta - \alpha m\} \kappa(x; \theta) dx}$$

のように表現され、特に

$$(3.11) \quad \int_0^{t_e} \{\beta - \alpha m\} \kappa(x; \theta) dx = \sum_{i=1}^m \int_{t_{i-1}}^{t_i} \{\beta - \alpha(i-1)\} \kappa(x; \theta) dx + \int_{t_m}^{t_e} \{\beta - \alpha m\} \kappa(x; \theta) dx \\ = \beta \Lambda(t_e; \theta) - \alpha \left[m \Lambda(t_e; \theta) - \sum_{i=1}^m \Lambda(t_i; \theta) \right]$$

のような表現を得る。式 (3.10) の階乗部分は

$$(3.12) \quad \prod_{i=1}^m \{\beta - \alpha(i-1)\} \kappa(t_i; \theta) = \frac{\Gamma(\gamma_{\alpha, \beta} + 1)}{\Gamma(\gamma_{\alpha, \beta} - m)} \alpha^m \prod_{i=1}^m \kappa(t_i; \theta)$$

のように書けるので、 $\Gamma(k+1) = k\Gamma(k)$ に注意すると、対数尤度関数は

$$(3.13) \quad \ln L_{GBP}(\alpha, \beta, \theta) = \left[\sum_{i=1}^m \ln \kappa(t_i; \theta) + \sum_{i=1}^m \alpha \Lambda(t_i; \theta) \right] + \ln[(\gamma_{\alpha, \beta})! - (\gamma_{\alpha, \beta} - m - 1)!] \\ + m \ln(\alpha) + [(\alpha m - \beta) \Lambda(t_e; \theta)]$$

となる。グループデータ (τ_i, n_i) ($i = 0, 1, \dots, m$) の場合、式 (2.18) から、対数尤度関数を

$$(3.14) \quad \ln L_{GBP}(\alpha, \beta, \theta) = \sum_{i=1}^m \ln \{(\gamma_{\alpha, \beta} - n_{i-1})! - (n_i - n_{i-1})! - (\gamma_{\alpha, \beta} - n_i)!\} \\ - \sum_{i=1}^m \{\gamma_{\alpha, \beta} - n_i\} \times \alpha \{\Lambda(\tau_i; \theta) - \Lambda(\tau_{i-1}; \theta)\} \\ + \sum_{i=1}^m (n_i - n_{i-1}) \times \ln[1 - e^{-\alpha \{\Lambda(\tau_i; \theta) - \Lambda(\tau_{i-1}; \theta)\}}]$$

のように求めることができる。

次に GPP に対する対数尤度関数について考える。時間領域データが与えられたとき、式 (3.13) から、直ちに

$$(3.15) \quad \ln L_{GPP}(\alpha, \beta, \theta) = \left[\sum_{i=1}^m \ln \kappa(t_i; \theta) + \sum_{i=1}^m \alpha \Lambda(t_i; \theta) \right] + \ln[(\gamma_{\alpha, \beta})! - (\gamma_{\alpha, \beta} - m - 1)!] \\ + (m \ln(\alpha) - [(\alpha m + \beta) \Lambda(t_e; \theta)])$$

を得る。グループデータ (τ_i, n_i) ($i = 0, 1, \dots, m$) の場合、GPP の対数尤度関数は

$$\begin{aligned}
 (3.16) \quad \ln L_{GPP}(\alpha, \beta, \theta) = & \sum_{i=1}^m \ln \{(n_i + \gamma_{\alpha, \beta} - 1)! - (n_i - n_{i-1})! - (n_{i-1} + \gamma_{\alpha, \beta} - 1)!\} \\
 & - \sum_{i=1}^m \{\gamma_{\alpha, \beta} + n_{i-1}\} \alpha \{\Lambda(\tau_i; \theta) - \Lambda(\tau_{i-1}; \theta)\} \\
 & + \sum_{i=1}^m (n_i - n_{i-1}) \ln [1 - e^{-\alpha \{\Lambda(\tau_i; \theta) - \Lambda(\tau_{i-1}; \theta)\}}]
 \end{aligned}$$

のように求めることが可能である。

4. ソフトウェアメトリクスの利用

4.1 一般化線形ソフトウェア信頼性モデル

ここまでソフトウェア信頼性モデルでは、フォールト検出時間やフォールト検出数のみを扱ってきた。しかし、ソフトウェアの信頼性は、ソフトウェアの設計やテスト工程における情報に大きく影響すると考えられ、信頼性評価の結果に基づいて設計や開発プロセスの改善を行うことも必要とされる。このように、ソフトウェアの設計やテスト工程における情報をソフトウェアメトリクスと呼び、これらをソフトウェア信頼性モデルに組み込む拡張モデルが提案されている。ソフトウェアメトリクスには、コード行数や複雑性のようなソフトウェアそのものの特徴を表すプロダクトメトリクス(あるいはデザインメトリクス)と、開発工程において計測されるレビュー回数、テストケース数、カバレッジのような開発プロセスの情報を表すプロセスマトリクスの2種類がある。これらのメトリクスはそれぞれ、ソフトウェアの信頼性に影響を与えるが、影響を与える範囲が異なる。Okamura and Dohi (2014)では、プロダクトメトリクスをソフトウェア信頼性モデルに組み込むことで、ソフトウェアの信頼性評価を行っている。また、Okamura et al. (2010)は、プロセスマトリクスを組み込んだソフトウェア信頼性モデルを提案している。いずれのモデルにおいても、基本的なアイデアは一般化線形モデル(Generalized Linear Model: GLM)と呼ばれる回帰モデルの活用である。

一般化線形モデルとは、正規分布による線形回帰(重回帰)モデルを一般化したモデルの枠組みであり、正規分布以外の分布や、多様なリンク関数を用いて独立変数と従属変数の因果関係を表すモデルである。いま、確率変数(従属変数) Y_1, \dots, Y_N とそれらに対応する確率変数(独立変数)ベクトル X_1, \dots, X_N の観測値が与えられるとする。ここで、 $X_i = (X_{i,1}, \dots, X_{i,L})$ であり、 $X_{i,l}$ ($l = 1, 2, \dots, L$) は Y_i に対する要因 l の値を表す。このとき、 Y と X の関係を表すモデルとして次の仮定を考える。

- Y_1, \dots, Y_N は指数分布族に属する同一の分布に従う。ただし、パラメータが異なる。
- 従属変数の平均は $g(E[Y_i]) = \beta x_i$ で与えられる。 x_i は X_i の観測値であり、 $g(\cdot)$ はリンク関数と呼ばれる。

上記の枠組みは Y_1, \dots, Y_N の分布とリンク関数を選ぶことにより、既存の回帰モデルに帰着させることができる。分布を正規分布、リンク関数を $g(\mu) = \mu$ とすると、重回帰モデルに帰着される。また、分布をポアソン分布、リンク関数を $g(\mu) = \log(\mu)$ とするとポアソン回帰、分布をベルヌーイ分布、リンク関数をロジット関数 $g(\mu) = \log \mu / (1 + \mu)$ とすると、ロジスティック回帰へそれぞれ帰着される。

上記の枠組みを利用してソフトウェアメトリクスと NHPP に基づいたソフトウェア信頼性モデルの融合を考える。いま、 j 個のモジュールからなるソフトウェアを考え、以下の仮定を設ける。

- ・ソフトウェアテストは逐次的かつ離散的に実施され、各テスト毎で検出されたフォールトは次のテストまでに修正される。
- ・モジュール i において 単一のフォールトが k 番目のテスト区間で検出されるフォールト検出確率 $p_{i,k}$ ($0 \leq p_{i,k} \leq 1$) は、モジュール i において k 番目のテスト区間にに関するプロセスメトリクスベクトル $\mathbf{x}_{i,k} = (x_{i,k,1}, \dots, x_{i,k,r_i})$ を用いて

$$(4.1) \quad l_p(p_{i,k}) = \alpha_{0,i} + \alpha_i \mathbf{x}_{i,k}$$

のように表現される。ここで $l_p(\cdot)$ はフォールト検出確率のリンク関数である。

- ・各モジュール $i = 1, \dots, j$ に含まれる総フォールト数(初期フォールト数) M_1, \dots, M_j はモジュール固有のプロダクトメトリクス \mathbf{s}_i に依存した平均 ξ_i のポアソン分布に従い、その平均は以下で与えられる。

$$(4.2) \quad l_\xi(\xi_i) = \beta_0 + \beta \mathbf{s}_i.$$

ここで、 $l_\xi(\cdot)$ は総フォールト数のリンク関数である。

フォールト検出確率のリンク関数としては、ロジット関数、プロビット関数および complementary log-log 関数などが適用できる。また、総フォールト数のリンク関数としては対数関数や一次関数が適用できる。いま、

$$(4.3) \quad l_p(p) = \text{logit}(p) = \log \frac{p}{1-p}, \quad l_\xi(\xi) = \log(\xi)$$

とすると、モジュール i において k 番目のテスト区間のフォールト検出確率は

$$(4.4) \quad p_{i,k} = \frac{\exp(\alpha_{0,i} + \alpha_i \mathbf{x}_{i,k})}{1 + \exp(\alpha_{0,i} + \alpha_i \mathbf{x}_{i,k})}$$

となる。さらに、モジュール i の初期フォールト数は平均 $\exp(\xi_i)$ のポアソン分布に従うため、モジュール i の k 番目のテスト区間までに検出されるフォールト数 $Y_{i,k}$ は以下の確率関数で与えられる。

$$(4.5) \quad \begin{aligned} P(Y_{i,k_i} = y_{i,k}) &= \sum_{m_i=y_{i,k}}^{\infty} P(Y_{i,k_i} = y_{i,k} | M_i = m_i) P(M_i = m_i) \\ &= \exp(-\xi_i \lambda_{i,k}) \frac{(\xi_i \lambda_{i,k})^{y_{i,k}}}{y_{i,k}!}. \end{aligned}$$

ここで、 $\lambda_{i,n_i} = 1 - \prod_{k=1}^{n_i} (1 - p_{i,k})$ である。このモデルは、モジュール i に含まれる総フォールト数に対するポアソン回帰と k 番目のテスト区間におけるフォールト検出確率に対するロジスティック回帰を同時に適用している。

4.2 EM アルゴリズム

次に効率的な推定を行う目的で EM(Expectation-Maximization) アルゴリズムの適用を考える。ここで考える一般化線形モデルの場合、未観測データを各モジュールの総フォールト数 N_1, \dots, N_j とし、モジュール m のフォールト検出時刻列を $T_{m,1} < T_{m,2} < \dots < T_{m,N_m}$ とする。このとき、完全対数尤度は

$$(4.6) \quad \begin{aligned} \log \mathcal{L}(\beta_0, \beta, \alpha_{0,1}, \alpha_1, \dots, \alpha_m) &= \sum_{m=1}^j N_m (\beta_0 + \beta \mathbf{s}_j) \\ &\quad - \sum_{m=1}^j \exp(\beta_0 + \beta \mathbf{s}_j) + \sum_{m=1}^j \sum_{k=1}^{N_m} \log(\lambda_{m,k} - \lambda_{m,k-1}) \end{aligned}$$

となる。つまり、最初の二項が N_1, \dots, N_m を従属変数としたポアソン回帰となり、 $\lambda_{m,k}$ に関する項は $N_m - \sum_{u=1}^{k-1} y_u$ を従属変数としたロジスティック回帰となる。一方で、総フォールト数の期待値は与えられたパラメータ $\beta_0, \beta, \alpha_{0,m}, \alpha_m$ に対して

$$(4.7) \quad E[N_m] = \sum_{k=1}^{K_m} y_{m,k} + \exp(\beta_0 + \beta s_m) \prod_{k=1}^{K_m} \frac{\exp(\alpha_{0,m} + \alpha_m \mathbf{x}_{k,m})}{1 + \exp(\alpha_{0,m} + \alpha_m \mathbf{x}_{k,m})}$$

として計算できる。最終的に、EM 原理に基づいた推定アルゴリズムは次のようになる。

- Step 1: Compute $\omega_m \leftarrow \exp(\beta_0 + \beta s_m), m = 1, \dots, j$.

- Step 2: Compute

$$(4.8) \quad p_{m,k} \leftarrow \frac{\exp(\alpha_{0,m} + \alpha_m \mathbf{x}_{k,m})}{1 + \exp(\alpha_{0,m} + \alpha_m \mathbf{x}_{k,m})}, \quad m = 1, \dots, j, \quad k = 1, \dots, K_m.$$

- Step 3: For $i = 1, \dots, j$, compute

$$(4.9) \quad Z_m \leftarrow \sum_{k=1}^{K_m} y_{m,k} + \omega_m \prod_{k=1}^{K_m} (1 - p_{m,k}).$$

- Step 4: $(\beta_0, \beta) \leftarrow PR(s_1, \dots, s_j; Z_1, \dots, Z_m)$.

- Step 5: $(\alpha_{0,m}, \alpha_m) \leftarrow LR(\mathbf{x}_{m,1}, \dots, \mathbf{x}_{m,K_m}; Z_m, \dots, Z_m - \sum_{u=1}^{K_m-1} y_u)$.

ここで、 $PR()$ と $LR()$ は、それぞれ通常のポアソン回帰およびロジスティック回帰による回帰係数の推定を表す。上記の Step 1 から Step 5 を繰り返すことで、回帰係数に対する最尤推定値が得られる。

一般化線形ソフトウェア信頼性モデルでは要因数に応じた回帰係数が必要となる。一般的に、パラメータ数を多くすればするほどデータへの適合は良くなるがモデルの汎化能力を高めることはできないことが知られている。このような問題に対処するためには不要な要因を削除して要因を選別する必要がある。これは要因選択と呼ばれ、回帰をもとにしたモデルでは一般的に行われる手続きである。

要因選択手法には、Deviance を使う手法や検定による手法があるが、ここでは AIC (Akaike Information Criterion) を用いた手法を用いる。AIC の定義式を以下のように与える。

$$(4.10) \quad AIC = -2 \text{ 最大対数尤度} + 2 \text{ パラメータ数}.$$

これは最尤法における真のモデルとのバイアスをパラメータ数によって近似した指標であり、AIC を最小にするモデルが最良のモデルとなる。このことから、AIC を最小にする要因の組み合わせを求ることで、データへの過適合を解消し汎化能力の高いモデルを得ることができる。具体的な手続きでは、すべての要因の組み合わせを総当たり的に調べることは不可能であるため変数増減法が適用される。変数増減法では、変数を加えることによって AIC を最も低くする要因から優先的に選択する手続きを AIC が増加するまで行う。これによって、AIC を最も小さくする要因の組み合わせを近似的に求める。

4.3 有効性検証

実際の開発で計測されたテストデータおよびオープンソースプロジェクトにおけるデータを用いてモデルの有効性に関する分析を行う。ここでは、従来の手法である単純な NHPP モデルと、プロダクトメトリクスだけをポアソン回帰として利用するソフトウェア信頼性モデル、プロセスマトリクスをロジスティック回帰として利用するソフトウェア信頼性モデルとの比較を行う。特に、AIC を用いた評価を行うことで、単純なデータとの誤差比較ではなく、背後にあ

る(真の)モデルを適切に表現できるかどうかの検証を行う。

実験で用いるデータは Apache Software Foundation (ASF) で管理されている Tomcat プロジェクトを対象とした。また、バージョンの違いによる影響も見るため、Tomcat については幾つかのバージョンに対するデータを収集する。さらに、それぞれのプロジェクト毎のバグトラッキングシステム (ASF Bugzilla) からバグレポートを収集した。Tomcat プロジェクトのモジュール構成、バグレポート収集期間は次の通りである。

Tomcat: Apache Software Foundation が開発しているオープンソースの Web サーバであり、Java サーブレットを提供する。主として Java 言語で作成されており、Apache Web Server と組み合わせて利用されることが多い。Bugzilla の管理では以下のプログラム群(モジュール)で構成される。

- webapps (manager): 管理用ウェブアプリケーション
- catalina: Servlet container のコア
- connectors: Tomcat と Apache の連携
- jasper: JSP ページコンパイラとランタイムエンジン
- servlets: Servlet プログラム群

対象とする Tomcat のバージョンは 5, 6, 7 であり、メジャーバージョンアップについては別アプリケーション(別プロジェクト)と見なして分析する。それぞれのバージョンにおけるバグデータの収集期間は Tomcat5 が 2004/5 から 2009/1, Tomcat6 が 2006/1 から 2013/11, Tomcat7 が 2010/6 から 2013/11 となっている。

一般化線形ソフトウェア信頼性モデルでは、プロダクトメトリクスと総フォールト数の因果関係、プロセスマトリクスとフォールト検出確率の因果関係をそれぞれポアソン回帰モデル、ロジスティック回帰モデルで関連づけている。そのため、実データの分析においてもこれらのメトリクスの収集が必要となる。実際に、各プロジェクトのモジュールにおけるソースコードから、直接以下のプロダクトメトリクスを算出した。

- Files (Fl): ファイル数
- Lines (Ln): コード行数
- Statements (St): ステートメント数
- % Branches (Br): 分岐の出現率
- Calls (Ca): メソッドの呼び出し回数
- % Comments (Cm): コメントの出現率
- Classes (Cl): クラス数
- Methods/Class (Me/Cl): 単一クラスあたりのメソッド数
- Avg Stmt/Method (St/Me): 単一メソッドあたりのステートメント数
- Max Complexity (MCx): 最大複雑度
- Max Depth (MDp): 最大のネストの深さ
- Avg Depth (ADp): 平均のネストの深さ
- Avg Complexity (ACx): 平均複雑度

一方、プロジェクトのアクティビティを表す指標として考えられる以下のプロセスマトリクスをそれぞれのモジュールに対して計測した。

- time: 日数
- ctime: 累積日数

- comments: Bugzilla 上での該当モジュールに対するコメント数
- votes: Bugzilla 上での該当モジュールに対する投票回数
- workers: コメントを投稿した人の数(Tomcat5 のみ)

まずモジュール毎で独立に(1)NHPP モデルの推定, (2)ロジスティック回帰によるソフトウェア信頼性モデルの推定を行った。 (1)では標準的に使われる 11 種類のモデルすべてに対してパラメータを推定し最小の AIC を示すモデルを選択した。また, ロジスティック回帰によるソフトウェア信頼性モデルでは, 変数増減法による要因選択を行って最小の AIC を示す要因の組み合わせを求めている。さらに, 各プロジェクトのモジュールに対するプロダクトメトリクスを用いて, (3)ポアソン回帰によるソフトウェア信頼性モデルの推定と(4)一般化線形ソフトウェア信頼性モデルの推定を行った。ポアソン回帰によるソフトウェア信頼性モデルの推定では, 事前に各モジュールに適用する NHPP モデルを選ぶ必要がある。これには, (1)の手順で選ばれたモデルと同じモデルを採用した。同様に(4)では, 各モジュールに適用するロジスティック回帰によるソフトウェア信頼性モデルで使用する要因を選ぶ必要がある。これも(2)で单一モジュールに対して選択された要因を用いることとした。加えて, ポアソン回帰によるソフトウェア信頼性モデルおよび一般化線形ソフトウェア信頼性モデルでは, プロダクトメトリクスに対する要因選択を行う必要がある。これも, 変数増減法により最小の AIC を示す要因の組み合わせを選んだ。

表 2 から表 4 は各プロジェクトに対して NHPP モデルのみ(NHPP), ロジスティック回帰に

表 2. Tomcat5 に対する AIC.

Module	NHPP	Logit	Poi	GLSRM
catalina	268.96	271.04	-	-
connectors	145.10	128.20	-	-
jasper	159.96	147.70	-	-
servlets	116.99	102.08	-	-
webapps	150.51	137.69	-	-
Total (project)	841.52	786.71	838.87	784.58

表 3. Tomcat6 に対する AIC.

Module	NHPP	Logit	Poi	GLSRM
catalina	489.30	388.66	-	-
connectors	277.74	201.74	-	-
jasper	293.51	247.27	-	-
manager	163.15	125.58	-	-
servlets	181.54	181.58	-	-
Total (project)	1405.24	1144.83	1403.82	1144.84

表 4. Tomcat7 に対する AIC.

Module	NHPP	Logit	Poi	GLSRM
catalina	247.91	210.87	-	-
connectors	168.82	129.44	-	-
jasper	167.34	137.57	-	-
manager	117.07	90.49	-	-
servlets	148.33	110.57	-	-
Total (project)	849.47	678.94	847.40	675.86

よるソフトウェア信頼性モデル(Logit), ポアソン回帰によるソフトウェア信頼性モデル(Poi), 一般化線形ソフトウェア信頼性モデル(GLSRM)それぞれを適用したときの AIC を示している。表からいずれの場合でも

$$(4.11) \quad \text{NHPP} \geq \text{Poi} \geq \text{Logit} \geq \text{GLSRM}$$

の関係が得られる。つまり、適合性の観点からは GLSRM が最も良いことがわかった。また、プロダクトメトリクスを扱うポアソン回帰によるソフトウェア信頼性モデルと、プロセスマトリクスを扱うロジスティック回帰によるソフトウェア信頼性モデルを比較すると、プロセスマトリクスを扱うロジスティック回帰によるソフトウェア信頼性モデルの方が適合性が高い。これは初期の総フォールト数を決めるための情報量よりも、観測されたバグ報告数から得られる情報量の方が多いことを示唆している。ここでは、モジュール数が比較的少ないためこのような結果になったものと考えられる。つまり、モジュール数が多い状況では、プロダクトメトリクスからの情報による影響がより大きくなるものと予想される。

また表 5 から表 7 は、各プロジェクトにおけるモジュールに対して AIC により選択された動的メトリクス、静的メトリクスを表している。選択されたメトリクスをみると、プロダクトメトリクスでは最大複雑度(MCx)が比較的多く、プロセスマトリクスではコメント数が多くデータで選択される結果となった。これは、プロダクトメトリクスでは従来から重要とされてきた、コード行数や複雑度が大きく影響することを意味している。また、コメント数などのプロジェクトのアクティビティを表す指標がフォールト検出確率を決定するのに利用可能であることがわかる。

表 5. Tomcat5 において選択されたメトリクス。

Module	Logit	Poi	GLSRM
catalina	comments, ctime	St, MCx	St, MCx
connectors	worker, ctime		
jasper	worker, ctime		
servlets	comments, ctime		
webapps	comments, ctime		

表 6. Tomcat6 において選択されたメトリクス。

Module	Logit	Poi	GLSRM
catalina	comments, votes, ctime	Fl, Cl, MCx	Fl, Ln, Cm, Cl
connectors	comments, votes, ctime		
jasper	time, comments		
manager	comments, ctime		
servlets	comments, ctime		

表 7. Tomcat7 において選択されたメトリクス。

Module	Logit	Poi	GLSRM
catalina	comments, votes	Fl, Ca, Cm	St, Cm
connectors	comments, votes		
jasper	comments		
manager	comments, votes		
servlets	comments, votes, ctime		

5.まとめと今後の課題

本稿では、ソフトウェア信頼性モデルに関する主要な進展を体系的に整理し、特に非一様マルコフ過程(NHMP)を中心とした統一的なアプローチを提案した。従来のHMPおよびNHPPに基づくソフトウェア信頼性モデルを一般化し、GBPやGPPといった柔軟性の高い新たなモデルの特性を分析した。また、ソフトウェアメトリクスの利用による信頼性予測モデルの有用性を実証的に検証し、プロダクトおよびプロセスマトリクスの重要性を明らかにした。これらの成果は、実際のソフトウェア開発プロジェクトにおける信頼性向上に寄与することが期待できる。

今後の課題として、提案モデルをさらに広範なデータセットや産業分野で検証し、特に商用ソフトウェアへの適用によってモデルの汎用性を評価する必要がある。また、ソフトウェアメトリクスの収集には手間がかかるため、自動化や標準化された計測手法を開発し、実務での利用可能性を高める必要がある。これらの課題への取り組みによって、ソフトウェア信頼性モデリングの理論的および実践的貢献がさらに進展することが期待できる。

参考文献

- Abdel-Ghaly, A. A., Chan, P. Y. and Littlewood, B. (1986). Evaluation of competing software reliability predictions, *IEEE Transactions on Software Engineering*, **SE-12**(9), 950–967.
- Achcar, J. A., Dey, D. K. and Niverthi, M. (1998). A Bayesian approach using nonhomogeneous Poisson processes for software reliability models, *Frontiers in Reliability* (eds. A. P. Basu, K. S. Basu and S. Mukhopadhyay), 1–18, World Scientific, Singapore.
- Boland, P. J. and Singh, H. (2003). A birth-process approach to Moranda's geometric software-reliability model, *IEEE Transactions on Reliability*, **52**(2), 168–174.
- Cretois, E. and Gaudoin, O. (1998). New results on goodness-of-fit tests for the power-law process and application to software reliability, *International Journal of Reliability, Quality and Safety Engineering*, **5**(3), 249–267.
- Dohi, T., Li, S. and Okamura, H. (2024). Local polynomial software reliability models and their application, *Information and Software Technology*, **166**, 107366.
- Duane, J. T. (1964). Learning curve approach to reliability monitoring, *IEEE Transactions on Aerospace*, **2**(2), 563–566.
- Erto, P., Giorgio, M. and Lepore, A. (2020). The generalized inflection S-shaped software reliability growth model, *IEEE Transactions on Reliability*, **69**(1), 228–244.
- Gat, Y. L. (2016). *Recurrent Event Modeling based on the Yule Process: Application to Water Network Asset Management*, **2**, John Wiley & Sons, London.
- Goel, A. L. (1985). Software reliability models: assumptions, limitations and applicability, *IEEE Transactions on Software Engineering*, **SE-11**(12), 1411–1423.
- Goel, A. L. and Okumoto, K. (1979). Time-dependent error-detection rate model for software reliability and other performance measures, *IEEE Transactions on Reliability*, **R-28**(3), 206–211.
- Gokhale, S. S. and Trivedi, K. S. (1998). Log-logistic software reliability growth model, *Proceedings of The 3rd IEEE International Symposium on High-Assurance Systems Engineering (HASE 1998)*, 34–41, IEEE CPS, Los Alamitos, California, USA.
- Ikemoto, S., Dohi, T. and Okamura, H. (2013). Quantifying software test process and product reliability simultaneously, *Proceedings of The 24th International Symposium on Software Reliability Engineering (ISSRE 2013)*, 108–117, IEEE CPS, Piscataway, New Jersey, USA.
- Jelinski, Z. and Moranda, P. B. (1972). Software reliability research, *Statistical Computer Performance Evaluation* (ed. W. Freiberger), 465–484, Academic Press, New York.

- Kendall, D. G. (1948). On the generalized “birth-and-death” process, *Annals of Mathematical Statistics*, **19**, 1–5.
- Konno, H. (2010). On the exact solution of a generalized Polya process, *Advances in Mathematical Physics*, **2010**, 504267.
- Kuwa, D. and Dohi, T. (2013a). Generalized Cox proportional hazards regression-based software reliability modeling with metrics data, *Proceedings of The 19th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2013)*, 328–337, IEEE CPS, Piscataway, New Jersey, USA.
- Kuwa, D. and Dohi, T. (2013b). Generalized logit-based software reliability modeling with metrics data, *Proceedings of The 37th Annual International Computer Software and Applications Conference (COMPSAC 2013)*, 246–255, IEEE CPS, Piscataway, New Jersey, USA.
- Lawless, J. F. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association*, **82**(399), 808–815.
- Li, S., Dohi, T. and Okamura, H. (2022). A comprehensive analysis of proportional intensity-based software reliability models with covariates, *Electronics*, **11**, 2353.
- Li, S., Dohi, T. and Okamura, H. (2023). Non-homogeneous Markov process modeling for software reliability assessment, *IEEE Transactions on Reliability*, **72**(4), 1540–1555.
- Littlewood, B. (1980). Theories of software reliability: How good are they and how can they be improved?, *IEEE Transactions on Software Engineering*, **SE-6**(5), 489–500.
- Littlewood, B. (1981). Stochastic reliability growth: A model for fault removal in computer-program and hardware-designs, *IEEE Transactions on Reliability*, **R-30**(4), 313–320.
- Littlewood, B. (1984). Rationale for a modified Duane model, *IEEE Transactions on Reliability*, **R-33**(2), 157–159.
- Lyu, M. R. (ed.) (1996). *Handbook of Software Reliability Engineering*, McGraw-Hill, New York.
- Moranda, P. B. (1979). Event-altered rate models for general reliability analysis, *IEEE Transactions on Reliability*, **R-28**(5), 376–381.
- Musa, J. D. and Okumoto, K. (1984). A logarithmic Poisson execution time model for software reliability measurement, *Proceedings of The 7th International Conference on Software Engineering (ICSE 1984)*, 230–238, ACM, New York, USA.
- Musa, J. D., Iannino, A. and Okumoto, K. (1987). *Software Reliability: Measurement, Prediction, Application*, McGraw-Hill, New York.
- Ohba, M. (1984). Inflection S-shaped software reliability growth model, *Stochastic Models in Reliability Theory* (eds. S. Osaki and Y. Hatoyama), 144–165, Springer-Verlag, Berlin.
- Ohishi, K., Okamura, H. and Dohi, T. (2009). Gompertz software reliability model: Estimation algorithm and empirical validation, *Journal of Systems and Software*, **82**(3), 535–543.
- Okamura, H. and Dohi, T. (2008). Software reliability modeling based on mixed Poisson distributions, *International Journal of Reliability, Quality and Safety Engineering*, **15**(1), 9–32.
- Okamura, H. and Dohi, T. (2013). SRATS: Software reliability assessment tool on spreadsheet, *Proceedings of The 24th International Symposium on Software Reliability Engineering (ISSRE 2013)*, 100–117, IEEE CPS, Piscataway, New Jersey, USA.
- Okamura, H. and Dohi, T. (2014). A novel framework of software reliability evaluation with software reliability growth models and software metrics, *Proceedings of The 15th IEEE International Symposium on High Assurance Systems Engineering (HASE 2014)*, 97–104, IEEE CPS, Piscataway, New Jersey, USA.
- Okamura, H. and Dohi, T. (2015). Towards comprehensive software reliability evaluation in open source software, *Proceedings of The 26th International Symposium on Software Reliability Engineering (ISSRE 2015)*, 121–129, IEEE CPS, Piscataway, New Jersey, USA.
- Okamura, H. and Dohi, T. (2016). Phase-type software reliability model: Parameter estimation algorithms with grouped data, *Annals of Operations Research*, **244**(1), 177–208.

- Okamura, H., Etani, Y. and Dohi, T. (2010). A multi-factor software reliability model based on logistic regression, *Proceedings of The 21st International Symposium on Software Reliability Engineering (ISSRE 2010)*, 31–40, IEEE CPS, Los Alamitos, California, USA.
- Okamura, H., Dohi, T. and Osaki, S. (2013). Software reliability growth models with normal failure time distributions, *Reliability Engineering and System Safety*, **116**, 135–141.
- Rinsaka, K., Shibata, K. and Dohi, T. (2006). Proportional intensity-based software reliability modeling with time-dependent metrics, *Proceedings of The 30th Annual International Computer Software and Applications Conference (COMPSAC 2006)*, 405–410, IEEE CPS, Los Alamitos, California, USA.
- Schick, G. J. and Wolverton, R. W. (1978). An analysis of competing software reliability models, *IEEE Transactions on Software Engineering*, **SE-4**(2), 104–120.
- Shaked, M., Spizzichino, F. and Suter, F. (2002). Nonhomogeneous birth processes and l_∞ spherical densities, with applications in reliability theory, *Probability in the Engineering and Informational Sciences*, **16**(3), 271–288.
- Shanthikumar, J. G. (1981). A general software reliability model for performance prediction, *Microelectronics and Reliability*, **21**(5), 671–682.
- Shanthikumar, J. G. (1983). Software reliability models: A review, *Microelectronics and Reliability*, **23**(5), 903–943.
- Shibata, K., Rinsaka, K. and Dohi, T. (2006). Metrics-based software reliability models using non-homogeneous Poisson processes, *Proceedings of The 17th International Symposium on Software Reliability Engineering (ISSRE 2006)*, 52–61, IEEE CPS, Los Alamitos, California, USA.
- Wagoner, W. L. (1973). The final report on a software reliability measurement study, Aerospace Corporation Report, TOR-0074 (4112)-1, Aerospace Corporation.
- Xiao, X. (2015). NHPP-based software reliability model with Marshall-Olkin failure time distribution, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **E98-A**(10), 2060–2068.
- Xie, M. (1989). On a generalization of the J-M model, *Proceedings of Reliability 1989*, 5 Ba/3/1–5 Ba/3/7.
- Xie, M. (1991). *Software Reliability Modeling*, World Scientific, Singapore.
- Yamada, S., Ohba, M. and Osaki, S. (1983). S-shaped reliability growth modeling for software error detection, *IEEE Transactions on Reliability*, **R-32**(5), 475–478.
- Zhao, M. and Xie, M. (1996). On maximum likelihood estimation for a general non-homogeneous Poisson process, *Scandinavian Journal of Statistics*, **23**(4), 597–607.

Some Aspects in Software Reliability Modeling from the Viewpoint of Non-homogeneous Markov Processes

Tadashi Dohi and Hiroyuki Okamura

Graduate School of Advanced Science and Engineering, Hiroshima University

This paper systematically summarizes the advancement of software reliability modeling and proposes a unified approach based on non-homogeneous Markov processes (NHMP). We generalize the conventional homogeneous Markov process (HMP) and non-homogeneous Poisson process (NHPP) based models, and introduces more flexible models such as generalized binomial process (GBP) and generalized Poisson process (GPP). Furthermore, we introduce a reliability prediction model using software product and process metrics, and verify its usefulness through an empirical study. These results contribute to improving software reliability assessment in actual software development, and expand the possibilities for future model development and applicability in industry.

相関した競合リスクを伴う左側切断・右側打ち切りデータのベイズ流解析

道前 洋史[†]

(受付 2024 年 9 月 28 日；改訂 12 月 20 日；採択 12 月 24 日)

要　　旨

生存時間解析は、ある基準となる時点から特定の事象(イベント)が発生するまでの時間を分析する手法であり、医学や疫学はもとより、経済学や社会科学など幅広い分野で活用されている。打ち切り、競合リスク、左側切断などは生存時間解析における古典的な話題である。本稿では、競合リスクを伴う左側切断・右側打ち切りの生存時間データに焦点を当て、その代表的な4つの解析モデルを紹介する(独立モデル、コピュラモデル、ベイズ流独立モデル、ベイズ流コピュラモデル)。特に、競合リスクにおけるコピュラを用いた潜在故障時間(周辺分布)の同時モデル化、同時モデルの条件付きによる左側切断のモデル化、尤度関数の構築、そしてベイズ推定量による周辺分布パラメータとコピュラパラメータの推定については詳細に説明する。シミュレーション実験では、現実的な条件設定(異なる周辺分布と左側切断割合の組み合わせ)で実施し、上記4つの解析モデルで各種パラメータを推定した。考察では、パラメータの推定性能を比較検討し、ベイズ流コピュラモデルにおける課題と今後の展望について述べる。

キーワード：コピュラ、事後平均、周辺生存関数、潜在故障時間、同時生存関数、ワイルブル分布。

1. はじめに

縦断研究や介入研究においては、研究者が注目しているイベント(リスク)が、他の先行するイベントの発生によって観察されないことがある。つまり、競合するイベントによって観察が打ち切られることがある。この現象を競合リスク問題と呼ぶ(西川, 2008; 斎藤・室谷, 2023, 2024)。競合リスク問題は生存時間解析における古典的な話題であり、主に3つのアプローチで対処されることが多い。医学や疫学分野では、原因別ハザードモデル(Prentice et al., 1978)や部分分布ハザードモデル(Fine and Gray, 1999)が伝統的に用いられているが、信頼性分野や保険数理分野では潜在故障時間モデル(Cox, 1959; Kaishev et al., 2007; 塚原, 2021)が広く活用されている。本稿では、潜在故障時間モデルに焦点を当てることとするが、原因別ハザードモデルや部分分布ハザードモデルに興味のある読者は西川(2008)や斎藤・室谷(2023, 2024)を参考されたい。

競合リスクにおける潜在故障時間モデルについて、原爆被爆者疫学調査(Bockwoldt et al., 2021; Spusto et al., 2023)を例に挙げて説明する(図1)。原爆の投下は1945年、この疫学調査の観察開始は1958年、観察終了は2009年である。研究者の興味のある調査項目は、登録者に

[†] 北里大学 薬学部：〒175-0082 東京都港区白金 5-9-1; michimae@pharm.kitasato-u.ac.jp

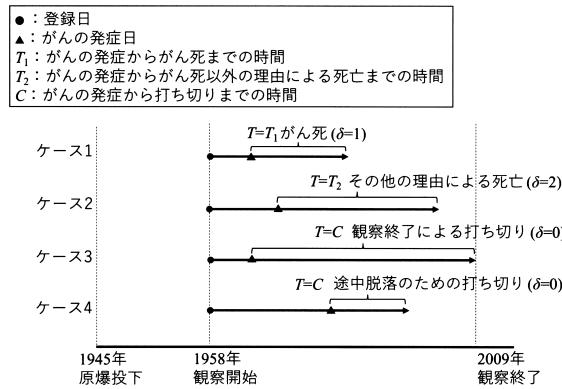


図 1. 競合リスクの模式図.

おけるがんの発症からがんによる死亡までであるとする。潜在故障時間モデルの仮定の下では、例えば、 T_1 をがんの発症からがん死までの時間、 T_2 をがんの発症からがん以外による死亡までの時間とした場合、観察時間は最小値 $T = \min(T_1, T_2)$ として記録される(図 1)。これは、がんの発症からそれぞれのイベント発生までの時間を潜在変数とみなし、実際に観察できるのは先に発生したイベントまでの観察時間であるという仮定に基づいている(Cox, 1959; 宮村, 1982)。また、通常縦断研究ではあらかじめ観察終了時点が設定されるため、その時点まで生存した対象は右側打ち切りとして扱われる(図 1 のケース 3)。観察期間中でも途中脱落のため観察打ち切りとなる対象も想定されるが、これも独立・無情報打ち切りの仮定の下、右側打ち切りとして扱うことが可能である(図 1 のケース 4)。結局、競合リスクの状況下では、 C を右側打ち切り時間として、 $T = \min(T_1, T_2, C)$ とイベント指示変数 $\delta \in (0, 1, 2)$ が各対象について観察されることになる。 δ は $T = T_1$ が観察された場合 $\delta = 1$ 、 $T = T_2$ が観察された場合 $\delta = 2$ 、 $T = C$ が観察された場合 $\delta = 0$ の値をとる指示変数である。古典的な潜在故障時間モデルでは、イベント間の独立性、すなわち確率変数 T_1 と T_2 の独立性の仮定に加えて、右側打ち切り時間 C との独立性も仮定されている(Cox, 1959; 宮村, 1982)。

左側切断も生存時間解析ではよく知られた概念である(Klein and Moeschberger, 2003; 江村・道前, 2021)。上述の原爆被爆者疫学調査を想定して左側切断データを説明する。ただし、研究者の興味のある調査項目はがんの発症から死因を問わない全ての死亡までの時間であるとする。左側切断データで解析上の取り扱いの注意を要する対象は、観察開始である 1958 年の時点で既にがんを発症している対象(図 2 のケース 2 とケース 4)と、既にがんを発症し死亡している対象(図 2 のケース 5)の 3 つである。1958 年の時点で死亡している対象は観察できないため、データとしては収集できず、切り捨てとして扱われる。問題とすべきは、1958 年の時点でがんを発症しており、観察期間中に死亡または右側打ち切りとなる対象である。これらの対象については、がんの発症日を情報として収集可能なのであれば、がんの発症から 1958 年までの時間を切断時間 τ として記録することで、切断時間をもつ、すなわち観察開始まで生存したという条件付きの対象として解釈する。図 2 は切断時間をもつ対象ともたない対象の混在データを示している。

信頼性分野であれば、電力変圧器の寿命データが良い左側切断の例である(Hong et al., 2009; Ranjan et al., 2021; Mitra et al., 2021; 江村・道前, 2021)。観察開始は 1980 年、観察終了は 2008 年である。電力変圧器は観察開始以前から設置され続け 2008 年まで設置が続いた。研究者の興味のある調査項目は電力変圧器の設置から故障までの期間である(上述の例における、が

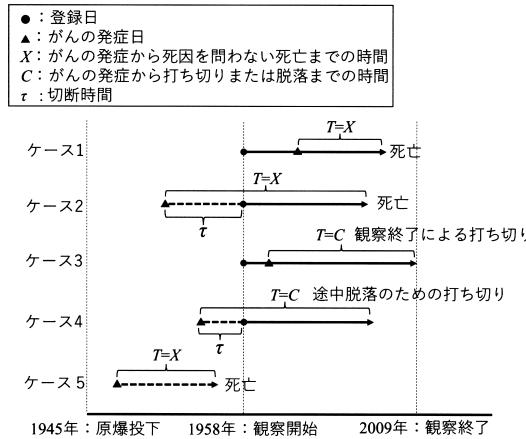


図 2. 左側切断・右側打ち切りデータの模式図.

んの発症日から死亡までの期間). 観察開始以前に設置され, かつ故障した電力変圧器はデータとして収集されないが(図 2 のケース 5 に相当), 故障しなかった電力変圧器は, 設置から観察開始までの期間を切断時間としてもつデータとして収集される(図 2 のケース 2 とケース 4 に相当). また, 観察開始以降に設置された電力変圧器は, 追跡対象となり観察期間中に故障または打ち切りとなる(図 2 のケース 1 とケース 3 に相当). 江村・道前(2021)では電力変圧器の寿命データの解析において, 複数のパラメトリック分布(例えばワイブル分布, ガンマ分布, 対数正規分布等)に基づく尤度関数の構成やパラメータの推定方法等について詳しく紹介している. 左側切断・右側打ち切りデータについての詳細な解説は, 江村・道前(2021)や Emura and Michimae(2022)を参照されたい.

現実のデータは複雑で複合的である. Kundu et al. (2017)は電力変圧器の寿命データを想定して競合リスクを伴う左側切断・右側打ち切りデータの解析のために新たなモデルを提案した. 先述の Hong et al. (2009)や Ranjan et al. (2021)では, 注目しているイベントが電力変圧器の故障だけであった. しかし, 競合リスク問題で説明したように, 2つの故障原因を同時に考慮するのであれば, これは競合リスクを伴う左側切断である. この競合リスクを伴う左側切断・右側打ち切りデータに対して, Kundu et al. (2017)は, 2つのイベントの潜在故障時間分布として, それぞれにワイブル分布を仮定し, かつ互いに独立であるという仮定下で, ワイブルパラメータ(すなわち生存時間分布)の推定法を提案した. 本稿では, Kundu et al. (2017)の提案手法を独立モデルと呼ぶ. しかし, 道前(2023)は, 独立モデルの仮定が成立しない場合, すなわちイベント間に相関がある場合, 独立モデルではワイブルパラメータを正確に推定できないことを示した.

近年, 競合リスク問題における独立性の仮定を緩めるため, イベント間の依存性をマーシャルオルキン分布やコピュラでモデリングする解析手法が提案されている(Wang et al., 2022; Michimae and Emura, 2022). Wang et al. (2022)は周辺分布をワイブル分布とした2変量マーシャルオルキン・ワイブル分布でイベント間の依存構造をモデリングし, 潜在故障時間の同時分布を構成, 最尤法によるパラメータ推定法を提案した. しかし, このモデリングでは依存構造と周辺分布は固定されており, 柔軟性という点で問題がある. 一方, Michimae and Emura (2022)は潜在故障時間の同時分布をコピュラで構成したコピュラモデルを提案した. コピュラを用いる大きな利点は周辺分布とは独立に依存関係を柔軟にモデリングできることであり(Nelsen, 2006), 言い換えれば周辺分布の選択が自由である. Michimae and Emura (2022)は,

相関した競合リスクを伴う左側切断・右側打ち切りデータにおいて、ワイブル分布を周辺分布としたコピュラ同時モデルに基づき尤度関数を構築し、最尤法で分布パラメータやコピュラパラメータを推定した。しかし、依存性のパラメータであるコピュラパラメータの推定は不安定で最尤法による推定の限界が示された。この問題に対して、近年 Michimae et al. (2024)は、より安定したパラメータの推定法として、ベイズ流コピュラモデルによる推定法を提案した。

本稿では、競合リスクを伴う左側切断・右側打ち切りデータにおいて、4つの解析モデルである独立モデル、コピュラモデル、ベイズ流独立モデル、そしてベイズ流コピュラモデルに基づくパラメータの推定性能を比較した研究結果を紹介する。特に、Michimae et al. (2024)で検討されていなかった競合リスクにおける各イベント発生割合の違いが、パラメータ推定性能にどの程度影響するのかを、モンテカルロ・シミュレーション実験を通じて4つの解析モデルで比較した。本稿の構成は次の通りである。2節では、競合リスクを伴う左側切断・右側打ち切りデータにおける、Kundu et al. (2017)の独立モデル、Michimae and Emura (2022)のコピュラモデル、Michimae et al. (2024)のベイズ流独立モデルとベイズ流コピュラモデルを紹介する。3節では、2節で紹介した4つのモデルのパラメータ推定性能を検証するためのシミュレーション方法とその結果について説明する。最後の4節では、考察として、今後の展望について意見を述べる。

2. 方法

第2節では、初めに、競合リスクを伴う左側切断・右側打ち切りデータの構造について、記法と共に説明する。続いて、2つの潜在故障時間の分布をパラメトリックなワイブル分布とする仮定下で、独立モデルとコピュラモデルにおける同時分布と尤度関数の構築方法について説明し、最後に最尤法とベイズ法を用いたワイブルパラメータ(形状と尺度パラメータ)とコピュラパラメータの推定法について説明する。

2.1 標本データ

1節で取り上げた競合リスクを伴う左側切断・右側打ち切りデータについて、上述の原爆被爆者疫学調査を例に挙げて説明する(図3)。研究者は対象 $i(1, \dots, n)$ について、がんの発症からがん死(イベント 1; $\delta_i = 1$)までの生存時間に興味があるが、がん以外の理由による死亡(イベント 2; $\delta_i = 2$)や打ち切り($\delta_i = 0$)等の理由のため、必ずしもがん死を観察できない。つまり、 T_{ji} を対象 i がイベント j で死亡する潜在故障時間 ($i = 1, \dots, n; j = 1, 2$), C_i を対象 i の観察終了による右側打ち切り時間(本稿では途中脱落を想定しない)とすると、研究者が対象 i について実際に観察できる時間 t_i は $t_i = \min(T_{1i}, T_{2i}, C_i)$ となる。また、観察開始である 1958 年の時点で既にがんを発症している対象(図3のケース 2, 4, 7)のうち、死亡している対象(ケース 7)は切り捨てとして扱われるが、ケース 2 と 4 は、がんの発症から 1958 年までの切断時間をもつ対象となる。以上、対象 i について左側切断時間 $\tau_i(\geq 0)$ の有無を示す指示変数を ν_i (左側切断; $\nu_i = 0$, 非切断; $\nu_i = 1$)とすると、競合リスクを伴う左側切断・右側打ち切りでの標本データは $\{(t_i, \tau_i, \delta_i, \nu_i); i = 1, \dots, n\}$ として記録される。

2.1節では、原爆被爆者疫学調査を想定し、“死亡”という事象の生起で説明したが、以降の節では“故障”という事象の生起で説明することとする。

2.2 独立モデル

Kundu et al. (2017)は、2つの潜在故障時間 T_1 と T_2 について、それぞれ独立にワイブル分布に従うと仮定し、各イベントにおけるワイブル分布の密度関数を $f_j(t) = \alpha_j \lambda_j t^{\alpha_j - 1} \exp(-\lambda_j t^{\alpha_j})$,

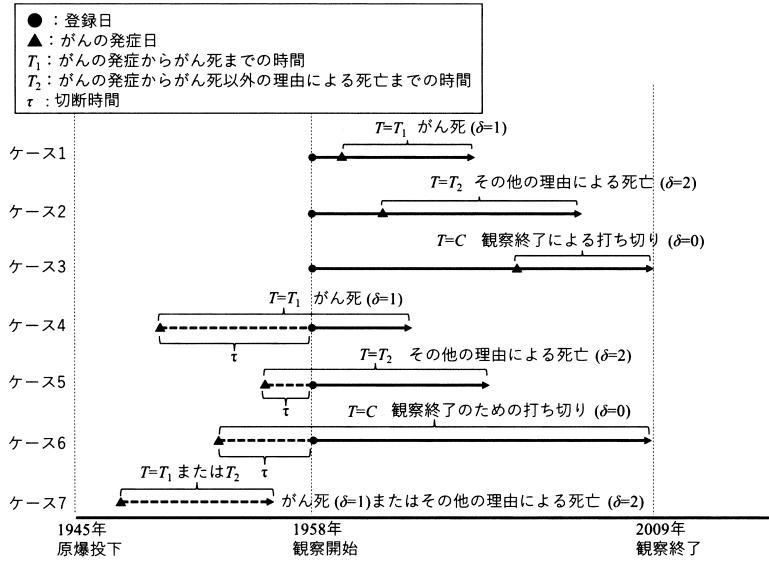


図 3. 競合リスクを伴う左側切断・右側打ち切りデータの模式図。

累積分布関数を $F_j(t) = 1 - \exp(-\lambda_j t^{\alpha_j})$, 生存関数を $S_j(t) = \exp(-\lambda_j t^{\alpha_j})$ とした。ただし, $t \geq 0$, $\alpha_j (> 0)$ と $\lambda_j (> 0)$ はそれぞれ形状パラメータと尺度パラメータである。これらパラメトリック分布の仮定の下, 独立モデルにおける尤度関数の構築方法を以下に説明する。

競合リスクを伴う左側切断・右側打ち切りの状況下において(図 3), 標本が得られるパターンとその尤度関数への貢献は次のケース 1 からケース 6 としてまとめられる (Kundu et al., 2017)。

- ケース 1 (イベント 1 で故障; $\delta_i = 1$, 切断時間なし; $\nu_i = 1$)

$$\begin{aligned} P(T_{1i} = t_i, T_{2i} > t_i) &= f_1(t_i) S_2(t_i) \\ &= \alpha_1 \lambda_1 t_i^{\alpha_1-1} \exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2}), i \in I_1 \end{aligned}$$

- ケース 2 (イベント 2 で故障; $\delta_i = 2$, 切断時間なし; $\nu_i = 1$)

$$\begin{aligned} P(T_{2i} = t_i, T_{1i} > t_i) &= f_2(t_i) S_1(t_i) \\ &= \alpha_2 \lambda_2 t_i^{\alpha_2-1} \exp(-\lambda_2 t_i^{\alpha_2}) \exp(-\lambda_1 t_i^{\alpha_1}), i \in I_2 \end{aligned}$$

- ケース 3 (打ち切り; $\delta_i = 0$, 切断時間なし; $\nu_i = 1$)

$$\begin{aligned} P(T_{1i} > t_i, T_{2i} > t_i) &= S_1(t_i) S_2(t_i) \\ &= \exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2}), i \in I_0 \end{aligned}$$

- ケース 4 (イベント 1 で故障; $\delta_i = 1$, 切断時間あり; $\nu_i = 0$)

$$\begin{aligned} P(T_{1i} = t_i, T_{2i} > t_i | T_{1i} > \tau_i, T_{2i} > \tau_i) &= f_1(t_i | T_i > \tau_i) S_2(t_i | T_i > \tau_i) \\ &= \frac{f_1(t_i)}{S_1(\tau_i)} \frac{S_2(t_i)}{S_2(\tau_i)} \\ &= \frac{\alpha_1 \lambda_1 t_i^{\alpha_1-1} \exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2})}{\exp(-\lambda_1 \tau_i^{\alpha_1}) \exp(-\lambda_2 \tau_i^{\alpha_2})}, i \in I_1 \end{aligned}$$

- ケース 5 (イベント 2 で故障 ; $\delta_i = 2$, 切断時間あり ; $\nu_i = 0$)

$$\begin{aligned}
 P(T_{2i} = t_i, T_{1i} > t_i | T_{1i} > \tau_i, T_{2i} > \tau_i) &= f_2(t_i | T_i > \tau_i) S_1(t_i | T_i > \tau_i) \\
 &= \frac{f_2(t_i)}{S_2(\tau_i)} \frac{S_1(t_i)}{S_1(\tau_i)} \\
 &= \frac{\alpha_2 \lambda_2 t_i^{\alpha_2-1} \exp(-\lambda_2 t_i^{\alpha_2}) \exp(-\lambda_1 t_i^{\alpha_1})}{\exp(-\lambda_1 \tau_i^{\alpha_1}) \exp(-\lambda_2 \tau_i^{\alpha_2})}, i \in I_2
 \end{aligned}$$

- ケース 6 (打ち切り ; $\delta_i = 0$, 切断時間あり ; $\nu_i = 0$)

$$\begin{aligned}
 P(T_{1i} > t_i, T_{2i} > t_i | T_{1i} > \tau_i, T_{2i} > \tau_i) &= S_1(t_i | T_i > \tau_i) S_2(t_i | T_i > \tau_i) \\
 &= \frac{S_1(t_i)}{S_1(\tau_i)} \frac{S_2(t_i)}{S_2(\tau_i)} \\
 &= \frac{\exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2})}{\exp(-\lambda_1 \tau_i^{\alpha_1}) \exp(-\lambda_2 \tau_i^{\alpha_2})}, i \in I_0
 \end{aligned}$$

ただし, $I_1 = \{i; T_{1i} \leq T_{2i}, T_{1i} \leq C_i\}$ をイベント 1 で故障した対象の集合, $I_2 = \{i; T_{2i} \leq T_{1i}, T_{2i} \leq C_i\}$ をイベント 2 で故障した対象の集合, $I_0 = \{i; C_i \leq T_{1i}, C_i \leq T_{2i}\}$ を右側打ち切りされた対象の集合とする. 切断時間をもつ対象 ($\nu_i = 0$) は, 観察開始まで生存したという条件の下で $t_i = \min(T_{1i}, T_{2i}, C_i)$ が観察される(図 3). すなわち, 条件付きの生存確率を求めていくことになる. 上述全てのケースについて考慮した尤度関数は以下の通りである.

$$\begin{aligned}
 (2.1) \quad L_{indep}(\alpha_1, \alpha_2, \lambda_1, \lambda_2) &= \prod_{i \in I_1} \{\alpha_1 \lambda_1 t_i^{\alpha_1-1} \exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2})\}^{\nu_i} \\
 &\quad \times \left\{ \frac{\alpha_1 \lambda_1 t_i^{\alpha_1-1} \exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2})}{\exp(-\lambda_1 \tau_i^{\alpha_1}) \exp(-\lambda_2 \tau_i^{\alpha_2})} \right\}^{1-\nu_i} \\
 &\quad \times \prod_{i \in I_2} \{\alpha_2 \lambda_2 t_i^{\alpha_2-1} \exp(-\lambda_2 t_i^{\alpha_2}) \exp(-\lambda_1 t_i^{\alpha_1})\}^{\nu_i} \\
 &\quad \times \left\{ \frac{\alpha_2 \lambda_2 t_i^{\alpha_2-1} \exp(-\lambda_2 t_i^{\alpha_2}) \exp(-\lambda_1 t_i^{\alpha_1})}{\exp(-\lambda_1 \tau_i^{\alpha_1}) \exp(-\lambda_2 \tau_i^{\alpha_2})} \right\}^{1-\nu_i} \\
 &\quad \times \prod_{i \in I_0} \{\exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2})\}^{\nu_i} \\
 &\quad \times \left\{ \frac{\exp(-\lambda_1 t_i^{\alpha_1}) \exp(-\lambda_2 t_i^{\alpha_2})}{\exp(-\lambda_1 \tau_i^{\alpha_1}) \exp(-\lambda_2 \tau_i^{\alpha_2})} \right\}^{1-\nu_i}.
 \end{aligned}$$

2.3 コピュラモデル

Michimae and Emura (2022) は, T_1 と T_2 の独立性の仮定を緩めるため, これら 2 つの潜在故障時間の依存性をコピュラでモデル化し, それに基づいて同時分布と尤度関数を構築した.

Michimae and Emura (2022) は, T_1 と T_2 の 2 変量同時生存関数をコピュラ C で以下のように定義した.

$$(2.2) \quad P(T_1 > t_i, T_2 > t_i) = C(S_1(t_i), S_2(t_i)).$$

実際のモデリングでは, 以下のクレイトンコピュラで実施したが, 他のコピュラでもモデリングは可能である.

$$(2.3) \quad C_\phi(u, v | \phi) = (u^{-\phi} + v^{-\phi} - 1)^{-\frac{1}{\phi}}, \phi \geq 0.$$

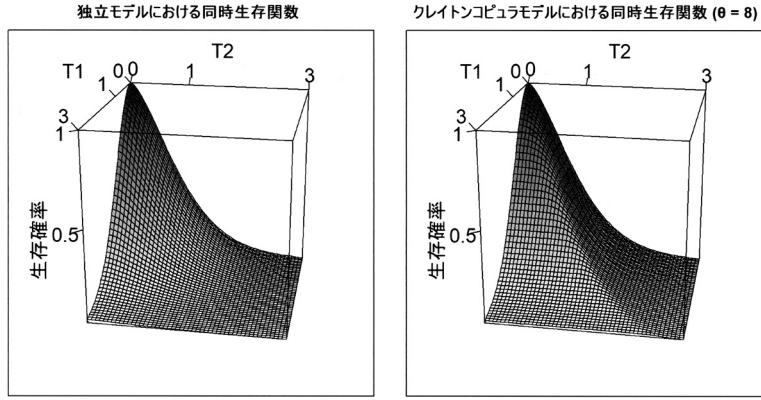


図 4. 独立モデルとコピュラモデルにおける同時生存関数.

ここに, ϕ はコピュラパラメータである. 参考までに, 周辺生存関数を $S_1(t) = \exp(-1.5t^{2.0})$ と $S_2(t) = \exp(-t^{1.5})$ とした場合の独立モデルとクレイトンコピュラモデル(コピュラパラメータ = 8)における同時生存関数を図 4 に示した. 図 4 では同時生存確率 $P(T_1 > t_1, T_2 > t_2)$ が曲面として示されている. 競合リスク下では, 先に起こったイベントだけが観察されるため, 同時生存確率は, $t_1 = t_2 = t$ を結ぶ尾根にあたる曲線上の確率 $P(T_1 > t, T_2 > t)$ となる. T_1 と T_2 の間に依存性があると, 独立な場合と比較して, 周辺生存関数は等しいが, 同時生存確率が高くなっている.

相関する競合リスクを伴う左側切断・右側打ち切りの状況下において, 標本が得られるパターンと周辺分布をワイブル分布としたときの尤度関数への貢献は, 次のケース 1 からケース 6 としてまとめられる.

- ケース 1 (イベント 1 で故障; $\delta_i = 1$, 切断時間なし; $\nu_i = 1$)

$$\begin{aligned} P(T_{1i} = t_i, T_{2i} > t_i) &= -\frac{\partial(u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}}}{\partial u_i} \frac{du_i}{dt_i} \\ &= \{\alpha_1 \lambda_1 t_i^{\alpha_1-1} u_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})}\} \end{aligned}$$

- ケース 2 (イベント 2 で故障; $\delta_i = 2$, 切断時間なし; $\nu_i = 1$)

$$\begin{aligned} P(T_{2i} = t_i, T_{1i} > t_i) &= -\frac{\partial(u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}}}{\partial v_i} \frac{dv_i}{dt_i} \\ &= \{\alpha_2 \lambda_2 t_i^{\alpha_2-1} v_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})}\} \end{aligned}$$

- ケース 3 (打ち切り; $\delta_i = 0$, 切断時間なし; $\nu_i = 1$)

$$P(T_{1i} > t_i, T_{2i} > t_i) = (u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}}$$

- ケース 4 (イベント 1 で故障; $\delta_i = 1$, 切断時間あり; $\nu_i = 0$)

$$\begin{aligned} P(T_{1i} = t_i, T_{2i} > t_i | T_{1i} > \tau_i, T_{2i} > \tau_i) &= -\frac{1}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}} \frac{\partial(u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}}}{\partial u_i} \frac{du_i}{dt_i} \\ &= \frac{\alpha_1 \lambda_1 t_i^{\alpha_1-1} u_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})}}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}} \end{aligned}$$

- ケース 5 (イベント 2 で故障 ; $\delta_i = 2$, 切断時間あり ; $\nu_i = 0$)

$$\begin{aligned} P(T_{2i} = t_i, T_{1i} > t_i | T_{1i} > \tau_i, T_{2i} > \tau_i) &= -\frac{1}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}} \frac{\partial (u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}}}{\partial v_i} \frac{dv_i}{dt_i} \\ &= \frac{\alpha_2 \lambda_2 t_i^{\alpha_2-1} v_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})}}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}} \end{aligned}$$

- ケース 6 (打ち切り ; $\delta_i = 0$, 切断時間あり ; $\nu_i = 0$)

$$P(T_{1i} > t_i, T_{2i} > t_i | T_{1i} > \tau_i, T_{2i} > \tau_i) = \frac{(u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}}}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}}$$

ここに, $u_i = \exp(-\lambda_1 t_i^{\alpha_1})$, $v_i = \exp(-\lambda_2 t_i^{\alpha_2})$, $x_i = \exp(-\lambda_1 \tau_i^{\alpha_1})$, $y_i = \exp(-\lambda_2 \tau_i^{\alpha_2})$ である.

上述全てのケースについて考慮した尤度関数は以下の通りである.

$$\begin{aligned} (2.4) \quad L_{copula}(\alpha_1, \alpha_2, \lambda_1, \lambda_2, \phi) &= \prod_{i \in I_1} \{ \alpha_1 \lambda_1 t_i^{\alpha_1-1} u_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})} \}^{\nu_i} \\ &\quad \times \left\{ \frac{\alpha_1 \lambda_1 t_i^{\alpha_1-1} u_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})}}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}} \right\}^{1-\nu_i} \\ &\quad \times \prod_{i \in I_2} \{ \alpha_2 \lambda_2 t_i^{\alpha_2-1} v_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})} \}^{\nu_i} \\ &\quad \times \left\{ \frac{\alpha_2 \lambda_2 t_i^{\alpha_2-1} v_i^{-\phi} (u_i^{-\phi} + v_i^{-\phi} - 1)^{-(1+\frac{1}{\phi})}}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}} \right\}^{1-\nu_i} \\ &\quad \times \prod_{i \in I_0} \{ (u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}} \}^{\nu_i} \left\{ \frac{(u_i^{-\phi} + v_i^{-\phi} - 1)^{-\frac{1}{\phi}}}{(x_i^{-\phi} + y_i^{-\phi} - 1)^{-\frac{1}{\phi}}} \right\}^{1-\nu_i}. \end{aligned}$$

2.4 ベイズモデル

Michimae et al. (2024)ではコピュラモデル (Michimae and Emura, 2022)をベイズ流コピュラモデルへと発展させた. ベイズ流コピュラモデルによるパラメータ推定には, 尤度関数 L_{copula} に加え, サイブルパラメータとコピュラパラメータの事前分布の設定が必要である. 本稿では, サイブルパラメータとコピュラパラメータの事前分布について弱情報の半コーシー分布 $\alpha_1, \alpha_2, \lambda_1, \lambda_2, \phi \sim Cauchy(0, 5)$ と弱情報の逆ガンマ分布 $\alpha_1, \alpha_2, \lambda_1, \lambda_2, \phi \sim Inv-Gamma(0.001, 0.001)$ を設定した(図 5). 半コーシー分布の密度関数は, μ を位置パラメータ, $\gamma (> 0)$ を尺度パラメータとして,

$$Cauchy(\mu, \gamma): = f(x) = \begin{cases} \frac{1}{\pi} \frac{\gamma}{(x - \mu)^2 + \gamma^2}, & x \geq \mu \\ 0, & \text{それ以外} \end{cases}$$

として与えられる. また逆ガンマ分布の密度関数は, $a (> 0)$ を形状パラメータ, $b (> 0)$ を尺度パラメータとして,

$$Inv-Gamma(a, b): = f(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right)$$

で与えられる. ただし, $\Gamma(\cdot)$ はガンマ関数である.

その他の事前分布の詳細な設定については, Michimae et al. (2024)を参照されたい. ベイズ

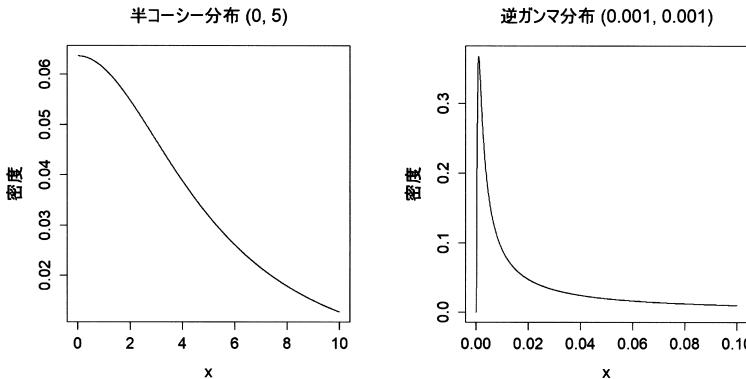


図 5. 本稿の事前分布として用いた半コーシー分布と逆ガンマ分布.

流コピュラモデルにおけるパラメータの同時事後分布は、上記密度関数で与えられる事前分布の下、以下の通りである。

$$(2.5) \quad f(\alpha_1, \alpha_2, \lambda_1, \lambda_2, \phi) \propto L_{copula}(\alpha_1, \alpha_2, \lambda_1, \lambda_2, \phi) f(\alpha_1) f(\alpha_2) f(\lambda_1) f(\lambda_2) f(\phi).$$

また、ベイズ流独立モデルにおけるパラメータの同時事後分布は以下の通りである。

$$(2.6) \quad f(\alpha_1, \alpha_2, \lambda_1, \lambda_2) \propto L_{indep}(\alpha_1, \alpha_2, \lambda_1, \lambda_2) f(\alpha_1) f(\alpha_2) f(\lambda_1) f(\lambda_2).$$

3. シミュレーション実験

競合リスクにおける各イベント発生割合の違いが、パラメータ推定性能にどの程度影響するのかを調べるために、潜在故障時間 T_1 と T_2 の相関、周辺分布(ワイブル分布)のパラメータ、切斷時間をもつ対象の割合を細かく設定したシミュレーションデータを用いた。シミュレーション実験では、これらシミュレーションデータにおいて、独立モデル、コピュラモデル、ベイズ流独立モデル、そしてベイズ流コピュラモデルのパラメータ推定性能を比較する。

3.1 シミュレーション設定

サンプルサイズを $n = 100$ と $n = 500$ の 2 パターン設定した。各サンプルサイズについて、ワイブルパラメータの真値の組み合わせについて 3 パターン、 T_1 と T_2 の相関について 2 パターン、切斷時間をもつ対象($\nu = 0$)と、もたない対象($\nu = 1$)の割合の組み合わせを 4 パターン、合計 24 (=3×2×4) パターンを設定した(表 1)。シミュレーションのデータ生成は、縦断研究を参考に観察開始日を s 、観察終了日を e として(図 3)、シミュレーションパターンごとに、以下の 3 ステップで実施した。

• ステップ 1

$\nu_i = 0$ のときデータ発生点 B_i (例えば発症日)を一様分布 $B_i \sim Uniform(0, s)$ で発生させ、 $\nu_i = 1$ のとき $B_i \sim Uniform(s, e)$ で発生させた。 $\nu_i = 0$ のとき切斷時間 $\tau_i = s - B_i$ 、 $\nu_i = 1$ のとき $\tau_i = 0$ とした。また、 $C_i = e - B_i$ である。この打ち切りは観察終了による打ち切り例であり、本稿のシミュレーションでは途中脱落例を想定していない。

• ステップ 2

独立した潜在故障時間を生成する場合(ケンドール順位相関係数 0 に対応)は、一様乱数 $u_{1i} \sim Uniform(0, 1)$ 、 $u_{2i} \sim Uniform(0, 1)$ をそれぞれ独立に n 例発生させ、ステップ 3 に進ん

表 1. シミュレーションの詳細設定。

ワイブルパラメータの真値	観察開始時間 s	コピュラパラメータ 観察打ち切り時間 e の真値 (相関係数)	切断割合
			0.2
		$\phi = 0 (0)$	0.4
			0.6
設定1			0.8
$\alpha_1 = 1.5, \lambda_1 = 1.0, \alpha_2 = 1.0, \lambda_2 = 1.0$	$s = 3, e = 4$		0.2
		$\phi = 2 (0.5)$	0.4
			0.6
			0.8
			0.2
		$\phi = 0 (0)$	0.4
			0.6
設定2			0.8
$\alpha_1 = 0.5, \lambda_1 = 1.0, \alpha_2 = 1.0, \lambda_2 = 1.0$	$s = 6, e = 7$		0.2
		$\phi = 2 (0.5)$	0.4
			0.6
			0.8
			0.2
		$\phi = 0 (0)$	0.4
			0.6
設定3			0.8
$\alpha_1 = 1.5, \lambda_1 = 1.0, \alpha_2 = 0.5, \lambda_2 = 1.0$	$s = 3, e = 4$		0.2
		$\phi = 2 (0.5)$	0.4
			0.6
			0.8

だ。相関した潜在故障時間を生成する場合は 2 変量クレイトンコピュラを用いた。コピュラパラメータ $\phi = 2$ (ケンドール順位相関係数 0.5 に対応) の設定下で、相関のある乱数の組 (u_{1i}, u_{2i}) を n 例発生させ、ステップ 3 へ進んだ。コピュラやコピュラを用いた乱数生成についての解説は本稿の目的ではないので、興味のある読者は、次の参考文献を参照されたい (戸坂・吉羽, 2005; Nelsen, 2006; 塚原, 2012; 江村・道前, 2020)。

• ステップ 3

逆関数法で T_{1i} を $T_{1i} \sim S_1^{-1}(u_{1i})$, T_{2i} を $T_{2i} \sim S_2^{-1}(u_{2i})$ として潜在故障時間を生成した。すなわち, $T_{1i} = (-\frac{\log(u_{1i})}{\lambda_1})^{\frac{1}{\alpha_1}}$ または $T_{2i} = (-\frac{\log(u_{2i})}{\lambda_2})^{\frac{1}{\alpha_2}}$ として生成される。ワイブル分布のハザード関数では、 $\alpha_1 = 0.5, \lambda_1 = 1.0$ の場合は減少ハザード, $\alpha_1 = 1.0, \lambda_1 = 1.0$ の場合は一定ハザード, $\alpha_1 = 1.5, \lambda_1 = 1.0$ の場合は増加ハザードに対応している。ステップ 1 で計算された C_i 及び潜在故障時間 T_{1i} と T_{2i} から $t_i = \min(T_{1i}, T_{2i}, C_i)$ を観察時間とした。 $t_i = T_{1i}$ なら $\delta_i = 1$, $t_i = T_{2i}$ なら $\delta_i = 2$, $t_i = C_i$ なら $\delta_i = 0$ である。

以上、3 つのステップから目的のシミュレーションデータ $\{(t_i, \tau_i, \delta_i, \nu_i); i = 1, \dots, n\}$ が生成される。本稿では、 T_{1i} と T_{2i} が互いに異なるワイブル分布に従う場合を想定したが、同一の分布に従う場合の結果については、Michimae et al. (2024) を参照されたい。

3.2 パラメータ推定と評価

シミュレーションパターン(表 1)ごとに、シミュレーションデータの生成からパラメータ推定までの作業を 1000 回繰り返した。パラメータは、最尤法とベイズ法で推定した。独立モデルにおける最尤推定量 $\hat{\theta}_{indep, MLE} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\lambda}_1, \hat{\lambda}_2)$ は、式(2.1)から、

$$(3.1) \quad \hat{\theta}_{indep, MLE} = \operatorname{argmax}_{(\alpha_1, \alpha_2, \lambda_1, \lambda_2)} \{ \log L_{indep}(\alpha_1, \alpha_2, \lambda_1, \lambda_2) \}$$

であり、同様に、コピュラモデルにおける最尤推定量 $\hat{\theta}_{copula, MLE} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\phi})$ は、式(2.4)から、

$$(3.2) \quad \hat{\theta}_{copula, MLE} = \operatorname{argmax}_{(\alpha_1, \alpha_2, \lambda_1, \lambda_2, \phi)} \{ \log L_{copula}(\alpha_1, \alpha_2, \lambda_1, \lambda_2, \phi) \}$$

となる。これらパラメータの推定等には optim 関数(パラメータの初期値は全て 0)で実施した。ただし、ワイルパラメータとコピュラパラメータは正であるため、例えば $\alpha_1 = \exp(\cdot)$ とすることにより、正の制約の条件下でパラメータを推定した。また、パラメータの 95% 信頼区間は、 α_1 であれば、

$$\hat{\alpha}_1 \pm 1.96 \times \operatorname{SE}(\hat{\alpha}_1)$$

として構成される。ここに、 $\operatorname{SE}(\hat{\alpha}_1)$ は $\hat{\alpha}_1$ の標準誤差である。95% 信頼区間の被覆確率は 1000 個のシミュレーションデータに基づき算出した。

一方で、ベイズ流独立モデルにおけるベイズ推定量 $\hat{\theta}_{indep, Bayes}$ とベイズ流コピュラモデルにおけるベイズ推定量 $\hat{\theta}_{copula, Bayes}$ は、それぞれ式(2.5)と(2.6)で与えられる同時事後分布から得られる周辺事後平均となる。最尤推定と同様に、95% 信頼区間とその被覆確率も算出した。パラメータの推定には RStan (Stan Development Team, 2017; 2018) を用いた。Stan は MCMC (マルコフ連鎖モンテカルロ法) アルゴリズムの改良型である No-U-Turn サンプラー (Hoffman and Gelman, 2014) を使用しており、複雑な事後分布から効率的にサンプリングすることが可能である。同時事後分布の推定では、マルコフ連鎖の数を 4、各連鎖の反復回数を 4000、各連鎖のワームアップ反復回数を 2000 に設定した。MCMC アルゴリズムが収束したかどうかについては、Rhat 基準 (Gelman and Rubin, 1992) で判定した(1.10 以下で収束)。

独立モデル、コピュラモデル、ベイズ流独立モデル、ベイズ流コピュラモデルのパラメータ推定性能を評価するため、平均二乗誤差(MSE: Mean Squared Error)をパラメータごとに以下の式で算出した。例えば α_1 を真値として、 α_1 の MSE であれば、

$$\operatorname{MSE}(\alpha_1) = \frac{1}{1000} \sum_{r=1}^{1000} (\hat{\alpha}_{1r} - \alpha_1)^2$$

である。ここで $\hat{\alpha}_{1r}$ は r 回目の推定値である。他のパラメータについても同様に MSE を算出した。さらに、時点 0.5, 1, および 2 における平均生存確率を以下の式を用いて推定した。例えば、各シミュレーションパターンにおけるイベント 1 と 2 の時点 0.5 の生存確率は、それぞれ、

$$\begin{aligned} \hat{S}_1(0.5) &= \frac{1}{1000} \sum_{r=1}^{1000} \exp(-\hat{\lambda}_{1r} 0.5^{\hat{\alpha}_{1r}}) \\ \hat{S}_2(0.5) &= \frac{1}{1000} \sum_{r=1}^{1000} \exp(-\hat{\lambda}_{2r} 0.5^{\hat{\alpha}_{2r}}) \end{aligned}$$

として推定される。

3.3 シミュレーション結果

3.3.1 イベントが独立の場合($\phi = 0$)

最尤推定, ベイズ推定(逆ガンマ分布と半コーシー分布)にかかわらず, 独立モデルのワイルパラメータに関する MSE はコピュラモデルと比較して小さい値を示した(表 2-4). この結果は, 独立データ($\phi = 0$)の場合は, 独立モデルの推定性能がコピュラモデルと比較して優れていることを示している. また, 小標本サイズ($n = 100$)と比較して大標本サイズ($n = 500$)では, 推定方法にかかわらず, MSE は小さい値を示し(表 2-4), ワイルパラメータ(コピュラモデルであればとコピュラパラメータも含め)の推定精度が改善することも示された.

独立モデルにおける最尤推定とベイズ推定の比較において, 小標本サイズでは, ベイズ推定(逆ガンマ分布もしくは半コーシー分布)による MSE が小さい値を示したが, 大標本サイズではほぼ同じ値を示し, 最尤推定とベイズ推定は同様の推定性能であることが示された. 一方, コピュラモデルでは, 小標本サイズと大標本サイズの両方において, 最尤推定と比較してベイズ推定のワイルパラメータに関する MSE は小さい値を示し, 特に逆ガンマ分布による MSE 値が小さく, その推定性能の高さを示した. この傾向はコピュラパラメータの推定においても同様であった.

独立モデルまたはコピュラモデルにかかわらず, おおよその傾向として, 切断割合の増加に伴ってワイルパラメータの MSE 値は増加し, パラメータの推定性能が低下する傾向にあった(表 2-4). 一方, コピュラパラメータの推定については, そのような傾向はなかった.

サンプルサイズにかかわらず, シミュレーション設定 1 では切断割合が増加するにつれ, イベント 1 の割合が増加し, イベント 2 の割合はほとんど変化せず, 打ち切り割合が減少した(表 2). 設定 2 では切断割合が増加するにつれ, イベント 1 の割合が減少し, イベント 2 の割合が増加する一方で, 打ち切り割合の変化は小さかった(表 3). 設定 3 では切断割合が増加するにつれ, イベント 1 の割合が増加し, イベント 2 の割合が減少するとともに, 打ち切り割合も減少した(表 3).

3.3.2 イベントが相關している場合($\phi = 2$)

独立モデルと比較してコピュラモデルのワイルパラメータに関する MSE は小さい値を示した(表 2-4). この結果は, 最尤推定, ベイズ推定にかかわらず, 相関データ($\phi = 2$)の場合は, コピュラモデルの推定性能が優れていることを示している. また, 独立モデルでは, 最尤推定, ベイズ推定にかかわらず, 標本サイズが大きくなても MSE 値はほとんど差を示さなかつたが, コピュラモデルでは MSE 値は小さい値を示し, パラメータの推定性能の改善が示された.

独立モデルでは, 標本サイズにかかわらず, 最尤推定と比較してベイズ推定のワイルパラメータに関する MSE 値は小さく, ベイズ推定が推定性能の点で優れていることが示された. 同様の傾向はコピュラモデルにおいても示された. コピュラモデルにおけるコピュラパラメータの推定においても, ベイズ推定, 特に逆ガンマ分布の優位性が示された. しかし, 小標本サイズと比較して大標本サイズでは, 最尤推定とベイズ推定の MSE 値の差は小さくなり, パラメータの推定性能の差は縮まった.

切断割合と各種パラメータの MSE 値との関係は, イベントが独立している場合と同様の傾向を示した(表 2-4). さらに, 切断割合と各イベントの発生割合の関係についても, 同様の傾向が見られた.

3.3.3 パラメータと生存確率の推定値

パラメータ推定値とその被覆確率の結果を Supplementary Information 表 S1-S3 に示した. イベントが独立であれば, 独立モデルは最尤推定やベイズ推定にかかわらず, 真値に近い推定値を与え, 大標本サイズでは, さらに真値に近くなる傾向があった. また, 被覆確率もほぼ

表 2. ワイブルパラメータの真値の設定 1 における MSE の比較結果。表の数値は 1000 回のモンテカルロ・シミュレーションの平均値。

n	切断割合	コピュラパラメータ (相関係数)	モデル	MSE					打ち切り割合	イベント1割合	イベント2割合
				$\lambda_1 = 1.0$	$\alpha_1 = 1.5$	$\lambda_2 = 1.0$	$\alpha_2 = 1.0$	ϕ			
0.2	$\phi=0 (0)$	独立モデル	最尤推定	0.0375	0.0277	0.0371	0.0171				
			逆ガンマ	0.0369	0.0273	0.0363	0.0166				
		半コーシー	0.0383	0.0270	0.0391	0.0171			0.39	0.29	0.32
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1297	0.0313	0.1647	0.0427	10.2013			
			逆ガンマ	0.0471	0.0280	0.0492	0.0202	0.1048			
		半コーシー	0.1546	0.0260	0.1956	0.0534	1.6800				
0.4	$\phi=0 (0)$	独立モデル	最尤推定	0.1103	0.0481	0.1233	0.0486				
			逆ガンマ	0.1135	0.0471	0.1274	0.0511				
		半コーシー	0.0995	0.0461	0.1080	0.0423			0.48	0.26	0.26
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0584	0.0391	0.0569	0.0233	163.9347			
			逆ガンマ	0.0502	0.0388	0.0574	0.0248	8.5221			
		半コーシー	0.0432	0.0263	0.0354	0.0166	7.4765				
100	$\phi=0 (0)$	独立モデル	最尤推定	0.0354	0.0250	0.0319	0.0154				
			逆ガンマ	0.0349	0.0250	0.0317	0.0152				
		半コーシー	0.0354	0.0241	0.0324	0.0152			0.31	0.37	0.32
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1297	0.0301	0.1311	0.0457	1.9243			
			逆ガンマ	0.0430	0.0270	0.0408	0.0202	0.0595			
		半コーシー	0.1477	0.0252	0.1711	0.0610	1.2453				
0.6	$\phi=0 (0)$	独立モデル	最尤推定	0.1172	0.0601	0.1645	0.0796				
			逆ガンマ	0.1166	0.0609	0.1689	0.0827				
		半コーシー	0.1040	0.0557	0.1492	0.0728			0.40	0.37	0.23
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0561	0.0401	0.0600	0.0239	78.4922			
			逆ガンマ	0.0484	0.0432	0.0643	0.0286	3.5651			
		半コーシー	0.0584	0.0266	0.0367	0.0160	4.0288				
0.8	$\phi=0 (0)$	独立モデル	最尤推定	0.0435	0.0289	0.0345	0.0162				
			逆ガンマ	0.0427	0.0292	0.0340	0.0162				
		半コーシー	0.0430	0.0272	0.0351	0.0158			0.22	0.45	0.33
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1182	0.0355	0.1253	0.0531	8.7483			
			逆ガンマ	0.0519	0.0321	0.0442	0.0221	0.0266			
		半コーシー	0.1745	0.0290	0.1794	0.0718	1.2248				
0.8	$\phi=0 (0)$	独立モデル	最尤推定	0.1473	0.0935	0.1861	0.1092				
			逆ガンマ	0.1426	0.0949	0.1891	0.1119				
		半コーシー	0.1273	0.0826	0.1693	0.1020			0.32	0.48	0.21
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0605	0.0425	0.0591	0.0314	104.8503			
			逆ガンマ	0.0575	0.0557	0.0667	0.0356	2.7513			
		半コーシー	0.0400	0.0281	0.0365	0.0208	3.0425				
0.8	$\phi=0 (0)$	独立モデル	最尤推定	0.0501	0.0316	0.0497	0.0245				
			逆ガンマ	0.0498	0.0316	0.0488	0.0244				
		半コーシー	0.0513	0.0287	0.0497	0.0231			0.14	0.53	0.33
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1534	0.0381	0.1422	0.0925	8.5129			
			逆ガンマ	0.0602	0.0346	0.0588	0.0354	0.0158			
		半コーシー	0.2206	0.0276	0.1936	0.1035	1.0503				
0.8	$\phi=0 (0)$	独立モデル	最尤推定	0.1949	0.1549	0.1914	0.1522				
			逆ガンマ	0.1827	0.1543	0.1889	0.1531				
		半コーシー	0.1622	0.1305	0.1687	0.1436			0.24	0.59	0.18
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0784	0.0564	0.0752	0.0409	69.1492			
			逆ガンマ	0.0803	0.0839	0.0841	0.0446	2.5775			
		半コーシー	0.0538	0.0377	0.0514	0.0240	2.5293				

表 2. (つづき)

0.2	$\phi=0$ (0)	独立モデル	最尤推定	0.0076	0.0052	0.0072	0.0035		0.39	0.29	0.32	
		逆ガンマ	0.0075	0.0052	0.0072	0.0035						
		半コーシー	0.0076	0.0052	0.0073	0.0035						
		独立モデル	最尤推定	0.0377	0.0131	0.0297	0.0102	0.1626				
		逆ガンマ	0.0098	0.0059	0.0105	0.0047	0.0071					
	$\phi=2$ (0.5)	半コーシー	0.0276	0.0077	0.0366	0.0129	0.1629					
		独立モデル	最尤推定	0.0920	0.0109	0.1143	0.0441					
		逆ガンマ	0.0926	0.0107	0.1151	0.0447						
		半コーシー	0.0898	0.0108	0.1110	0.0427						
		独立モデル	最尤推定	0.0106	0.0074	0.0123	0.0051	1.3426	0.48	0.26	0.26	
0.4	$\phi=0$ (0)	コピュラモデル	逆ガンマ	0.0097	0.0073	0.0118	0.0049	0.8844		0.31	0.37	0.32
		半コーシー	0.0095	0.0063	0.0100	0.0044	1.0392					
		独立モデル	最尤推定	0.0069	0.0045	0.0064	0.0029					
		逆ガンマ	0.0069	0.0045	0.0064	0.0029						
		半コーシー	0.0069	0.0045	0.0064	0.0029						
	$\phi=2$ (0.5)	独立モデル	最尤推定	0.0190	0.0073	0.0259	0.0098	0.2351		0.40	0.37	0.23
		逆ガンマ	0.0090	0.0054	0.0095	0.0043	0.0063					
		半コーシー	0.0255	0.0075	0.0335	0.0134	0.1337					
		独立モデル	最尤推定	0.0983	0.0206	0.1538	0.0763					
		逆ガンマ	0.0983	0.0206	0.1547	0.0769						
500	$\phi=0$ (0)	半コーシー	0.0957	0.0202	0.1507	0.0749			0.40	0.37	0.23	
		独立モデル	最尤推定	0.0118	0.0079	0.0132	0.0055	1.6119				
		逆ガンマ	0.0105	0.0077	0.0123	0.0052	0.7758					
		半コーシー	0.0103	0.0065	0.0107	0.0046	0.9240					
		独立モデル	最尤推定	0.0077	0.0047	0.0070	0.0032					
	$\phi=2$ (0.5)	逆ガンマ	0.0077	0.0047	0.0070	0.0032			0.22	0.45	0.32	
		半コーシー	0.0077	0.0046	0.0070	0.0032						
		独立モデル	最尤推定	0.0203	0.0075	0.0243	0.0119	0.0603				
		逆ガンマ	0.0102	0.0056	0.0104	0.0054	0.0061					
		半コーシー	0.0317	0.0076	0.0359	0.0180	0.1751					
0.6	$\phi=0$ (0)	独立モデル	最尤推定	0.1243	0.0396	0.1775	0.1085			0.32	0.48	0.20
		逆ガンマ	0.1235	0.0398	0.1781	0.1090						
		半コーシー	0.1203	0.0385	0.1740	0.1069						
		独立モデル	最尤推定	0.0136	0.0097	0.0140	0.0066	1.0558				
		逆ガンマ	0.0126	0.0097	0.0134	0.0063	0.8365					
	$\phi=2$ (0.5)	半コーシー	0.0114	0.0077	0.0111	0.0055	0.8393					
		独立モデル	最尤推定	0.0107	0.0062	0.0100	0.0045					
		逆ガンマ	0.0107	0.0062	0.0099	0.0045						
		半コーシー	0.0108	0.0061	0.0100	0.0044						
		独立モデル	最尤推定	0.0418	0.0099	0.0368	0.0257	0.4157	0.14	0.53	0.33	
0.8	$\phi=0$ (0)	コピュラモデル	逆ガンマ	0.0142	0.0073	0.0131	0.0085	0.0232		0.24	0.59	0.18
		半コーシー	0.0664	0.0093	0.0543	0.0349	0.5085					
		独立モデル	最尤推定	0.1729	0.0792	0.1734	0.1536					
		逆ガンマ	0.1704	0.0792	0.1730	0.1537						
		半コーシー	0.1658	0.0760	0.1688	0.1518						
	$\phi=2$ (0.5)	独立モデル	最尤推定	0.0155	0.0107	0.0156	0.0105	2.5819		0.24	0.59	0.18
		逆ガンマ	0.0149	0.0117	0.0151	0.0097	0.7558					
		半コーシー	0.0126	0.0086	0.0126	0.0085	0.9076					

95% 程度となった(最尤推定で一部 80% 程度). しかし, コピュラモデルでは, 推定値は過大推定される傾向にあり, 大標本サイズでは, やや真値に近くはなるが過大推定の傾向であった. また, 被覆確率については逆ガンマでは 95% 程度であったが, 最尤推定と半コーシー分布のベイズ推定では, 95% を下回ることが多かった. コピュラパラメータについては, 逆ガンマによ

表 3. ワイブルパラメータの真値の設定 2 における MSE の比較結果。表の数値は 1000 回のモンテカルロ・シミュレーションの平均値。

n	切断割合	コピュラパラメータ (相関係数)	モデル	MSE					打ち切り割合	イベント1割合	イベント2割合
				$\lambda_1 = 1.0$	$\alpha_1 = 0.5$	$\lambda_2 = 1.0$	$\alpha_2 = 1.0$	ϕ			
0.2	$\phi=0 (0)$	独立モデル	最尤推定	0.0313	0.0046	0.0376	0.0118				
			逆ガンマ	0.0308	0.0044	0.0370	0.0117				
		半コーシー	0.0331	0.0048	0.0384	0.0115					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.5779	0.0097	286.7544	0.0132	4.8704	0.34	0.36	0.30
			逆ガンマ	0.0401	0.0055	0.0501	0.0115	0.4962			
		半コーシー	0.1175	0.0134	0.1201	0.0105	1.4928				
0.4	$\phi=0 (0)$	独立モデル	最尤推定	0.0679	0.0079	0.1451	0.0330				
			逆ガンマ	0.0694	0.0081	0.1496	0.0318				
		半コーシー	0.0565	0.0066	0.1320	0.0316					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0423	0.0067	0.0666	0.0195	49.6640	0.44	0.33	0.23
			逆ガンマ	0.0377	0.0056	0.0565	0.0202	4.6454			
		半コーシー	0.0337	0.0058	0.0502	0.0117	6.0326				
100	$\phi=0 (0)$	独立モデル	最尤推定	0.0302	0.0045	0.0359	0.0108				
			逆ガンマ	0.0297	0.0044	0.0354	0.0109				
		半コーシー	0.0317	0.0045	0.0356	0.0104					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0936	0.0115	0.0761	0.0120	2.3475	0.32	0.32	0.36
			逆ガンマ	0.0397	0.0057	0.0424	0.0112	0.1828			
		半コーシー	0.1349	0.0164	0.1043	0.0101	0.7445				
0.6	$\phi=0 (0)$	独立モデル	最尤推定	0.0972	0.0134	0.1458	0.0406				
			逆ガンマ	0.0998	0.0140	0.1458	0.0411				
		半コーシー	0.0832	0.0116	0.1299	0.0378					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0553	0.0074	0.0731	0.0199	41.7721	0.43	0.26	0.31
			逆ガンマ	0.0459	0.0063	0.0569	0.0215	5.1855			
		半コーシー	0.0411	0.0057	0.0573	0.0126	6.1253				
0.8	$\phi=0 (0)$	独立モデル	最尤推定	0.0395	0.0052	0.0481	0.0120				
			逆ガンマ	0.0392	0.0051	0.0473	0.0123				
		半コーシー	0.0402	0.0051	0.0481	0.0114					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1114	0.0164	0.1023	0.0131	1.1302	0.30	0.27	0.43
			逆ガンマ	0.0501	0.0072	0.0586	0.0126	0.0860			
		半コーシー	0.1582	0.0234	0.1428	0.0109	0.6192				
0.0	$\phi=0 (0)$	独立モデル	最尤推定	0.1478	0.0210	0.1653	0.0553				
			逆ガンマ	0.1514	0.0219	0.1606	0.0571				
		半コーシー	0.1290	0.0184	0.1415	0.0494					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0654	0.0106	0.0753	0.0204	72.9473	0.42	0.19	0.39
			逆ガンマ	0.0624	0.0082	0.0628	0.0254	6.5301			
		半コーシー	0.0462	0.0074	0.0582	0.0135	6.7793				
1.0	$\phi=0 (0)$	独立モデル	最尤推定	0.0611	0.0090	0.0663	0.0171				
			逆ガンマ	0.0600	0.0090	0.0650	0.0174				
		半コーシー	0.0610	0.0088	0.0659	0.0153					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1655	0.0403	0.1834	0.0193	12.4432	0.29	0.22	0.49
			逆ガンマ	0.0728	0.0129	0.0767	0.0178	0.0186			
		半コーシー	0.2046	0.0422	0.2133	0.0140	0.7279				
1.2	$\phi=0 (0)$	独立モデル	最尤推定	0.2005	0.0345	0.2060	0.0853				
			逆ガンマ	0.2031	0.0354	0.1913	0.0868				
		半コーシー	0.1717	0.0304	0.1636	0.0702					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0913	0.0194	0.0964	0.0261	141.5546	0.40	0.12	0.48
			逆ガンマ	0.0929	0.0134	0.0825	0.0377	10.4212			
		半コーシー	0.0606	0.0117	0.0704	0.0169	8.2485				

る推定値が真値に近かった。

イベントが相関している場合、独立モデルでは最尤推定やベイズ推定にかかわらず、推定値は過小評価または過大評価され、大標本サイズでは被覆確率が 0% 近くまで低下した。しかし、

表 3. (つづき)

0.2	$\phi=0 (0)$	独立モデル	最尤推定	0.0064	0.0009	0.0069	0.0022	0.34	0.36	0.30	
		逆ガンマ	最尤推定	0.0064	0.0009	0.0068	0.0022				
		半コーシー	最尤推定	0.0064	0.0009	0.0069	0.0022				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0110	0.0014	0.0124	0.0026	0.0150	0.44	0.33	0.23
		逆ガンマ	最尤推定	0.0079	0.0011	0.0089	0.0023	0.0059			
		半コーシー	最尤推定	0.0142	0.0018	0.0156	0.0025	0.0276			
	$\phi=0 (0)$	独立モデル	最尤推定	0.0559	0.0059	0.1292	0.0123				
		逆ガンマ	最尤推定	0.0561	0.0059	0.1303	0.0120				
		半コーシー	最尤推定	0.0533	0.0056	0.1267	0.0122				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0090	0.0012	0.0122	0.0037	0.8400	0.44	0.33	0.23
		逆ガンマ	最尤推定	0.0090	0.0012	0.0115	0.0037	0.6887			
		半コーシー	最尤推定	0.0086	0.0012	0.0113	0.0031	0.8018			
0.4	$\phi=0 (0)$	独立モデル	最尤推定	0.0065	0.0009	0.0070	0.0020				
		逆ガンマ	最尤推定	0.0065	0.0009	0.0070	0.0020				
		半コーシー	最尤推定	0.0066	0.0009	0.0070	0.0020				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0138	0.0020	2.8552	0.0026	0.0107	0.32	0.32	0.36
		逆ガンマ	最尤推定	0.0086	0.0012	0.0084	0.0022	0.0030			
		半コーシー	最尤推定	0.0176	0.0025	0.0139	0.0025	0.0162			
	$\phi=0 (0)$	独立モデル	最尤推定	0.0889	0.0120	0.1244	0.0194				
		逆ガンマ	最尤推定	0.0894	0.0121	0.1246	0.0194				
		半コーシー	最尤推定	0.0858	0.0116	0.1214	0.0191				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0097	0.0014	0.0117	0.0037	0.6679	0.43	0.26	0.31
		逆ガンマ	最尤推定	0.0094	0.0014	0.0109	0.0037	0.5576			
		半コーシー	最尤推定	0.0092	0.0014	0.0114	0.0033	0.6778			
500	$\phi=0 (0)$	独立モデル	最尤推定	0.0072	0.0010	0.0084	0.0022				
		逆ガンマ	最尤推定	0.0072	0.0010	0.0084	0.0022				
		半コーシー	最尤推定	0.0073	0.0010	0.0084	0.0022				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0153	0.0022	0.0147	0.0026	0.0072	0.31	0.27	0.43
		逆ガンマ	最尤推定	0.0095	0.0014	0.0101	0.0023	0.0021			
		半コーシー	最尤推定	0.0203	0.0030	0.0166	0.0026	0.0135			
	$\phi=0 (0)$	独立モデル	最尤推定	0.1300	0.0205	0.1368	0.0279				
		逆ガンマ	最尤推定	0.1307	0.0207	0.1361	0.0282				
		半コーシー	最尤推定	0.1260	0.0200	0.1321	0.0274				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0130	0.0018	0.0133	0.0038	0.6365	0.42	0.19	0.39
		逆ガンマ	最尤推定	0.0128	0.0018	0.0128	0.0041	0.5451			
		半コーシー	最尤推定	0.0119	0.0017	0.0129	0.0035	0.6312			
0.6	$\phi=0 (0)$	独立モデル	最尤推定	0.0116	0.0015	0.0124	0.0029				
		逆ガンマ	最尤推定	0.0116	0.0015	0.0123	0.0029				
		半コーシー	最尤推定	0.0116	0.0015	0.0125	0.0028				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0357	0.0087	0.4281	0.0042	0.1669	0.29	0.22	0.49
		逆ガンマ	最尤推定	0.0149	0.0023	0.0154	0.0030	0.0027			
		半コーシー	最尤推定	0.0315	0.0057	0.0294	0.0033	0.0443			
	$\phi=0 (0)$	独立モデル	最尤推定	0.1776	0.0337	0.1722	0.0450				
		逆ガンマ	最尤推定	0.1782	0.0339	0.1692	0.0453				
		半コーシー	最尤推定	0.1716	0.0329	0.1631	0.0432				
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0187	0.0034	0.0191	0.0046	0.7163	0.40	0.12	0.48
		逆ガンマ	最尤推定	0.0186	0.0031	0.0183	0.0051	0.5780			
		半コーシー	最尤推定	0.0170	0.0032	0.0182	0.0042	0.7296			

コピュラモデルでは、大標本サイズにおける推定値は真値に近くなり、被覆確率もほぼ 95% 程度となった。ただし、最尤推定によるコピュラパラメータについては、大標本サイズでも被覆確率は改善されなかった。

時点ごとの生存確率の推定値を Supplementary Information 表 S4–S6 に示した。イベントが独立の場合、シミュレーション設定、サンプルサイズ、さらにはモデルにも関係なく、真の生

表 4. ワイブルパラメータの真値の設定 3 における MSE の比較結果。表の数値は 1000 回のモンテカルロ・シミュレーションの平均値。

n	切断割合	コピュラパラメータ (相関係数)	モデル	MSE					打ち切り割合	イベント1割合	イベント2割合
				$\lambda_1 = 1.0$	$\alpha_1 = 1.5$	$\lambda_2 = 1.0$	$\alpha_2 = 0.5$	ϕ			
0.2	$\phi=0 (0)$	独立モデル	最尤推定	0.0450	0.0282	0.0333	0.0054				
			逆ガンマ	0.0440	0.0280	0.0329	0.0053				
		半コーシー	0.0461	0.0275	0.0363	0.0059					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1065	0.0291	0.0538	0.0077	0.7369	0.33	0.28	0.38
		コピュラモデル	逆ガンマ	0.0665	0.0270	0.0397	0.0063	0.2431			
		半コーシー	0.1512	0.0247	0.0811	0.0118	1.1009				
0.4	$\phi=0 (0)$	独立モデル	最尤推定	0.2028	0.1673	0.0429	0.0057				
		逆ガンマ	0.2048	0.1669	0.0437	0.0058					
		半コーシー	0.1822	0.1542	0.0363	0.0049					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0796	0.0662	0.0374	0.0063	9.8568	0.41	0.22	0.36
		コピュラモデル	逆ガンマ	0.0785	0.0888	0.0340	0.0054	7.5370			
		半コーシー	0.0607	0.0397	0.0370	0.0064	5.6235				
100	$\phi=0 (0)$	独立モデル	最尤推定	0.0364	0.0244	0.0352	0.0055				
		逆ガンマ	0.0357	0.0246	0.0347	0.0054					
		半コーシー	0.0366	0.0231	0.0371	0.0058					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0673	0.0251	0.0603	0.0089	0.1550	0.27	0.40	0.33
		コピュラモデル	逆ガンマ	0.0478	0.0245	0.0429	0.0067	0.0655			
		半コーシー	0.1002	0.0222	0.0925	0.0143	0.3651				
0.6	$\phi=0 (0)$	独立モデル	最尤推定	0.2127	0.1959	0.0636	0.0087				
		逆ガンマ	0.2087	0.1995	0.0652	0.0090					
		半コーシー	0.1872	0.1748	0.0538	0.0073					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0663	0.0555	0.0437	0.0074	7.5950	0.35	0.35	0.29
		コピュラモデル	逆ガンマ	0.0683	0.0792	0.0405	0.0062	4.2102			
		半コーシー	0.0552	0.0363	0.0414	0.0074	4.0709				
0.8	$\phi=0 (0)$	独立モデル	最尤推定	0.0410	0.0280	0.0386	0.0059				
		逆ガンマ	0.0403	0.0283	0.0383	0.0058					
		半コーシー	0.0408	0.0260	0.0395	0.0062					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0667	0.0280	0.0657	0.0118	0.1340	0.21	0.51	0.28
		コピュラモデル	逆ガンマ	0.0485	0.0275	0.0446	0.0081	0.0386			
		半コーシー	0.1056	0.0239	0.1031	0.0211	0.2758				
0.8	$\phi=0 (0)$	独立モデル	最尤推定	0.2362	0.2348	0.0904	0.0146				
		逆ガンマ	0.2268	0.2369	0.0927	0.0150					
		半コーシー	0.2034	0.2039	0.0771	0.0125					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0640	0.0508	0.0460	0.0104	20.2983	0.29	0.48	0.23
		コピュラモデル	逆ガンマ	0.0690	0.0799	0.0448	0.0086	5.5181			
		半コーシー	0.0553	0.0352	0.0422	0.0100	4.7468				
100	$\phi=0 (0)$	独立モデル	最尤推定	0.0551	0.0334	0.0510	0.0112				
		逆ガンマ	0.0552	0.0332	0.0503	0.0111					
		半コーシー	0.0579	0.0305	0.0527	0.0114					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.1006	0.0326	0.0920	0.0277	0.2350	0.15	0.62	0.23
		コピュラモデル	逆ガンマ	0.0723	0.0324	0.0632	0.0162	0.0264			
		半コーシー	0.1879	0.0304	0.1564	0.0489	0.3506				
100	$\phi=0 (0)$	独立モデル	最尤推定	0.2687	0.2905	0.1311	0.0245				
		逆ガンマ	0.2519	0.2871	0.1316	0.0246					
		半コーシー	0.2254	0.2441	0.1110	0.0208					
	$\phi=2 (0.5)$	独立モデル	最尤推定	0.0723	0.0530	0.0628	0.0201	22.3929	0.23	0.61	0.15
		コピュラモデル	逆ガンマ	0.0810	0.0917	0.0661	0.0148	6.9486			
		半コーシー	0.0645	0.0377	0.0566	0.0192	5.2917				

存確率に近い確率を示した(ただし、コピュラモデルの半コーシー分布ではやや過小評価)。しかし、イベントが相関している場合は、コピュラモデルが真の生存確率に近い確率を示す一方、独立モデルでは全て過大評価されていた。

表4. (つづき)

0.2	$\phi=0$ (0)	独立モデル	最尤推定	0.0075	0.0051	0.0061	0.0010		0.33	0.28	0.38
		逆ガンマ	0.0074	0.0051	0.0060	0.0010					
		半コーン	0.0075	0.0051	0.0062	0.0010					
		独立モデル	最尤推定	0.0200	0.0070	2.9763	0.0021	0.0162			
		逆ガンマ	0.0096	0.0052	0.0071	0.0011	0.0063				
	$\phi=2$ (0.5)	半コーン	0.0178	0.0053	0.0108	0.0015	0.0295				
		独立モデル	最尤推定	0.1841	0.0840	0.0267	0.0034				
		逆ガンマ	0.1846	0.0836	0.0269	0.0034					
		半コーン	0.1801	0.0828	0.0250	0.0032					
		独立モデル	最尤推定	0.0158	0.0137	0.0070	0.0012	0.6915	0.42	0.22	0.36
0.4	$\phi=0$ (0)	逆ガンマ	0.0163	0.0152	0.0070	0.0012	0.6145		0.27	0.40	0.33
		半コーン	0.0143	0.0118	0.0071	0.0012	0.6404				
		独立モデル	最尤推定	0.0075	0.0046	0.0066	0.0010				
		逆ガンマ	0.0075	0.0046	0.0066	0.0010					
		半コーン	0.0076	0.0045	0.0067	0.0011					
	$\phi=2$ (0.5)	独立モデル	最尤推定	0.0143	0.0052	0.0106	0.0016	0.0115		0.35	0.35
		逆ガンマ	0.0094	0.0047	0.0076	0.0012	0.0027				
		半コーン	0.0157	0.0048	0.0122	0.0018	0.0137				
		独立モデル	最尤推定	0.1929	0.1144	0.0511	0.0074				
		逆ガンマ	0.1922	0.1148	0.0515	0.0074					
500	$\phi=0$ (0)	半コーン	0.1878	0.1120	0.0488	0.0070		0.21	0.51	0.28	
		独立モデル	最尤推定	0.0124	0.0101	0.0080	0.0014	0.4175			
		逆ガンマ	0.0126	0.0115	0.0079	0.0013	0.3738				
		半コーン	0.0115	0.0092	0.0077	0.0014	0.3855				
		独立モデル	最尤推定	0.0087	0.0048	0.0075	0.0012		0.30	0.48	0.22
	$\phi=2$ (0.5)	逆ガンマ	0.0087	0.0048	0.0075	0.0012					
		半コーン	0.0088	0.0048	0.0075	0.0013					
		独立モデル	最尤推定	220.9184	0.0070	0.1126	0.0023	0.0057			
		逆ガンマ	0.0105	0.0049	0.0089	0.0016	0.0021				
		半コーン	0.0168	0.0050	0.0146	0.0027	0.0107				
0.6	$\phi=0$ (0)	独立モデル	最尤推定	0.2147	0.1467	0.0823	0.0136		0.15	0.62	0.23
		逆ガンマ	0.2127	0.1470	0.0828	0.0137					
		半コーン	0.2079	0.1426	0.0793	0.0131					
		独立モデル	最尤推定	0.0117	0.0092	0.0095	0.0019	0.3525			
		逆ガンマ	0.0122	0.0105	0.0097	0.0018	0.3271				
	$\phi=2$ (0.5)	半コーン	0.0113	0.0086	0.0091	0.0019	0.3379		0.23	0.48	0.22
		独立モデル	最尤推定	0.0107	0.0060	0.0093	0.0018				
		逆ガンマ	0.0107	0.0060	0.0093	0.0018					
		半コーン	0.0109	0.0059	0.0094	0.0018					
		独立モデル	最尤推定	0.0186	0.0066	0.0159	0.0040	0.0084			
0.8	$\phi=0$ (0)	逆ガンマ	0.0131	0.0060	0.0109	0.0025	0.0022		0.15	0.62	0.23
		半コーン	0.0211	0.0059	0.0181	0.0046	0.0107				
		独立モデル	最尤推定	0.2502	0.1975	0.1121	0.0239				
		逆ガンマ	0.2467	0.1970	0.1122	0.0239					
		半コーン	0.2411	0.1903	0.1078	0.0231					
	$\phi=2$ (0.5)	独立モデル	最尤推定	0.0142	0.0100	0.0130	0.0034	0.3346		0.23	0.61
		逆ガンマ	0.0145	0.0115	0.0133	0.0033	0.3116				
		半コーン	0.0137	0.0094	0.0125	0.0035	0.3358				

4. 考察

本稿では、相関した競合リスクを伴う左側切断・右側打ち切りデータにおけるワイルパラメータとコピュラパラメータの推定について検討した。本稿で得られた結果から重要な点として、初めに、最尤推定よりベイズ推定のこれらパラメータの推定性能が、特に小標本サイズで

高かったことがあげられる。ただし、ベイズ推定では事前分布の違いによりパラメータの推定性能が異なり、最尤推定より推定性能が劣る事前分布も存在する。この点を踏まえると、実データ解析においては、複数の事前分布を設定して解析結果の頑健性を確認することが望ましい。

次に、独立データでは、独立モデルのパラメータ推定性能がコピュラモデルと比較して高く、大標本サイズではその推定性能も高くなつた。しかし、相関データでは、独立モデルのパラメータの推定性能がコピュラモデルより低く、大標本サイズでもパラメータの推定性能が改善されないことがあげられる。競合リスクが存在する状況下において、リスク(イベント)間の相関から生じる問題については以前から指摘されてきた(宮村, 1982; 西川, 2008; 道前, 2023)。独立モデルを実データに適用する場合には、そのパラメータ推定性能が独立データにのみ担保されている点に注意する必要がある。

相関データでは、コピュラモデル、特にベイズ流コピュラモデルがパラメータの推定において優れていた。ただし、注意点として、コピュラモデルは全てのシミュレーション条件でパラメータの推定性能が優れているわけではなく、独立データの場合は、上述したように独立モデルの推定性能が優れている。実データ解析においては、2つのイベントのうちどちらか1つのイベントしか観察されないことから、実際の観察時間からイベント間に相関があるのかどうか判断し難いが、イベントが独立であると推測できる場合はベイズ流独立モデル、相関が疑われる場合は複数の事前分布を設定したベイズ流コピュラモデルで解析を行い、その結果を比較検討することが勧められる。

切断割合のパラメータ推定性能への影響は、推定に用いたモデルにかかわらず、特にワイブルパラメータで大きかった(表2-4)。全てのシミュレーション設定で、切断割合の増加に伴って、パラメータの推定性能が下がる傾向にあり、実データの解析では、観察された各イベントの発生割合も考慮に入れた結果の解釈が必要であることを示している。しかし、このパラメータ推定性能の低下はイベントの発生割合とは無関係であると推測される。例えば、シミュレーション設定1と3では切断割合の増加と共にイベント1の発生割合は増加、イベント2の発生割合は減少したが、 α_1 と λ_1 および α_2 と λ_2 のMSE値は大きくなる傾向があった。一方、シミュレーション設定2では切断割合の増加と共にイベント1の発生割合は減少、イベント2の発生割合は増加したが、 α_1 と λ_1 および α_2 と λ_2 のMSE値は大きくなる傾向があった。このような切断割合の増加によるパラメータの推定性能の低下の原因として、切断割合の増加に伴う両イベント発生割合の乖離の増大が考えられる。いずれのシミュレーション設定においても、切断割合の増加に伴って、イベント発生割合の乖離が増大しており、その増大に伴いパラメータの推定性能が低下している。したがって、実データの解析では、観察された各イベントの発生割合ではなく、発生割合の乖離を考慮に入れた結果の解釈が必要である。

以上は全てパラメータの推定性能に関する考察であったが、生存確率の推定でも、独立イベントが想定される場合は独立モデル、相関イベントが想定される場合はコピュラモデルでの推定が推奨される。

今後の展望として、次の3点があげられる。初めに、周辺分布の選択である。本稿では、同時潜在故障時間分布の周辺分布としてワイブル分布を用いたが、実データへの適用拡大を考えると、ガンマ分布や対数正規分布など他のパラメトリック分布(武富・山本, 2023)も検討する必要がある。この場合、周辺分布の実データへの適合度の検討が必要である(Bakker et al., 2023)。次に、コピュラモデルにおけるコピュラの選択である。本稿では、クレイトンコピュラを用いたが、コピュラを用いる大きな利点は周辺分布とは独立に依存関係を柔軟にモデリングできることである(Nelsen, 2006)。したがって、コピュラの選択が可能になれば多様な同時潜在故障時間分布の構成が可能となり、実データへの適用範囲も広がることが期待できる。課

題として、周辺分布だけでなく依存性の実データへの適合度の検討が必要である (Escarela et al., 2023)。最後に、共変量についてのモデル化である。特に医療データ等の実データ解析では、研究目的の観点から、分布パラメータやコピュラパラメータの推定より共変量の効果を推定することが重要である。この場合、基準ハザード関数にパラメトリック分布(指数分布やワイブル分布等)を仮定し、共変量の効果を乗法的に作用させたパラメトリックモデルの回帰係数として、共変量の効果を推定する方法が有望である。

謝 辞

江村剛志先生及び太田修平先生には、本稿の執筆の機会を頂いたことに心より感謝申し上げます。また、本稿の改稿にあたり、有益なコメントを頂いた4名の査読者に感謝申し上げます。本研究は科研費(JP21K12127)からの助成金を受けて実施されました。

参 考 文 献

- Bakker, L. J., Thielen, F. W., Redekop, W. K., Uyl-de Groot, C. A. and Blommestein, H. M. (2023). Extrapolating empirical long-term survival data: The impact of updated follow-up data and parametric extrapolation methods on survival estimates in multiple myeloma, *BMC Medical Research Methodology*, **23**, 132, <https://doi.org/10.1186/s12874-023-01952-2>.
- Bockwoldt, B., Sugiyama, H., Tsai, K., Bhatti, P., Brenner, V. A., Hu, A., Kerr, F. K., Morenz, E., French, B. and Phipps, I. A. (2021). Gastrointestinal cancer survival and radiation exposure among atomic bomb survivors: The life span study, *Cancer Epidemiology, Biomarkers & Prevention*, **30**, 412–418.
- Cox, D. R. (1959). The analysis of exponentially distributed lifetimes with two types of failures, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **21**, 411–421.
- 江村剛志, 道前洋史 (2020). コピュラを用いた生存時間解析, 統計数理, **68**(1), 147–274.
- 江村剛志, 道前洋史 (2021). 左切断・右側打ち切りのあるフィールド寿命データ解析～パラメトリックモデルに基づく統計的推論と信頼性の予測～, 信学技報, **121**, 7–12.
- Emura, T. and Michimae, H. (2022). Left-truncated and right-censored field failure data: Review of parametric analysis for reliability, *Quality and Reliability Engineering International*, **38**, 3919–3934.
- Escarela, G., Vásquez, A. R., González-Farías, G. and Márquez-Urbina, J. U. (2023). Copula modeling for the estimation of measures of marker classification and predictiveness performance with survival outcomes, *Statistical Methods in Medical Research*, **32**, 1203–1216.
- Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496–509.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457–511.
- Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo, *Journal of Machine Learning Research*, **15**, 1593–1623.
- Hong, Y., Meeker, W. Q. and McCalley, J. D. (2009). Prediction of remaining life of power transformers based on left truncated and right censored lifetime data, *Annals of Applied Statistics*, **3**, 857–879.
- Kaishev, K. V., Dimitrova, S. D. and Haberman, S. (2007). Modelling the joint distribution of competing risks survival times using copula functions, *Insurance: Mathematics and Economics*, **41**, 339–361.

- Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York.
- Kundu, D., Mitra, D. and Ganguly, A. (2017). Analysis of left truncated and right censored competing risks data, *Computational Statistics & Data Analysis*, **108**, 12–26.
- 道前洋史 (2023). 競合リスクを伴う左側切断・右側打ち切りデータの解析～現状と今後の展望について～, 日本統計学会誌, **52**, 203–220.
- Michimae, H. and Emura, T. (2022). Likelihood inference for copula models based on left-truncated and competing risks data from field studies, *Mathematics*, **10**, 2163, <https://doi.org/10.3390/math10132163>.
- Michimae, H., Emura, T., Miyamoto, A. and Kishi, K. (2024). Bayesian parametric estimation based on left-truncated competing risks data under bivariate Clayton copula models, *Journal of Applied Statistics*, **51**, 2690–2708.
- Mitra, D., Kundu, D., and Balakrishnan, N. (2021). Likelihood analysis and stochastic EM algorithm for left truncated right censored data and associated model selection from the Lehmann family of life distributions, *Japanese Journal of Statistics and Data Science*, **4**, 1019–1048.
- 宮村鍛夫 (1982). 競合モデルに関する統計的手法, オペレーションズ・リサーチ, **27**, 871–876.
- Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer, New York.
- 西川正子 (2008). 生存時間解析における競合リスクモデル, 計量生物学, **29**, 141–170.
- Prentice, R. L., Kalbfleisch, J. D., Peterson Jr., A. V., Flurnoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in presence of competing risks, *Biometrics*, **34**, 541–554.
- Ranjan, R., Sen, R. and Upadhyay, S. K. (2021). Bayes analysis of some important lifetime models using MCMC based approaches when the observations are left truncated and right censored, *Reliability Engineering & System Safety*, **214**, <https://doi.org/10.1016/j.ress.2021.107747>.
- 齋藤哲雄, 室谷健太 (2023). マルチステートモデルの理論とがん臨床研究への応用, 日本統計学会誌, **52**, 221–267.
- 齋藤哲雄, 室谷健太 (2024). がん臨床試験と競合リスク・マルチステートモデル, 計量生物学, **45**, 37–65.
- Sposto, R., Sugiyama, H., Tsuruyama, T. and Brenner, A. V. (2023). Effect of radiation exposure on survival after first solid cancer diagnosis in A-bomb survivors, *Cancer Epidemiology*, **83**, <https://doi.org/10.1016/j.canep.2023.102341>.
- Stan Development Team (2017). Stan Modeling Language Users Guide and Reference Manual, <https://mc-stan.org> (最終アクセス日 2024年12月16日).
- Stan Development Team (2018). RStan: The R interface to Stan. R package version 2.32.6, <http://mc-stan.org> (最終アクセス日 2024年12月16日).
- 武富奈菜美, 山本和嬉 (2023). 生存時間解析・信頼性解析のための統計モデル, 日本統計学会誌, **52**, 69–112.
- 戸坂凡展, 吉羽要直 (2005). コピュラの金融実務での具体的な活用方法の解説, 金融研究, **24**, 115–162.
- 塚原英敦 (2012). 接合分布関数(コピュラ)の理論と応用, 『21世紀の統計科学 III: 数理・計算の統計科学』, 増補 HP版, 1–140, 東京大学出版会, 東京.
- 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, **51**, 101–121.
- Wang, L., Lio, Y., Tripathi, Y. M., Dey, S. and Zhang, F. (2022). Inference of dependent left-truncated and right-censored competing risks data from a general bivariate class of inverse exponentiated distributions, *Statistics*, **56**, 347–374.

Bayesian Analysis of Left-truncated Dependent Competing Risks Data

Hirofumi Michimae

School of Pharmacy, Kitasato University

Survival analysis examines the time from a reference point to the occurrence of a specific event and is widely applied in fields such as medicine, epidemiology, economics, and the social sciences. Key topics in survival analysis include censoring, left truncation, and competing risks. Thus, this study focuses on left-truncated and right-censored competing risks data and introduces four representative analytical models (independent, copula, Bayesian independent, and Bayesian copula models). In particular, the study provides a detailed exploration of the following: the joint modeling of latent failure times (marginal distributions) using copulas in competing risks; the modeling of left truncation conditional on joint models; the construction of likelihood functions; and the estimation of marginal distribution parameters and copula parameters using Bayesian estimators. Simulation experiments were conducted under realistic conditions (different combinations of marginal distributions and left-truncation proportions), and the parameters were estimated using the above four analytical models. Furthermore, our discussion compares the parameter estimation performance of the models and discusses the challenges and future prospects of the Bayesian copula model.

従属打ち切りの下での治療効果推定のための 要因計画・要因分析 —コピュラに基づく方法—

江村 剛志¹・室谷 健太²

(受付 2024 年 10 月 15 日；改訂 2025 年 2 月 11 日；採択 2 月 12 日)

要 旨

治療などの要因が生存時間に与える効果を推定・検定するとき、古典的な正規分布に基づく推測法は適切でない。生存時間の解析では、分布型に依存しないノンパラメトリック法が近年注目を集めている。本論文では、まず独立打ち切りの仮定下での生存時間の要因計画・要因分析のためのノンパラメトリック法を紹介し、それをコピュラモデルに基づいた従属打ち切りの仮定下での解析に拡張した方法を紹介する。続いて、この手法を R 関数 `surv.factorial()` を用いて実行する方法を解説する。これまで調べられていなかった、大域的検定(オムニバス検定)と局所的検定の性能を調べるシミュレーションを行い、これらの検出力特性を明らかにする。最後に、コピュラモデルに基づく手法を、結腸がんの切除手術を受けた被験者のデータに適用し、結果を吟味する。データ解析のための R コードをウェブ上の Supplement で提供する。

キーワード：アルキメデスコピュラ、カプラン・マイヤー推定量、コピュラ・グラフィック推定量、実験計画法、マン・ホイットニー検定、マン・ホイットニー効果。

1. はじめに

実験計画法における要因計画法・要因分析法は、治療などの要因(Factor)が個体に与える効果を推定・検定するための手法である(Montgomery, 2019)。単純に 2 つの治療法を比較する無作為化比較試験に加え、一元配置実験、二元配置実験など、様々な要因の複雑な効果を評価するための統一的なフレームワークが実験計画法に集約されている。例えば、3 つの治療を無作為に割り付ける無作為化比較試験では、全ての治療が等しい効果という大域的仮説の検定に加え、ある 2 つの治療のペアを比較する検定も考えうる。また、治療と性別で 2 つの要因の効果を調べる二元配置実験は、治療の主効果に加え、治療と性別の交互作用効果を推定・検定する(松井, 2023)。

要因計画・要因分析でよく知られている手法は古典的な分散分析(Analysis of Variance, ANOVA)法である(東京大学教養学部統計学教室, 1992)。これは、個体の応答変数が正規分布に従う場合に、治療効果の推定・検定を行う手法である。分散分析において最も基礎的な検定法は、全ての治療群の応答(の平均値)が等しいという大域的な帰無仮説を検証する F -検定で

¹ 広島大学 情報科学部：〒739-8511 広島県東広島市鏡山一丁目 3 番 2 号; takeshiemura@gmail.com

² 久留米大学 医学部：〒830-0011 福岡県久留米市旭町 67; kmurotani@med.kurume-u.ac.jp

ある。この要因全体に効果が無いことを示すための大域的な検定は“オムニバス検定”とも呼ばれる (Mishra et al., 2019; Futschik et al., 2019)。オムニバス検定の棄却は、考慮している治療群間に何らかの差が内在していることを示唆するため、その後に2群間のペアワイズ比較や、主効果や交互作用などの詳細な推定・検定を行う根拠となる。

がんの臨床試験では、通常は2つの治療の無作為化比較試験が行われるが、3つの治療の優劣を比較する試験が行われることもある (Moertel et al., 1995)。また、がんの臨床試験における治療効果の検証において、2要因×2水準 (= 4つの治療)の要因計画・要因分析が行われることがある。例えば、進行子宮頸がん患者の生存に対する治療の効果を調べた試験 (Tewari et al., 2014) や、転移性前立腺がん患者を対象とした進行中の第III相臨床試験 (Fizazi et al., 2022) を参照されたい。2017年の調査では、第III相がん治療の2×2実験計画の臨床試験が30件特定された (Freidlin and Korn, 2017)。

がん臨床試験のように、応答変数が被験者の生存時間の場合、応答変数の正規性に基づく平均・分散を基にした古典的な手法の利用は適切でない。生存時間データは右に歪んだ形状を持つ場合も多く、正規分布に従わないだけでなく、平均・分散は存在しないこともある (武富・山本, 2023)。このため、生存時間解析における多群の治療効果比較における要因分析では、分布型に依存しないノンパラメトリックな方法が近年注目を集めている。例えば、マン・ホイットニー効果 (Dobler and Pauly, 2020)、標本分割を用いたペアワイズ効果 (Gorfine et al., 2020)、生存期間中央値に対する効果 (Ditzhaus et al., 2021)、累積ハザードに対する効果 (Ditzhaus et al., 2023) などで治療効果を定義し、ノンパラメトリック推定・検定を行う。

生存時間解析で難しい点は“打ち切り”の扱いである (Klein and Moeschberger, 2003)。上述のノンパラメトリック要因計画・要因分析法で提案されている生存時間の統計手法は、打ち切り発生が生存リスクに影響しないという“独立打ち切り”の仮定の基に構築されている。とりわけこれら手法は、独立打ち切りの仮定を基に導出された Kaplan-Meier (KM) 推定量を適用しているものもある。しかしながら、KM 推定量は独立打ち切りが成立しないとき、生存関数の推定にバイアスをもたらす (Rivest and Wells, 2001; Emura and Hsu, 2020; 江村, 2023; 斎藤・室谷, 2024)。

独立でない打ち切り (従属打ち切りと呼ばれる) は、被験者追跡を要する医学データの分析で現れやすい (Andersen and Perme, 2010; Staplin et al., 2015; Xu et al., 2018; Moradian et al., 2019; Collett, 2023; Li et al., 2023; Das et al., 2024)。典型的な例は、健康状態の悪化による被験者の脱落 (ドロップアウト) による打ち切りである。例えば、Staplin et al. (2015) によると、追跡の打ち切りは被験者が肝移植の待機リストから削除されたときに発生する。この例において、死期が近い被験者は移植を受ける可能性が高くなるため、“従属打ち切り”をもたらす可能性がある。より一般に、病状悪化のために追跡不能となった被験者がその後に死亡している疑いのあるデータ (6節を参照) では、死亡時間と打ち切り時間が正相関している可能性がある。

独立打ち切りの仮定に立脚しない代替手法の開発も数多く行われている (Lu and Zhang, 2012; Staplin et al., 2015; Dettoni et al., 2020; Deresa and Van Keilegom, 2020, 2025; Deresa and Van Keilegom, 2020; Schneider et al., 2020, 2022)。例えば、従属打ち切り下での KM 推定量はバイアスを持つが、Andersen and Perme (2010) は、観測可能な共変量の情報を用いて、バイアスを削減する手法を提案している。Emura and Chen (2016) は、従属打ち切りと生存時間の従属構造をクレイトンコピュラでモデル化し、単変量 Cox 回帰法による変数選択の代替手法を提案している。同様に、Emura and Hsu (2020) もコピュラモデルを用いて、従属打ち切り下でのバイアスを修正したマン・ホイットニー効果量に基づく推定・検定法を提案している。他にも、従属打ち切り問題への対処のための様々な手法は、Emura and Chen (2018) と Collett (2023) の書籍にまとめられているが、その後多くの手法が提案されている (Jo et al., 2023; Gares et

al., 2024; Gottselig et al., 2025; Delhelle and Van Keilegom, 2025). ただし, 要因計画・要因分析における従属打ち切りの問題への対処法を提案した手法は Emura et al. (2024) に限られる.

本論文では, まず Dobler and Pauly (2020) が提案した生存時間解析のための要因計画・要因分析手法を紹介する(2節). その後, コピュラモデルに基づき従属打ち切りを考慮することができる要因分析法 (Emura et al., 2024) を詳説する(3節). この手法では, 打ち切り時間と生存時間の従属構造を想定した感度分析を行う手法である. 4節では, これら解析を R 関数 `surv.factorial()` を用いて実行する方法を解説する. 5節ではシミュレーションを行い, これまで調べられていなかった, 大域的検定(オムニバス検定)と局所的検定の性能を調べるシミュレーションを行い, これらの検出力特性を明らかにする. また本論文では, 新たなデータ解析例として, 結腸がんの切除手術を受けた被験者のデータへの適用例を紹介する(6節). 7節を本論文の結びとする. 尚, データ解析に使用した R コードはオンライン Supplement から入手可能である.

2. 生存時間の要因計画・要因分析

本節では, 生存時間データに基づく要因計画・要因分析の手法 (Dobler and Pauly, 2020) の概要を紹介する. まず, 一元配置, 二元配置, その他の様々な実験計画において統一的に定義可能な治療効果を定式化する. その後, 推定・検定法を紹介する.

2.1 治療効果

治療などの要因が生存時間に与える影響を“治療効果”と呼ぶ. ただし, 要因は性別や環境要因などの実験的に制御出来ないものでも良い. 生存時間のデータは正規分布に従うことが稀であるため, 平均の等価性を検定するための古典的な分散分析の利用は適切でない. 故に, 治療効果の定式化には, ノンパラメトリックな治療効果指標を要する. 以下で紹介する手法は, 2群のノンパラメトリック検定で良く知られている「マン・ホイットニー検定」の考え方を拡張した治療効果の推定・検定法 (Dobler and Pauly, 2020; Brunner et al., 2017; Dobler et al., 2020) である.

治療が d 通りあるとし, 各治療下での生存時間を T_i , その生存関数を $S_i(t) = P(T_i > t)$, $i = 1, 2, \dots, d$, とする. i 番目と ℓ 番目を比較した“ペアワイズ効果”は, 次のようになる.

$$(2.1) \quad w_{i\ell} = P(T_i > T_\ell) + \frac{1}{2}P(T_i = T_\ell) = - \int S_i^\pm(t)dS_\ell(t),$$

ここで $S_i^\pm(t) = \{S_i(t+) + S_i(t-)\}/2$, $S_i(t+) = \lim_{s \downarrow t} S_i(s)$, $S_i(t-) = \lim_{s \uparrow t} S_i(s)$ である. この $w_{i\ell}$ は, i 番目の治療を受けた被験者が, ℓ 番目の治療を受けた被験者より長く生存する確率である. もし $w_{i\ell} > 1/2$ であれば, i 番目の治療が ℓ 番目の治療に比べて有益であることを示す(もし $w_{i\ell} < 1/2$ であれば, その逆である). なお, $w_{ii} = 1/2$ である.

2標本($d = 2$)の場合, ペアワイズ効果 w_{12} のみを用いて推定・検定を行えば良い (Efron, 1967; Koziol and Jia, 2009; Dobler and Pauly, 2018; Emura and Hsu, 2020; Nowak et al., 2022) が, 本論文で注目するのは主に $d \geq 3$ の場合である. このように 3種類以上の治療法がある場合, ペアワイズ効果ではなく, 一般的な治療と対比した“相対的な効果”で治療効果を定式化すると良い. この定式化は, 古典的な分散分析における治療効果を, 群平均と一般平均の差で定義していることと同じ考え方である. 以下でその方法を記述する.

ノンパラメトリック分散分析 (Brunner and Puri, 2001; Brunner et al., 2017) の枠組みを用い, Dobler and Pauly (2020) は, i 番目の治療の相対的な治療効果を次式で定義した.

$$p_i = \frac{1}{d} \sum_{\ell=1}^d w_{i\ell} = - \int S_i^\pm(t) d\bar{S}(t) = P(T_i > \bar{T}) + \frac{1}{2} P(T_i = \bar{T}),$$

ここに $\bar{S} = \sum_{\ell=1}^d S_\ell/d$ と定義し, これは集団全体での生存関数である. この p_i は, i 番目の治療を受けた被験者が, 被験者全体の集団より長く生存する割合と解釈できる. 治療が d 通りあるため, これらの治療効果をベクトルでまとめると次のようになる:

$$(2.2) \quad \mathbf{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_d \end{pmatrix} = \mathbf{A}\mathbf{w}, \quad \mathbf{w} = \left(w_{11}, \dots, w_{1d} : w_{21}, \dots, w_{2d} : \dots : w_{d1}, \dots, w_{dd} \right)',$$

ここで $\mathbf{A} = \mathbf{I}_d \otimes (\mathbf{1}'_d/d)$ であり, \mathbf{I}_d は $d \times d$ 単位行列, $\mathbf{1}'_d = (1, \dots, 1)$ は長さ d のベクトル, \otimes はクロネッカー積である. クロネッカー積の定義と式(2.2)の証明は付録 1 を参照されたい.

治療効果 \mathbf{p} に関する帰無仮説・対立仮説は, 対比行列 \mathbf{C} を用いて

$$H_0: \mathbf{C}\mathbf{p} = \mathbf{0} \quad \text{vs.} \quad H_1: \mathbf{C}\mathbf{p} \neq \mathbf{0}$$

と表せる. 最も単純な仮説検定の例は, $\mathbf{C} = \mathbf{P}_d = \mathbf{I}_d - \mathbf{1}_d\mathbf{1}'_d/d$ で定まる大域的な検定問題

$$H_0: p_1 = p_2 = \dots = p_d \quad \text{vs.} \quad H_1: p_i \neq p_\ell, \quad \exists(i, \ell).$$

である(付録 1). 他の検定問題は, 例えば局所仮説 $H_0: p_1 = p_d$ は $\mathbf{C} = (1, 0, \dots, 0, -1)$ など, 対比行列を適宜指定することによって定式化できる.

少し複雑な例として, 二元配置を考える. a 個の水準を持つ“因子 A”と b 個の水準を持つ“因子 B”を考える. 因子計画では, 全ての治療を組み合わせた $a \times b$ 個の治療効果を考えるため,

$$\mathbf{p} = \left(p_{11}, \dots, p_{1b} : p_{21}, \dots, p_{2b} : \dots : p_{a1}, \dots, p_{ab} \right)',$$

と書け, ここに p_{jk} は因子 A の j 番目の水準と因子 B の k 番目の水準での治療効果である. 例えば, 因子 A の主効果の有無を検定する場合は $\mathbf{C} = \mathbf{P}_a \otimes \mathbf{1}'_b/b$ を指定すると,

$$H_0: p_{1\cdot} = p_{2\cdot} = \dots = p_{a\cdot} \quad \text{vs.} \quad H_1: p_{i\cdot} \neq p_{\ell\cdot} \text{ for } \exists(i, \ell), i \neq \ell,$$

となり, ここに $p_{i\cdot} = \sum_{\ell=1}^b p_{i\ell}$ である. 同様に, 因子 B の主効果の有無の検定は $\mathbf{C} = \mathbf{1}'_a/a \otimes \mathbf{P}_b$, 交互作用の有無の検定 $\mathbf{C} = \mathbf{P}_a \otimes \mathbf{P}_b$ と指定すれば良い.

2.2 独立打ち切り下での要因分析法

Dobler and Pauly (2020) は, 生存時間データに基づき, 治療効果の推定・検定を行うノンパラメトリックな手法を提案した. この手法では, 独立打ち切りの仮定の下, 治療効果の推定量を

$$\hat{\mathbf{p}} = \begin{pmatrix} \hat{p}_1 \\ \vdots \\ \hat{p}_d \end{pmatrix} = \mathbf{A}\hat{\mathbf{w}}, \quad \hat{\mathbf{w}} = \left(\hat{w}_{11}, \dots, \hat{w}_{1d} : \hat{w}_{21}, \dots, \hat{w}_{2d} : \dots : \hat{w}_{d1}, \dots, \hat{w}_{dd} \right)'$$

としており, ここにペアワイズ効果の推定量 $\hat{\mathbf{w}}$ は生存関数 $S_i(\cdot)$ の KM 推定量 $\hat{S}_i(\cdot)$ を用い,

$$\hat{w}_{i\ell} = - \int \hat{S}_i^\pm(t) d\hat{S}_\ell(t),$$

としており, ここに $\hat{S}_i^\pm(t) = \{\hat{S}_i(t+) + \hat{S}_i(t-)\}/2$ である. ただし, $\hat{S}_i(t)$ が $S_i(t)$ の一致推定量

となる範囲は $t \in [0, \tau]$ である ($\tau > 0$ はある定数) から, 上式の積分に修正が必要となる (3 節で説明する). 全群のサンプルサイズ(被験者数)を N とすると, $N \rightarrow \infty$ のとき漸近正規近似 $\sqrt{N}(\hat{p} - p) \sim \mathcal{N}(\mathbf{0}, V)$ が導出されており, ここに V は共分散行列であり, その推定量 \hat{V} も提案されている. これより \hat{p}_i の標準誤差(SE) および p_i の信頼区間が求まる.

帰無仮説 $H_0: C\mathbf{p} = \mathbf{0}$ からの逸脱は統計量 $C\hat{p}$ と $\mathbf{0}$ の差で調べることが出来る. Dobler and Pauly (2020) は次の F -統計量を提案している:

$$F_N = \frac{N\hat{p}'T\hat{p}}{\text{tr}(T\hat{V})} = \frac{N(C\hat{p})'(CC')^+(C\hat{p})}{\text{tr}[(CC')^+\hat{V}CC']},$$

ここに, $T = C'(CC')^+C$ は射影行列, $\text{tr}(\cdot)$ はトレース, $(CC')^+$ は CC' の一般化逆行列(ムーア・ペンローズ逆行列)である(一般化逆行列の計算は, 例えば R であれば関数 `ginv(.)` を用いれば良い).

統計量 F_N の $N \rightarrow \infty$ のときの漸近帰無分布は, 重みづけカイ二乗分布の和であることが示されており, 漸近帰無分布の平均は 1 である (Dobler and Pauly, 2020). 対立仮説の下で, F_N は無限大に確率収束することも示されるため, 条件 $F_N > c_{N,\alpha}$ で帰無仮説を棄却する検定方式が採用され, ここに棄却限界値 $c_{N,\alpha}$ は数値的に求める (Dobler and Pauly, 2020). この検定は独立打ち切りの仮定の下で妥当であり, 従属打ち切り下では妥当性は無いことに注意する.

2.3 大域的仮説と局所的仮説の関係

Dobler and Pauly (2020) の方法ならびに多くの要因分析では, 多数の群の間のどこに差があるか事前に分からぬいため, 最初に大域的仮説を行うことがある. もし, 大域的仮説が棄却されれば, より詳細な群間差を検出するための局所的仮説を検定して行く. 例えば, 3 群の場合では以下の 3 つの局所的検定問題と大域的検定問題が考えられる:

$$H_0: p_1 = p_2 \quad \text{vs.} \quad H_1: p_1 \neq p_2,$$

$$H_0: p_1 = p_3 \quad \text{vs.} \quad H_1: p_1 \neq p_3,$$

$$H_0: p_2 = p_3 \quad \text{vs.} \quad H_1: p_2 \neq p_3$$

$$H_0: p_1 = p_2 = p_3 \quad \text{vs.} \quad H_1: p_1 \neq p_2, \quad \text{or} \quad p_1 \neq p_3, \quad \text{or} \quad p_2 \neq p_3.$$

通常の手続きとしては, 最後の大域的仮説検定で棄却となった場合に限り, 最初の 3 つの局所的仮説を検定して行く. ここで問題になるのは, 大域的仮説検定と局所的仮説検定の検出力の違いである. 例えば, 大域的仮説が棄却された場合に局所的仮説を検定する手順で進める場合は, 大域的仮説検定の検出力が十分高い必要がある. この検出力の性能評価は生存時間の要因分析の文脈で議論されてこなかったため, 5 節で数値的に議論する.

3. 従属打ち切り下での推定・検定

まず, 従属打ち切り下での推定・検定を行うための生存コピュラモデルを定式化する. 次に, Rivest and Wells (2021) による生存関数の推定量を紹介する. その後, 要因分析のために提案された Emura et al. (2024) による治療効果の推定法・検定法を紹介する.

3.1 生存関数の推定

治療が d 通りあるとし, $i = 1, 2, \dots, d$ は治療群を表すとする. i 番目の治療に n_i 人の被験者が割り付けられるとし, $j = 1, \dots, n_i$ は被験者番号を表すとする. 各被験者の生存時間を T_{ij} , 打ち切り時間を U_{ij} とすると, 右打ち切り下で観測されるデータは

$\{(X_{ij}, \delta_{ij}); i = 1, \dots, d, j = 1, \dots, n_i\}$ であり, ここで $X_{ij} = \min(T_{ij}, U_{ij})$, $\delta_{ij} = 1\{T_{ij} \leq U_{ij}\}$, $1\{\cdot\}$ は指示関数である. 全被験者数(サンプルサイズ)は $N = \sum_{i=1}^d n_i$ である.

従属打ち切りの構造を定めるため, 打ち切り時間と生存時間の同時生存関数に次のモデルを仮定する:

$$(3.1) \quad P(T_{ij} > t, U_{ij} > u) = C_{i\theta_i}(S_i(t), G_i(u)), \quad i = 1, 2, \dots, d,$$

ここで $C_{i\theta_i}(\cdot, \cdot)$ はコピュラ (Nelsen, 2006; Durante and Sempi, 2016; 江村, 2025) で, θ_i は従属の正負や強さを表すパラメータ, $S_i(t) = P(T_{ij} > t)$, $G_i(u) = P(U_{ij} > u)$ は周辺生存関数である. クレイトンコピュラ (Clayton, 1978) を仮定する場合,

$$C_{i\theta_i}(u, v) = (u^{-\theta_i} + v^{-\theta_i} - 1)^{-\frac{1}{\theta_i}}, \quad \theta_i > 0, \quad i = 1, 2, \dots, d,$$

とし, ここに θ_i はケンドール順位相関(ケンドールのタウ)に変換すると $\tau_{\theta_i} = \theta_i/(\theta_i + 2)$ となる. 独立コピュラ $C_{i\theta_i}(u, v) = uv$, $i = 1, 2, \dots, d$, は独立打ち切りの仮定に相当する. 他のコピュラの例としては, グンベルコピュラやフランクコピュラがある (杉本・田中, 2023; 江村, 2025). 多くのコピュラでは, 極限 $\theta_i \rightarrow 0$ は $C_{i\theta_i}(u, v) \rightarrow uv$ となるため, モデル(3.1)は独立打ち切りのモデルを包含している. また, モデル(3.1)は, 周辺生存関数 $S_i(\cdot)$, $G_i(\cdot)$ にモデル化は行わず, 相関構造のみをモデル化している.

モデル(3.1)の下で, 観測データ $\{(X_{ij}, \delta_{ij}); i = 1, \dots, d, j = 1, \dots, n_i\}$ から, 周辺生存関数 $S_i(\cdot)$ のノンパラメトリック推定が可能となることが知られている. Rivest and Wells (2001) は, モデル(3.1)がアルキメデスコピュラである場合の, 具体的な推定量を提案している. アルキメデスコピュラとは, 生成素と呼ばれる関数 $\phi_{i\theta_i}$ を用いて次のように表せるモデルである:

$$(3.2) \quad P(T_{ij} > t, U_{ij} > u) = \phi_{i\theta_i}^{-1}[\phi_{i\theta_i}\{S_i(t)\} + \phi_{i\theta_i}\{G_i(u)\}], \quad i = 1, \dots, d,$$

ここで, 生成素 $\phi_{i\theta_i}$ は減少かつ Convex な連続関数であり $\phi_{i\theta_i}(0) = \infty$ かつ $\phi_{i\theta_i}(1) = 0$ を満足する必要がある. 同じ群内でタイが無い ($j \neq k$ に対し $X_{ij} \neq X_{ik}$) と仮定する. このとき, モデル(3.2)の仮定の下, Rivest and Wells (2001) は次の Copula-Graphic(CG) 推定量

$$\hat{S}_i^{\text{CG}}(t) = \phi_{i\theta_i}^{-1} \left[\sum_{j: X_{ij} \leq t, \delta_{ij}=1} \left\{ \phi_{i\theta_i} \left(\frac{\bar{Y}_i(X_{ij}) - 1}{n_i} \right) - \phi_{i\theta_i} \left(\frac{\bar{Y}_i(X_{ij})}{n_i} \right) \right\} \right], \quad 0 \leq t \leq \max_j(X_{ij}),$$

を提案しており, ここに $\bar{Y}_i(x) = \sum_{j=1}^{n_i} 1\{X_{ij} \geq x\}$ である. モデル(3.2)が正しく指定されている場合, ある正則条件下で推定量は $\hat{S}_i^{\text{CG}}(t)$ は $S_i(t)$ の一致推定量となる (Rivest and Wells, 2001).

クレイトンコピュラの下では, 生成素 $\phi_{i\theta_i}(t) = (t^{-\theta_i} - 1)/\theta_i$, $\theta_i > 0$, を用い,

$$\hat{S}_i^{\text{CG}}(t) = \left[1 + \sum_{j: X_{ij} \leq t, \delta_{ij}=1} \left\{ \left(\frac{\bar{Y}_i(X_{ij}) - 1}{n_i} \right)^{-\theta_i} - \left(\frac{\bar{Y}_i(X_{ij})}{n_i} \right)^{-\theta_i} \right\} \right]^{-\frac{1}{\theta_i}}$$

となる. グンベルコピュラの下では, 生成素 $\phi_{i\theta_i}(t) = (-\log t)^{\theta_i+1}$, $\theta_i \geq 0$, を用い

$$\hat{S}_i^{\text{CG}}(t) = \exp \left[- \left(\sum_{j: X_{ij} \leq t, \delta_{ij}=1} \left[\left\{ -\log \left(\frac{\bar{Y}_i(X_{ij}) - 1}{n_i} \right) \right\}^{\theta_i+1} - \left\{ -\log \left(\frac{\bar{Y}_i(X_{ij})}{n_i} \right) \right\}^{\theta_i+1} \right] \right)^{\frac{1}{\theta_i+1}} \right].$$

さらに, フランクコピュラの下で生成素 $\phi_{i\theta_i}(t) = -\log[(e^{-\theta_i t} - 1)/(e^{-\theta_i} - 1)]$ を適用すると,

$$\hat{S}_i^{\text{CG}}(t) = -\frac{1}{\theta_i} \log \left[1 + (e^{-\theta_i} - 1) \exp \left(\sum_{j: X_{ij} \leq t, \delta_{ij}=1} \log \left\{ \frac{\exp \left(-\theta_i \frac{\bar{Y}_i(X_{ij})-1}{n_i} \right) - 1}{\exp \left(-\theta_i \frac{\bar{Y}_i(X_{ij})}{n_i} \right) - 1} \right\} \right) \right]$$

となる。加えて、独立コピュラの下では $\phi_{i\theta_i}(t) = -\log(t)$ を用い、次の KM 推定量となる

$$\hat{S}_i^{\text{CG}}(t) = \hat{S}_i(t) = \prod_{j: X_{ij} \leq t, \delta_{ij}=1} \left[1 - \frac{1}{\bar{Y}_i(X_{ij})} \right], \quad 0 \leq t \leq \max_j(X_{ij}).$$

CG 推定量を用いると、さまざまな従属打ち切りの相関構造を想定した下での生存関数の推定および感度分析が可能となる。もし $\phi_{i\theta_i}$ が母集団の相関構造を正しく特定していれば、CG 推定量は一致推定となるが、誤特定している場合はバイアスが生じる。Rivest and Wells (2001) の感度分析によると、母集団のケンドール順位相関の大きさを θ_i が捉えている場合、 $\phi_{i\theta_i}$ の関数型に誤特定があっても CG 推定量のバイアスは KM 推定量のそれよりも小さくなると数値的に示している。

上述の CG 推定量を計算するには、R パッケージ *compound.Cox* (Emura et al., 2019) 内にある関数 *CG.Clayton()*, *CG.Frank()*, *CG.Gumbel()* を用いれば良い。

3.2 治療効果の推定

従属打ち切り下で生存関数の推定が行えることを説明したが、要因分析において推定・検定の対象としているのは治療効果 p である。前述の CG 推定量を用いると、独立打ち切り下での推定量 \hat{p} (2.2 節で定義した) は従属打ち切りの下に自然に拡張される。

まず、式(2.1)のペアワイズ効果 $w_{i\ell}$ を次のように推定する：

$$\hat{w}_{i\ell}^{\text{CG}} = - \int \hat{S}_i^{\text{CG}\pm}(t) d\hat{S}_\ell^{\text{CG}}(t).$$

次に、式(2.2)の $p = Aw$ を用い、 p の推定量を次式で定義する (Emura et al., 2024)。

$$\hat{p}^{\text{CG}} = \begin{pmatrix} \hat{p}_1^{\text{CG}} \\ \vdots \\ \hat{p}_d^{\text{CG}} \end{pmatrix} = A\hat{w}^{\text{CG}} = A\left(\hat{w}_{11}^{\text{CG}}, \dots, \hat{w}_{1d}^{\text{CG}} : \hat{w}_{21}^{\text{CG}}, \dots, \hat{w}_{2d}^{\text{CG}} : \dots : \hat{w}_{d1}^{\text{CG}}, \dots, \hat{w}_{dd}^{\text{CG}}\right)'$$

これは、生存関数に分布の仮定を課さないノンパラメトリックな推定量である。

ある正則条件(付録 2)の下、モデル(3.2)が正しければ推定量は漸近正規性を満足するため、 $N \rightarrow \infty$ としたとき近似 $\sqrt{N}(\hat{p}^{\text{CG}} - p) \sim N_d(\mathbf{0}, \mathbf{V})$ が可能となり、ここに \mathbf{V} は共分散行列である。未知の \mathbf{V} はジャックナイフ推定量を N 倍してスケーリングした $\hat{\mathbf{V}} = N \times \mathbf{V}(\hat{p}^{\text{CG}})$ で推定可能であり (Emura et al., 2024)，ここにジャックナイフ推定量は

$$\mathbf{V}(\hat{p}^{\text{CG}}) = \begin{pmatrix} \hat{\sigma}_1^2 & \dots & \hat{\sigma}_{d1} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{d1} & \dots & \hat{\sigma}_d^2 \end{pmatrix} = \frac{N-1}{N} \sum_{i=1}^d \sum_{j=1}^{n_i} (\hat{p}^{\text{CG}(-ij)} - \hat{p}^{\text{CG}(.)})(\hat{p}^{\text{CG}(-ij)} - \hat{p}^{\text{CG}(.)})',$$

ここで $\hat{p}^{\text{CG}(-ij)}$ は i 番目の群の j 番目の被験者を除いて計算した推定量、 $\hat{p}^{\text{CG}(.)} = \sum_{i=1}^d \sum_{j=1}^{n_i} \hat{p}^{\text{CG}(-ij)}/N$ はそれらの平均である。原論文の Emura et al. (2024) ではジャックナイフ分散推定量の式の $\frac{N-1}{N}$ の部分を $\frac{N}{N-1}$ としたが、これはタイポであり、本稿では $\frac{N-1}{N}$ と修正した。R 関数 *surv.factorial()* 上での計算プログラムに誤りはなく、修正の必要は無い。

CG 推定量はタイを許容しない都合上、通常のブートストラップ法のような復元抽出による

方法は適用が難しい。また、標準誤差(SE)は $SE(\hat{p}_i^{CG}) = \hat{\sigma}_i$ となり、 p_i の $(1 - \alpha) \times 100\%$ 信頼区間(CI)は

$$[\hat{p}_i^{CG} - z_{\alpha/2} \times SE(\hat{p}_i^{CG}), \quad \hat{p}_i^{CG} + z_{\alpha/2} \times SE(\hat{p}_i^{CG})],$$

ここに $z_{\alpha/2}$ は標準正規分布の上側 $(\alpha/2) \times 100\%$ 点である。信頼区間は上限・下限が p_i の範囲 $[0, 1]$ を超える可能性があるが、その場合必要に応じて、logit 変換などで区間 $[0, 1]$ に収まるようにも良い (Brunner et al., 2017; Perme and Manevski, 2019) が、本論文では特に用いない。

以上の手法において、独立コピュラを適用すると Dobler and Pauly (2020) の手法に帰着する。しかしながら、彼らは、ジャックナイフ法とは異なり、Multiplier ブートストラップ法で SE と信頼区間を構成している。Multiplier ブートストラップ法は、独立打ち切り下では有用とみられるが、上記のコピュラモデル下での適用は難しい。

3.3 治療効果の検定

ある対比行列 C に対する検定問題 $H_0: Cp = \mathbf{0}$ vs. $H_1: Cp \neq \mathbf{0}$ を定式化する。独立打ち切りの仮定下で定義した F_N の式(2.2 節で定義した)に \hat{p}^{CG} を適用すると、次の統計量を得る：

$$F_N^{CG} = \frac{N(\hat{p}^{CG})' \mathbf{T} \hat{p}^{CG}}{\text{tr}(\mathbf{T} \hat{V}^{CG})}$$

ここで、 $\mathbf{T} = C'(CC')^+C$ は射影行列、 $\text{tr}(\cdot)$ はトレース、 $\hat{V}^{CG} = N \times \mathbf{V}(\hat{p}^{CG})$ はジャックナイフ推定量を N 倍して共分散行列 \mathbf{V} の推定量にスケーリングしたものである。

統計量 F_N^{CG} の漸近分布について次のような結果が得られている。いま、 $\mathbf{T} \mathbf{V}$ の固有値を $\lambda_i, i = 1, \dots, d$ とすると、 $\sum_{i=1}^d \lambda_i = \text{tr}(\mathbf{T} \mathbf{V})$ である。帰無仮説 H_0 の下で $N \rightarrow \infty$ としたときの F_N^{CG} 漸近分布は、重み付きカイ二乗分布の和

$$F_N^{CG} \sim \frac{\sum_{i=1}^d \lambda_i \chi_i^2(1)}{\sum_{i=1}^d \lambda_i} = \frac{\sum_{i=1}^d \lambda_i \chi_i^2(1)}{\text{tr}(\mathbf{T} \mathbf{V})},$$

であり、ここに $\chi_i^2(1)$ は自由度 1 のカイ二乗分布を表す (Emura et al., 2024)。この重みづけカイ二乗分布の和の上側 $\alpha \times 100\%$ 点を $c_{N,\alpha}$ とする。帰無仮説の下で漸近分布の期待値は、上式から明らかに 1 である。また、対立仮説の下で F_N^{CG} は無限大に確率収束する。よって、漸近的に有意水準 α である検定は棄却領域を $F_N^{CG} > c_{N,\alpha}$ とすれば良い。

この棄却限界 $c_{N,\alpha}$ は未知の行列 \mathbf{V} に依存するため、データから推定する必要がある。次に示すよう、推定法として「シミュレーション法」と「解析的方法」の 2 つが提案されているが、どちらを用いてもほぼ同じような数値結果が得られる(5 節、6 節)。

シミュレーション法による $c_{N,\alpha}$ の推定

打ち切りの無い場合に Brunner et al. (2017) で提案された方法を適用する。まず λ_i の推定値 $\hat{\lambda}_i$ を $\mathbf{T} \hat{V}^{CG}$ の固有値で定める。次に自由度 1 のカイ二乗分布に従う乱数 $\chi_{i,(r)}^2(1)$, $i = 1, \dots, d$, $r = 1, \dots, R$ を生成し、ここに R は大きな数 $R = 1000$ などとする。最後に、

$$F_{N,(r)}^{CG} = \frac{\sum_{i=1}^d \hat{\lambda}_i \chi_{i,(r)}^2(1)}{\text{tr}(\mathbf{T} \hat{V}^{CG})}, \quad r = 1, \dots, R$$

のヒストグラムの上側 $\alpha \times 100\%$ 点として $c_{N,\alpha}$ を推定する。検定の P 値は、ヒストグラムの上側確率 $\sum_{r=1}^R 1\{F_{N,(r)}^{CG} > F_N^{CG}\}/R$ で定める。

解析的方法による $c_{N,\alpha}$ の推定

打ち切りの無い場合に Brunner and Puri (2001) と Brunner et al. (2017) で提案された方法を

適用する。これは重み付きカイ二乗分布の和(漸近分布)をカイ二乗分布で近似する手法である。ある正の定数 f と g に対し、以下の近似

$$\frac{\sum_{i=1}^d \lambda_i \chi_i^2(1)}{\text{tr}(\mathbf{T}\mathbf{V})} \approx g\chi^2(f)$$

を考える。ここで、 f と g は上式の右辺と左辺の期待値と分散が同じになるよう定めると、

$$\frac{\sum_{i=1}^d \lambda_i}{\text{tr}(\mathbf{T}\mathbf{V})} = gf, \quad 2 \frac{\sum_{i=1}^d \lambda_i^2}{[\text{tr}(\mathbf{T}\mathbf{V})]^2} = 2g^2f$$

である。最初の方程式左辺は 1 であるから、 $1 = gf$ となる。これと 2 番目の方程式より f と g が求まり、さらに未知の \mathbf{V} に推定値を代入すると以下となる。

$$c_{N,\alpha} = \frac{\chi^2(1 - \alpha, \hat{f})}{\hat{f}}, \quad \hat{f} = \frac{[\sum_{i=1}^d \hat{\lambda}_i]^2}{\sum_{i=1}^d \hat{\lambda}_i^2} = \frac{[\text{tr}(\mathbf{T}\hat{\mathbf{V}}^{\text{CG}})]^2}{\text{tr}(\mathbf{T}\hat{\mathbf{V}}^{\text{CG}}\mathbf{T}\hat{\mathbf{V}}^{\text{CG}})}.$$

ここで、 $\chi^2(1 - \alpha, \hat{f})$ は自由度 \hat{f} のカイ二乗分布の上側 $100 \times \alpha\%$ 点である。検定の P 値は、カイ二乗分布の上側確率 $1 - F_{\chi^2(\hat{f})}(\hat{f}F_N^{\text{CG}})$ で定め、ここに $F_{\chi^2(\hat{f})}$ は $\chi^2(\hat{f})$ の累積分布関数である。

3.4 フォローアップ期間の制約

定まったフォローアップ期間内で収集されたデータを基にした生存関数の推定は不完全であるから、その結果として治療効果 p の推定も不完全となる。ある上限 $\tau > 0$ を超えた生存時間が確実に打ち切られる状況下において、生存関数 $S_i(t)$ が推定可能な範囲は $t \in [0, \tau]$ に制約され、ここに

$$\tau < \min_i [\sup_u \{u: G_i(u)S_i(u) > 0\}]$$

である (Emura et al., 2024)。これより、生存関数 $S_i(t)$ は $t \in [0, \tau]$ においてのみ推定可能となる。以下、治療効果 p を推定可能な生存関数 $S_i(t)$ の範囲 $t \in [0, \tau]$ を用いたものに修正する。

まずペアワイズ効果を次のように修正する。

$$\begin{aligned} w_{i\ell} &= P[\min(T_i, \tau) > \min(T_\ell, \tau)] + \frac{1}{2}P[\min(T_i, \tau) = \min(T_\ell, \tau)] \\ &= - \int_0^\tau S_i^\pm(t) dS_\ell(t) + \frac{S_i(\tau)S_\ell(\tau)}{2}. \end{aligned}$$

右辺の最後の項は、積分を切斷したことによるペアワイズ効果の減少量を補正している。なお、 $w_{ii} = 1/2$ である。これより、修正したペアワイズ効果 $w_{i\ell}$ の推定量は次のようになる

$$\hat{w}_{i\ell}^{\text{CG}} = - \int_0^\tau \hat{S}_i^{\text{CG}\pm}(t) d\hat{S}_\ell^{\text{CG}}(t) + \frac{\hat{S}_i^{\text{CG}}(\tau)\hat{S}_\ell^{\text{CG}}(\tau)}{2}.$$

これより、治療効果 $\hat{p}^{\text{CG}} = \mathbf{A}\hat{\mathbf{w}}^{\text{CG}}$ が求まる。これと同様に検定法も修正すれば良い。

3.5 感度分析

治療効果の推定・検定を実行するためには、コンピュータの型とパラメータ θ_i を指定する必要がある。しかしながら、従属打ち切りのモデルが同定困難 (Tsiatis, 1975) であるため、CG 推定量に用いるコンピュータもデータから推定することは困難である。このため、コンピュータとそのパラメータの候補をいくつか選択し、推定結果がどう変化するか調べる「感度分析法」が採用されてきた (Rivest and Wells, 2001; De Uña-Álvarez and Veraverbeke, 2013, 2017; Emura and Hsu, 2020; Yeh et al., 2023; Emura et al., 2024)。例えば、コンピュータの型やパラメータ θ_i の選択に関

わらず検定が有意になる場合、この有意性の結論は様々な従属打ち切り構造下でも妥当といえる。一方、特定の θ_i の場合のみ有意になる結果などは、決定的なものにはならない。このような、様々な従属打ち切りのモデル下での感度分析は、従来の独立打ち切りの仮定下のみでの分析よりも有意性について厳しい条件を要求する。

感度分析を行う上で認識すべき点として、推定目標である治療効果 p は、打ち切りの分布や従属構造に依存せず生存関数のみで定まる母数である（式(2.1)および式(2.2)を参照）ことである。よって、様々なコピュラの仮定で変わる推定結果は、推定値（推定法）が変わる結果であり、母数が変わるわけではない。

4. ソフトウェア

3 節で紹介した手法は R パッケージ *compound.Cox* (Emura et al., 2019) 内の R 関数 *surv.factorial()* で実行可能である。この関数は、推定値 \hat{p}^{CG} とその SE、および p の 95% 信頼区間を算出する。また、検定統計量 F_N^{CG} とその棄却限界値 $c_{N,\alpha}$ 、および P 値も算出する。棄却限界はシミュレーション法と解析的手法の 2 つで計算され、それぞれ有意水準 $\alpha = 0.10, 0.05, 0.01$ の値を算出する。仮説 $H_0: C\mathbf{p} = \mathbf{0}$ における対比行列 C は自由に指定できるが、指定がなれば大域的仮説 $H_0: p_1 = p_2 = \dots = p_d$ を検定する。また、最大フォローアップ期間 τ も指定できるが、指定がなければ CG 推定量 $\hat{S}_i^{\text{CG}}(t)$ が全ての i と $t \in [0, \tau]$ で定義可能となるよう

$$\tau = \min_i \left[\max_j (X_{ij}) \right] < \min_i [\sup \{u: G_i(u) S_i(u) > 0\}]$$

をデフォルトにしている。これはいずれかの群で生存者が残っている最大の時間である。しかしながら、推定量 $\hat{S}_i^{\text{CG}}(t)$ が時間 τ の近傍で不安定になりやすいことから、 τ をさらに小さくとり、

$$(4.1) \quad \tau = \min_i \left[\max_{j: \delta_{ij}=1} (X_{ij}) \right] \leq \min_i \left[\max_j (X_{ij}) \right]$$

と指定でも良い（6 節の例を参照）。これはいずれかの群で死亡例が残っている最大の時間である。

入出力系統などの詳細については、オンラインマニュアルを参照頂きたい。利便性の都合上、*surv.factorial()* 関数では群間でコピュラが同一である条件下で計算する。即ち、 $C_{1\theta_1} = \dots = C_{d\theta_d}$ かつ $\theta_1 = \dots = \theta_d$ を仮定するため、コピュラの型 1 つとパラメータ 1 つを指定すれば十分である。

R コードの説明のために、R の *survival* パッケージ内にある *colon* データの解析コードの例を下に示す。まず入力すべき変数を準備し、“*surv.factorial(t.vec, d.vec, group, copula=CG, Clayton, theta)*” を実行すれば良い。ここでは “*copula=CG.Clayton*” でクレイトンコピュラを指定し “*theta*” でパラメータを指定している。3 群のデータの例なので、クレイトンコピュラのパラメータ $\theta_1 = \theta_2 = \theta_3 = 2$ の下で解析が行われる。局所的仮説の検定を行うには、対比行列 “*C=C12*” などと設定すれば良い。

```
library(survival)
library(compound.Cox)
data(cancer)
dat=subset(colon,etype==1)
## Treatment effects ##
```

```

> surv.factorial(t.vec,d.vec,group,copula=CG.Clayton,theta,t.upper=tau)
$copula.parameter
[1] 2

$p
  estimate      se    lower    upper      P
[1,] 0.4693207 0.01281679 0.4442003 0.4944412 9.916598e-01
[2,] 0.4701347 0.01283573 0.4449771 0.4952922 9.900103e-01
[3,] 0.5616860 0.01243152 0.5373207 0.5860513 3.487346e-07

$Var
      [,1]      [,2]      [,3]
[1,] 1.642701e-04 -8.720548e-05 -7.708513e-05
[2,] -8.720548e-05 1.647558e-04 -7.749042e-05
[3,] -7.708513e-05 -7.749042e-05 1.545426e-04

$F
[1] 11.65897

$c.simu
  0.10      0.05      0.01
2.418888 3.072035 4.776394

$c.anal
  0.10      0.05      0.01
2.303544 2.997522 4.609088

$p.value
  simu      anal
0.000000e+00 8.766784e-06

```

図 1. `surv.factorial()` 関数で `colon` データを解析した R コードのアウトプット.

```

t.vec=dat$time
d.vec=dat$status
trt=dat$rx
C12=matrix(c(1,-1,0), 1,3,byrow=TRUE)
C13=matrix(c(1,0,-1), 1,3,byrow=TRUE)
C23=matrix(c(0,1,-1), 1,3,byrow=TRUE)
group=as.numeric(trt) # 1=Obs; 2=Lev; 3=Lev+5FU
tau=min(tapply(t.vec[d.vec==1],group[d.vec==1], max)) ## maximum follow-up date
theta=2 ## copula parameter
surv.factorial(t.vec,d.vec,group,copula=CG.Clayton,theta,t.upper=tau)
surv.factorial(t.vec,d.vec,group,copula=CG.Clayton,theta,C=C12,t.upper=tau)
surv.factorial(t.vec,d.vec,group,copula=CG.Clayton,theta,C=C13,t.upper=tau)
surv.factorial(t.vec,d.vec,group,copula=CG.Clayton,theta,C=C23,t.upper=tau)

```

R コード出力は図 1 のようになる。推定値 $\hat{p}_i^{\text{CG}}, i = 1, 2, 3$ とそれらの SE, および p_i , $i = 1, 2, 3$ の 95% 信頼区間が表示されているのがわかる。加えて、ジャックナイフ法による $(\hat{p}_1^{\text{CG}}, \hat{p}_2^{\text{CG}}, \hat{p}_3^{\text{CG}})$ の共分散行列 $V(p^{\text{CG}})$ もある。また、統計量 F_N^{CG} と棄却限界値 $c_{N,\alpha}$ (シミュレーション法と解析的方法での 10%, 5%, 1% 棄却限界), および P 値が出力されているのがわかる。

5. シミュレーション

Emura et al. (2024) のシミュレーション研究により, `surv.factorial()` 関数を用いた推定・検定の性能はある程度調べられている。この既出のシミュレーションではまだ検証されていなかった局所的仮説の検定に関する性能, および大域的仮説検定と局所的仮説検定の関連を調べるために, 新たなシミュレーション実験を以下で行う。

5.1 シミュレーション設計

3つの治療群($d = 3$)を考え、各群 $i \in 1, 2, 3$ のサンプルサイズは $n_i = 50$ または $n_i = 100$ とした。生存時間と打ち切り時間のペア (T_{ij}, U_{ij}) , $i = 1, 2, 3$, $j = 1, 2, \dots, n_i$ をクレイトンコピュラモデルから生成し、コピュラパラメータを $\theta_i = 2$ ($\tau_{\theta_i} = 0.5$; 中程度の正相関)とした。周辺分布は T_{ij} が強度パラメータ λ_i の指数分布, U_{ij} が強度パラメータ μ_i の指数分布とし、フォローアップの最大時刻 τ に対して、 $U_{ij} = \min(U_{ij}, \tau)$ と再定義した。打ち切り時刻の強度パラメータを $\mu_1 = \mu_2 = \mu_3 = 1$, フォローアップ最大時刻を $\tau = 1$ とした。これより右打ち切り下での観測データ $\{(X_{ij}, \delta_{ij}); i = 1, 2, \dots, d, j = 1, 2, \dots, n_i\}$, $X_{ij} = \min(T_{ij}, U_{ij})$, $\delta_{ij} = 1_{T_{ij} \leq U_{ij}}$ を得た。

3.3 節で紹介した検定が有用となるのは、3つの群の差がKM推定量のグラフでは視覚的に判定しにくい状況である。シミュレーションではそのようなシナリオを2つ考えた：

- シナリオ 1 ($\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1.25$)：群1と群2に差が無いが、群1と群3および群2と群3に小さい差がある場合。このときの治療効果は、 $p_1 = 0.517$, $p_2 = 0.517$, $p_3 = 0.467$ 。
- シナリオ 2 ($\lambda_1 = 0.75$, $\lambda_2 = 1$, $\lambda_3 = 1.25$)：全ての群に小さい差がある場合。このときの治療効果は、 $p_1 = 0.556$, $p_2 = 0.497$, $p_3 = 0.447$ 。

値 1.25 や 0.75 は弱い治療効果の差を表す。例えば、強度 1.25 の群は強度 1.00 の群からハザードが 25% 増えるが、ハザード比 1.25 は非劣勢試験のマージンとして設定される典型的な量である (Yamada et al., 2018; Derkx et al., 2022)。

打ち切り割合 $P(U_{ij} < T_{ij})$ は1つ目のシナリオで、0.63(群1), 0.63(群2), 0.54(群3)であり、2つ目のシナリオで、0.73(群1), 0.63(群2), 0.54(群3)である。

以下の4つの検定問題を考える：

$$H_0: p_1 = p_2 \quad \text{vs.} \quad H_1: p_1 \neq p_2,$$

$$H_0: p_1 = p_3 \quad \text{vs.} \quad H_1: p_1 \neq p_3,$$

$$H_0: p_2 = p_3 \quad \text{vs.} \quad H_1: p_2 \neq p_3,$$

$$H_0: p_1 = p_2 = p_3 \quad \text{vs.} \quad H_1: p_1 \neq p_2, \quad \text{or} \quad p_1 \neq p_3, \quad \text{or} \quad p_2 \neq p_3.$$

最初の3つの局所的な検定では、対比行列を $\mathbf{C} = (1, -1, 0)$, $\mathbf{C} = (1, 0, -1)$, $\mathbf{C} = (0, 1, -1)$ とし、最後の大域的な検定では $\mathbf{C} = \mathbf{I}_3 - \mathbf{1}_3 \mathbf{1}'_3 / 3$ とした。各設定下で、データから統計量 F_N^{CG} と上側 5% 点 $c_{N,\alpha=0.05}$ を算出した。これを 1000 回繰り返し、 F_N^{CG} の平均値と検定の棄却率を計算した。

5.2 シミュレーション結果

表1にシミュレーションの結果を示す。帰無仮説が成立している $H_0: p_1 = p_2$ の検定では、 F_N^{CG} の平均値は 1 に近くなっ。これは漸近帰無分布の期待値が 1 であることと整合している。また、シナリオ 1 における $H_0: p_1 = p_2$ の検定では棄却率が名目値の 0.05 に近くなり、これは棄却限界値 $c_{N,\alpha}$ の設定法が妥当で、検定のサイズが正しく保たれていることを示している。帰無仮説が成立していない下では、 F_N^{CG} の平均値は 1 より大きくなり、近棄却率は 0.05 より高くなっ。特に、サンプルサイズが大きい($n_i = 100$)場合は、小さい場合($n_i = 50$)に比べ、棄却率は高くなっ。棄却限界値 $c_{N,\alpha}$ の 2 とおりの計算法(シミュレーション法と解析的方法)の間に差は殆ど見られなかっ。これは、解析的方法で用いた漸近帰無分布をカイ二乗分布で近似する手法が適切に働いていることを意味する。

シナリオ 1 では、大域的仮説($H_0: p_1 = p_2 = p_3$)は、局所的仮説($H_0: p_1 = p_3$ と $H_0: p_2 = p_3$)

表 1. シミュレーション 1000 回による統計量 F_N^{CG} の平均値および名目上の有意水準 $\alpha = 0.05$ の下での仮説 H_0 の棄却率.

シナリオ 1*	$p_1 = p_2$	n_i	$E[F_N^{\text{CG}}]$	棄却率:	棄却率:	棄却率の解釈
				シミュレーション	解析的	
シナリオ 1*	$p_1 = p_2$	50	1.082	0.061	0.064	有意水準
		100	0.962	0.044	0.047	有意水準
シナリオ 2**	$p_1 = p_3$	50	1.749	0.130	0.128	検出力
		100	2.032	0.172	0.175	検出力
シナリオ 2**	$p_2 = p_3$	50	1.687	0.131	0.131	検出力
		100	1.988	0.167	0.167	検出力
シナリオ 2**	$p_1 = p_2 = p_3$	50	1.467	0.118	0.121	検出力
		100	1.634	0.158	0.159	検出力
シナリオ 2**	$p_1 = p_2$	50	1.855	0.146	0.144	検出力
		100	2.553	0.233	0.230	検出力
シナリオ 2**	$p_1 = p_3$	50	4.036	0.381	0.382	検出力
		100	6.210	0.628	0.623	検出力
シナリオ 2**	$p_2 = p_3$	50	1.680	0.133	0.131	検出力
		100	1.992	0.165	0.167	検出力
シナリオ 2**	$p_1 = p_2 = p_3$	50	2.447	0.312	0.308	検出力
		100	3.540	0.515	0.514	検出力

*シナリオ 1: $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1.25; p_1 = 0.517, p_2 = 0.517, p_3 = 0.467$;

**シナリオ 2: $\lambda_1 = 0.75, \lambda_2 = 1, \lambda_3 = 1.25; p_1 = 0.556, p_2 = 0.497, p_3 = 0.447$

よりも、わずかに棄却されにくくなった. これは、大域的仮説がその一部に検出不能な $p_1 = p_2$ を含むからである. 逆に、シナリオ 2 では、大域的仮説は、局所仮説 $H_0: p_1 = p_2$ と $H_0: p_2 = p_3$ よりも棄却されやすかったが、これは、 p_1, p_2, p_3 の 3 つとも異なる値を持つため、大域的仮説が棄却されやすいのは自然な結果である. 一方、シナリオ 2 における局所仮説 $H_0: p_1 = p_3$ は、大域的仮説よりもさらに棄却されやすかったが、これは p_1 と p_3 のペアは最も差が大きいペアだからである.

以上のシミュレーションの結果により、3.3 節で紹介した検定は、3 群の比較において妥当な有意水準(サイズ)と自然な検出力特性を持つことが示された. 検出力の性能では、群間差の大きいペアの局所的検定の検出力は、大域的検定の検出力より高まることが示された. しかしながら、通常の検定手順としては、最初に大域的検定を行って棄却された後に局所的検定を行うべきである.

6. データ解析

結腸がんのデータ解析例を通じて、3 節で紹介した手法を解説する. 解析は `surv.factorial()` 関数を用いて行う. 本節の解析結果を再現する R コードはオンライン Supplement を参照されたい.

Moertel et al. (1995) は、 $N = 929$ 人のステージ 3 結腸がんの切除手術を受けた被験者を次の 3 つの治療群($d = 3$)に無作為割付けし、各治療のがん再発に与える効果を検証した.

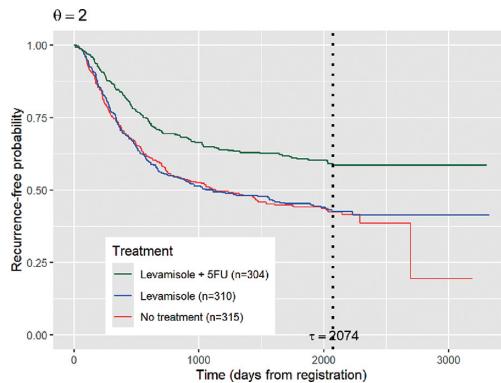


図 2. 結腸がんの切除手術を受けた 929 人の被験者の各治療における無再発確率の CG 推定値. クレイトンコピュラに $\theta_i = 2$ を適用. 経過観察(No adjuvant treatment), レバミゾール投与(Levamisole), レバミゾールと 5FU の併用(Levamisole + 5FU)の 3 群比較.

- 経過観察群 ($i = 1$) : 術後経過観察のみ,
- レバミゾール群 ($i = 2$) : レバミゾール投与による術後補助化学療法,
- 併用群 ($i = 3$) : レバミゾールと 5FU を併用した術後補助化学療法.

これらの治療が“無再発期間(Recurrence free interval)”に与える効果を推定・検定する. 無再発期間とは, 登録日から再発が確認されるまでの日数(生存時間 T_i に対応する数値)であるので, 長い期間ほど良い治療効果と見做す. 打ち切り時間は再発が観測されないまま観察が終わるまでの日数(打ち切り時間 U_i に対応する数値)であるが, 観察終了の理由は死亡によるものも含む(Moertel et al., 1995, Fig. 1). 再発時間と死亡時間は正に相関している可能性は十分あり, この独立打ち切りの仮定は妥当でない可能性がある(7 節で再び議論する). このような理由から, コピュラに基づく要因分析の手法(3 節)を用いて, 従属打ち切りを想定した解析を以下で行う. データは R パッケージ *survival* のデータセット *colon* から入手可能である.

図 2 に各群における無再発確率の CG 推定値(生存曲線)を図示した. ここで CG 推定値は, クレイトンコピュラにパラメータ $\theta_i = 2$ ($\tau_{\theta_i} = 0.5$; 中程度の正相関)を仮定した下で計算した. 図より, 併用群は他の 2 つの群に比べて無再発率が高い(治療効果が高い)ことが分かる. メディアン生存時間で見ると, レバミゾール投与群と経過観察群で 1200 日(3.3 年)ほどであるが, 併用群では 3000 日(8.2 年)以上もある. 一方, レバミゾール群と経過観察群での優劣は図から判定困難であり, 検定による判断に委ねることにする.

表 2 に群間差を検定した結果を示す. 3 つの群の生存率が比較可能な期間を $t \in [0, \tau]$ とし, ここに日数 $\tau = 2074$ は式(4.1)を用いた. 大域的な帰無仮説 $H_0: p_1 = p_2 = p_3$ の検定統計量は $F_N^{CG} = 11.659$ となり, これは帰無平均 1 から大きく離れた値であり, 有意水準 1% で有意となった($P < 0.001$). また, 併用群との比較に焦点を当てた局所的な帰無仮説($H_0: p_1 = p_3$ と $H_0: p_2 = p_3$)も 1% で有意となった($P < 0.001$). この有意差は, 図 2 から観測できる無再発率の大きな差から導かれる自然な結果である. 一方, 図 2 から予想できるよう, 経過観察群とレバミゾール群の同等性($H_0: p_1 = p_2$)は棄却されなかった. 以上の検定結果は, コピュラパラメータを $\theta_i = 2$ と仮定した下で行ったが, 他のコピュラパラメータ($\theta_i = 0.001, \theta_i = 4, \theta_i = 8$)の下での感度分析の結果も同じ結論となった(詳細は付録 3). 以下で個々の治療効果の推定を行う.

表 2. 結腸がんの切除手術を受けた 929 人の被験者の治療効果の同等性の検定結果. クレイトンコピュラにパラメータを $\theta_i = 2 (\tau_{\theta_i} = 0.5$; 中程度の正相関)とした下で計算.

比較する群	帰無仮説 H_0	検定統計量:		P 値:	P 値:
		F_N^{CG}	シミュレーション法		
経過観察とレバミゾール	$p_1 = p_2$	0.001	0.976	0.971	
経過観察と併用	$p_1 = p_3$	18.037	<0.001	<0.001	
レバミゾールと併用	$p_2 = p_3$	17.672	<0.001	<0.001	
3 つの群全て	$p_1 = p_2 = p_3$	11.659	<0.001	<0.001	

表 3. 結腸がんの切除手術を受けた 929 人の被験者の治療効果の推定値. 母数 p_i は, i 番目の治療の相対的な治療効果を表し, 値が大きいほど高い治療効果を表す.

コピュラパラメータ	治療群	母数	推定値	SE	95%信頼区間
$\theta_i = 0.001 (\tau_{\theta_i} = 0.0)$	経過観察	p_1	0.469	0.013	0.444, 0.494
	レバミゾール	p_2	0.470	0.013	0.446, 0.495
	併用	p_3	0.561	0.012	0.538, 0.585
$\theta_i = 2 (\tau_{\theta_i} = 0.5)$	経過観察	p_1	0.469	0.013	0.444, 0.494
	レバミゾール	p_2	0.470	0.013	0.445, 0.495
	併用	p_3	0.562	0.012	0.537, 0.586
$\theta_i = 4 (\tau_{\theta_i} = 0.67)$	経過観察	p_1	0.469	0.013	0.444, 0.495
	レバミゾール	p_2	0.470	0.013	0.445, 0.496
	併用	p_3	0.562	0.013	0.537, 0.586
$\theta_i = 8 (\tau_{\theta_i} = 0.80)$	経過観察	p_1	0.469	0.013	0.443, 0.495
	レバミゾール	p_2	0.471	0.013	0.446, 0.497
	併用	p_3	0.561	0.013	0.536, 0.586

表 3 に様々な θ_i の値の下で推定した治療効果 \hat{p}_i^{CG} をまとめた. 例えば, $\theta_i = 2$ での結果を見ると, 併用群の治療効果は $\hat{p}_3 = 0.562$ (CI: 0.537, 0.586) であるから, 併用群の被験者集団が全被験者集団よりも再発を遅らせられる確率は 56.2% である. これより, 併用治療は良い治療と結論できる. これより併用群の無再発期間は他の 2 つの群のそれに比べて長いことが分かり, これは図 2 から得られる印象と整合する. 一方, レバミゾール群と経過観察群の治療効果の推定値を見ると, これらの群の被験者集団が全被験者集団より無再発期間を長く出来る確率は 47% ほどであり, これらは良くない治療といえる. 以上のような結果と解釈は, θ_i を他の値に変えて解析してもほぼ同じであり(表 3), さらにゲンベルコピュラの下でも同じであった(結果は割愛した). 故に, 併用群が高い治療効果, レバミゾール群と経過観察群が低い治療効果を示す, という結論は従属打ち切りの構造に影響を受けない頑健なものであることが示された. 以上の結果と解釈は, 独立打ち切りの仮定下で得られた Moertel et al. (1995) の報告結果をより強固にするものである.

7. 結論と考察

本論文では、打ち切りが独立でない状況における生存時間の群間比較を行うための要因計画・要因分析法 (Emura et al., 2024) を詳説した。この方法はコピュラを用いて、打ち切り時間と生存時間の従属構造を特定した下で、治療効果の推定・検定を実行する。コピュラの型や相関パラメータの特定は、実際のデータからは困難であるため、コピュラとパラメータの候補をいくつか用意し、それらの下での解析結果を吟味する感度分析法を紹介した(3.3 節, 6 節)。また、本論文では新たなシミュレーション実験を行い、原著では考察されていなかった局所的検定の検出力特性および大域的検定との関連を明らかにした。最後に紹介した結腸がんデータの解析例も、原著論文には無い新たな例であり、提案手法の有用性を再確認した。スペースの制限により、本論文の数値例ではクレイトンコピュラの結果のみを報告したが、グンベルコピュラやフランクコピュラなどの他のアルキメデスコピュラでも同様の解析を行うことが可能となる。本稿ではクレイトンコピュラ下での結果のみ報告したが、グンベルやフランクコピュラ下で解析する場合は、Supplement の R コードで `copula=CG.Clayton` を `copula=CG.Gumbel` に変更などすれば良い。ただし、紹介した手法はアルキメデスコピュラによるモデルに限定されるため、アルキメデスコピュラでないコピュラ(例えば、正規コピュラ、FGM コピュラ、プラケットコピュラなどがある)をどう活用するのか考察することは今後の課題となる。

結腸がんのデータ解析では、再発時間と死亡時間は正に相関しているという考え方の下、独立打ち切りの仮定が妥当でない可能性を考慮した解析を行った。死亡と増悪(再発・転移など)が相関しているという考えは、がん領域では一般的であり、がん臨床データの解析では独立打ち切りの仮定を必要としない競合リスクモデルによる解析が推奨されることもある(齋藤・室谷, 2024)。競合リスク解析では、“原因別ハザードモデル”または“部分分布ハザードモデル”などを用いれば、リスク間の独立性が成立するかどうかに関わらず解析を進めることができある。一方、本論文の従属打ち切りのコピュラモデルは、競合リスクモデルの一種である“潜在故障時間モデル”(宮村, 1982; 武富・山本, 2023; 道前, 2023, 2025)と位置付けられる。潜在故障時間モデルにおいてコピュラを用いる場合、コピュラのパラメータをデータから推定する方法 (Emura and Chen, 2016; Deresa and Van Keilegom, 2020, 2022; Czado and Van Keilegom, 2023; Lo et al., 2024; Michimae et al., 2024; Delhelle and Van Keilegom, 2025) と、そうでない仮定されたコピュラの方法 (Zheng and Klein, 1994, 1995; De Uña-Álvarez and Veraverbeke, 2013, 2017; Emura and Michimae, 2017; Emura and Hsu 2020; Rivest and Wells, 2001; Michimae and Emura, 2022; Emura et al., 2024) がある。本論文の要因実験・要因分析の方法論では、Rivest and Wells (2001) に従い、仮定されたコピュラの下で CG 推定法を適用していることから、感度分析法(3.3 節, 6 節)を採用した。コピュラモデルは、がん臨床試験において無増悪生存期間と全生存期間の相関の推定に用いられることも多く(江村・大庭, 2024; Simon et al., 2025), 本論文のデータ解析で考えた再発時間と打ち切り時間の相関のコピュラモデルもそれに準じたものと理解できる。

謝 辞

改稿に当たり、有益なコメントを下さった 2 名の査読者と編集委員の方に感謝申し上げます。Emura et al. (2024) に共著者として重要な貢献を頂いた Marc Ditzhaus 氏は、本論文を執筆中の 2024 月 9 月に永眠されました。謹んで感謝と哀悼の意を表し、ご冥福をお祈りします。本研究は科研費(22K11948; 20H04147; 24K13473)の助成を受けました。

付録 1. クロネッカー積と対比行列による帰無仮説

任意の $m \times n$ 行列 \mathbf{D} と, 任意の行列 \mathbf{E} に対して, クロネッカー積は次のように定義される:

$$\mathbf{D} \otimes \mathbf{E} = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{m1} & \dots & d_{mn} \end{pmatrix} \otimes \mathbf{E} = \begin{pmatrix} d_{11}\mathbf{E} & \dots & d_{1n}\mathbf{E} \\ \vdots & \ddots & \vdots \\ d_{m1}\mathbf{E} & \dots & d_{mn}\mathbf{E} \end{pmatrix}.$$

これより,

$$\mathbf{A} = \mathbf{I}_d \otimes \left(\frac{\mathbf{1}'_d}{d} \right) = \frac{1}{d} \begin{pmatrix} 1, \dots, 1 & 0, \dots, 0 & \dots & 0, \dots, 0 \\ 0, \dots, 0 & 1, \dots, 1 & \dots & 0, \dots, 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0, \dots, 0 & 0, \dots, 0 & \dots & 1, \dots, 1 \end{pmatrix}$$

であるから,

$$\begin{aligned} \mathbf{A}\mathbf{w} &= \frac{1}{d} \begin{pmatrix} 1, \dots, 1 & 0, \dots, 0 & \dots & 0, \dots, 0 \\ 0, \dots, 0 & 1, \dots, 1 & \dots & 0, \dots, 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0, \dots, 0 & 0, \dots, 0 & \dots & 1, \dots, 1 \end{pmatrix} \\ &\quad \times (w_{11}, \dots, w_{1d} : w_{21}, \dots, w_{2d} : \dots : w_{d1}, \dots, w_{dd})' \\ &= \frac{1}{d} \begin{pmatrix} w_{11} + \dots + w_{1d} \\ \vdots \\ w_{d1} + \dots + w_{dd} \end{pmatrix} = \begin{pmatrix} p_1 \\ \vdots \\ p_d \end{pmatrix} \end{aligned}$$

となり, $\mathbf{p} = \mathbf{A}\mathbf{w}$ が示された. ちにみに, クロネッカー積は可換でなく, $\mathbf{D} \otimes \mathbf{E} = \mathbf{E} \otimes \mathbf{D}$ は一般に成立しないため, 積の順序に注意する必要がある.

また, 対比行列を $\mathbf{C} = \mathbf{P}_d = \mathbf{I}_d - \mathbf{1}_d\mathbf{1}'_d/d$ とおくと, 帰無仮説 $H_0: \mathbf{C}\mathbf{p} = \mathbf{0}$ は次のようになる:

$$\mathbf{C}\mathbf{p} = \begin{pmatrix} 1 - 1/d & -1/d & \dots & -1/d \\ -1/d & 1 - 1/d & \dots & -1/d \\ \vdots & \ddots & \ddots & \vdots \\ -1/d & -1/d & \dots & 1 - 1/d \end{pmatrix} \times \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_d \end{pmatrix} = \begin{pmatrix} p_1 - \bar{p} \\ p_2 - \bar{p} \\ \vdots \\ p_d - \bar{p} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

これより, 帰無仮説は $H_0: p_1 = p_2 = \dots = p_d$ と書き換えられる.

付録 2. 漸近論の正則条件

Emura et al. (2024) の付録 A.3 にある 3 つの条件を要する. 1 つ目は, $N \rightarrow \infty$ のとき, ある定数 $\lambda_i \in (0, 1)$ が存在し,

$$\frac{\mathbf{n}'}{N} = \frac{(n_1, n_2, \dots, n_d)}{N} \longrightarrow \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_d), \quad i = 1, 2, \dots, d,$$

となることであり, 2 つ目は $\pi(\tau) = P(X_{ij} \geq \tau) > 0$ となる $\tau > 0$ が存在すること, 3 つ目は関数 $\phi'_{i\theta_i}(t) = d\phi_{i\theta_i}(t)/dt$ と $\psi'_{i\theta_i}(t) = d\psi_{i\theta_i}(t)/dt$ が $(\pi(\tau), 1)$ 上で有界であることであり, ここに

$\psi_{i\theta_i}(t) = -u\phi'_{i\theta_i}(t)$ である。これら3つは緩やかな条件であり、特に3つ目の条件は、クレイトン・フランク・グンベルコピュラの全てで満足している。

付録3. 結腸がんのデータ解析の感度分析

結腸がんのデータ解析例における感度分析の結果を示す。すなわち、クレイトンコピュラのパラメータを $\theta_i = 0.001, \theta_i = 2, \theta_i = 4, \theta_i = 8$ と変化させたときの解析結果を調べる。図3に従属の強さを変えたときの生存関数のCG推定値を与えた。この結果によると、 θ_i の値が大きくなるほど、再発率の推定値は、どの群でも少しずつ減少していくことが分かる。特に、クレイトンコピュラの下側裾従属性からか、2000日以降の生存時間に大きな群間差がみられる。よって、再発率の推定値は、打ち切りの従属の影響を少なからず受けていることが分かる。一方で、群間比較の検定・推定の結果は θ_i の値の影響を殆ど受けないことが分かる(表4)。この結果より、従属打ち切りは、再発率自体には影響しても、再発率の群間差にはあまり影響を受けないと結論できる。

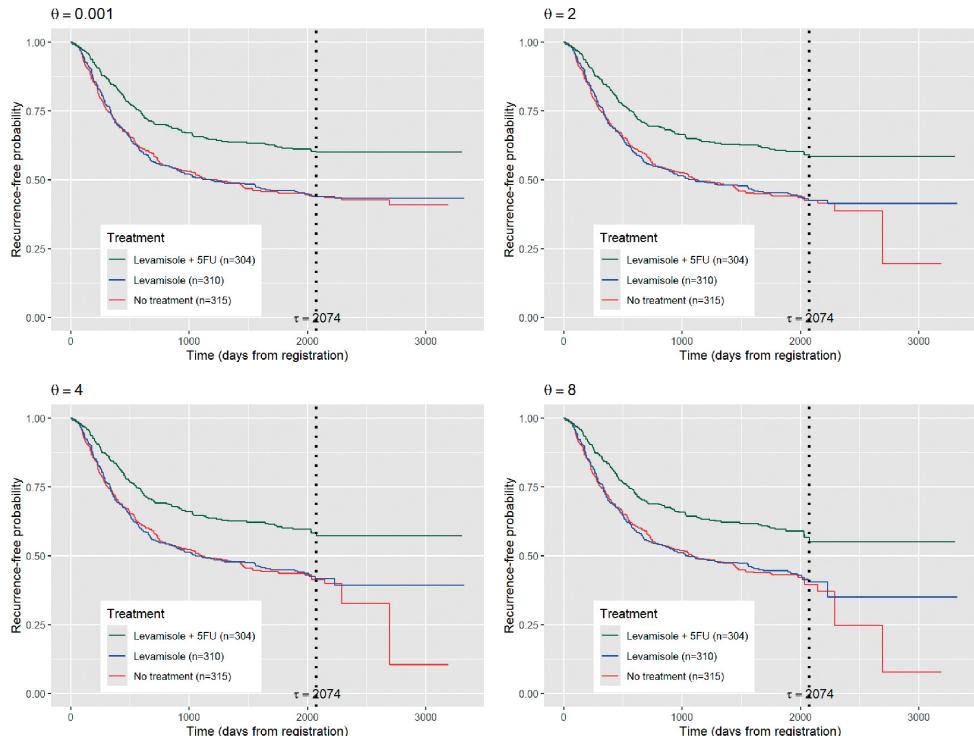


図3. 結腸がんの切除手術を受けた929人の被験者の各治療における無再発確率のCG推定値。クレイトンコピュラのパラメータ θ_i に4つの値を適用。術後経過観察(No adjuvant treatment), レバミゾール投与(Levamisole), レバミゾールと5FUの併用(Levamisole + 5FU)の比較。

表 4. 結腸がんの切除手術を受けた 929 人の被験者の治療効果の差の検定結果. クレイトン
コピュラのパラメータ θ_i に 4 つの値を与え感度分析を行った.

θ_i	比較する群	帰無仮説 H_0	統計量:	P 値:	P 値:
			F_N^{CG}	シミュレーション	解析的
0.001	経過観察とレバミゾール	$p_1 = p_2$	0.003	0.959	0.958
	経過観察と併用	$p_1 = p_3$	18.566	<0.001	<0.001
	レバミゾールと併用	$p_2 = p_3$	18.069	<0.001	<0.001
	3 つ全て	$p_1 = p_2 = p_3$	11.921	<0.001	<0.001
2	経過観察とレバミゾール	$p_1 = p_2$	0.001	0.976	0.971
	経過観察と併用	$p_1 = p_3$	18.037	<0.001	<0.001
	レバミゾールと併用	$p_2 = p_3$	17.672	<0.001	<0.001
	3 つ全て	$p_1 = p_2 = p_3$	11.659	<0.001	<0.001
4	経過観察とレバミゾール	$p_1 = p_2$	0.003	0.951	0.957
	経過観察と併用	$p_1 = p_3$	17.616	<0.001	<0.001
	レバミゾールと併用	$p_2 = p_3$	17.139	<0.001	<0.001
	3 つ全て	$p_1 = p_2 = p_3$	11.381	<0.001	<0.001
8	経過観察とレバミゾール	$p_1 = p_2$	0.011	0.905	0.915
	経過観察と併用	$p_1 = p_3$	16.715	<0.001	<0.001
	レバミゾールと併用	$p_2 = p_3$	15.939	<0.001	<0.001
	3 つ全て	$p_1 = p_2 = p_3$	10.759	<0.001	<0.001

参考文献

- Andersen, P. K. and Perme, M. P. (2010). Pseudo-observations in survival analysis, *Statistical Methods in Medical Research*, **19**(1), 71–99.
- Brunner, E. and Puri, M. L. (2001). Nonparametric methods in factorial designs, *Statistical Papers*, **42**(1), 1–52.
- Brunner, E., Konietzschke, F., Pauly, M. and Puri, M. L. (2017). Rank-based procedures in factorial designs: Hypotheses about non-parametric treatment effects, *Journal of the Royal Statistical Society: Series B*, **79**(5), 1463–1485.
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65**(1), 141–151.
- Collett, D. (2023). *Modelling Survival Data in Medical Research*, 4th ed., Chapman & Hall, Oxfordshire, UK.
- Czado, C. and Van Keilegom, I. (2023). Dependent censoring based on parametric copulas, *Biometrika*, **110**(3), 721–738.
- Das, S., Bhattacharya, R. and Shome, M. (2024). A multi-treatment two stage allocation design for survival outcomes with nonrandom censoring: A copula-based approach, *Sequential Analysis*, **43**(4), 461–476.
- De Uña-Álvarez, J. and Veraverbeke, N. (2013). Generalized copula-graphic estimator, *Test*, **22**(2), 343–360.
- De Uña-Álvarez, J. and Veraverbeke, N. (2017). Copula-graphic estimation with left-truncated and right-censored data, *Statistics*, **51**(2), 387–403.

- Delhelle, M. and Van Keilegom, I. (2025). Copula based dependent censoring in cure models, *Test*, <https://doi.org/10.1007/s11749-024-00961-7>.
- Deresa, N. W. and 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.
- Deresa, N. W. and Van Keilegom, I. (2025). Semiparametric transformation models for survival data with dependent censoring, *Annals of the Institute of Statistical Mathematics*, <https://doi.org/10.1007/s10463-024-00921-w>.
- Deresa, N. W., Van Keilegom, I. and Antonio, K. (2022). Copula-based inference for bivariate survival data with left truncation and dependent censoring, *Insurance: Mathematics and Economics*, **107**, 1–21.
- Derkzen, J. W., Smit, K. C., May, A. M. and Punt, C. J. (2022). Systematic review and non-inferiority meta-analysis of randomised phase II/III trials on S-1-based therapy versus 5-fluorouracil-or capecitabine-based therapy in the treatment of patients with metastatic colorectal cancer, *European Journal of Cancer*, **166**, 73–86.
- Dettoni, R., Marra, G. and Radice, R. (2020). Generalized link-based additive survival models with informative censoring, *Journal of Computational and Graphical Statistics*, **29**(3), 503–512.
- Ditzhaus, M., Dobler, D. and Pauly, M. (2021). Inferring median survival differences in general factorial designs via permutation tests, *Statistical Methods in Medical Research*, **30**(3), 875–891.
- Ditzhaus, M., Genuneit, J., Janssen, A. and Pauly, M. (2023). CASANOVA: Permutation inference in factorial survival designs, *Biometrics*, **79**(3), 203–215.
- Dobler, D. and Pauly, M. (2018). Bootstrap- and permutation-based inference for the Mann-Whitney effect for right-censored and tied data, *Test*, **27**(3), 639–658.
- Dobler, D. and Pauly, M. (2020). Factorial analyses of treatment effects under independent right-censoring, *Statistical Methods in Medical Research*, **29**(2), 325–343.
- Dobler, D., Friedrich, S. and Pauly, M. (2020). Nonparametric MANOVA in meaningful effects, *Annals of the Institute of Statistical Mathematics*, **72**(4), 997–1022.
- Durante, F. and Sempi, C. (2016). *Principles of Copula Theory*, CRC Press, Boca Raton, Florida.
- Efron, B. (1967). The two sample problem with censored data, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol.5.4, 831–853.
- 江村剛志 (2023). 生存時間の2標本問題—コピュラに基づく従属打ち切り問題への対処—, 日本統計学会誌, **52**(2), 295–317.
- 江村剛志 (2025). 『コピュラ理論の基礎』, コロナ社, 東京.
- Emura, T. and Chen, Y. H. (2016). Gene selection for survival data under dependent censoring, a copula-based approach, *Statistical Methods in Medical Research*, **25**(6), 2840–2857.
- Emura, T. and Chen, Y. H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*, Springer, Singapore.
- Emura, T. and 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. and 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), 151–173.
- 江村剛志, 大庭幸治 (2024). 生存時間変数に対する代替性評価—メタアナリシスアプローチ—, 計量生物学, **45**(1), 67–85.
- Emura, T., Matsui, S. and Chen, H. Y. (2019). compound. Cox: Univariate feature selection and compound covariate for predicting survival, *Computer Methods and Programs in Biomedicine*, **168**, 21–37.

- Emura, T., Ditzhaus, M., Dobler, D. and Murotani, K. (2024). Factorial survival analysis for treatment effects under dependent censoring, *Statistical Methods in Medical Research*, **33**(1), 61–79.
- Fizazi, K., Foulon, S., et al. (2022). Abiraterone plus prednisone added to androgen deprivation therapy and docetaxel in de novo metastatic castration-sensitive prostate cancer (PEACE-1): A multicentre, open-label, randomised, phase 3 study with a 2×2 factorial design, *The Lancet*, **399**(10336), 1695–1707.
- Freidlin, B. and Korn, E. (2017). Two-by-Two factorial cancer treatment trials: is sufficient attention being paid to possible interactions?, *Journal of the National Cancer Institute*, **109**(9), djx146.
- Futschik, A., Taus, T. and Zehetmayer, S. (2019). An omnibus test for the global null hypothesis, *Statistical Methods in Medical Research*, **28**(8), 2292–2304.
- Gares, V., Hudson, M., Manuguerra, M. and Gebski, V. (2024). Effect of a correlated competing risk on marginal survival estimation in an accelerated failure time model, *Communications in Statistics—Simulation and Computation*, <https://doi.org/10.1080/03610918.2024.2380005>.
- Gorfine, M., Schlesinger, M. and Hsu, L. (2020). K -sample omnibus non-proportional hazards tests based on right-censored data, *Statistical Methods in Medical Research*, **29**, 2830–2850.
- Gottselig, M., Schneider, S., Demarqui, F. N. (2025). Cure rate regression models for dependent censoring under a copula-based approach, *Japanese Journal of Statistics and Data Science*, <https://doi.org/10.1007/s42081-025-00294-1>.
- Jo, J. H., Gao, Z., Jung, I., et al. (2023). Copula graphic estimation of the survival function with dependent censoring and its application to analysis of pancreatic cancer clinical trial, *Statistical Methods in Medical Research*, **32**(5), 944–962.
- Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York.
- Koziol, J. A. and Jia, Z. (2009). The concordance index C and the Mann-Whitney parameter $\Pr(X > Y)$ with randomly censored data, *Biometrical Journal*, **51**(3), 467–474.
- Li, D., Hu, X. J. and Wang, R. (2023). Evaluating association between two event times with observations subject to informative censoring, *Journal of the American Statistical Association*, **118**(542), 1282–1294.
- Lo, S. M., Wilke, R. A. and Emura, T. (2024). A semiparametric model for the cause-specific hazard under risk proportionality, *Computational Statistics & Data Analysis*, **195**, 107953.
- Lu, Z. and Zhang, W. (2012). Semiparametric likelihood estimation in survival models with informative censoring, *Journal of Multivariate Analysis*, **106**, 187–211.
- 松井茂之 (2023). 要因デザイン, 『臨床試験の事典』(丹後俊郎, 松井茂之 編), 5章 22節, 朝倉書店, 東京.
- 道前洋史 (2023). 競合リスクを伴う左側切断・右側打ち切りデータの解析—現状と今後の展望について, 日本統計学会誌, **52**(2), 203–220.
- 道前洋史 (2025). 相関した競合リスクを伴う左側切断・右側打ち切りデータのペイズ流解析, 統計数理, **73**(2), 167–188.
- Michimae, H. and Emura, T. (2022). Likelihood inference for copula models based on left-truncated and competing risks data from field studies, *Mathematics*, **10**(13), 2163.
- Michimae, H., Emura, T., Miyamoto, A. and Kishi, K. (2024). Bayesian parametric estimation based on left-truncated competing risks data under bivariate Clayton copula models, *Journal of Applied Statistics*, **51**(13), 2690–2708.
- Mishra, P., Singh, U., Pandey, C. M., Mishra, P. and Pandey, G. (2019). Application of student's t-test, analysis of variance and covariance, *Annals of Cardiac Anaesthesia*, **22**(4), 407–411.
- 宮村鍛夫 (1982). 競合モデルに関する統計的手法, オペレーションズ・リサーチ, **27**, 871–876.
- Moertel, C. G., Fleming, T. R., Macdonald, J. S., et al. (1995). Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: A final report, *Annals of Internal*

- Medicine*, **122**(5), 321–326.
- Montgomery, D. C. (2019). *Design and Analysis of Experiments*, John Wiley & Sons, New Jersey.
- Moradian, H., Larocque, D. and Bellavance, F. (2019). Survival forests for data with dependent censoring, *Statistical Methods in Medical Research*, **28**(2), 445–461.
- Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer, New York.
- Nowak, C. P., Mtze, T. and Konietschke, F. (2022). Group sequential methods for the Mann-Whitney parameter, *Statistical Methods in Medical Research*, **31**(10), 2004–2020.
- Perme, M. P. and Manevski, D. (2019). Confidence intervals for the Mann-Whitney test, *Statistical Methods in Medical Research*, **28**(12), 3755–3768.
- Rivest, L. P. and Wells, M. T. (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring, *Journal of Multivariate Analysis*, **79**(1), 138–155.
- 齋藤哲雄, 室谷健太 (2024). がん臨床試験と競合リスク・マルチステートモデル, 計量生物学, **45**(1), 37–65.
- Schneider, S., Demarqui, F. N., Colosimo, E. A. and Mayrink, V. D. (2020). An approach to model clustered survival data with dependent censoring, *Biometrical Journal*, **62**(1), 157–174.
- Schneider, S., Demarqui, F. and de Freitas Costa, E. (2022). Free-ranging dogs' lifetime estimated by an approach for long-term survival data with dependent censoring, *Environmental and Ecological Statistics*, **29**(4), 869–911.
- Simon, F., Ligtvoet, R., Robrecht, S., et al. (2025). Endpoint surrogacy in first-line chronic lymphocytic leukemia, *Journal of Clinical Oncology*, **43**(4), 381–391.
- Staplin, N., Kimber, A., Collett, D. and Roderick, P. (2015). Dependent censoring in piecewise exponential survival models, *Statistical Methods in Medical Research*, **24**(3), 325–341.
- 杉本知之, 田中健太 (2023). 2変量生存時間モデルにおけるコピュラとその利用, 日本統計学会誌, **52**(2), 153–176.
- 武富奈菜美, 山本和嬉 (2023). 生存時間解析・信頼性解析のための統計モデル, 日本統計学会誌, **52**(2), 69–112.
- Tewari, K. S., Sill, M. W., et al. (2014). Improved survival with bevacizumab in advanced cervical cancer, *New England Journal of Medicine*, **370**(8), 734–743.
- 東京大学教養学部統計学教室 (1992). 『自然科学の統計学』, 第3章, 東京大学出版会, 東京.
- Tsiatis, A. A. (1975). Nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Sciences*, **72**, 20–22.
- Xu, J., Ma, J., Connors, M. H. and Brodaty, H. (2018). Proportional hazard model estimation under dependent censoring using copulas and penalized likelihood, *Statistics in Medicine*, **37**(14), 2238–2251.
- Yamada, Y., Denda, T., Gamoh, M., et al. (2018). S-1 and irinotecan plus bevacizumab versus mFOLFOX6 or CapeOX plus bevacizumab as first-line treatment in patients with metastatic colorectal cancer (TRICOLORE): A randomized, open-label, phase III, noninferiority trial, *Annals of Oncology*, **29**(3), 624–631.
- Yeh, C. T., Liao, G. Y. and Emura, T. (2023). Sensitivity analysis for survival prognostic prediction with gene selection: A copula method for dependent censoring, *Biomedicines*, **11**(3), 797.
- Zheng, M. and Klein, J. P. (1994). A self-consistent estimator of marginal survival functions based on dependent competing risk data and an assumed copula, *Communications in Statistics—Theory and Methods*, **23**(8), 2299–2311.
- Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula, *Biometrika*, **82**(1), 127–138.

A Review of Factorial Survival Analysis for Treatment Effects
under Dependent Censoring
—A Copula-based Approach—

Takeshi Emura¹ and Kenta Murotani²

¹School of Informatics and Data Science, Hiroshima University

²School of Medical Technology, Kurume University

For estimating treatment effects on survival, it is not suitable to apply the classical analysis of variance method under the normal distribution model. Nonparametric methods are desirable for defining and estimating treatment effects, which do not impose any parametric model. In this article, we review a nonparametric method for factorial survival analysis under dependent censoring, which utilizes copula-graphic estimators. To implement the method, we introduce an R function, `surv.factorial()`. To investigate the performance of the global and local tests for no treatment effects, we conduct simulation studies and report the results. We also analyze a dataset on colon cancer patients and discuss the consequences.

天文学における生存時間解析：銀河の光度関数 推定

竹内 努^{1,2}

(受付 2024 年 11 月 27 日；改訂 2025 年 3 月 13 日；採択 3 月 14 日)

要　　旨

天文学の観測データは例外なく観測装置の検出限界に対応する切断を受けている。このようなデータから天体の統計量の分布関数を推定する際には、切断データの生存時間解析を用いるのが適切である。しかし、統計学で発展した生存時間解析は天文学分野には 1980 年代までまったく知られておらず、長らく天文学者が独立に考案した推定法が用いられてきた。天文学者による方法の多くは最終的には生存時間解析に帰着するものの数学的に未整理であり、統計的に体系だった議論がなされるようになったのは 21 世紀に入ってからである。本稿では、天文学における打ち切りデータ解析の応用として、銀河の光度関数推定について議論する。まず 1 変数の場合について天文データの特徴を詳しく紹介し、生存時間解析との対応を述べる。次に 2 変数の場合を導入する。2 変数行動関数の推定は天体の標本の取り方(サンプリング)によっては切断に加え打ち切りも入り、一般的な議論は複雑になる。ここでは 2 変数光度関数推定法を最も一般的な形で構築することを試みる。

キーワード：多波長光度、銀河進化、光度関数、生存時間解析、切断、打ち切り。

1. はじめに

1.1 銀河とその形成進化

銀河とは、星と星間物質(ガスと塵の混合流体)、暗黒物質からなる巨大な天体である。その質量は 10^4 – $10^{12} M_\odot$ (M_\odot : 太陽質量) と広い範囲にわたる。そして観測可能な宇宙の範囲に $\sim 10^{11}$ 個もの銀河が存在している。宇宙は 138 億年前に誕生したが、初期の宇宙の物質はほぼ一様に分布しており、銀河のような天体は存在していなかった。つまり、銀河は形成し、現在の姿に時間的に進化してきたのである。すなわち時間進化が銀河の本質であり、これを定量化する銀河形成進化の研究は Tinsley (1980) 以来半世紀近くにわたり銀河研究の中心であり続けている。

単体としての銀河の中では、暗黒物質が形作る重力場の中で星間物質から星が誕生し、そして死滅して持っていた物質の一部は星間空間に戻る。さらに、星は中心部での核融合によって重元素を合成し、その死とともに合成した元素も星間空間に還元する。銀河誕生から連続と繰り返されるこの過程を物質循環といい、その結果星間物質の元素組成は刻々と変化してきた。

¹ 名古屋大学 素粒子宇宙物理学専攻：〒 464–8602 愛知県名古屋市千種区不老町; tsutomu.takeuchi.ttt@gmail.com

² 統計数理研究所 客員：〒 190–8562 東京都立川市緑町 10–3

1.2 天文学観測データの普遍的問題

1.2.1 天文学における「明るさ」

天文学の観測は、その長い歴史の中で常に可能な限り暗い天体を検出する挑戦の連続である。大口径の光学望遠鏡や大規模電波干渉計が建造され、最新の観測データを提供してきた。どのような波長の天文観測データでも、我々観測者から遠ければ天体からの放射は検出限界を下回ってしまい、観測サンプルから落ちてしまう。このことを厳密に理解するため、いくつか重要な物理量を導入する。天体の本来の明るさである光度(luminosity: L)、すなわち天体が単位時間あたりに放射するエネルギーを考える。単位振動数当たりの光度を単色光度(monochromatic luminosity)と呼び、 L_ν と表す。光度は天体が持つ本来の性質であり、観測者に依存しない量である。これに対し、天体から遠く離れた観測者が単位時間あたりに受ける検出器の単位面積当たりのエネルギーを放射流束(flux)と呼ぶ。ここでは放射流束を S で表記する。光度と同じく、単位振動数当たりの放射流束を考えることができる。これを放射流束密度(flux density)と呼び、 S_ν と表記する。放射流束密度は天体からの距離に依存する量で、遠ければ遠いほど値が小さくなる。この性質は遠くの光が暗く見えるという我々の直感と一致する。定量的に理解するため最も単純な定常ユークリッド空間の場合を考える¹⁾。距離を d と置くと

$$S = \frac{L}{4\pi d^2}$$

が成り立つ。すなわち放射流束は距離の 2 乗に反比例する。よって、天文学で「明るい」「暗い」というときには光度の意味なのか放射流束の意味なのかに常に注意を払う必要がある。光度が高いことは天体本来の(intrinsic)性質であるのに対し、放射流束が高いことは光度が高いか距離が近いのかの 2 通りの理由が考えられる。

1.2.2 切断による天文学データ特有の偏り

このことを踏まえ、天文学データの持つ共通の面倒な性質について考える。あらゆる天文学探査データに共通するのが、データは観測装置の検出限界(detection limit)よりも明るい(放射流束が高い)天体しか含まれないことである。すなわち $S > S^{\lim}$ となる天体しかサンプルとして検出されない。振動数 ν での天体の等級(magnitude) m_ν は

$$(1.1) \quad m_\nu \equiv -2.5 \log_{10} S_\nu + \text{定数}$$

と定義される(詳しくは Supplement 第 B 章を参照)²⁾。等級を用いれば、観測データに含まれるのは $m_\nu < m_\nu^{\lim}$ の天体のみである。これが等級選択効果(magnitude selection effect)である。放射流束(あるいは放射流束密度)では、検出限界のあるデータは生存時間解析の左側切断データに対応する。同じデータを等級単位で扱った場合には係数 -2.5 のため右側切断データとなる。いずれにしても、電磁波で観測される天文データは必然的に切断データとならざるを得ない。この切断を図 1 に示す。

1.3 光度関数推定の歴史

銀河の光度関数とは、基本的には銀河の光度の確率密度関数で、慣習的に $\phi(L)$ と表記される。ただし、 n_{gal} を銀河の宇宙空間における個数密度(単位 [個 Mpc^{-3}])と定義すると、多くの天文学的応用において光度関数は

$$\int_L \phi(L) dL = n_{\text{gal}},$$

すなわち全光度範囲で積分すると空間個数密度となるよう規格化される。銀河の光度関数は、銀河系外天文学と観測宇宙論にとって本質的に重要な役割を果たす。まず、銀河の基本的な統

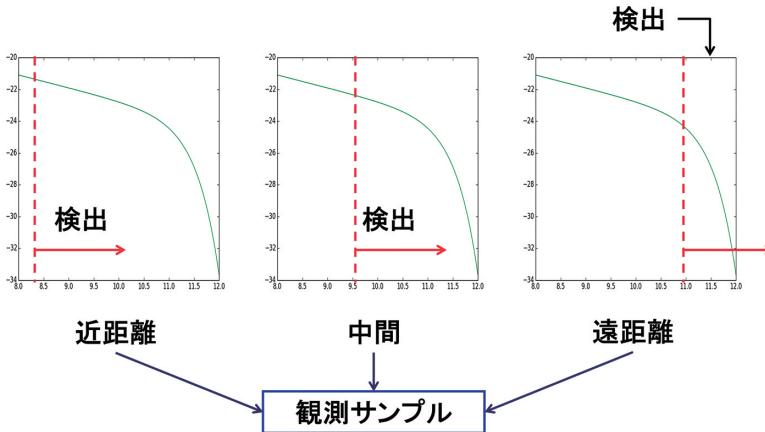


図 1. 銀河の 1 変数光度関数に対する放射流束選択効果. この図は(対数)光度を横軸に, (対数)空間密度を縦軸に取っている.

計的記述の 1 つであり, 色, 形態, あるいは周囲の銀河密度への依存性や, あるいは銀河の相関関数解析³⁾などを通じて銀河の様々な性質を解析するのに用いられている. また光度関数は銀河形成理論の試金石としても重要である (たとえば Takeuchi et al., 2000, およびその参考文献を参照).

第 1.1 節で述べたように, 銀河は時間的に進化する (たとえば Tinsley, 1980). 銀河進化によって, 銀河の光度関数も時間進化してゆく (たとえば Tinsley and Danly, 1980; Cowie et al., 1996; Sawicki et al., 1997; Pascarelle et al., 1998). よって, 銀河の光度関数推定は銀河進化研究の最も重要かつ基本的な統計的推定の一つである. しかし, 観測された銀河カタログからの行動関数推定はまったく自明な作業ではない. 上記のように, 赤方偏移探査の検出限界によってデータは切断されており, これを扱う適切な統計手法が必要となる. 銀河系外天文学の初期段階では, 考えている宇宙空間体積 V 内の銀河の数密度 N/V という古典的推定量が光度関数推定に用いられた (たとえば Hubble, 1936). もちろん, これは詳細な研究には不十分であり, 多くの専門家が様々な方法を提案してきた.

Schmidt (1968) は, クエーサーの統計的研究の中で, 有名な $1/V_{\max}$ 推定量を提唱した. そして Eales (1993) は宇宙年齢による光度関数の進化を調べるために, $1/V_{\max}$ 推定量を進化のあるサンプルに適用するための拡張を提案した. この推定量は, ある銀河 i が検出可能な最大距離を半径とする「最大観測可能体積」 $V_{\max}(i)$ を計算する. 高光度の銀河ほど遠くまで観測できる, すなわち V_{\max} が大きいことからその逆数を重みとして分布関数を求めるのが $1/V_{\max}$ 法である. $1/V_{\max}$ 推定量の詳細については Supplement 第 C 章で述べている. $1/V_{\max}$ 法では, 銀河は最大観測可能体積 V_{\max} のどこに存在していてもよい, すなわち銀河の分布が空間的に均一であると仮定していることが本質的である. しかし現在では銀河が大規模構造内で強くクラスタリングしていることが分かっており, これの仮定は $1/V_{\max}$ 法の短所と見なされている. この欠点にもかかわらず, $1/V_{\max}$ 推定量は現在でも銀河研究に頻繁に使用されている (たとえば Lilly et al., 1996; Ellis et al., 1996). これは主に計算が簡単であるという理由による.

その後, 主として可視光と周辺波長での光度関数を適切に表現する関数が Schechter (1976) によって提示され, これにより光度関数の関数形を仮定したパラメトリック推定の方法が提案されるようになった. Schechter 関数の仮定のもと, Sandage et al. (1979) は最尤法を導入

し, 密度の非一様性による影響を取り除いた光度関数推定法を提案した. 彼らの方法(STY 法)は爆発的に流行し, 現在では光度関数のパラメトリック推定の標準的手法となっている (de Lapparent et al., 1989; Lin et al., 1996). Lin et al. (1999) は光度関数のパラメータ進化を扱う拡張を行った. また Heyl et al. (1997) は光度関数と銀河密度のパラメトリックな同時推定法を開発した. これとは違った方法として Marshall et al. (1983) は, 等級-赤方偏移空間上での天体のポアソン分布を仮定し, 光度関数と進化パラメータの両方を同時に扱うことができる別のパラメトリック推定量を提示した. そして近代的な多波長銀河探査が主流となった現在では, 光度関数推定も多変数への拡張が不可欠となっている. 多波長での光度関数推定はしかし, 複数の観測波長における選択効果が入りこむためより複雑な作業となる. 天文学者はこれまで多波長あるいは一般的の物理量を含む多変数への拡張に挑んできたが, 天文分野においては統計学の方法論の知識は極めて限られていたため, 往々にして提案手法は *ad hoc* であり拡張性に欠けていた (たとえば Chołoniewski, 1985). 系統的な光度関数の多波長化と選択効果の適切な議論は 2000 年代に入り, コピュラ多変数光度関数の導入によって可能となった (Takeuchi, 2010; Takeuchi et al., 2013; Takeuchi and Kono, 2020).

銀河の光度関数の詳細な解析が進むにつれ, 光度関数が銀河環境や色, 形態など実に様々な物理量に依存することが分かってきた (たとえば Binggeli et al., 1988). このような詳細な特徴を記述するには, 解析的なフィッティング式では不十分である. このような背景から, 本稿は光度関数のノンパラメトリック推定に焦点を絞って議論する. パラメトリックな推定法については稿を改めて紹介したい. 天文学的サンプルの偏りの定量評価と銀河の光度関数推定法について, 2011 年までの歴史と応用例の最も完備な解説は Johnston (2011) にある. 本章の内容がより具体的に示されているので, 天文学的内容に興味を持たれた読者は Johnston (2011) を参照されたい.

なお, これらの方法で推定する分布関数としては, 銀河の持つ様々な種類の天体の総質量 M の分布, いわゆる質量関数も重要な研究対象である. 質量は原理的にはある波長での銀河の光度から換算して得られる量であり, 統計的解析方法は光度関数の場合と本質的に同一である. このため, 以降の議論では光度関数と質量関数は区別せず, 光度関数で統一する.

本論文の構成は以下のようになっている. 第 2 章では, Lynden-Bell (1971) の C^- 法と呼ばれる推定法を導出する. 続いて, Woodroffe (1985) による生存時間解析の切断(truncated)データ解析の観点からの再定式化を示す. この章の後半では Woodroffe (1985) の議論と, 2 変数への拡張のための一般化を展開する. 第 3 章では 2 変数光度関数推定問題を議論する. 天体探査データの場合, 2 変数への拡張, すなわちある探査データに別の波長での観測を組み合わせる際, 銀河の放射流束測定に切断に加えて打ち切り(censoring)が生じる. まず伝統的に天文学で用いられてきた 2 変数光度関数推定の手続きについて述べる. 次に van der Laan (1996) によって導入された, 切断データの 2 変数分布関数の最尤推定量を導く. これを踏まえ, 最終的にを目指す天文学観測データの完全な 2 変数光度関数推定法を示す. 第 4 章では, 本論文の内容のまとめを示し, 天文学における光度関数推定で今後検討すべき潜在的課題について言及する. Dabrowska (1988) の与えた 2 変数打ち切りデータにおける分布関数推定法について, 本論文で用いる式とその導出を付録 A に示す.

本論文は天文学・宇宙物理学および生存時間解析の横断的研究であるため, 必要な基礎知識を Supplement にまとめている. Supplement A 章では宇宙論の基礎的な量を紹介する. 天文学で長く用いられている等級という量については混乱しがちな点が多いため, 第 B 章で解説する. 本文では簡単に紹介している Schmidt の $1/V_{\max}$ 推定量については, 第 C 章で正確な定義を示す. 第 D 章では, 打ち切りデータの分布の最尤推定量であるカプラン-マイヤー (Kapkan-Meier) 推定量の導出を示す. 一般に生存関数のノンパラメトリック最尤推定量であ

る積極限推定量の分散は影響関数(influence function)で表せる。影響関数の基礎的説明は第E章で述べる。

明示的には扱わないが、本論文では宇宙年齢の算出などにおいて宇宙論パラメータ $h = H_0/(100 \text{ [km s}^{-1} \text{ Mpc}^{-1}]) = 0.7$, $\Omega_{\Lambda 0} = 0.7$, $\Omega_{M0} = 0.3$, 曲率パラメータ $\Omega_{K0} = 0$ を採用している(Supplement D章参照)⁴⁾。

2. 1 変数光度関数推定

2.1 銀河のクラスタリングと $1/V_{\max}$ 法の限界

既に議論したように、 $1/V_{\max}$ の適用限界を決めている短所は、天体の分布が空間的に一様であるという仮定である。銀河を含むほんやりとした天体である“星雲(nebula)”の天球上での分布が一様ではなくゆらぎを示すという事実は、それらが銀河系外天体であると証明されるずっと前、19世紀からウィリアム・ハーシェル(William Herschel)やシャルル・メシエ(Charles Messier)の研究によって知られていた。Shane and Wirtanen (1967)はパロマーハーモー天文台のシュミット望遠鏡で撮像された写真乾板を用いて19等より明るい銀河の天球上での地図を作製し、銀河の広範囲にわたるクラスタリングを初めて明らかにした。このような研究を通じて、銀河の空間分布における銀河群、銀河団、超銀河団⁵⁾といった銀河分布の階層構造が明らかになった。そして de Lapparent et al. (1986) 以降、宇宙論的赤方偏移を距離として用い、銀河の3次元空間分布を直接探査する赤方偏移探査(redshift survey)プロジェクトが大規模におこなわれ、銀河の3次元クラスタリングの性質は精密に測定されている。銀河分布の非一様性を無視して $1/V_{\max}$ 法を用いると、光度関数推定に顕著なバイアスが生じることがよく知られている。たとえば Takeuchi et al. (2000) はこのバイアスが極めて深刻になり得ることを数値実験によって明確に示した。クラスタリングは $1/V_{\max}$ の枠組みでは極めて厄介な問題だが、意外にも宇宙物理学においてこの問題を回避できる代替手法が開発されるまでにさほど時間はかからなかつた。これが Lynden-Bell (1971) によって導入された“ C^- 法”である。

2.2 Lynden-Bell の C^- 法

Lynden-Bell (1971) は統計学分野での生存時間解析の存在を知らないまま、天文学分野で独立にその一種である切断データ解析の方法を構築した。彼はこの方法を Schmidt (1967) のクエーサー探査データに適用してクラスタリングの影響を受けない光度関数を求めるに成功した。近代的な視点から見ると、Lynden-Bell の方法は切断(truncated)データに対する生存時間解析の応用例となっている。統計分野では、Woodroffe (1985) による Lynden-Bell の方法の紹介以降研究が活発になり、議論が大きく発展した。本章ではオリジナルの議論と生存時間解析からの再構成を詳しく紹介する。なお、天文学的表記に不慣れな読者は本節(第2.2節)の式(2.1)–(2.6)の議論は飛ばして次節(第2.3節)に進まれてもよい。

まずいくつかの基本的な量を定義する。銀河の絶対等級を M 、見かけの等級を m 、赤方偏移 z に対応する光度距離を $d_L(z)$ (単位 [Mpc]) とする。

$$M = m - 5 \log d_L(z) - 25$$

である。また、銀河の光度距離は非常に大きな範囲を取るため、その対数を用いた距離の指標である距離指数(distance modulus)

$$D(z) \equiv m - M = 5 \log d_L(z) + 25$$

を導入する。赤方偏移 z と距離指標 $D(z)$ は等価な情報であり、往々にして引数 z は省略して

単に D と表記する。距離指数の単位は等級 [mag] である(詳細は Supplement 第 B 章を参照)。絶対等級で表現した光度関数を $\phi(M)$ [$\text{Mpc}^{-3} \text{mag}^{-1}$] と書く。また N_{obs} を考えている銀河探査で検出された銀河の総数とする。 C^- 法は積分分布関数を Dirac の δ 関数 δ_D の重み付き和として表現する。すなわち、等級の測定誤差は仮定しない。見かけの等級の誤差を考慮するためには、たとえば等級に対してスムージングカーネルを導入することもできる(例えば Caditz and Petrosian, 1993; SubbaRao et al., 1996)。

仮想的に観測装置の検出限界に制約されない状況を考えると、Hubble (1936) で用いられた初期の議論そのものに帰着し、積分光度関数 $\Phi(M)$ の構築は容易である。しかし、現実では検出限界による見かけの等級 m への制約により、現実には偏りのある部分サンプルのみが観測される。具体的には、みかけの等級が限界等級 m^{lim} よりも明るい銀河のみがサンプルに含まれる。これを $X(M)$ と表記すると、

$$(2.1) \quad \frac{d\Phi(M)}{\Phi(M)} > \frac{dX(M)}{X(M)}$$

となる。後述のように、限界等級 m^{lim} への依存性は絶対等級 M を通じて $X(M)$ に入っている。

次に、観測された銀河の分布を絶対等級 M と距離指標 D の平面 $M-D$ で考え、 M と D は変数分離可能と仮定する(図 2)。図 2において、探査の検出限界 m^{lim} は斜めの線で示されている。観測された母集団での M と D の確率密度関数を次のように書くことができる。

$$dP = \rho(D)dD \phi(M)dM \Theta(m^{\text{lim}} - m),$$

すなわち

$$\frac{d^2P}{dDdM} = \rho(D)\phi(M)\Theta(m^{\text{lim}} - m).$$

ここで $\Theta(m^{\text{lim}} - m)$ は等級による選択効果(打ち切り)を記述する Heaviside 関数で、

$$\Theta(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

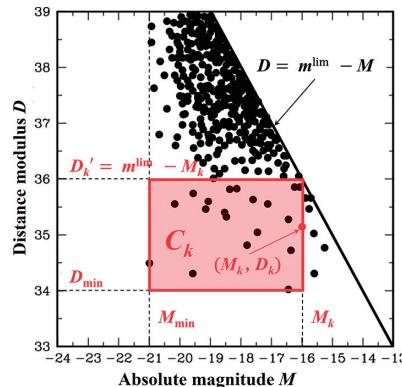


図 2. Lynden-Bell (1971) によって導入された C^- 法の概念図。黒丸シンボルがひとつひとつつの銀河を表す。サンプル内の各銀河 (M_k, D_k) に対して領域 C_k を定義する。この領域内において、 (M_k, D_k) 以外の全ての銀河をカウントし、計数を $C^-(M_k)$ と表す。 M_{min} はサーベイデータ中で最も明るい銀河によって決定される。同様に、 D_{min} は最も近い赤方偏移 z_{min} の銀河によって定義される最小距離指数である。

と定義される。

観測された $X(M)$ の部分領域 $C^-(M)$ を

$$\frac{d\Phi}{\Phi} = \frac{dX}{C^-}$$

となるように取れば、累積光度関数(累積分布関数)はこれを積分することにより

$$\Phi(M) = A \exp \left[\int_{-\infty}^M \frac{dX(M')}{C^-(M')} \right]$$

という形で求められる。ここで A は規格化定数、 $X(M)$ は $C^-(M)$ に存在する部分サンプルである。

しかし、実際の解析では微分光度関数(確率密度関数に対応)および空間数密度が必要であり、これはそれぞれ Dirac の δ 関数 δ_D の系列として表すことができる。実際に観測されたサンプルを $\{(M_i, D_i)\}$ ($i = 1, \dots, N$) とする。データは明るい順(絶対等級 M_i の数値が小さい順)にソートされているとする。光度関数と空間密度はそれぞれ次のように表される。

$$(2.2) \quad \phi(M) = \sum_{i=1}^N \psi_i \delta_D(M - M_i),$$

$$(2.3) \quad \rho(D) = \sum_{i=1}^N \rho_i \delta_D(D - D_i),$$

ここで、 ψ_i と ρ_i はそれぞれのステップサイズ、すなわちサンプル i と $i+1$ の積分光度関数および空間数密度の差である。次に、積分光度関数を構築するために以下の量を定義する。得られた観測サンプルに含まれるある銀河 k のデータを (M_k, D_k) とし、この銀河が仮想的に検出される最大の距離指標

$$D'_k = m^{\lim} - M_k$$

を考える。関数 $C^-(M_k)$ を、図 2 の影つきの領域で示された領域 C_k

$$\begin{cases} M_{\min} \leq M_i < M_k, \\ D_{\min} \leq D_i \leq D'_k \end{cases}$$

内に存在する銀河の数として定義する。

$$C^-(M_k) \equiv \#\{(M_i, D_i) \in C_k\} \quad (k = 1, \dots, N).$$

Jackson (1974) によると、上付きの $-$ は $C^-(M_k)$ を評価するときに (M_k, D_k) の点が含まれないことを強調するためのものである。光度関数の係数は次の漸化式から決定される。

$$(2.4) \quad \psi_{k+1} = \psi_k \frac{C^-(M_k) + 1}{C^-(M_{k+1})}.$$

以上より、積分光度関数は

$$(2.5) \quad \Phi(M_k) = \int_{M_{\min}}^{M_k} \phi(M) dM = \psi_1 \prod_{k: M_k < M} \frac{C^-(M_k) + 1}{C^-(M_k)}$$

によって推定できる。これが Lynden-Bell の公式と呼ばれる推定量である。ここで、式(2.5)に

おいて、

$$(2.6) \quad \frac{C^-(M_1) + 1}{C^-(M_1)} = 1$$

とし、積が $k = 2$ から始まるように設定することを注意しておく (Chołoniewski, 1987; Takeuchi et al., 2000).

Lynden-Bell の C^- 法は、データセット内の天体の空間分布に関する仮定を必要としない点において従来の $1/V_{\max}$ 法よりも本質的に優れている。天文学者の考案したその他のノンパラメトリックな光度関数推定法は、本質的には全て各ビン当たりの銀河数を $\{0, 1\}$ にした極限を取れば C^- 法に帰着する (たとえば Takeuchi et al., 2000; Johnston, 2011, および参考文献)。

Chołoniewski (1987) はオリジナルの C^- 法を再検討し、難解な原論文を簡略化しただけでなく、光度関数を適切に規格化して銀河の密度進化を同時に推定できるよう改良した。Chołoniewski の方法の証明は Takeuchi et al. (2000) によって与えられた。しかしこの改良版 C^- 法はこれまでのところ、応用上はごく希にしか用いられていない (たとえば Takeuchi et al., 2000, 2006)。

2.3 Woodrooffe (1985) による C^- 法の再定式化

Woodrooffe (1985) は Lynden-Bell の C^- 法を生存時間解析の立場から再構成し、推定量の漸近的性質を導いた。これにより、Lynden-Bell (1971) の業績は Woodrooffe によって統計学分野に知られることとなった。本節では Woodrooffe の定式化について述べる。

仮想的な母集団サンプル $(X_1, Y_1), \dots, (X_N, Y_N)$ を考える。 X と Y の分布関数をそれぞれ $X \sim F$, $Y \sim G$ とし、 X と Y は独立であるという基本的仮定を置く。サンプル $i (i \leq N)$ が観測されるのは次の条件

$$(2.7) \quad Y_i \leq X_i$$

が満たされる場合に限られるとする。観測されたサンプルのデータの集合を $(x_1, y_1), \dots, (x_n, y_n)$ と表し、 $\{x_i\} (i = 1, \dots, n)$ および $\{y_j\} (j = 1, \dots, n)$ の経験分布関数を、それぞれ F_n^* と G_n^* とする。すなわち、 $0 \leq t < \infty$ について次のように定義する。

$$(2.8) \quad F_n^*(t) = \mathbb{P}(X \leq t | Y \leq X) = \frac{1}{n} \# \{i \leq n : x_i \leq t\},$$

$$(2.9) \quad G_n^*(t) = \mathbb{P}(Y \leq t) = \frac{1}{n} \# \{j \leq n : y_j \leq t\}.$$

これらの量の直感的意味を図 3 に示す。 F_n^* および G_n^* は条件 (2.7) の下での分布関数 F^* および G^* の標本推定量となる (すなわち $n \rightarrow \infty$ において $F_n^* \rightarrow F^*$, $G_n^* \rightarrow G^*$)。

ここで次の量

$$C_n(t) \equiv G_n^*(t) - F_n^*(t-) \quad (0 \leq t < \infty)$$

を定義する。すると $\forall i \leq n$ について

$$C_n(x_i) \geq \frac{1}{n}$$

となる。分布 F の累積ハザード関数は次で推定される。

$$(2.10) \quad \hat{\Lambda}_n(t) \equiv \int_0^t \frac{dF_n^*(x)}{C_n(x)} = \sum_{i: x_i \leq t} \frac{1}{n C_n(x_i)} \quad (0 \leq t < \infty).$$

式 (2.10) を用いると、分布関数 F の標本推定量 $\hat{F}_n(t)$ は次のようにになる。

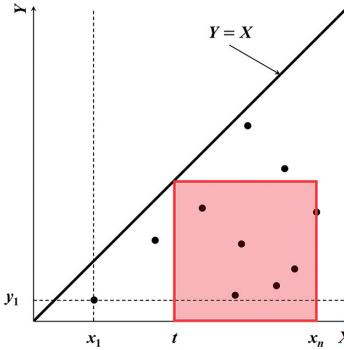


図 3. Woodroffe (1985) の方法におけるデータ構造の概略図。

$$(2.11) \quad \hat{F}_n(t) = 1 - \prod_{i: x_i \leq t} \left[1 - \frac{r(x_i)}{nC_n(x_i)} \right] \quad (0 \leq t < \infty),$$

ここで $r(x_i) \equiv \#\{k \leq n : x_k = x_i\}$ である。また

$$(2.12) \quad \hat{G}_n(t) = \prod_{j: y_j > t} \left[1 - \frac{s(y_j)}{nC_n(y_j)} \right] \quad (0 \leq t < \infty)$$

が成り立つ。ただし $s(y_j) \equiv \#\{k \leq n : y_k = y_j\}$ 。

さらに

$$(2.13) \quad \alpha \equiv \mathbb{P}\{Y \leq X\}$$

の最尤推定量は

$$(2.14) \quad \hat{\alpha}_n = \int_0^\infty \hat{G}_n d\hat{F}_n$$

で与えられる。この $\hat{\alpha}_n$ を用いて、母集団サイズ N の推定量は

$$(2.15) \quad \hat{N}_n \equiv \frac{n}{\hat{\alpha}_n}$$

のように求められる。

データに同値(タイ)がない場合、次が成り立つ。

$$(2.16) \quad \hat{G}_n(t) = \prod_{j: y_j > t} \left[\frac{nC_n(y_j) - 1}{nC_n(y_j)} \right] \equiv \prod_{j: y_j > t} \left(\frac{n_j - 1}{n_j} \right).$$

天文学における実践では同値が生じることは基本的にないため、式(2.16)の形が最尤推定量として用いられている。これは Woodroffe の公式、あるいは Lynden-Bell-Woodroffe の公式と呼ばれる。 \hat{G}_n の漸近分散推定量は

$$(2.17) \quad \hat{\mathbb{V}}[\hat{G}_n(t)] = \hat{G}_n(t)^2 \sum_{j: y_j > t} \frac{1}{n_j(n_j - 1)}$$

で与えられる (Wang et al., 1986)。式(2.14), (2.15), (2.16)は最尤推定量であることが示されている (Woodroffe, 1985)。また収束性、一致性、および漸近正規性の証明は Woodroffe (1985), Wang et al. (1986), Keiding and Gill (1990), および van der Vaart (1991) によって

与えられている。

ここで Lynden-Bell (1971) と Woodroffe (1985) の光度関数推定量の関係について考察する。 Lynden-Bell の公式(2.5)の積の部分は次のように表される。

$$(2.18) \quad \prod_{k:M_k \leq M} \frac{C^-(M_k) + 1}{C^-(M_k)} = \frac{1}{\prod_{k:M_k > M} \frac{C^-(M_k) + 1}{C^-(M_k)}} = \prod_{k:M_k > M} \frac{C_k^-(M)}{C^-(M_k) + 1}.$$

ここで $n_k = C^-(M_k) + 1$ と置くと, Woodroffe の公式(2.16)が得られる。

この関係は視覚的に理解できる。極限等級 m^{\lim} より明るい銀河が実際の観測サンプルに含まれることを想起する。検出の境界に対応する絶対等級 M^{\lim} は次のように表される。

$$(2.19) \quad M^{\lim}(D) = -D + m^{\lim}.$$

等級の定義 [式(1.1)] により, 光度が高い銀河ほど絶対等級 M は小さくなる。したがって, 図 2 の横軸は銀河の光度の順序としては反転していることになる。

ここで新しい変数 W を

$$W \equiv -(M - m^{\lim})$$

と定義すると, 限界等級を表す線[式(2.19)]は

$$D = W$$

と表される。この変換によって $(W, D) = (X, Y)$ とみなせ, Woodroffe (1985) の定式化が得られることが分かる。

2.4 2変数推定量への拡張のための一般的定式化

ここでは, 推定量の 2変数への拡張を目指して, 上記の議論をより一般的な形で記述する。確率変数 X と Y を考え, 分布関数を $X \sim F$ および $Y \sim G$ とする。前節と同様, X と Y は独立と仮定する。それぞれの経験分布関数を前節同様 F_n^*, G_n^* とし, その極限形式を F^*, G^* とおく。なお, 前節の Woodroffe (1985) の議論では絶対等級の符号を変えることで左側切断データとして扱ったが, 天文学データでは混乱のもととなるためこれ以降は符号は変えず, 右側切断データとして取り扱う。即ち $X \leq Y$ のときにサンプルとして観測されるとする。 X_i および Y_i ($i = 1, \dots, n$) を観測したとき, 生存関数 $S(x)$,

$$(2.20) \quad S(x) \equiv 1 - F(x)$$

のノンパラメトリック最尤推定量(NPMLE)は積極限推定量(product limit estimator)として次の式(2.21)で定式化される。

$$(2.21) \quad S_{\text{PL}}(x) \equiv \prod_{(0, x]} [1 - \Lambda_n(ds)].$$

ここで,

$$(2.22) \quad \Lambda_n(ds) = \frac{\sum_{i=1}^n \mathbb{1}(X_i \in ds)}{\sum_{i=1}^n \mathbb{1}(X_i \geq s, Y_i \geq s)}$$

はハザード確率

$$(2.23) \quad \Lambda(ds) \equiv \mathbb{P}(X \in ds | X \geq s, X \leq Y)$$

の推定量である。積極限推定量は, 漸近線型かつ有効推定量であることが示されている。漸近

線型性とは

$$(2.24) \quad \sqrt{n}[S_{\text{PL}}(x) - S(x)] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbb{IF}_{\text{PL}}(X_i, Y_i | F, G, x) + o_p(1)$$

を意味する。ここで、 $\mathbb{IF}_{\text{PL}}(X_i, Y_i | F, G, x)$ は推定量の影響関数である。積極限推定量の影響関数は次で与えられる。

$$(2.25) \quad \begin{aligned} \mathbb{IF}_{\text{PL}}(X_i, Y_i | F, G, x) \\ = -S(x) \left[\frac{\mathbb{1}(X \leq s | X \leq Y)}{(G^* - F^*)(X)} - \alpha \int_0^{x \wedge X} \frac{F(\text{d}s)}{S^2(s)G(s)} \alpha \int_0^{c \wedge Y} \frac{F(\text{d}s)}{S^2(s)G(s)} \right]. \end{aligned}$$

(Bickel et al., 1993)。ここで α は前節で定義した式(2.13)であり、

$$(2.26) \quad (G^* - F^*)(s) = \mathbb{P}(X \geq s, Y \leq s | X \leq Y)$$

である。また $a \wedge b \equiv \min\{a, b\}$ と表記している。したがって、 \hat{S}_{PL} の分散は式(2.24)を用いて評価できる。一般に、NPMLE の分散は影響関数を用いることで評価可能である。影響関数の基本的事項は Supplement 第 E 章で述べている。

さらに、右側打ち切りデータが存在する場合について議論する。具体的には、 Y^* を右側打ち切り変数、 X を関心のある変数とし、 $\Delta \equiv \mathbb{1}(X \leq Y^*)$ とすると、条件 $X \leq Y$ のもとで $(\tilde{X} = X \wedge Y^*, Y, \Delta)$ を観測する。この場合の S の積極限推定量は次で与えられるハザード関数推定量を式(2.21)に代入することで得られる。

$$(2.27) \quad \Lambda_n(\text{d}s) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_i \in \text{d}s, \Delta_i = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_i \geq s, Y_i \geq s)},$$

ここで $\Delta_i = \mathbb{1}(X_i \leq Y_i^*)$ である。上記の影響関数は、この場合にも拡張可能である。

3. 2 変数光度関数推定

3.1 2 変数データにおける選択効果

現在、大規模天文サーベイはすべて多波長(マルチバンド: multiband)で行われている。ここでは 2 変量の場合(すなわち、2 つのバンドで選択されたサンプル)を考え、その 2 変数光度関数を $\phi^{(2)}(M_1, M_2)$ で表す。より多くのバンドの場合(あるいは、より一般的に任意の物理量で選択される場合)への拡張は直接的である。放射流束によるセレクションは、光度-光度($L_1 - L_2$) 平面上に下限 L^{lim} を設定することとして記述される(図 4 を参照)。同様に、絶対等級を用いれば $M_1 - M_2$ 平面上に上限 M^{lim} を設定することとして表される。上限絶対等級 M^{lim} は、距離指數の関数としての限界等級 m^{lim} によって定義される[式(2.19)]。一般的なサーベイでは、ある特定の波長を主選択バンドと設定する。例えば B -バンド(可視光)、 Ks -バンド(近赤外)、 $60 \mu\text{m}$ (遠赤外)などである。

バンド 1 で天体(この場合は銀河)のサンプルを選択すると、距離指數 D において $M_1 \geq M^{\text{lim}}_1(D)$ である天体はサンプルに含まれない。したがって、検出される天体は $M_1 < M^{\text{lim}}_1(D)$ かつ $M_2 < M^{\text{lim}}_2(D)$ である必要がある。したがって、 $M_1 - M_2$ 平面上での検出された天体の 2 次元分布

$$(3.1) \quad \begin{aligned} \Sigma^{\text{det}}(M_1, M_2, D) \\ \equiv \int_0^D \frac{\text{d}^2 V}{\text{d}D' \text{d}\Omega} \phi^{(2)}(M_1, M_2) \{1 - \Theta[M^{\text{lim}}_1(D')]\} \{1 - \Theta[M^{\text{lim}}_2(D')]\} \text{d}D' \end{aligned}$$

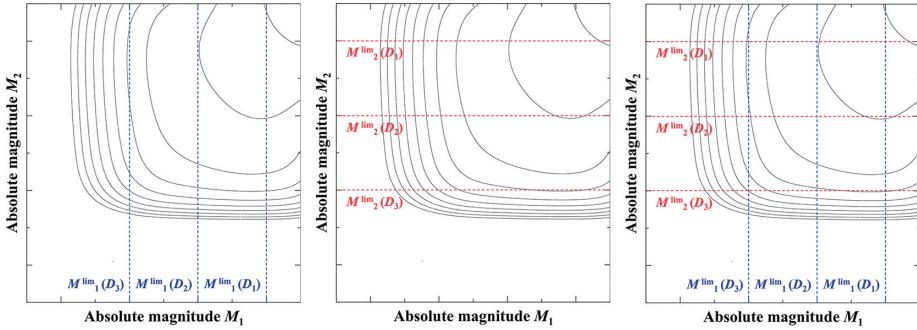


図4. 2バンドサーベイにおける選択効果の概略図。等高線は銀河の2変数光度関数のモデルを表す。ここでは銀河の光度は絶対等級で表している。 M_1 , M_2 がそれぞれバンド1, 2での絶対等級, $M^{\lim 1}(D)$, $M^{\lim 2}(D)$ が距離指数(D)でのそれぞれのバンドにおける検出限界光度を表す。左: バンド1に基づくサンプルセレクションの場合、中央: バンド2に基づくサンプルセレクションの場合、右: バンド1および2の2つのセレクションがかかる場合をそれぞれ示している。バンド1では垂直線よりも左側、バンド2では水平線よりも下側にある銀河が検出される。

のように表される。ここで Ω は立体角である。 Σ^{\det} は $M_1 - M_2$ 平面上で両バンドで検出された天体の面密度に比例する量である。

バンド1を主選択バンドとすると、バンド1で検出されているがバンド2では検出されない天体が存在することになる。

この場合、これらの天体については見かけの等級の下限値(放射流束の上限値)のみが得られる。バンド2における等級の下限値の2次元分布も同様に次のように定式化される。

$$(3.2) \quad \Sigma^{\text{LL2}}(M_1, M_2, D) \equiv \int_0^D \frac{d^2V}{dD'd\Omega} \phi^{(2)}(M_1, M_2) \{1 - \Theta[M^{\lim 1}(D')] \} \Theta[M^{\lim 2}(D')] dD'.$$

上付き添字 LL2 は‘バンド2における下限値(lower limit at band 2)’を表す。このように、天体の存在は確定しているが特定の量の上限(または下限)のみが得られる場合は統計学では‘打ち切り’(censored)と呼ばれる。式(3.2)で分布 $\Sigma^{\text{LL2}}(M_1, M_2)$ を定義できるが、このカテゴリに属するサンプル天体はプロット上で下限値としてのみ現れる。一方、バンド1で天体を選択しているため、バンド1には下限値は存在しない。下限値以上の等級に天体が存在するかどうかはわからないからである。この場合は統計学の‘切断(truncated)’に対応する。

バンド2で天体を選択する場合も、バンド1で選択されたサンプルと全く同様に、検出された天体と下限値の2次元分布を定式化できる。この場合も両バンドで検出された天体の2次元分布は式(3.1)で表される。バンド2で検出され、バンド1では検出されない天体は

$$(3.3) \quad \Sigma^{\text{LL1}}(M_1, M_2, D) \equiv \int_0^D \frac{d^2V}{dD'd\Omega} \phi^{(2)}(M_1, M_2) \Theta[M^{\lim 1}(D')] \{1 - \Theta[M^{\lim 2}(D')] \} dD'$$

のように表される。天文学の多波長サーベイでは、様々な D におけるサンプルが同時に含まれたデータが得られる。このデータから2変数光度関数を推定する問題は複雑だが、2変数生存時間分析の手法を用いることで統一的に扱うことができる。

3.2 生存時間解析の方法による2変数光度関数の推定

3.2.1 天文学で用いられてきた手法

2変量光度関数の推定は、長らく天文学者の関心を集めてきた重要な問題である。にも拘ら

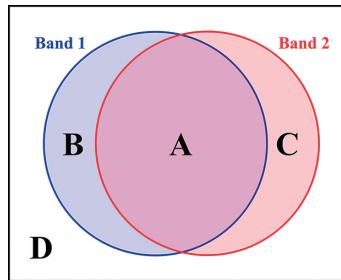


図 5. 2 変数光度選択でのデータのカテゴリを示すヴェン図。それぞれ、領域 A: バンド 1 と 2 の両方で検出された天体、領域 B: バンド 1 で検出されるがバンド 2 では検出されていない天体、領域 C: バンド 2 で検出されるがバンド 1 では検出されていない天体、領域 D: バンド 1 と 2 のどちらでも検出されていない天体に対応する。バンド 1 で選択されたサンプルは A と B の和集合に対応し、バンド 2 で選択されたサンプルは A と C の和集合に対応する。

ず、実際にはこの問題におけるデータと推定量の構造は天文学および天体物理学では適切に理解されてこなかった。天文学における膨大な数の論文が‘2 変数分布関数’に言及しているが、データが 1 つの波長バンドで選択されたと暗黙のうちに仮定している。つまり、天文学で‘2 変数分布関数推定’として知られるほぼすべての手法は、主選択バンドで測定された観測量の周辺分布が完全に既知であるか推定可能であり、上限値または下限値を含む第 2 の観測量がある場合の問題について扱っている (e.g., Mobasher et al., 1993; Keres et al., 2003; Wall and Jenkins, 2003)。

しかし、図 4 からも理解できるように、これは 2 変数分布推定のうちの 1 つの特殊な場合に過ぎない。図 5 は、2 つの波長バンド(バンド 1 とバンド 2)で選択されたサンプルのカテゴリを示している。

バンド 1 で選択されたサンプルのバンド 2 光度関数は通常次のように推定される。

- (1) 何らかの推定法、たとえば C^{-} 法を用いてバンド 1 の 1 変数光度関数を推定する。
- (2) サンプルのバンド 1 光度を bin に分割する。
- (3) 各バンド 1 光度 bin について、打ち切りデータ解析(例: Kaplan-Meier 法: Kaplan and Meier 1958)を用いてバンド 2 の 1 変数光度関数を推定する。
- (4) バンド 1 光度関数と条件付きバンド 2 光度関数を組み合わせ、最終的な 2 変数光度関数を構築する。

この方法の応用例としては、例えば Buat et al. (2009) がある。Kaplan-Meier 推定量については Supplement 第 D 章で説明されている。ここで注意すべき点は、この方法で得られる 2 変数光度関数は、天文学者が本来求めているものとは一致しないということである。この混乱は、いくつかの天文学文献において銀河の母集団の物理的特性について偏った結論をもたらした。以下ではこの問題を明確に定義し、より適切に扱うための方法について議論する。

3.2.2 切断(truncated)データの 2 変数生存時間分析

まず van der Laan (1996) の提案した、2 バンド切断データの分布関数推定法を構築する。van der Laan (1996) は最尤法のスコア方程式の解として推定量を導出したが、ここでは Huang et al. (2001) に従い、構成的に推定量を求める。このために第 2.4 節の議論を直接拡張すると、2 変数切断データは以下のようないわゆるモデルで表現できる。解析対象の確率変数を (X_1, X_2) 、切断を決め

る確率変数を (Y_1, Y_2) とする. それぞれの分布関数を $(X_1, X_2) \sim F(x_1, x_2)$, $(Y_1, Y_2) \sim G(y_1, y_2)$ と置く. ここでも (X_1, X_2) と (Y_1, Y_2) は独立と仮定する.

第2.4節の議論と同様, データには $X_1 \leq Y_1$ かつ $X_2 \leq Y_2$ の場合のみが含まれる. 観測された (X_1, X_2) および (Y_1, Y_2) の同時分布は次で与えられる.

$$(3.4) \quad \mathbb{P}(X_1 \leq x_1, Y_2 \leq x_2, Y_1 \leq y_1, Y_2 \leq y_2 | X_1 \leq Y_1, X_2 \leq Y_2).$$

$x_1 \geq 0$ および $x_2 \geq 0$ が固定されているとする. 観測された (X_1, X_2) の2変数分布を次のように考える.

$$(3.5) \quad F^*(x_1, x_2) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2 | X_1 \leq Y_1, X_2 \leq Y_2),$$

観測された(換算)絶対等級 (y_1, y_2) の分布関数を $G^*(x_1, x_2)$ とすると, 次が成り立つ.

$$\begin{aligned} (3.6) \quad G^*(x_1, x_2) &= \mathbb{P}(Y_1 \leq x_1, Y_2 \leq x_2 | X_1 \leq Y_1, X_2 \leq Y_2) \\ &= \frac{1}{\alpha^{(2)}} \mathbb{P}[(Y_1 \leq x_1, Y_2 \leq x_2), (X_1 \leq Y_1, X_2 \leq Y_2)] \\ &= \frac{1}{\alpha^{(2)}} \int_{x_1}^{\infty} \int_{x_2}^{\infty} F(y_1, y_2) dG(y_1, y_2). \end{aligned}$$

したがって,

$$(3.7) \quad dG^*(x_1, x_2) = \frac{1}{\alpha^{(2)}} F(x_1, x_2) dG(y_1, y_2)$$

が成り立つことより, 式(3.5)は次のように書き換えられる.

$$\begin{aligned} (3.8) \quad F^*(x_1, x_2) &= \frac{1}{\alpha^{(2)}} \int \mathbb{P}(X_1 \leq x_1 \wedge y_1, X_2 \leq x_2 \wedge y_2) dG(y_1, y_2) \\ &= \int \frac{F(x_1 \wedge y_1, x_2 \wedge y_2)}{F(y_1, y_2)} dG^*(y_1, y_2) \\ &= G^*(x_1, x_2) + \Xi_1^*(x_1, x_2) + \Xi_2^*(x_1, x_2) + F(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{dG^*(y_1, y_2)}{F(y_1, y_2)}, \end{aligned}$$

ここで

$$\begin{aligned} \Xi_1^*(x_1, x_2) &= \int_{x_1}^{\infty} \int_0^{x_2} \frac{dG^*(y_1, y_2)}{F(y_1, y_2)} \\ &= \mathbb{P}(Y_1 > x_1, X_1 \leq x_1, Y_2 \leq x_2 | X_1 \leq Y_1, X_2 \leq Y_2), \end{aligned}$$

および

$$\begin{aligned} \Xi_2^*(x_1, x_2) &= \int_0^{x_1} \int_{x_2}^{\infty} \frac{dG^*(y_1, y_2)}{F(y_1, y_2)} \\ &= \mathbb{P}(Y_2 > x_2, X_2 \leq x_2, Y_1 \leq x_1 | X_1 \leq Y_1, X_2 \leq Y_2) \end{aligned}$$

である. 式(3.8)を評価するため, 次の量を定義する.

$$(3.9) \quad K^*(x_1, x_2) \equiv F^*(x_1, x_2) - G^*(x_1, x_2) - \Xi_1^*(x_1, x_2) - \Xi_2^*(x_1, x_2).$$

(X_1, X_2) および (Y_1, Y_2) が独立であるという仮定のもと, 式(3.8)を用いると次が得られる.

$$\begin{aligned}
(3.10) \quad K^*(X_1, X_2) &= \mathbb{P}(X_1 \leq x_1, x_1 < Y_1, X_2 \leq x_2, x_1 < Y_1 | X_1 \leq Y_1, X_2 \leq Y_2) \\
&= \frac{1}{\alpha^{(2)}} \mathbb{P}(X_1 \leq x_1, x_1 < Y_1, X_2 \leq x_2, x_2 < Y_2) \\
&= \frac{1}{\alpha^{(2)}} F(x_1, x_2) \mathbb{P}(x_1 < Y_1, x_2 < Y_2) \\
&= \frac{1}{\alpha^{(2)}} F(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} dG(y_1, y_2).
\end{aligned}$$

式(3.10)と式(3.7)を組み合わせると

$$(3.11) \quad K^*(x_1, x_2) = F(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{dG(y_1, y_2)}{F(y_1, y_2)}$$

が得られる。 $(X_{1,i}, X_{2,i}, Y_{1,i}, Y_{2,i})$ ($i = 1, \dots, n$) を、条件 $X_{1,i} \leq Y_{1,i}$ および $X_{2,i} \leq Y_{2,i}$ を満たす観測データとする。2変数関数 K^* は、その経験的関数

$$(3.12) \quad K_n^*(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_{1,i} \leq x_1 < Y_{1,i}, X_{2,i} \leq x_2 \leq Y_{2,i})$$

によって直接推定できる。 G^* の標本推定量 $G_n^*(x_1, x_2)$ は、観測された (Y_1, Y_2) の単純な経験分布関数として次で与えられる。

$$(3.13) \quad G_n^*(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(Y_{1,i} < x_1, Y_{2,i} \leq x_2).$$

したがって、ノンパラメトリック推定量 $F_n^{(0)}$ は次のように導入できる。

$$(3.14) \quad K_n^*(x_1, x_2) = F_n^{(0)}(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{dG_n^*(y_1, y_2)}{F_n^{(0)}(y_1, y_2)}.$$

しかし、式(3.14)には一意な解が存在しない。もし $F_n^{(0)}$ が式(3.14)を満たすならば、定数 k に対して $kF_n^{(0)}$ もまたこの式を満たす。 $F_n^{(0)}$ が次の点

$$(X_{1,n+1}, X_{2,n+1}) \equiv \left(\max_{1 \leq i \leq n} \{X_{1,i}\} + \frac{1}{n}, \max_{1 \leq i \leq n} \{X_{2,i}\} + \frac{1}{n} \right)$$

で 1 となるように指定する。つまり

$$F_n^{(0)}(X_{1,n+1}, X_{2,n+1}) = 1.$$

繰り返し演算のアルゴリズム

推定量 $K^*(x_1, x_2)$ は、 $2n$ 個の点 $(X_{1,i}, X_{2,i})$, $i = 1, \dots, n$ および $(Y_{1,i}, Y_{2,i})$ ($i = 1, \dots, n$) で決定される。

Step 1. (1)次を定義する。

$$(3.15) \quad (X_{1,n+1}, X_{2,n+1}, Y_{1,n+1}, Y_{2,n+1}) = \left(0, 0, \max_{1 \leq i \leq n} \{X_{1,i}\} + \frac{1}{n}, \max_{1 \leq i \leq n} \{X_{2,i}\} + \frac{1}{n} \right)$$

そして、これを $(n+1)$ 番目の観測点として追加する。

(2)この新しい観測点における $F_n^{(0)}$ の値を割り当てる。

$$(3.16) \quad F_n^{(0)}(X_{1,n+1}, X_{2,n+1}) = 0,$$

$$(3.17) \quad F_n^{(0)}(Y_{1,n+1}, Y_{2,n+1}) = 0.$$

Step 2. 式(3.12)を使用して, $(s_1, s_2) = (Y_{1,i}, Y_{2,i})$, $(i=1, \dots, n)$ および $(X_{1,i}, X_{2,i})$, $(i=1, \dots, n)$ における $K_n^*(s_1, s_2)$ を計算する.

$$K_n^*(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_{1,i} \leq x_1 < Y_{1,i}, X_{2,i} \leq x_2 \leq Y_{2,i}).$$

Step 3. (1) $\{(Y_{1,i}, Y_{2,i}), (X_{1,i}, X_{2,i}) \mid i = 1, \dots, n\}$ の中から, 次の条件を満たす点 (x_1, x_2) を探索する.

$\{(Y_{1,i}, Y_{2,i}) : Y_{1,i} > x_1, Y_{2,i} > x_2, i = 1, \dots, n+1\}$ におけるすべての $F_n^{(0)}$ の値が既知である.

条件を満たさない場合, $\{(Y_{1,i}, Y_{2,i}) : Y_{1,i} > x_1, Y_{2,i} > x_2, i = 1, \dots, n+1\}$ 内の点 (s_1, s_2) で $F_n^{(0)}$ の値が未知であるものが存在する. (s_1, s_2) を条件に合致するか確認する.

$F_n^{(0)}(Y_{1,n+1}, Y_{2,n+1}) = 1$ が既知であり, $(Y_{1,n+1}, Y_{2,n+1}) = 1$ も既知であるため, 上記の手順で必要な (x_1, x_2) が得られる.

(2) (x_1, x_2) における $F_n^{(0)}$ の値は次のようにして求められる.

$$(3.18) \quad F_n^{(0)}(x_1, x_2) = \frac{(n+1)K_n^*(x_1, x_2)}{1 + \sum_{\substack{(Y_{1,i}, Y_{2,i}) : \\ Y_{1,i} > x_1, Y_{2,i} > x_2}} \frac{1}{nF_n^{(0)}(Y_{1,i}, Y_{2,i})}}.$$

Step 4. $F_n^{(0)}$ の値が $\forall 2n$ 点 $(Y_{1,i}, Y_{2,i})$ および $(X_{1,i}, X_{2,i})$, $i = 1, \dots, n$ に対して既知になるまで, Step 1-3 を繰り返す.

Step 5. 最後に, F_n およびその周辺分布 F_{1n} と F_{2n} を次で計算する.

$$F_n(x_1, x_2) = \frac{1}{n} \sum_{\substack{X_{1,i} \leq x_1, \\ X_{2,i} \leq x_2}} \frac{F_n^{(0)}(X_{1,i}, X_{2,i})}{K_n^*(X_{1,i}, X_{2,i})},$$

$$F_{1n}(x_1) = F_n(x_1, \infty),$$

$$F_{2n}(x_2) = F_n(\infty, x_2).$$

このようにして, 2 変数右切断データにおける分布関数とその周辺分布関数を推定できる. van der Laan (1996) や Quale and van der Laan (2000) は F_n を

$$(3.19) \quad F_n(x_1, x_2) = \int_0^{x_1} \int_0^{x_2} \frac{dF_n^*(x'_1, x'_2)}{\int_{x'_1}^{\infty} \int_{x'_2}^{\infty} \frac{dF_n^*(y_1, y_2)}{F_n(y_1, y_2)}}$$

という反復解で表現している. 2 変数の場合も推定量の分散は影響関数を用いて評価できるが (van der Laan, 1996), この場合は分散が閉じた式で表現されないため, 例えばブートストラップサンプリングを用いる方が簡単である(たとえば Quale and van der Laan, 2000). 変数 (X_1, X_2) を (M_1, M_2) に変換することで, 図 6 の領域 A に属する銀河サンプルから目的の 2 変数光度関数を得ることができる.

3.2.3 天文学における 2 変数選択問題への拡張

しかし, 上で求めた推定量について注意しておかねばならない重要な問題がある. 第 3.2.1 節で概略を見たように, van der Laan (1996) および Huang et al. (2001) で扱われているデータ選択は, そのままでは天文データの 2 バンド選択とは異なっている. これを理解するため, ある距離指標 D での銀河の 2 バンド選択を考える(図 6). van der Laan (1996) の 2 変数切断は, $M_{1,i} \leq M^{\lim}_1(D)$, $M_{2,i} \leq M^{\lim}_2(D)$ という条件が満たされている場合であり, 図 6 における領

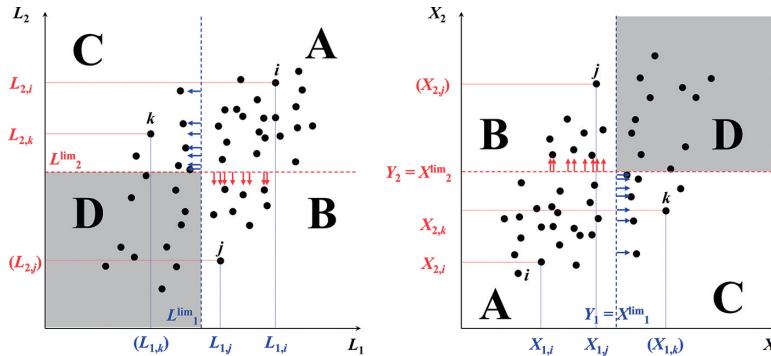


図 6. 光度 L および(換算)絶対等級 X で表した 2 変数光度選択。左パネルが光度による表現、右パネルが換算絶対等級による表現である。領域 A, B, C, D は図 5 に対応している。領域 D は天体は本来は存在するものの、ある距離指標 D での検出限界によって切断されており、データには含まれないため影をつけて示している。銀河 i はバンド 1, 2 とともに検出され、測定値を持つ。銀河 j はバンド 1 では検出されているが、バンド 2 の検出限界を下回るため、得られるのは上限値 $L^{\lim}_2(D)$ ないし下限値 $Y_2 = X^{\lim}_2(D)$ のみである。一方銀河 k はバンド 2 で検出されているがバンド 1 では上限値 $L^{\lim}_1(D)$ ないし下限値 $Y_1 = X^{\lim}_1(D)$ のみが得られる。即ち銀河 j, k は打ち切り(censored)データである。

域 A に対応する。しかし、前節で述べたように、多波長銀河探査ではあるバンドで検出されている銀河が別のバンドでは検出されないことも多い。この場合は銀河が存在することは分かっているが、別のバンドでは光度の上限値しか得られない。これは切断ではなく打ち切りデータに対応する状況である。図 5 および図 6 の領域 B および C に入るデータ点が打ち切りを受けている。しかし、これらのデータも天体の光度に関する情報を持っており、推定から除外してしまうことでデータに偏りが生じる懼れがある。前節で紹介した天文学で用いられる方法では、主選択バンドを設定し、もう 1 つのバンドの上限値を含む情報を付加する。これは Quale and van der Laan (2000) の示した以下の手順で定式化できる。

まず、データが主選択バンド 1 でのみ切断されていると仮定する。これは $Y_2 = \infty$ を設定することと同等である。このとき、測度 dG_n^* は

$$dG_n^* = g_1(y_1) \delta_D(\infty - y_2) dy_1 dy_2$$

のようく表される。これにより次が導かれる。

$$dF_n(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{dG_n^*(y_1, y_2)}{F_n(y_1, y_2)} = dF_n(x_1, x_2) \int_{x_1}^{\infty} \frac{g_1(y_1) dy_1}{F_n(y_1, \infty)} = dF_n^*(x_1, x_2).$$

したがって

$$dF_n(x_1, \infty) dF_n(x_1, x_2) \int_{x_1}^{\infty} \frac{g_1(y_1) dy_1}{F_n(y_1, \infty)} = dF_n^*(x_1, \infty).$$

これは 1 変数の場合と全く同じであり、解は積極限推定量である。これは生存関数 S を通じて導出できる。

周辺サンプル (X_1, X_1) , $(X_1 \leq Y_1)$ に対する積極限推定量を $S_n^{\text{PL}}(x_1, \infty)$ とすると、

$$F_n(x_1, \infty) = 1 - S_n^{\text{PL}}(x_1, \infty),$$

より

$$\begin{aligned}
 (3.20) \quad F_n(x_1, x_2) &= \int_0^{x_1} \int_0^{x_2} \frac{dF_n^*(x_1, x_2)}{\int_0^{x_1} \int_0^{x_2} \frac{dG_{n1}^*(y_1)}{1 - S_n^{\text{PL}}(y_1, \infty)}} \\
 &= \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}(X_{1,i} \leq x_1, X_{2,i} \leq x_2, X_{1,i} \leq Y_{1,i}, X_{2,i} \leq Y_{2,i})}{\frac{1}{n} \sum_{j=1}^n \frac{\mathbb{1}(X_{1,i} \leq Y_{1,j}, X_{1,i} \leq Y_{1,i})}{1 - S_n^{\text{PL}}(Y_{1,j}, \infty)}} \\
 &= \sum_{i=1}^n \frac{\mathbb{1}(X_{1,i} \leq x_1, X_{2,i} \leq x_2, X_{1,i} \leq Y_{1,i}, X_{2,i} \leq Y_{2,i})}{\sum_{j=1}^n \frac{\mathbb{1}(X_{1,i} \leq Y_{1,j}, X_{1,i} \leq Y_{1,i})}{1 - S_n^{\text{PL}}(Y_{1,j}, \infty)}}.
 \end{aligned}$$

が得られる。1 バンド切断の場合に注目すべきは、2 バンド切断の場合とは対照的に、ノンパラメトリック最尤推定量 (NPMLE) が陽に得られることである。

天文学における応用では、主バンドの切断と二次バンドの打ち切りを考慮する必要がある。この場合、以前に研究された天文学の例では、打ち切りはバンド 2 での二次選択でのみ発生することに注意する。推定量は式(3.20)における S_n^* に、打ち切りデータに対応する生存関数 S を単純に代入することで実現できる (e.g., van der Laan, 1996; Quale and van der Laan, 2000)。打ち切りデータを扱うためのプラグイン推定量としては、例え Dąbrowska 推定量 S_n^{Dab} が適切である (Dąbrowska, 1988) [付録 A の式(A.18)を参照]。打ち切りはバンド 2 における Y_2 を通じて生じる。変数 X_1 と X_2 が分離可能であるという事実により、この方法は第 3.2.1 節での天文学的 2 変数分布の古典的推定法の厳密な導出となっている。

3.2.4 2 変数光度関数の最終的な「真の」推定量

ここでついに 2 変数光度関数の「真の」推定量をどのように推定するかについて議論し、本論を締めくくる。「真の」2 変数光度関数とは何を意味するかを明確にする。図 5 および図 6 に戻ると、van der Laan 推定量ではこれらの図における領域 A のみを考慮することになる。つまり、バンド 1 選択サンプルとバンド 2 選択サンプルの和集合に含まれる領域 B および C の標本を落とすことで、情報量を大幅に損失している。一方で、古典的な天文学的手法を採用した場合、バンド 1 またはバンド 2 選択サンプルのどちらかに依存することになる。この場合、領域 B または C のサンプル情報を失わないが、サンプル選択は構造上バイアスがかかっており、その結果として天文学的結論も必然的に偏ったものとなる。

2 バンド選択データから偏りのないサンプルを構築しようとする場合、図 6 の領域 A, B, C の和集合から得られる最大限の情報を利用する推定量が必要である。これを達成するには、両方の手法の長所を取り入れるべきである。先の議論から明らかのように、2 変数切断データに対する van der Laan 推定量を採用し、2 変数生存関数に対して Dąbrowska 推定量

$$(3.21) \quad S_n^{\text{Dab}}(x_1, x_2) = S_n^{\text{Dab}}(x_1, 0) S_n^{\text{Dab}}(0, x_2) \prod_{\substack{0 < x'_1 \leq x_1 \\ 0 < x'_2 \leq x_2}} [1 - Q_n(\Delta x'_1, \Delta x'_2)]$$

を代入することでこれを実現できる。推定量の構築および関連する式は長くなるため、付録 A に提示する。式(3.21)では、打ち切り指標の集合 $(\Delta_{1,i}, \Delta_{2,i})$ ($i = 1, \dots, n$) を使用する。実用上は、最終的な NPMLE の陽な表現がないため取り扱いは容易ではないが、式(3.19)および式(3.21)のセットが、2 バンド選択下での銀河の 2 変数光度関数の最終的な推定量である。

データの情報を完備にするため、式(3.21)において適切な打ち切り指標の集合 $(\Delta_{1,i}, \Delta_{2,i})$ ($i = 1, \dots, n$) を割り当てる。データの元の構造は次の通りである

$$[\text{バンド 1 の等級}, \text{バンド 2 の等級}, \text{赤方偏移}] = [m_{\lambda_1, i}, m_{\lambda_2, i}, z_i],$$

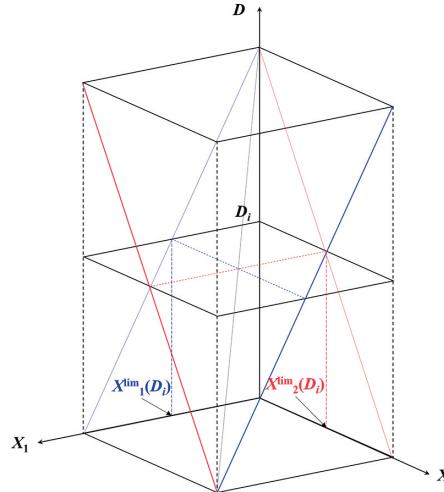


図 7. 2 変数光度選択でのデータ選択効果における距離指標 D への依存性. この図では銀河の光度および検出限界は換算絶対等級 $(X_1, X_2), (X^{\lim}_1, X^{\lim}_2)$ で表示している. 底面で示した観測可能範囲は図 5 および図 6 における領域 A に対応し, バンド 1, 2 両方で検出され, 測定値を持つ. 距離 D_1 ではその範囲が狭まり, 領域 B, C に対応する打ち切りデータが含まれる. 天体の多波長探査データでは距離方向に積分されているため, 異なった検出限界のデータが混在している.

対応する限界等級は m^{\lim}_1 および m^{\lim}_2 である. 光度関数を構築するために, まずデータを

$$[\text{バンド 1 の光度, バンド 2 の光度, 光度距離}] = [L_{\lambda_1, i}, L_{\lambda_2, i}, d_L(z_i)],$$

あるいはまったく等価であるが

$$[\text{バンド 1 の絶対等級, バンド 2 の絶対等級, 距離指数}] = [M_{\lambda_1, i}, M_{\lambda_2, i}, D_i]$$

という形式に変換する. 検出限界も (L^{\lim}_1, L^{\lim}_2) または (M^{\lim}_1, M^{\lim}_2) に変換される. さらにこれらを簡略化のために, 換算絶対等級として $[X_{1, i}, X_{2, i}, D_i]$ に変換し, 関連する検出限界 (X^{\lim}_1, X^{\lim}_2) を付け加えて, 詳細な統計的議論の記述を行う.

換算絶対等級単位では, 切断変数 $(Y_{1, i}, Y_{2, i})$ は次のように定義される

$$(3.22) \quad Y_{1, i} = X^{\lim}_{1, i}(D_i),$$

$$(3.23) \quad Y_{2, i} = X^{\lim}_{2, i}(D_i).$$

$X_1 < Y_1$ または $X_2 < Y_2$ (「かつ」ではない)を満たす銀河が観測され, 2 バンド選択サンプルに含まれる. これは図 7 で概略的に示されている. 図 6 の領域 A では, $X_1 \leq Y_1$ かつ $X_2 \leq Y_2$ を満たす銀河が含まれている. この場合, データに打ち切りは存在しない. 領域 B では, バンド 1 での検出があるが, バンド 2 では検出がない. つまり, $X_1, X_2 = Y_2, \Delta_1 = 1, \Delta_2 = 0$ として打ち切りデータが存在する. 同様に領域 C では, $X_1 = Y_1, X_2, \Delta_1 = 0, \Delta_2 = 1$ としてデータが存在する. したがって, 次の構造

$$(3.24) \quad [X_{1, i}, X_{2, i}, \Delta_{1, i}, \Delta_{2, i}]$$

を持つデータセットを準備する. この形式で記述されたデータと式(3.19)および式(3.21)を用

いることで、本来の 2 変数光度関数を推定することができる。

4.まとめと議論

本研究では、天文サーベイから銀河の光度関数を推定する手法を紹介した。この種の解析は、切断データ解析として知られている。まず最初に、1 変数光度関数の場合を導入した。切断データから 1 変数光度関数を推定する手法は、天文学コミュニティにおいて Lynden-Bell (1971) によって導入されたのが最初である。Woodroffe (1985) は Lynden-Bell の推定量を再定式化した。彼はこれを統計学コミュニティに導入し、その収束特性について数学的に厳密な証明を行った。

しかし、この問題の 2 変数版である 2 変数光度関数の厳密な方法論は、天文学で適切な定式化に基づいて議論されることはほとんどなかった。天文学における 2 変数光度関数は、主選択バンドから推定された 1 変数光度関数に基づいて構築される。次に、二次選択バンドからの条件付き光度関数をカプラン-マイヤー推定量で推定し、それを組み合わせる。しかし、この手法は単一バンド選択に基づくことが原因で、2 変数関数に対して偏った推定をもたらしてしまう。一方、van der Laan (1996) は 2 変数切断データの場合の 2 変数分布関数の推定量を提案した。この手法は 2 変数切断には適しているものの、2 バンド選択の場合には避けられない打ち切りデータの情報を取りこぼしてしまう。そこで我々は、本論文で打ち切り情報も取り入れることができる統合手法を提案した。これは、切断変数と打ち切り変数を設定することで達成できる。この推定法の性能評価と実データへの応用は、今後の研究で数値的検討とともに提示する予定である (Takeuchi et al. 2025, in preparation)。

最後に、天文光度関数の推定についての注意を述べる。すべての関連する議論において、 X と Y の独立性が仮定されている。しかし、図 2 および図 7 に示されるように、この独立性が完全に成立するかどうかは慎重に検討されるべきである。特に、最近の深いサーベイでは、サンプルにおける時間進化の影響は無視できない。この問題は一部の著者によって指摘され、検討されるようになってきている (たとえば Efron and Petrosian, 1992; Chiou et al., 2018; Chiou et al., 2019)。今後の巨大サーベイプロジェクトでは、この問題がますます深刻になるため、より数学的に厳密な議論が必要となると考えられる。

注.

- 1) 定常ユークリッド空間とは、宇宙膨張のように空間自体の計量 (metric) が時間変化することのないユークリッド空間を意味する宇宙論の用語である。定常ユークリッド空間での議論では、単位振動数当たりの量 L_ν および S_ν で考えても全く同じ関係が成り立つ。
- 2) 宇宙物理学分野では、常用対数を \log 、自然対数を \ln で表すのが習慣である。本稿でもこれ以降は底の 10 を省略する。
- 3) 銀河の集団化の度合いをクラスタリングと呼ぶ。クラスタリングを n 次 factorial cumulant (宇宙論では n 点相関関数と呼ぶ) を用いて定量的に評価する方法を相関関数解析という。
- 4) 天文学で用いられる長さの単位 Mpc (メガパーセク) は 3.26×10^{24} cm に対応する。
- 5) 超銀河団と呼ばれる構造は長く伸びた形状の銀河の集中を指し、銀河団の集団ではない。現在はフィラメント構造と呼ぶのが一般的になっている。
- 6) Dorota M. Dąbrowska は論文中で自身の名前を Dabrowska と綴っているが、本稿ではオリジナルの綴りを採用する。

謝 辞

本研究は日本学術振興会(JSPS)科学研究費補助金(21H01128 および 24H00247)の補助を受けて行っている。また本研究の一部は統計数理研究所共同研究費「機械学習的宇宙構造論：構造形成から銀河進化へ」(2024-ISMCRP-2025)の支援も受けて行った。

本論文は天文学と統計学両方にまたがる内容を扱っており、両分野の慣習の違いのため初稿は未整理な部分やミスが多数残っていた。辛抱強く丁寧に議論や導出を追い、極めて有用なコメントをしていただいた2人の査読者に深く感謝申し上げる。統計数理研究所の江村 剛志氏には本特集号にご招待いただき、また内容についても大変重要な示唆を頂いた。氏の大きな貢献に心から感謝する。

付録 A. 2変数打ち切りデータに対する Dabrowska 推定量

本章では、2変数打ち切りデータの分布関数に対する Dabrowska 推定量⁶⁾を紹介する(Dabrowska, 1988)。ここでは本論で必要な最低限の議論に留めたが、Dabrowska 推定量の入門的解説はたとえば杉本・田中 (2023, 第 2.1 節)を参照されたい。また漸近的性質等の詳細は Dabrowska (1989)を参照されることをお勧めする。

2変数打ち切りデータ (X_1, X_2) と打ち切り変数 (Y_1, Y_2) のセットを考える。打ち切りデータ (X_1, X_2) と (Y_1, Y_2) の分布関数をそれぞれ F と G とする。本文での議論に即し、左側打ち切り問題を扱う。すなわち、 $X_1 \leq Y_1$ かつ $X_2 \leq Y_2$ のときにのみ (X_1, X_2) が観測される。2変数生存関数 $S(x_1, x_2)$ を次のように定義する。

$$S(x_1, x_2) \equiv \mathbb{P}(X_1 > x_1, X_2 > x_2).$$

これより、次が成り立つ。

$$F(x_1, x_2) = 1 - S(x_1, 0) - S(0, x_2) + S(x_1, x_2).$$

同様に、分布 G に対して生存関数 $R(x_1, x_2)$ を定義する。さらに次を定義する。

$$\begin{aligned} (\tilde{X}_1, \tilde{X}_2) &\equiv (X_1 \wedge Y_1, X_2 \wedge Y_2), \\ \Delta_1 &\equiv \mathbb{1}(X_1 \leq Y_1), \\ \Delta_2 &\equiv \mathbb{1}(X_2 \leq Y_2). \end{aligned}$$

ここで (X_1, X_2) と (Y_1, Y_2) が独立であると仮定する。

まず以下の関数を定義する。

$$\begin{aligned} H(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2), \\ K_{11}(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_1 = 1, \Delta_2 = 1), \\ K_{10}(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_1 = 1), \\ K_{01}(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_2 = 1). \end{aligned}$$

$H(x_1, x_2) > 0$ を満たす (x_1, x_2) に対して次が成り立つ。

$$(A.1) \quad H(x_1, x_2) = R(x_1, x_2)S(x_1, x_2),$$

$$(A.2) \quad K_{11}(\mathrm{d}x_1, \mathrm{d}x_2) = R(x_1-, x_2-)S(\mathrm{d}x_1, \mathrm{d}x_2),$$

$$(A.3) \quad K_{10}(\mathrm{d}x_1, x_2) = R(x_1-, x_2)S(\mathrm{d}x_1, x_2),$$

$$(A.4) \quad K_{01}(x_1, \mathrm{d}x_2) = R(x_1, x_2-)S(x_1, \mathrm{d}x_2).$$

2変数累積ハザード関数を次のように定義する。

$$(A.5) \quad \Lambda_{11}(x_1, x_2) = \int_0^{x_1} \int_0^{x_2} \frac{K_{11}(\mathrm{d}x'_1, \mathrm{d}x'_2)}{H(x'_1-, x'_2-)} ,$$

$$(A.6) \quad \Lambda_{10}(x_1, x_2) = - \int_0^{x_1} \frac{K_{10}(\mathrm{d}x'_1, x_2)}{H(x'_1-, x_2)} ,$$

$$(A.7) \quad \Lambda_{01}(x_1, x_2) = - \int_0^{x_2} \frac{K_{01}(x_1, \mathrm{d}x'_2)}{H(x_1, x'_2-)} .$$

式(A.1)–(A.7)の標本推定量は次のように表される。

$$(A.8) \quad H_n(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} > x_1, \tilde{X}_{2,i} > x_2) ,$$

$$(A.9) \quad K_{n11}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_{1,i} = 1, \Delta_{2,i} = 1) ,$$

$$(A.10) \quad K_{n10}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_{1,i} = 1) ,$$

$$(A.11) \quad K_{n01}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_{2,i} = 1) ,$$

$$(A.12) \quad \Lambda_{n11}(x_1, x_2) = \int_0^{x_1} \int_0^{x_2} \frac{K_{n11}(\mathrm{d}x'_1, \mathrm{d}x'_2)}{H_n(x'_1-, x'_2-)} ,$$

$$(A.13) \quad \Lambda_{n10}(x_1, x_2) = - \int_0^{x_1} \frac{K_{n10}(\mathrm{d}x'_1, x_2)}{H_n(x'_1-, x_2)} ,$$

$$(A.14) \quad \Lambda_{n01}(x_1, x_2) = - \int_0^{x_2} \frac{K_{n01}(x_1, \mathrm{d}x'_2)}{H_n(x_1, x'_2-)} .$$

さらに、記号を簡略化するために、任意の2変数関数 $f(s, t)$ に対して次の記号を定義する。

$$(A.15) \quad f(\Delta s, \Delta t) = f(s, t) - f(s, t-) - f(s-, t) + f(s-, t-) ,$$

$$(A.16) \quad f(\Delta s, t) = f(s, t) - f(s-, t) ,$$

$$(A.17) \quad f(s, \Delta t) = f(s, t) - f(s, t-) .$$

式(A.7)–(A.17)を組み合わせることで、 $S(x_1, x_2)$ の自然な推定量が次の式で与えられる。

$$(A.18) \quad S_n^{\mathrm{Dab}}(x_1, x_2) = S_n^{\mathrm{Dab}}(x_1, 0) S_n^{\mathrm{Dab}}(0, x_2) \prod_{\substack{0 < x'_1 \leq x_1 \\ 0 < x'_2 \leq x_2}} [1 - Q_n(\Delta x'_1 \Delta x'_2)] ,$$

$$(A.19) \quad Q_n(\Delta x'_1 \Delta x'_2) \equiv \frac{\Lambda_{n10}(\Delta x'_1, x_2-) \Lambda_{n01}(x_1-, \Delta x_2) - \Lambda_{n11}(\Delta x'_1 \Delta x'_2)}{[1 - \Lambda_{n10}(\Delta x'_1, x_2-)] [1 - \Lambda_{n01}(x'_1-, \Delta x_2)]} ,$$

$$(A.20) \quad S_n^{\mathrm{Dab}}(x_1, 0) \equiv \prod_{x'_1 \leq x_1} [1 - \Lambda_{n10}(\Delta x'_1, 0)] ,$$

$$(A.21) \quad S_n^{\mathrm{Dab}}(0, x_2) \equiv \prod_{x'_2 \leq x_2} [1 - \Lambda_{n01}(0, \Delta x'_2)] .$$

これらの推定量(式(A.18)–(A.21))を実用的な形に変形すると以下のようになる。

$$(A.22) \quad \Lambda_{n10}(\Delta x'_1, 0) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} = x'_1) \mathbb{1}(\Delta_{1,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x'_1)} ,$$

$$(A.23) \quad \Lambda_{n10}(0, \Delta x'_2) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{2,i} = x'_2) \mathbb{1}(\Delta_{2,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{2,i} \geq x'_2)}.$$

同様に、 Q_n の式を次のように導出する。

$$(A.24) \quad \Lambda_{n11}(\Delta x'_1, \Delta x'_2) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} = x'_1, \tilde{X}_{2,i} = x'_2) \mathbb{1}(\Delta_{1,i} = 1, \Delta_{2,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x'_1, \tilde{X}_{2,i} \geq x'_2)},$$

$$(A.25) \quad \Lambda_{n10}(\Delta x'_1, x'_2-) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} = x'_1, \tilde{X}_{2,i} \geq x'_2) \mathbb{1}(\Delta_{1,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x'_1, \tilde{X}_{2,i} \geq x'_2)},$$

$$(A.26) \quad \Lambda_{n01}(x'_1-, \Delta x'_2-) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x'_1, \tilde{X}_{2,i} = x'_2) \mathbb{1}(\Delta_{2,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x'_1, \tilde{X}_{2,i} \geq x'_2)}.$$

以上の一連の式を使用して、Dąbrowska 推定量を計算することができる。

参 考 文 献

- Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins Series in the Mathematical Sciences, Springer, New York.
- Binggeli, B., Sandage, A. and Tammann, G. A. (1988). The luminosity function of galaxies, *Annual Review of Astronomy and Astrophysics*, **26**, 509–560, <http://dx.doi.org/10.1146/annurev.aa.26.090188.002453>.
- Buat, V., Takeuchi, T. T., Burgarella, D., Giovannoli, E. and Murata, K. L. (2009). The infrared emission of ultraviolet-selected galaxies from $z = 0$ to $z = 1$, *Astronomy & Astrophysics*, **507**(2), 693–704, <http://dx.doi.org/10.1051/0004-6361/200912024>.
- Caditz, D. and Petrosian, V. (1993). Smoothed nonparametric estimation of the luminosity function for flux-limited samples, *The Astrophysical Journal*, **416**, 450–457, <http://dx.doi.org/10.1086/173250>.
- Chiou, S. H., Qian, J., Mormino, E. and Betensky, R. A. (2018). Permutation tests for general dependent truncation, *Computational Statistics & Data Analysis*, **128**(C), 308–324.
- Chiou, S. H., Austin, M. D., Jing, Q. and Betensky, R. A. (2019). Transformation model estimation of survival under dependent truncation and independent censoring, *Statistical Methods in Medical Research*, **28**(12), 3785–3798.
- Choloniewski, J. (1985). Bivariate luminosity function of E and S0 galaxies, *Monthly Notices of the Royal Astronomical Society*, **214**, 197–202, <http://dx.doi.org/10.1093/mnras/214.2.197>.
- Choloniewski, J. (1987). On Lynden-Bell's method for the determination of the luminosity function, *Monthly Notices of the Royal Astronomical Society*, **226**, 273–280, <http://dx.doi.org/10.1093/mnras/226.2.273>.
- Cowie, L. L., Songaila, A., Hu, E. M. and Cohen, J. G. (1996). New insight on galaxy formation and evolution from Keck spectroscopy of the Hawaii deep fields, *The Astronomical Journal*, **112**, 839–864, <http://dx.doi.org/10.1086/118058>.
- Dąbrowska, D. M. (1988). Kaplan-Meier estimate on the plane, *The Annals of Statistics*, **16**(4), 1475–1489, <http://dx.doi.org/10.1214/aos/1176351049>.
- Dąbrowska, D. M. (1989). Kaplan-Meier estimate on the plane: Weak convergence, LIL, and the bootstrap, *Journal of Multivariate Analysis*, **29**(2), 308–325, [http://dx.doi.org/https://doi.org/10.1016/0047-259X\(89\)90030-4](http://dx.doi.org/https://doi.org/10.1016/0047-259X(89)90030-4).
- de Lapparent, V., Geller, M. J. and Huchra, J. P. (1986). A slice of the universe, *Astrophysical Journal*, **302**, L1–L4.
- de Lapparent, V., Geller, M. J. and Huchra, J. P. (1989). The luminosity function for the CfA redshift survey slices, *The Astrophysical Journal*, **343**, 1–17, <http://dx.doi.org/10.1086/167679>.

- Eales, S. (1993). Direct construction of the galaxy luminosity function as a function of redshift, *The Astrophysical Journal*, **404**, 51–62, <http://dx.doi.org/10.1086/172257>.
- Efron, B. and Petrosian, V. (1992). A simple test of independence for truncated data with applications to redshift surveys, *The Astrophysical Journal*, **399**, 345–352, <http://dx.doi.org/10.1086/171931>.
- Ellis, R. S., Colless, M., Broadhurst, T., Heyl, J. and Glazebrook, K. (1996). Autofib redshift survey—I. Evolution of the galaxy luminosity function, *Monthly Notices of the Royal Astronomical Society*, **280**(1), 235–251, <http://dx.doi.org/10.1093/mnras/280.1.235>.
- Heyl, J., Colless, M., Ellis, R. S. and Broadhurst, T. (1997). Autofib redshift survey—II. Evolution of the galaxy luminosity function by spectral type, *Monthly Notices of the Royal Astronomical Society*, **285**(3), 613–634, <http://dx.doi.org/10.1093/mnras/285.3.613>.
- Huang, J., Vieland, V. J. and Wang, K. (2001). Nonparametric estimation of marginal distributions under bivariate truncation with application to testing for age-of-onset anticipation, *Statistica Sinica*, **11**(4), 1047–1068.
- Hubble, E. P. (1936). *Realm of the Nebulae*, Yale University Press, New Haven.
- Jackson, J. C. (1974). The analysis of quasar samples, *Monthly Notices of the Royal Astronomical Society*, **166**, 281–296, <http://dx.doi.org/10.1093/mnras/166.2.281>.
- Johnston, R. (2011). Shedding light on the galaxy luminosity function, *Astronomy & Astrophysics Review*, **19**, 41, <http://dx.doi.org/10.1007/s00159-011-0041-9>.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**(282), 457–481.
- Keiding, N. and Gill, R. D. (1990). Random truncation models and Markov processes, *The Annals of Statistics*, **18**(2), 582–602, <http://dx.doi.org/10.1214/aos/1176347617>.
- Keres, D., Yun, M. S. and Young, J. S. (2003). CO luminosity functions for far-infrared- and B-band-selected galaxies and the first estimate for $\Omega_{\text{HI}+\text{H}_2}$, *The Astrophysical Journal*, **582**(2), 659–667, <http://dx.doi.org/10.1086/344820>.
- Lilly, S. J., Le Fevre, O., Hammer, F. and Crampton, D. (1996). The Canada-France redshift survey: The luminosity density and star formation history of the universe to $z \sim 1$, *The Astrophysical Journal*, **460**, L1–L4, <http://dx.doi.org/10.1086/309975>.
- Lin, H., Kirshner, R. P., Shectman, S. A., Landy, S. D., Oemler, A., Tucker, D. L. and Schechter, P. L. (1996). The luminosity function of galaxies in the las campanas redshift survey, *The Astrophysical Journal*, **464**, 60–78, <http://dx.doi.org/10.1086/177300>.
- Lin, H., Yee, H. K. C., Carlberg, R. G., Morris, S. L., Sawicki, M., Patton, D. R., Wirth, G. and Shepherd, C. W. (1999). The CNOC2 field galaxy luminosity function. I. A description of luminosity function evolution, *The Astrophysical Journal*, **518**(2), 533–561, <http://dx.doi.org/10.1086/307297>.
- Lynden-Bell, D. (1971). A method of finding distances to double galaxies, *Monthly Notices of the Royal Astronomical Society*, **155**, 95–118, <http://dx.doi.org/10.1093/mnras/155.1.95>.
- Marshall, H. L., Tananbaum, H., Avni, Y. and Zamorani, G. (1983). Analysis of complete quasar samples to obtain parameters of luminosity and evolution functions, *The Astrophysical Journal*, **269**, 35–41, <http://dx.doi.org/10.1086/161016>.
- Mobasher, B., Sharples, R. M. and Ellis, R. S. (1993). A complete galaxy redshift survey-V. Infrared luminosity functions for field galaxies, *Monthly Notices of the Royal Astronomical Society*, **263**, 560–574, <http://dx.doi.org/10.1093/mnras/263.3.560>.
- Pascarelle, S. M., Lanzetta, K. M. and Fernández-Soto, A. (1998). The ultraviolet luminosity density of the universe from photometric redshifts of galaxies in the hubble deep field, *The Astrophysical Journal*, **508**(1), L1–L4, <http://dx.doi.org/10.1086/311708>.
- Quale, C. and van der Laan, M. J. (2000). Inference with bivariate truncated data, *Lifetime Data Analysis*, **6**, 391–408, <http://dx.doi.org/10.1023/A:1026513500285>.
- Sandage, A., Tammann, G. A. and Yahil, A. (1979). The velocity field of bright nearby galaxies. I. The

- variation of mean absolute magnitude with redshift for galaxies in a quiet velocity field, *The Astrophysical Journal*, **232**, 352–364, <http://dx.doi.org/10.1086/157295>.
- Sawicki, M. J., Lin, H. and Yee, H. K. C. (1997). Evolution of the galaxy population based on photometric redshifts in the Hubble deep field, *The Astronomical Journal*, **113**, 1–12, <http://dx.doi.org/10.1086/118231>.
- Schechter, P. (1976). An analytic expression for the luminosity function for galaxies, *The Astrophysical Journal*, **203**, 297–306, <http://dx.doi.org/10.1086/154079>.
- Schmidt, M. (1967). Space distribution of quasi-stellar radio sources, *Publications of the Astronomical Society of the Pacific*, **79**(470), 437–438, <http://dx.doi.org/10.1086/128576>.
- Schmidt, M. (1968). Space distribution and luminosity functions of quasi-stellar radio sources, *The Astrophysical Journal*, **151**, 393–409, <http://dx.doi.org/10.1086/149446>.
- Shane, C. D. and Wirtanen, C. A. (1967). *Publications of Lick Observatory*, **22**, part 1.
- SubbaRao, M. U., Connolly, A. J., Szalay, A. S. and Koo, D. C. (1996). Luminosity functions from photometric redshifts I: Techniques, *The Astronomical Journal*, **112**, 929–936, <http://dx.doi.org/10.1086/118066>.
- 杉本知之, 田中健太 (2023). 2 変量生存時間モデルにおけるコピュラとその利用, *日本統計学会誌*, **52**(2), 153–176, <http://dx.doi.org/10.11329/jjssj.52.153>.
- Takeuchi, T. T. (2010). Constructing a bivariate distribution function with given marginals and correlation: Application to the galaxy luminosity function, *Monthly Notices of the Royal Astronomical Society*, **406**(3), 1830–1840, <http://dx.doi.org/10.1111/j.1365-2966.2010.16778.x>.
- Takeuchi, T. T. (2025). in preparation.
- Takeuchi, T. T. and Kono, K. T. (2020). Constructing a multivariate distribution function with a vine copula: Towards multivariate luminosity and mass functions, *Monthly Notices of the Royal Astronomical Society*, **498**(3), 4365–4378, <http://dx.doi.org/10.1093/mnras/staa2558>.
- Takeuchi, T. T., Yoshikawa, K. and Ishii, T. T. (2000). Tests of statistical methods for estimating galaxy luminosity function and applications to the Hubble deep field, *The Astrophysical Journal Supplement Series*, **129**(1), 1–31, <http://dx.doi.org/10.1086/313409>.
- Takeuchi, T. T., Ishii, T. T., Dole, H., Dennefeld, M., Lagache, G. and Puget, J. L. (2006). The ISO 170 μm luminosity function of galaxies, *Astronomy & Astrophysics*, **448**(2), 525–534, <http://dx.doi.org/10.1051/0004-6361:20054272>.
- Takeuchi, T. T., Sakurai, A., Yuan, F.-T., Buat, V. and Burgarella, D. (2013). Far-ultraviolet and far-infrared bivariate luminosity function of galaxies: Complex relation between stellar and dust emission, *Earth, Planets, and Space*, **65**(3), 281–290, <http://dx.doi.org/10.5047/eps.2012.06.008>.
- Tinsley, B. M. (1980). Evolution of the stars and gas in galaxies, *Fundamentals of Cosmic Physics*, **5**, 287–388.
- Tinsley, B. M. and Danly, L. (1980). On the density of star formation in the universe, *The Astrophysical Journal*, **242**, 435–442, <http://dx.doi.org/10.1086/158477>.
- van der Laan, M. J. (1996). Nonparametric estimation of the bivariate survival function with truncated data, *Journal of Multivariate Analysis*, **58**(1), 107–131, <http://dx.doi.org/https://doi.org/10.1006/jmva.1996.0042>.
- van der Vaart, A. (1991). On differentiable functionals, *The Annals of Statistics*, **19**(1), 178–204, <http://dx.doi.org/10.1214/aos/1176347976>.
- Wall, J. and Jenkins, C. (2003). *Practical Statistics for Astronomers*, Cambridge Observing Handbooks for Research Astronomers, Cambridge University Press, Cambridge.
- Wang, M.-C., Jewell, N. P. and Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation, *The Annals of Statistics*, **14**(4), 1597–1605, <http://dx.doi.org/10.1214/aos/1176350180>.
- Woodroffe, M. (1985). Estimating a distribution function with truncated data, *The Annals of Statistics*, **13**(1), 163–177.

Survival Analysis in Astrophysics: Estimation of Galaxy Luminosity Function

Tsutomu T. Takeuchi^{1,2}

¹Division of Particle and Astrophysical Science, Nagoya University

²The Institute of Statistical Mathematics

Observational data in astronomy are invariably subject to truncation due to the detection limits of observation instruments. When estimating distribution functions of astronomical statistical quantities from such data, it is appropriate to apply survival analysis of truncated data. However, survival analysis, which has been developed in the field of statistics, was practically unknown in the field of astronomy until the 1980s, and astronomers have developed estimation methods independently of the framework in mathematical statistics. Although many of the methods devised by astronomers finally converge with survival analysis, they have been mathematically unorganized, and statistical discussions became systematic only after the 21st century. In this paper, we discuss the estimation of galaxy luminosity functions as an application of survival analysis of truncated data in astronomy. First, we introduce the characteristics of astronomical data in the univariate case and explain the correspondence with survival analysis. Next, we introduce the bivariate case. The estimation of a bivariate luminosity function can become more complex, since it involves both truncation and censoring, depending on the sampling method of astronomical objects. Here, we make attempt to construct the estimation method of bivariate luminosity function as generally as possible.

COM-Poisson 分布について

—過小分散も過大分散も表現するポアソン分布—

長塚 豪己[†]

(受付 2024 年 11 月 28 日; 改訂 2025 年 2 月 7 日; 採択 2 月 12 日)

要　　旨

Poisson 分布は、カウントデータの統計モデルとして、最もよく知られた分布である。しかし、Poisson 分布には、平均と分散が等しいという意味での等分散性(equi-dispersion)が課されている。応用の場では、多くのカウントデータに、分散が平均より大きい過分散性(over-dispersion)，時には、分散が平均より小さい過小分散性(under-dispersion)が見られる。それぞれの場合において、負の二項分布や幾何分布等がモデルとして適宜検討される。Conway と Maxwell によって生み出された Conway-Maxwell-Poisson (COM-Poisson) 分布は、Poisson 分布の一般化分布の一つである。生存時間解析と信頼性解析の分野においては、治癒率モデルに関する研究において、近年、COM-Poisson 分布を用いた報告が多く見られ、注目を集めている。COM-Poisson 分布は、パラメータ数が少なく(2 パラメータ)シンプルなモデルでありながら、広い範囲の過分散、過小分散を柔軟に表現でき、かつよく知られた幾何分布を含み、極限分布としてベルヌイ分布を持つ点が特徴的である。これらの特徴が、他のポアソン分布の一般化モデルと一線を画す。しかし、それらメリットの代償として、確率関数の規格化定数に無限級数を含み、推定、検定においては困難をきたす。これまで、規格化定数の各種近似法や、規格化定数の計算を回避する推定が研究されてきた。本稿では、COM-Poisson 分布、並びに近年提案された手法も含む代表的な COM-Poisson 分布の推定法について紹介する。

キーワード：カウントデータ，Poisson 分布，最尤法，規格化定数。

1. はじめに

既定の期間、あるいは空間内におけるイベントの発生数を表すデータであるカウントデータは、古くから広い分野で利用してきた。古典的な例としては、伝染病や負傷による兵士の死亡数 (Nightingale, 1858)，コレラによる死者数 (Snow, 1858) が挙げられる。近年でもその有用性は衰えることなく、その利用例は、実店舗やオンラインストアでの来客数や購入数、SNS の利用数や閲覧数、TV やオンライン動画サービスにおける視聴数、等々と枚挙にいとまがない。

カウントデータのモデリングに最も広く用いられてきた分布はポアソン分布であるといえる。ポアソン分布、あるいは関連分布によるカウントデータのモデリングの応用範囲は広く、古典的な例は、プロイセン騎兵隊における 20 年間にわたる馬蹴りによる年間死亡者数にポアソン分布をあてはめた例 (von Bortkewitsch, 1898; Hanleya and Bhatnagar, 2022) であり、近年でも、円形の競馬場におけるスタート位置毎の勝った馬の数や斑点病にかかった植物の葉

[†] 中央大学 理工学部：〒112-8551 東京都文京区春日 1-13-27; hideki07@kc.chuo-u.ac.jp

のグレード毎の数にポアソン分布をあてはめて解析を行った例等例は無数にある (Wani and Ahmad, 2024). しかし, ポアソン分布はシンプルで極めてよく知られた分布であるが, 多くの実データはポアソン分布の基礎となる仮定, すなわち平均と分散が等しいという意味での等分散性を満たさない.

Conway と Maxwell によって生み出された Conway-Maxwell-Poisson (COM-Poisson) 分布 (Conway and Maxwell, 1962) は, Poisson 分布の一般化分布の一つである. 生存時間解析と信頼性解析の分野においては, 治癒率モデルに関する研究において, 近年, COM-Poisson 分布を用いた報告が多く見られ, 注目を集めている. COM-Poisson 分布は, パラメータ数が少なく (2 パラメータ) シンプルなモデルでありながら, 広い範囲の過分散, 過小分散を柔軟に表現でき, かつよく知られた幾何分布を含み, 極限分布としてベルヌーイ分布を持つ点が特徴的である. これらの特徴が, 他のポアソン分布の一般化モデルと一線を画す. しかし, それらメリットの代償として, 確率関数の規格化定数に無限級数を含み, 推定, 検定においては困難をきたす. これまで, 規格化定数の各種近似法や, 規格化定数の計算を回避する推定が研究されてきた. 本稿では, COM-Poisson 分布, 並びに近年提案された手法も含む代表的な COM-Poisson 分布の推定法について紹介する.

本稿は, 以下のように構成されている. 第 2 節では, COM-Poisson 分布について概説する. 第 3 節では, 代表的な推定法や近年提案された新しい推定法を紹介する. 第 4 節では, COM-Poisson 分布の応用例を紹介する. 第 5 節で, 本稿を締めくくる.

2. COM-Poisson 分布

2.1 分布の概要

(Ω, \mathcal{F}, P) を, 本論文で扱う確率変数が定義された確率空間とする. さらに, \mathbb{N} は自然数全体を表す集合, $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$ と定義する. 確率変数 X が, 次の確率質量関数を持つとき, X は, パラメータ λ, ν の COM-Poisson 分布 ($\text{CMP}(\lambda, \nu)$ で表す) にしたがうという:

$$f_{\text{CMP}}(x) := P(X = x) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)}, \quad x \in \mathbb{N}_0,$$

ただし, $\lambda > 0, \nu \geq 0$ はパラメータで,

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu},$$

である.

平均パラメータが λ であるポアソン分布 $\text{Poi}(\lambda)$ の確率質量関数

$$f_{\text{P}}(x) = \frac{\lambda^x}{x!} e^{-\lambda} \propto \frac{\lambda^x}{x!},$$

において, $x!$ に, 過分散, 過小分散を表現する形状パラメータ ν をべき乗し, あらためて標準化定数を $Z(\lambda, \nu)$ にとったのが, f_{CMP} と考えられる. よって, $\text{CMP}(\lambda, \nu)$ は, $\text{Poi}(\lambda)$ の一般化されたモデルであり, $\nu = 1$ のとき, $\text{Poi}(\lambda)$ となる. $\nu < 1$ のとき, 過分散モデル (分散が平均より大きいモデル, over-dispersion model), $\nu > 1$ のとき, 過小分散モデル (分散が平均より小さいモデル, under-dispersion model), $\nu = 1$ のとき, 平均と分散が等しいモデル (equi-dispersion model) になる. 図 1 に, 各パラメータ値における $\text{CMP}(\lambda, \nu)$ の確率質量関数のグラフを示す. $\text{CMP}(\lambda, \nu)$ は, 前述のように $\nu = 1$ のときポアソン分布, $\nu = 0, \lambda < 1$ のとき幾何分布, $\nu \leftarrow \infty$ のときベルヌーイ分布となる (詳細については, 表 1 を参照のこと). 表 2 に, $\text{CMP}(\lambda, \nu)$ の確率質量関数, 分布関数, k 次積率 ($k \in \mathbb{N}$), 積率母関数, 特性関数, 確率母

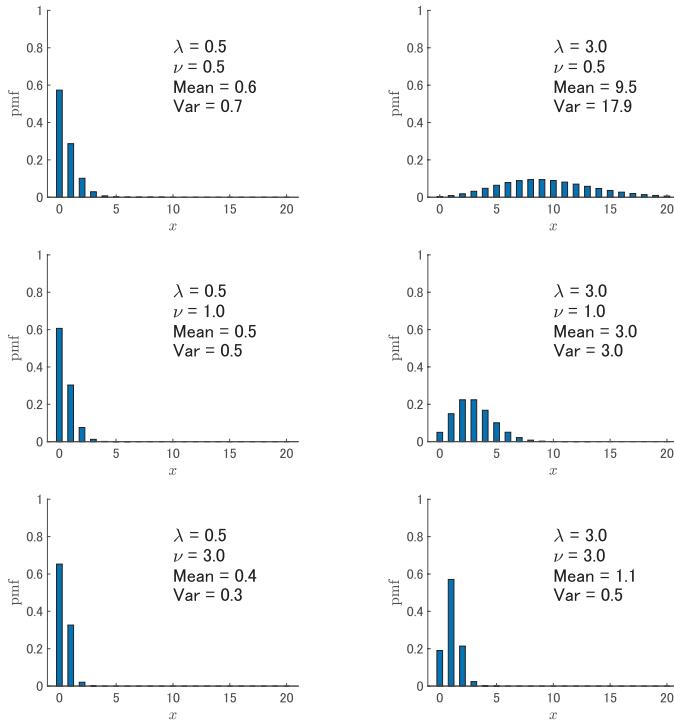
図 1. $\text{CMP}(\lambda, \nu)$ の確率質量関数(pmf)のグラフ (Mean は平均, Var は分散を表す).

表 1. COM-Poisson 分布の特殊例.

λ, ν	$Z(\lambda, \nu)$	確率質量関数	分布名(パラメータ)
$\nu = 1$	e^λ	$\frac{\lambda^x e^{-\lambda}}{x!}, x \in \mathbb{N}_0$	ポアソン (λ)
$\nu = 0, \lambda < 1$	$1/(1 - \lambda)$	$(1 - \lambda) \lambda^x, x \in \mathbb{N}_0$	幾何 ($p = 1 - \lambda$)
$\nu \rightarrow \infty$	$1 + \lambda$	$\begin{cases} \frac{1}{1+\lambda}, & x = 0 \\ \frac{\lambda}{1+\lambda}, & x = 1 \\ 0, & x \in \mathbb{N} \setminus \{1\} \end{cases}$	ベルヌーイ ($p = \frac{\lambda}{1+\lambda}$)

表 2. $\text{CMP}(\lambda, \nu)$ の確率質量関数(pmf), 分布関数(CDF), k 次積率(k 's moment), 積率母関数(MGF), 特性関数(CF), 確率母関数(PGF).

pmf	CDF	k 's moment	MGF	CF	PGF
$\frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)}$	$\sum_{k=0}^x \frac{\lambda^k}{(k!)^\nu} \frac{1}{Z(\lambda, \nu)}$	$\sum_{j=0}^{\infty} \frac{j^k \lambda^j}{(j!)^\nu Z(\lambda, \nu)}$	$\frac{Z(e^t \lambda, \nu)}{Z(\lambda, \nu)}$	$\frac{Z(e^{it} \lambda, \nu)}{Z(\lambda, \nu)}$	$\frac{Z(t \lambda, \nu)}{Z(\lambda, \nu)}$

関数を示す。なお, COM-Poisson 分布の由来と導出については, Sellers (2023, pp.22–24) に詳しい。

2.2 標準化定数の計算

COM-Poisson 分布は、平均と分散が等しいという意味での等分散性の制約に縛られず、過分散性も過小分散性も表現できる柔軟なモデルである一方、規格化定数は閉形式でないことから、計算困難の問題を持つ。この規格化定数の計算困難を解決するために、いくつかのアプローチが提案してきた。その一つは、以下の式で与えられる有限和近似である (Shmueli et al., 2005; Sellers et al., 2019) :

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \approx \sum_{j=0}^M \frac{\lambda^j}{(j!)^\nu}, \quad M \in \mathbb{N},$$

ただし、 M は、十分大きくとった適当な自然数である。

さらに、以下の漸近展開式を利用した近似も提案されている (Minka et al., 2003; Shmueli et al., 2005; Sellers et al., 2019) :

$$Z(\lambda, \nu) = \frac{\exp(\nu \lambda^{1/\nu})}{\lambda^{(\nu-1)/(2\nu)} (2\pi)^{(\nu-1)/2} \sqrt{\nu}} \left\{ 1 + O\left(\frac{1}{\lambda^{1/\nu}}\right) \right\}, \quad \nu \leq 1, \quad \text{あるいは, } \lambda > 10^\nu.$$

他の近似法も含め、 $Z(\lambda, \nu)$ の近似法については、Sellers (2023, pp.66–69) に詳しい。

3. 推定法の紹介

3.1 一般化最小二乗法

$X \sim \text{CMP}(\nu, \lambda)$ のとき、

$$\frac{P(X = x-1)}{P(X = x)} = \frac{x^\nu}{\lambda}, \quad x \in \mathbb{N},$$

であるから、両辺対数を取って、

$$(3.1) \quad \ln \frac{P(X = x-1)}{P(X = x)} = -\ln \lambda + \nu \ln x,$$

が得られる。(3.1)の右辺は、 $-\ln \lambda$ を切片、 ν を傾き、 $\ln x$ を説明変数としたときの単回帰式と見ることができる。左辺の $\ln \frac{P(X=x-1)}{P(X=x)}$ において、 $P(X = x-1)$, $P(X = x)$ を、それぞれに対応する相対頻度を用いてノンパラメトリックに推定することにより、正規確率紙や、ワイブル確率紙のような COM-Poisson 確率紙(図 2)と、一般化最小二乗法によるパラメータ推定法が提案されている (Shmueli et al., 2005)。

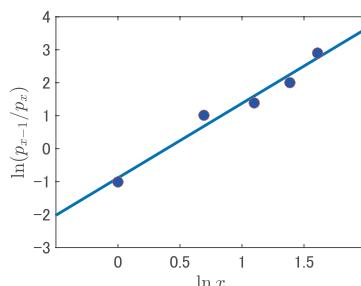


図 2. COM-Poisson 確率紙の例(図中の点は、 $X \sim \text{CMP}(\lambda, \nu)$ の観測値 x について、 $(\ln x, \ln \frac{p_{x-1}}{p_x})$ を座標とする点(ただし、 $x = 0$ は除く)、直線は、それらの点に対し、一般化最小二乗法を用いてあてはめた直線である)。

$X_i \stackrel{\text{i.i.d.}}{\sim} \text{CMP}(\nu, \lambda)$, $i \in \{1, \dots, n\}$, とする. このとき, $p_x := P(X = x)$, $x \in \mathbb{N}_0$, 及びその相対頻度を, \hat{p}_x とおく. つまり, $\hat{p}_x = \frac{\#\{i: X_i = x\}}{n}$, $x \in \mathbb{N}_0$, である(なお, 実データに適用する場合には, \hat{p}_x が 0 とならないよう補正が必要である. 著者が知る限り, 補正の方法について言及した論文はほとんどない).

すると, デルタ法により, $\ln \frac{\hat{p}_{x-1}}{\hat{p}_x}$ に関する漸近分散, 漸近共分散が以下のように得られる:

$$(3.2) \quad \text{var} \left(\ln \frac{\hat{p}_{x-1}}{\hat{p}_x} \right) \approx \frac{1}{n p_x} + \frac{1}{n p_{x-1}}, \quad \text{cov} \left(\ln \frac{\hat{p}_{x-1}}{\hat{p}_x}, \ln \frac{\hat{p}_x}{\hat{p}_{x+1}} \right) \approx -\frac{1}{n p_x}.$$

(3.2)において, p_x に, \hat{p}_x を代入し, さらに, (i, j) 成分を $\text{cov} \left(\ln \frac{\hat{p}_{i-1}}{\hat{p}_i}, \ln \frac{\hat{p}_j}{\hat{p}_{j+1}} \right)$ とする分散共分散行列を $\widehat{\mathbf{M}}$ とおく. さらに, $\mathbf{y} = \left(\ln \frac{\hat{p}_1}{\hat{p}_2}, \dots, \ln \frac{\hat{p}_{n-1}}{\hat{p}_n} \right)'$, $\mathbf{x}'_i = (1, \ln i)$, $\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_n)'$ とおくと, $\xi = (-\ln \lambda, \nu)$ の“分散共分散行列をその推定値で置き換えたときの一般化最小二乗推定量”(FGLS; feasible generalized least squares estimators) $\hat{\xi}$ が, 以下で得られる:

$$\hat{\xi}' = \arg \inf_{\theta \in \Theta} (\mathbf{y} - \mathbf{X} \xi)' \widehat{\mathbf{M}}^{-1} (\mathbf{y} - \mathbf{X} \xi) = (\mathbf{X}' \widehat{\mathbf{M}}^{-1} \mathbf{X})^{-1} \mathbf{X}' \widehat{\mathbf{M}}^{-1} \mathbf{y},$$

ただし, Θ はパラメータ空間であり, $\Theta = \{(\lambda, \nu) : \lambda > 0, \nu > 0, \text{あるいは } 0 < \lambda < 1 \text{ かつ } \nu = 0\}$ である. 以上より, θ の FGLS 推定量 $\hat{\theta}_{FGLS}$ が以下で得られる:

$$\hat{\theta}_{FGLS} = \begin{pmatrix} e^{-\hat{\xi}_{(1)}} & \hat{\xi}_{(2)} \end{pmatrix},$$

ただし, $\hat{\xi}_{(i)}$ は, $\hat{\xi}$ の第 i 成分である. 有限の n において $\ln \frac{\hat{p}_{x-1}}{\hat{p}_x}$ は正規分布に従わないので, $\hat{\theta}$ は, 最良線形不偏推定量(BLUE)といった最適性が保証された推定量ではない. 一見有用そうに見える手法であるが, 本手法の実データへの適用例は, 筆者の知る限り極めて少なく Shmueli et al. (2005) で示されているのみである. Shmueli et al. (2005) での適用例は, 頻度が 0 である数値を取り除いた作為データによる適用例しか示されていない. 実際, 相対頻度が 0 となる点を含むデータでは, 本手法はそのままでは適用できないので工夫が必要である.

3.2 最尤推定法

$X_i \stackrel{\text{i.i.d.}}{\sim} \text{CMP}(\nu, \lambda)$, $i \in \{1, \dots, n\}$, であり, $x_i, i \in \{1, \dots, n\}$, をこれらの実現値とする. このとき, $\mathbf{x} = (x_1, \dots, x_n)$ に基づく, $\theta = (\lambda, \nu)$ の対数尤度関数 $l_{\mathbf{x}}(\theta)$ は, 以下で与えられる:

$$(3.3) \quad l_{\mathbf{x}}(\theta) = \sum_{i=1}^n x_i \ln \lambda - \nu \sum_{i=1}^n \ln(x_i!) - n \ln Z(\lambda, \nu).$$

因子分解定理により,

$$\left(\sum_{i=1}^n x_i, \prod_{i=1}^n x_i! \right),$$

が, θ の十分統計量であることが, 直ちに導かれる.

$l_{\mathbf{x}}(\theta)$ の λ, ν それぞれについての 1 階偏導関数を 0 とおくことで, 以下の尤度方程式が得られる:

$$(3.4) \quad \begin{cases} \frac{\partial l_{\mathbf{x}}(\theta)}{\partial \lambda} = \frac{\sum_{i=1}^n x_i}{\lambda} - n \frac{\frac{\partial Z(\lambda, \nu)}{\partial \lambda}}{Z(\lambda, \nu)} = 0, \\ \frac{\partial l_{\mathbf{x}}(\theta)}{\partial \nu} = - \sum_{i=1}^n \ln(x_i!) - n \frac{\frac{\partial Z(\lambda, \nu)}{\partial \nu}}{Z(\lambda, \nu)} = 0, \end{cases}$$

ただし,

$$(3.5) \quad \frac{\partial Z(\lambda, \nu)}{\partial \lambda} = \sum_{j=0}^{\infty} \frac{j \lambda^{j-1}}{(j!)^{\nu}} = \frac{Z(\lambda, \nu)}{\lambda} \sum_{j=0}^{\infty} \frac{j \lambda^j}{(j!)^{\nu} Z(\lambda, \nu)} = \frac{Z(\lambda, \nu) E(X)}{\lambda},$$

$$(3.6) \quad \frac{\partial Z(\lambda, \nu)}{\partial \nu} = - \sum_{j=0}^{\infty} \frac{\lambda^j \ln(j!)}{(j!)^{\nu}} = -Z(\lambda, \nu) \sum_{j=0}^{\infty} \frac{\ln(j!) \lambda^j}{(j!)^{\nu} Z(\lambda, \nu)} = -Z(\lambda, \nu) E(\ln(X!)).$$

さらに, (3.5) と (3.6) を, (3.4) に代入して, $E(X)$ と $E(\ln X!)$ について解くと, 以下の式が得られる:

$$(3.7) \quad \begin{cases} E(X) = \lambda \bar{x}, \\ E(\ln X!) = \overline{\ln x!}, \end{cases}$$

ただし, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\overline{\ln x!} = \frac{1}{n} \sum_{i=1}^n \ln(x_i!)$ である. 尤度方程式 (3.7) は, 陽には解けないので, 数値的解法を用いて解く必要がある. 尤度方程式の解の存在条件は, Bedbur et al. (2023) において示されている.

3.3 Tomio and Nagatsuka (2022) の方法

$X_i \stackrel{\text{i.i.d.}}{\sim} \text{CMP}(\nu, \lambda)$, $i = 1, \dots, n$, であり, $x_i, i \in \mathbb{N}$, をこれらの実現値であるとする. このとき, $S_1 = s_1$ (ただし, $S_1 = \sum_{i=1}^n X_i$, $s_1 = \sum_{i=1}^n x_i$) が与えられた元での, $x = (x_1, \dots, x_n)$ に基づく ν の条件付き対数尤度関数は, 以下で与えられる.

$$l_{x|s_1}(\nu) = \nu \ln \left(\frac{s_1}{x_1, \dots, x_n} \right) - \ln \mathcal{C}_{s_1}(\nu),$$

ただし,

$$\mathcal{C}_{s_1}(\nu) = \sum_{z_1=0}^{s_1} \sum_{z_2=0}^{s_1-z_1} \cdots \sum_{z_{n-1}=0}^{s_1-\sum_{i=1}^{n-2} z_i} \left(\frac{s_1}{z_1, \dots, z_n} \right)^{\nu},$$

である.

また, $S_2 = s_2$ (ただし, $S_2 = \prod_{i=1}^n (X_i!)$, $s_2 = \prod_{i=1}^n (x_i!)$) が与えられた元での, $x = (x_1, \dots, x_n)$ に基づく ν の条件付き対数尤度関数は, 以下で与えられる.

$$l_{x|s_2}(\lambda) = \left(\sum_{i=1}^n x_i \right) \ln \lambda - \ln \sum_{z \in \Omega_n(s_2)} \lambda^{\sum_{i=1}^n z_i},$$

ただし, $\Omega_n(s) = \{(z_1, \dots, z_n) \in \mathbb{N}_0^n : \prod_{i=1}^n z_i! = s\}$, $s \in \mathbb{N}_0$, である.

$l_{x|s_1}(\nu)$ の ν についての 1 階偏導関数 $\frac{\partial}{\partial \nu} l_{x|s_1}(\nu)$ と, $l_{x|s_2}(\lambda)$ の λ についての 1 階偏導関数 $\frac{\partial}{\partial \lambda} l_{x|s_2}(\lambda)$ は,

(3.8)

$$\frac{\partial}{\partial \nu} l_{x|s_1}(\nu) = \ln \left(\frac{s_1}{x_1, \dots, x_n} \right) - \sum_{z_1=0}^{s_1} \cdots \sum_{z_{n-1}=0}^{s_1-\sum_{i=1}^{n-2} z_i} \left(\frac{s_1}{z_1, \dots, z_n} \right)^{\nu} \ln \left(\frac{s_1}{z_1, \dots, z_n} \right) \Big/ \mathcal{C}_{s_1}(\nu),$$

(3.9)

$$\frac{\partial}{\partial \lambda} l_{x|s_2}(\lambda) = - \frac{\sum_{z \in \Omega_n(s_2)} (\sum_{i=1}^n z_i - \sum_{i=1}^n x_i) \lambda^{\sum_{i=1}^n z_i - 1}}{\sum_{z \in \Omega_n(s_2)} \lambda^{\sum_{i=1}^n z_i}},$$

と求められ、これらをそれぞれ 0 とおいて、尤度方程式が得られる。尤度方程式の解の存在について、以下の定理が得られている (Tomio and Nagatsuka, 2022).

定理 1(Tomio and Nagatsuka, 2022). 任意の $\nu \in \mathbb{R}$ と、 $\mathbf{x} = (x_1, \dots, x_n)$, $x_i \in \mathbb{N}_0$, $n \in \mathbb{N} \setminus \{1\}$, について、 $s_1 = \sum_{i=1}^n x_i$ が与えられた元での ν の条件付き尤度方程式 $\frac{\partial}{\partial \nu} l_{\mathbf{x}|s_1}(\nu) = 0$ の解は、 \mathbf{x} の要素の内、少なくとも $n - 1$ 個の要素が 0 でないか、 $\mathbf{x} = \arg \max_{\mathbf{z} \in D(s_1)} \binom{s_1}{z_1, \dots, z_n}$ でない限り存在し、一意である。ただし、 $D(s) = \{(z_1, \dots, z_n) : \sum_{i=1}^n z_i = s, z_i \in \mathbb{N}_0\}$ である。 \mathbf{x} の要素の内、少なくとも $n - 1$ 個の要素が 0 の場合には、解は $-\infty$ に発散し、 $\mathbf{x} = \arg \max_{\mathbf{z} \in D(s_1)} \binom{s_1}{z_1, \dots, z_n}$ の場合には、解は ∞ に発散する。

定理 2(Tomio and Nagatsuka, 2022). 任意の $\lambda \in \mathbb{R}$ と、 $\mathbf{x} = (x_1, \dots, x_n)$, $x_i \in \mathbb{N}_0$, $n \in \mathbb{N} \setminus \{1\}$, について、 $s_2 = \prod_{i=1}^n (x_i!)$ が与えられた元での λ の条件付き尤度方程式 $\frac{\partial}{\partial \lambda} l_{\mathbf{x}|s_2}(\lambda) = 0$ の解は、 $\mathbf{x} = \arg \min_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ でないか、 $\mathbf{x} = \arg \max_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ でない限り存在し、一意である。 $\mathbf{x} = \arg \min_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ の場合には、解は 0 に収束し、 $\mathbf{x} = \arg \max_{\mathbf{z} \in \Omega_n(s_2)} \sum_{i=1}^n z_i$ の場合には、解は ∞ に発散する。

ただし、尤度方程式(3.8), (3.9)に含まれる和の計算は、NP 困難問題であり、 n がある程度大きくなると計算困難に陥る。この推定法が適用できるのは、多くとも n が 12 程度のデータであることが (Tomio and Nagatsuka, 2022) によって指摘されている。なお、同様の問題は、第 4.2 で紹介する sCOM-Poisson 分布の確率質量関数の計算においても起こる。

4. COM-Poisson 分布の応用

本節では、COM-Poisson 分布の応用例として、Cure rate モデルへの応用と、確率過程モデルへの応用について紹介する。

4.1 COM-Poisson 分布の治癒率(Cure rate)モデルへの応用

生存時間分析や信頼性データ解析の分野において、治癒率(cure rate)モデル(長期間生存モデル, long-term survival model とも呼ばれる)は、広く研究されてきたが、COM-Poisson 分布に基づく治癒率モデルは、Rodrigues et al. (2009)において提案された。そして、これまで多くの研究が行われてきた (Balakrishnan and Pal, 2013; Wiangnak and Pal, 2018; Majakwara and Pal, 2019; He and Emura, 2019; Balakrishnan et al., 2024).

ここでは、まず、治癒率モデルについて概説する。その後、COM-Poisson 分布に基づく治癒率モデルについて紹介する。

4.1.1 治癒率モデルについて

生存時間分析や信頼性工学では、対象についてあるイベント(死亡、疾病の発症、故障、等)が生じるまでの時間に興味があることが多い。しかし、このようなイベントが、一つの要因により生じるとは限らず、多くの場合、複数の要因に起因して生じる。例えば、ある患者を観察中、心疾患による死亡までの時間に興味があったにも拘らず、脳疾患により死亡する場合がある。このような時、興味があった心疾患による死亡までの時間のデータは、完全には得られない(このとき得られるのは、打切りデータ)。このように、興味がある要因によって生じるイベントの観察において、その他の要因に起因するイベントによって観察が打切りになってしまうリスクを、競合リスクと呼ぶ。

ある要因に起因するイベントが一度起きた後、二度と起こらないことがありうる。例えば、ある疾病の治療後、その疾病の再発が起こりえない場合や、あるバグによるソフトウェアの

エラーを修復後、そのバグによるエラーの再発が起りえない場合である。これらを「治癒(cure)」と表現し、治癒する割合を「治癒率(cure rate)」と呼ぶ。治癒率を取り入れた生存関数で表現される治癒率モデル(cure rate model)は、がん臨床試験のデータを解析する際に、これまでよく用いられてきた。これは、乳がん、非ホジキンリンパ腫、白血病、前立腺がん、メラノーマ、頭頸部がん等の疾病においては、患者のかなりの割合が治癒する(再発がない)とされているためである(Ibrahim et al., 2001)。「治癒」とは、生物医学分野においては、その名の通り治癒した、あるいは病気の再発がないことを意味し、工学分野においては、故障やエラーの再発がないことを意味する。

これまで研究してきた治癒率モデルは、混合治癒率モデル(mixture cure rate model)と促進時間治癒率モデル(promotion time cure rate model)に大別される。

混合治癒率モデル(mixture cure rate model)は、治癒した対象と、イベントが起りうる対象に関する混合モデルである(Boag, 1949; Berkson and Gage, 1952)。混合治癒率モデルでは、時点 y における累積分布関数と生存関数は、それぞれ以下のように表される(Boag, 1949; Berkson and Gage, 1952; Ibrahim et al., 2001)。

$$F_p(y) = (1 - p_0)F_s(y), \\ S_p(y) = 1 - (1 - p_0)F_s(y) = p_0 + (1 - p_0)S_s(y),$$

ただし、 $F_p(\cdot)$ と $S_p(\cdot)$ はそれぞれ、対象全体の分布関数と生存関数を示し、 $F_s(\cdot)$ と $S_s(\cdot)$ は、それぞれ、イベントが起りうる対象のみの分布関数と生存関数を示す。また、 p_0 は、全体に対する治癒する対象の割合で、治癒率を表す。

促進時間治癒率モデル(promotion time cure rate model)は、イベント発生に関連する M 個の競合リスク(競合する要因)を仮定し、さらに M はポアソン分布に従う確率変数とするモデルである(Chen et al., 1999)。 W_j を j 番目の競合リスクによるイベント発生までの時間(つまり「生存時間」とし、 $j = 1, \dots, M$ 、とする。 $M = m$ が与えられたとき、 W_j は独立かつ同一の分布にしたがい、分布関数を $F(\cdot)$ 、生存関数を $S(\cdot)$ とする。また、 $W_j, j = 1, \dots, m$ 、は M と独立であるとする。

競合リスクの数は、観測できない場合が多いので M は潜在変数であると考えるのが妥当である。また、対象の生存時間は、競合リスクにおける潜在故障時間モデルの仮定(Cox and Oakes, 1984)の下で、 $Y = \min(W_0, W_1, W_2, \dots, W_M)$ (W_0 については後で説明)で与えられるため、一つの競合リスク j における生存時間 W_j の観測値は得られるが、それ以外の観測値は得られない。よって、 $W_j, j = 1, \dots, m$ 、も潜在変数と考えるのが妥当である。なお、 W_0 は、治癒したイベントの生存時間を表し、これは、再発しないイベントの生存時間であるから、 $P(W_0 = \infty) = 1$ である。 $p_0 := P(M = 0)$ は、全体に対する治癒する対象の割合で、治癒率を表す。

M が平均 λ のポアソン分布にしたがう時、促進時間治癒率モデルにおける生存関数は次のように表される：

$$S_p(y) = \exp(-\lambda F(y)).$$

また、イベントが起りうる対象のみの生存関数 $S_s(\cdot)$ は、以下で与えられる：

$$S_s(y) = \frac{\exp(-\lambda F(y)) - \exp(-\lambda)}{1 - \exp(-\lambda)}.$$

なお、 $S_p(y)$ と $S_s(y)$ の導出については、Chen et al. (1999)を参照のこと。

このとき、 $S_p(y)$ と $S_s(y)$ の間には、以下の関係が成り立つ：

$$S_p(y) = \exp(-\lambda) + \{1 - \exp(-\lambda)\} S_s(y).$$

これは、混合治癒モデルの生存時間関数でもある。以上より、 M がポアソン分布にしたがうと仮定すると、促進時間治癒率モデルは、混合治癒率モデルにもなり、治癒率モデルを統一的に表現できる利点が生まれる。

4.1.2 COM-Poisson 分布に基づく治癒率モデル

前節で説明した促進時間治癒率モデルと混合治癒率モデルの統一モデルにおいて、 M が従う分布を COM-Poisson 分布に拡張した治癒率モデルが、Rodrigues et al. (2009) により提案された。このモデルでは、 M に、より柔軟なモデルを仮定している一方、促進時間治癒率モデルと混合治癒率モデルの統一モデルになるとの特徴は失われていることに注意されたい。以下にこのモデルについて、詳述する。

改めて定義を行う。 M は、競合リスクの種類数を表す確率変数で、 M の確率質量関数を $f_M(m) = P(M = m)$, $m \in \mathbb{N}_0$, とする。 $M = m$ が与えられたもとで、確率変数列 W_j , $j = 1, \dots, m$, を考える。 W_j は、 j 番目の競合リスクによる生存時間を表す確率変数で、互いに独立、かつ M とも独立で、同一の分布に従うとする。

競合リスクにおける潜在故障時間モデルの仮定 (Cox and Oakes, 1984) の下で、競合リスク数 M や、競合リスク j による生存時間 W_j は潜在変数であり、対象全体の生存時間 Y は以下で与えられる：

$$Y = \min\{W_0, W_1, \dots, W_M\},$$

ただし、 W_0 は、治癒したイベントの生存時間を表し、これは、再発しないイベントの生存時間であるから、 $P(W_0 = \infty) = 1$ である。このようなリスクの存在は、 $P(M = 0) > 0$ であることを示唆する。 $p_0 := P(M = 0)$ は治癒率である。 $M \sim \text{CMP}(\lambda, \nu)$ の仮定の下で、治癒率は、

$$p_0 = f_{\text{CMP}}(0) = \frac{1}{Z(\lambda, \nu)},$$

となる。さらに、 W_j , $j \in \mathbb{N}$, が従う分布の生存関数を $S(y) := P(W_j \geq y)$ とすると、 Y の生存関数 $S_p(\cdot)$ は、

$$S_p(y) = P(Y \geq y) = \sum_{m=0}^{\infty} P(M = m) S(y)^m = \frac{Z(\lambda S(y), \nu)}{Z(\lambda, \nu)},$$

とシンプルな形で書ける。さらに、 Y の確率密度関数 $f_p(\cdot)$ は、

$$f_p(y) = -\frac{\partial}{\partial y} S_p(y) = \frac{1}{Z(\lambda, \nu)} \frac{f(y)}{S(y)} \sum_{j=1}^{\infty} \frac{j(\lambda S(y))}{(j!)^{\nu}},$$

となる。なお、 $S(y)$ として、通常、パラメトリック分布の生存関数が選ばれ、特に生存時間分析での生存時間や信頼性工学における寿命を表すのによく用いられるワイブル分布、対数正規分布、ガンマ分布が良く用いられる (Balakrishnan and Pal, 2013, 2015b, 2016)。ワイブル分布とガンマ分布の一般化分布である一般化ガンマ分布を用いた例もみられる (Balakrishnan and Pal, 2015a)。 $\lambda, \nu, S(y)$ が持つパラメータ、並びに後述する共変量モデルにおける回帰係数の推定には、最尤法が用いられる場合がほとんどである (Rodrigues et al., 2009; Balakrishnan and Pal, 2013, 2015a, 2015b)。

4.1.3 共変量を持つ治癒率モデル

共変量を持つ治癒率モデルについても、これまで研究がなされてきた。治癒率モデルに共変

量を導入する方法として、最初に考えられるのは、GLIM(一般化線形モデル)の枠組みにおいてリンク関数を対数関数に取る方法、つまり、 $\mathbf{x}'_i = (1, x_{i1}, \dots, x_{ip})$ を i 番目の対象の共変量、 $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ として、パラメータ λ を、 $\exp(\mathbf{x}'_i \beta)$ におきかえた共変量モデルである (Balakrishnan and Pal, 2012; Cancho et al., 2012; Rodrigues et al., 2009)。しかし、 $\nu = 0$ のとき、COM-Poisson 分布は幾何分布となり、 $\lambda < 1$ とならなければいけないが、上記の方法だと、 $\exp(\mathbf{x}'_i \beta)$ が 1 以上になりうる。そこで、考え出されたのが、対象が i のときの治癒率 $p_0^{(i)}$ と共変量 x_i の間のリンク関数をロジット関数にとった

$$p_0^{(i)} = \frac{1}{1 + \exp(\mathbf{x}'_i \beta)} = \frac{1}{Z(\lambda, \nu)}$$

とするモデルである (Balakrishnan and Pal, 2013, 2015a, 2015b; Balakrishnan and Feng, 2018)。

COM-Poisson 分布に基づく、共変量を持つ治癒率モデルの適用例として、がんの再発に関するデータへの適用例 (Rodrigues et al., 2009) が良く知られている。この適用例では、悪性がんの一種である皮膚黒色腫を持つ患者に、薬剤(インターフェロンアルファ-2b)を利用したときの治療効果を評価するため、腫瘍の厚さを共変量とし、患者の生存時間データに対しモデルが適用されている。

4.2 COM-Poisson 分布の確率過程モデルへの応用

ポアソン分布は、再生性を持つためポアソン分布に基づくレヴィ過程である定常ポアソン過程を構築できる。一方、COM-Poisson 分布は、 $\nu = 1$ つまり、ポアソン分布の時を除いて再生性を持たないばかりか、 $\nu = 0$ 、あるいは $\nu = 1$ でない限り、無限分解可能(infinite divisible)ではない (Geng and Xia, 2022)。よって、特殊ケースを除き、COM-Poisson 分布に基づくレヴィ過程を構築することができない。

しかし、COM-Poisson 分布の代わりに、COM-Poisson 分布にしたがう *i.i.d.* 確率変数の和が従う分布である sCOM-Poisson 分布(Sum of COM-Poisson 分布) (Sellers et al., 2017) は、再生性を持ち、これに基づくレヴィ過程を構築することが可能である (Zhu et al., 2017)。

$X_i \stackrel{\text{i.i.d.}}{\sim} \text{CMP}(\nu, \lambda)$, $i = 1, \dots, n$, のとき、 $S = \sum_{i=1}^n X_i$ が従う分布は、sCOM-Poisson 分布と呼ばれ、以下の確率質量関数を持つ:

$$f_{\text{sCMP}}(s) = \frac{\lambda^s}{(s!)^\nu Z(\lambda, \nu)^n} \sum_{x_1 + \dots + x_n = s} \binom{s}{x_1 \dots x_n}, \quad s = 0, \dots, n.$$

パラメータは、 $\lambda > 0$, $\nu \geq 0$, $n \in \mathbb{N}$ であり、これらのパラメータを持つ sCOM-Poisson 分布を、本稿では、sCMP(λ, ν, n) で表す。

計数過程 $\{N(t), t \geq 0\}$ が以下の条件を満たす時、 $\{N(t), t \geq 0\}$ を、パラメータ λ, ν の定常 COM-Poisson 過程と呼ぶ (Zhu et al., 2017) :

- (1) $N(0) = 0$, *a.s.*
- (2) $N(s+t) - N(s) \sim \text{sCMP}(\lambda, \nu, n)$.
- (3) 独立、かつ定常増分を持つ。

パラメータ λ, ν の定常 COM-Poisson 過程は、 $\nu = 1$ のとき、パラメータ λ の定常ポアソン過程となるので、定常ポアソン過程の一般化モデルである。

ただし、sCOM-Poisson 分布の確率質量関数に含まれる $\sum_{x_1 + \dots + x_n = s} \binom{s}{x_1 \dots x_n}$ の計算は、 n がある程度大きくなると計算困難に陥る。このモデルを実データ解析に用いるには、この計算困難問題解決する必要があるが、著者の知る限り未解決である。

5. 最後に

本稿では、COM-Poisson 分布、並びに近年提案された手法も含む代表的な COM-Poisson 分布の推定法について紹介してきた。COM-Poisson 分布の区間推定や仮説検定については、Sellers (2023) でいくつか触れられているものの、点推定に比べると研究はまだ少ない。COM-Poisson 分布の区間推定、仮説検定において用いることができるピボットが見つかっていないため、推定量の漸近性質を用いた近似法、並びにブートストラップ法が有効であると考えられる。ちなみに、著者は現在、規格化定数 $Z(\lambda, \nu)$ に依存しない新たな条件付き尤度関数を導出し、それに基づく点推定、区間推定、並びに検定法を提案した論文を投稿中である。この手法は、サンプル数 n が大きくても実行可能であり、実データ解析に問題なく適用可能である。COM-Poisson 分布の区間推定や仮説検定の方法も含めて、別の機会があれば紹介するかもしれない。

謝 辞

有意義かつ適切な助言を頂いた3名の匿名査読者に深謝いたします。本研究の一部は、JSPS 科研費 基盤研究(C) (24K07952)の助成を受けました。ここに記して感謝の意を表します。

参 考 文 献

- Balakrishnan, N. and Feng, T. (2018). Proportional odds under Conway-Maxwell-Poisson cure rate model and associated likelihood inference, *Statistics, Optimization & Information Computing*, **6**, 305–334.
- Balakrishnan, N. and Pal, S. (2012). EM algorithm-based likelihood estimation for some cure rate models, *Journal of Statistical Theory and Practice*, **6**, 698–724.
- Balakrishnan, N. and Pal, S. (2013). Lognormal lifetimes and likelihood-based inference for flexible cure rate models based on COM-Poisson family, *Computational Statistics & Data Analysis*, **67**, 41–67.
- Balakrishnan, N. and Pal, S. (2015a). An EM algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood- and information-based methods, *Computational Statistics*, **30**, 151–189.
- Balakrishnan, N. and Pal, S. (2015b). Likelihood inference for flexible cure rate models with gamma lifetimes, *Communications in Statistics — Theory and Methods*, **44**, 4007–4048.
- Balakrishnan, N. and Pal, S. (2016). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes, *Statistical Methods in Medical Research*, **25**, 1535–1563.
- Balakrishnan, N., Barui, S. and Milienos, F. S. (2024). Piecewise linear approximations of baseline under proportional hazards based COM-Poisson cure models, *Communications in Statistics — Simulation and Computation*, **53**, 1135–1160.
- Bedbur, S., Kamps, U. and Imm, A. (2023). On the existence of maximum likelihood estimates for the parameters of the Conway-Maxwell-Poisson distribution, *Latin American Journal of Probability and Mathematical Statistics*, **20**, 561–575.
- Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association*, **47**, 501–515.
- Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society, Series B*, **11**, 15–53.
- Cancho, V. G., de Castro, M. and Rodrigues, J. (2012). A Bayesian analysis of the Conway-Maxwell-Poisson cure rate model, *Statistical Papers*, **53**, 165–176.

- Chen, M. H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association*, **94**, 909–919.
- Conway, R. W. and Maxwell, W. L. (1962). A queuing model with state dependent service rates, *Journal of Industrial Engineering*, **12**, 132–136.
- Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, London.
- Geng, X. and Xia, A. (2022). When is the Conway-Maxwell-Poisson distribution infinitely divisible?, *Statistics and Probability Letters*, **181**, <https://doi.org/10.1016/j.spl.2021.109264>.
- Hanleya, J. A. and Bhatnagar, S. (2022). The “Poisson” distribution: History, reenactments, adaptations, *The American Statistician*, **76**, 363–371.
- He, Z. and Emura, T. (2019). The COM-Poisson cure rate model for survival data-computational aspects, *Journal of the Chinese Statistical Association*, **57**, 1–42.
- Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian Survival Analysis*, Springer, New York.
- Majakwara, J. and Pal, S. (2019). On some inferential issues for the destructive COM-Poisson-generalized gamma regression cure rate model, *Communications in Statistics — Simulation and Computation*, **48**, 3118–3142.
- Minka, T. P., Shmueli, G., Kadane, J. B., Borle, S. and Boatwright, P. (2003). Computing with the COM-Poisson distribution, Technical Report 776, Department of Statistics, Carnegie Mellon University, Pittsburgh.
- Nightingale, F. (1858). *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army, Founded Chiefly on the Experience of the Late War*, Harrison of Sons, London.
- Rodrigues, J., de Castro, M., Cancho, V. G. and Balakrishnan, N. (2009). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data, *Journal of Statistical Planning and Inference*, **139**, 3605–3611.
- Sellers, K. F. (2023). *The Conway-Maxwell-Poisson Distribution*, Cambridge University Press, Cambridge.
- Sellers, K. F., Swift, A. W. and Weems, K. S. (2017). A flexible distribution class for count data, *Journal of Statistical Distributions and Applications*, **4**, <https://doi.org/10.1186/s40488-017-0077-0>.
- Sellers, K. F., Lotze, T. and Raim, A. (2019). COMPoissonReg: Conway-Maxwell-Poisson (COM-Poisson) regression, version 0.7.0., <https://cran.r-project.org/web/packages/COMPoissonReg/index.html> (最終アクセス日 2025 年 2 月 15 日).
- Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution, *The Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **54**, 127–142.
- Snow, J. (1858). *On Chloroform and Other Anaesthetics and Their Action and Administration*, John Churchill, London.
- Tomio, Y. and Nagatsuka, H. (2022). A conditional maximum likelihood estimation of the COM-Poisson distribution and its uniqueness and existence, *Total Quality Science*, **7**, 137–148.
- von Bortkewitsch, L. (1898). *Das Gesetz der Kleinen Zahlen*, Teubner, Leipzig.
- Wani, M. K. and Ahmad, P. B. (2024). One-inflated zero-truncated Poisson distribution: Statistical properties and real life applications, *Annals of Data Science*, **71**, <https://doi.org/10.1007/s40745-024-00526-3>.
- Wiangnak, P. and Pal, S. (2018). Gamma lifetimes and associated inference for interval-censored cure rate model with COM-Poisson competing cause, *Communications in Statistics — Theory and Methods*, **47**, 1491–1509.
- Zhu, L., Sellers, K. F., Morris, D. S. and Shmueli, G. (2017). A generalized stochastic process for count data, *The American Statistician*, **71**, 71–80.

On COM-Poisson Distribution

—An Extended Poisson Model for Under- and Over-dispersed Count Data—

Hideki Nagatsuka

Department of Data Science for Business Innovation, Chuo University

The Poisson distribution is one of the most popular distributions for modeling count data. While the Poisson distribution is a simple model and easy to treat, its applicability is often limited due to the equi-dispersion assumption. The Conway-Maxwell-Poisson (abbreviated as COM-Poisson) distribution, provided by Conway and Maxwell in 1962, is a more flexible model for modeling count data, which generalizes the Poisson distribution by adding a parameter to model over-dispersion and under-dispersion. In the survival analysis and reliability analysis, there have been many papers in recent years on cure rate models using the COM-Poisson distribution, which has attracted a lot of attention. The COM-Poisson distribution is a simple model with only two parameters, but it can flexibly be fitted a wide range of overdispersion and underdispersion phenomena and includes the Poisson and geometric distributions as special cases, and the Bernoulli distribution as a limiting case. These features distinguish it from other generalized models of the Poisson distribution. However, the price of these benefits is that the normalization constant of the probability function includes an infinite series, which makes estimation and testing difficult. So far, various approximation methods for the normalization constant or estimation methods that avoid calculating the normalization constant have appeared in the literature. In this paper, we introduce the COM-Poisson distribution and its prominent estimation methods for the COM-Poisson distribution, including recently proposed methods appeared in recent years.

共役ガンマ事前分布を用いた ワイブル分布のベイズ推定

作村 建紀¹・柳本 武美²

(受付 2024 年 11 月 29 日；改訂 2025 年 4 月 13 日；採択 4 月 14 日)

要　　旨

ベイズ推論における推定量の構築において、推定対象となる母数の適切な選択と事前密度の選択は重要な課題である。前者においては、指型分布族では自然母数が候補になる。後者については、弱い情報しか含まない場合はしばしば共役な事前密度が仮定される。強い客観性が望まれると無情報事前密度が仮定される。これは情報が極端に少ない共役事前分布の極限の状態として理解できる。本稿では、2母数のワイブル分布に対して、共役な事前分布としての共役ガンマ事前密度と無情報事前分布としての Jeffreys 事前密度を連続的に繋がる形で導入する。その上で、指型分布族における最適性を有するベイズ推定量に基づき、ワイブル分布のベイズ推定量を提案する。提案推定量の性能は二乗損失とカルバッカ・ライブラー損失のもとでのリスク、および推定量の偏りによって評価される。これらは数値実験によって最尤推定量と比較される。リスク比較の結果は提案推定量が有望な推定量であることを支持する。提案推定量の近似的な不偏性も観察される。また、実データへの適用の結果は推定量の選択の必要性を示唆する。

キーワード：自然母数、共役事前分布、事後平均、無情報事前分布、リスク比較。

1. はじめに

ワイブル分布は信頼性工学分野で重要な分布であり、特に形状パラメータの推定に関心を持たれる。ワイブル分布のハザード関数は形状パラメータの値によって柔軟に変化するためである。また、加速試験などの信頼性試験においても、寿命分布にワイブル分布を基礎とした拡張モデルを考えることができる (Nelson, 1990)。ワイブル分布は次の密度を持つ。

$$f(x|\mu, \tau) = \tau \mu^{-\tau} x^{\tau-1} e^{-(x/\mu)^\tau}, \quad (x > 0)$$

ここで、 $\mu > 0, \tau > 0$ はそれぞれ尺度パラメータ、形状パラメータと呼ばれる。この密度は、 τ を固定したとき指型分布族に属する。任意の実数 $a > 0$ について、 x^a の期待値は $\mu^a \Gamma(1 + a/\tau)$ となる。 $\tau = 1$ のとき、密度 $\mu^{-1} e^{-x/\mu}$ の指型分布になる。母数 (μ, τ) の推定にはこれまで多くの推定量が提案されている。中でも最尤推定量 (MLE) がよく知られているが、シンプルな形式では表せない (Johnson et al., 1994)。

ワイブル分布の母数推定は 3 つの極値分布族の一つであるゲンベル分布と強い関係があ

¹ 法政大学 理工学部：〒184-8584 東京都小金井市梶野町 3-7-2; sakumura@hosei.ac.jp

² 統計数理研究所 名誉教授：〒190-8562 東京都立川市緑町 10-3; yanagmt@ism.ac.jp

る。すなわち、確率変数 Y が期待値 $\mu_y - \psi(1)/\tau_y$ のゲンベル分布に従うとき、 e^{-Y} は期待値 $e^{-\mu_y} \Gamma(1+1/\tau_y)$ のワイブル分布に従う。ここで、 $\psi(\cdot)$ はディガンマ関数である。ワイブル分布の母数 (μ, τ) とは、 $(\mu_y, \tau_y) = (-\log \mu, \tau)$ の関係になる。ワイブル分布での母数推定の議論は、ゲンベル分布にも大いに関係することは興味深い。

本研究では、このワイブル分布の母数 (μ, τ) に関する推定をベイズ推論の立場で考える。ベイズ推定では事前分布の選択が重要である。ワイブル分布はその密度関数や分布関数が初等関数で書くことができるにも関わらず、事前分布の仮定は難しい。過去の十分なデータがあれば、安定した適切な事前分布が導かれる。一方そうでなければ、無情報事前分布が考えられる。たとえば、Jeffreys 事前分布は一つの候補である。また、参照事前分布についてのワイブル分布に関連する先行研究も存在する (Bernardo, 1979; Sun, 1997; Xu et al., 2015)。本研究では、無情報事前分布として Jeffreys 事前分布を導入する。また、 μ に対して共役な事前分布、 τ に対してガムマ事前分布とする弱情報事前分布との連結を試みる。

ベイズ推定で重要なもう一つの点は、推定対象の選択である。母数は (μ, τ) に限らず、 $(1/\mu, 1/\tau)$, (μ^τ, τ) などいろいろと考えられる。どれを対象とするのかによってその事後平均は異なる推定量を誘導する。指數型分布族に対しては自然母数を推定対象とし、その事後平均を考えることで最適性を満たす推定量を構築できる (Yanagimoto and Ohnishi, 2009)。たとえば、平均 u 、分散 $1/v$ の正規分布に従う標本 $\mathbf{w} = (w_1, \dots, w_n)$ を考える。標本平均を \bar{w} とする。参照事前分布 v^{-1} のもとで、自然母数 (uv, v) の事後平均は、 $\hat{w}v = \hat{v}\bar{w}$, $\hat{v} = (n-1)/\sum_i(w_i - \bar{w})^2$ になる。 u の推定量 \hat{u} は $\hat{u} = \hat{w}v/\hat{v} = \bar{w}$ として求めると、 (\hat{u}, \hat{v}) をプラグインした予測分布と $(\hat{w}v, \hat{v})$ による予測分布が共通になることが示される。 (\hat{u}, \hat{v}) をプラグインした予測子は最適性を有する。さらに、 $(\hat{u}, 1/\hat{v})$ は $(u, 1/v)$ の不偏推定量になる (Sakumura and Yanagimoto, 2024)。これは、適切な推定対象の選択が重要であることを示唆するものである。ワイブル分布は τ が未知のときには指數型分布族に属さないため、これは難しい問題になる。本研究では、自然母数に着想を得た推定量を考える。推定量の性能はリスクで評価される。

第 2 節では共役な事前分布と無情報事前分布を滑らかにつなげる事前分布の導出を行う。第 3 節では、その事前分布のもとで提案推定量を導出する。その性能は損失関数のもとでのリスク比較によって第 4 節で検討される。また、偏りについても検討する。第 5 節で実データの例が与えられる。第 6 節で本提案推定量についての包括的な考察を行う。

2. 事前分布の導出

母数 η によるサイズ n の指數型分布族の標本を $\mathbf{y} = (y_1, \dots, y_n)$ と書く。その確率密度関数は

$$p(\mathbf{y}|\eta) = \exp\{n(t(\mathbf{y})\eta - M(\eta))\}a(\mathbf{y})$$

で表される。 $t(\mathbf{y})$ は十分統計量の関数、 η は自然母数、 $a(\mathbf{y})$ は正規化項である。また、 $E[t(\mathbf{y})] = \frac{\partial}{\partial \eta} M(\eta)$ である。このとき、共役事前密度は適当な $\delta \geq 0$ によって、

$$\pi(\eta; \eta_0, \delta) = \exp\{-\delta D(\eta_0, \eta)\}b(\eta)e^{-k(\eta_0, \delta)}$$

と定義される。ここで、 η_0 は η の真値を想定した推測値である。 δ はその想定に対する信念の度合いを表している。また、 $D(\eta_0, \eta) = E[\log(p(\mathbf{y}|\eta_0)/p(\mathbf{y}|\eta)); p(\mathbf{y}|\eta_0)]$ であり、 $E(g; f)$ は密度 f に関する関数 g の期待値を意味する。この事前密度では η_0 , δ , $b(\eta)$ を選ぶ必要がある。 δ が小さくなると、事前分布として $b(\eta)$ を仮定した場合とほぼ同等である。通常は無情報な状態になる $\delta = 0$ の場合を、この事前分布族の極限として含む。関数 $b(\eta)$ は台測度関数である。したがって、 $b(\eta)$ は無情報事前分布と見なされる。

ワイブル分布からのサイズ n の標本を $\mathbf{x} = (x_1, \dots, x_n)$ と書く。前述の指數型分布族の表記では、固定した τ のもとで、 $t(\mathbf{x}) = -\sum_i x_i^\tau/n$, $\eta = 1/\mu^\tau$, $M(\eta) = -\log \eta$ と対応する。また、 $E[x_i^\tau] = \mu^\tau$ である。 $\hat{\mu}_m^\tau = \sum x_i^\tau/n$ とする。 $\hat{\mu}_m$ は μ の最尤推定量である。すると、 $t(\mathbf{x}) = -\hat{\mu}_m^\tau$ であるから、標本 \mathbf{x} の幾何平均を $\tilde{x} = \prod_i x_i^{1/n}$ として、標本密度は次式になる。

$$(2.1) \quad p(\mathbf{x}|\eta, \tau) = \eta^n \tau^n \tilde{x}^{n(\tau-1)} e^{-n\eta\hat{\mu}_m^\tau}.$$

自然母数は $1/\mu^\tau$ であり、対応する十分統計量は $\sum x_i^\tau$ である。

次に、 $\eta_0 = 1/m^\tau$ として、 $E[\hat{\mu}_m^\tau; p(\mathbf{x}|\eta_0)] = m^\tau$ より、 $D(\eta_0, \eta) = m^\tau \eta - \log(m^\tau \eta) - 1$ となるから、固定した τ のもとでの η の共役事前分布を次で定義する。 $\delta_1 > 0$ に対して、

$$\pi(\eta|\tau; m, \delta_1) = e^{-\delta_1(m^\tau \eta - \log(m^\tau \eta) - 1)} b(\eta) e^{-k(m^\tau, \delta_1)}$$

とする。ここで、 m は μ の真値と想定した推測値、 δ_1 はその想定に対する信念の度合いを表している。 $k(m^\tau, \delta_1)$ は正規化定数である。 $\delta_1 = 0$ のときは $b(\eta)$ に比例するとする。通常、 $b(\eta)$ には一様分布が仮定される (Diaconis and Ylvisaker, 1979) が、不变性を担保した無情報事前分布としては Jeffreys 事前分布があり、1母数の推論においては魅力的な事前分布である。よって、 $b(\eta)$ に Jeffreys 事前分布 $\pi_J(\eta) = 1/\eta$ を仮定する。

また、 τ に対してはガンマ事前分布として、次で定義する。

$$\pi(\tau; t, \delta_2) = (t\delta_2)^{\delta_2} \tau^{\delta_2-1} e^{-t\delta_2\tau} / \Gamma(\delta_2)$$

ここで、 t は $1/\tau$ の真値と想定した推測値、 δ_2 は δ_1 と同様に想定した値に対する信念の度合いを表す。 $\delta_2 = 0$ のときは $1/\tau$ に比例するとする。

以上から、 (η, τ) の事前分布を $\pi(\eta, \tau) = \pi(\eta|\tau; m, \delta_1)\pi(\tau; t, \delta_2)$ と仮定する。 $\delta_1 = 0$ の場合を考えると、 $\pi(\eta|\tau; m, \delta_1) \propto 1/\eta$ であるから、 (η, τ) の事前分布は、

$$(2.2) \quad \pi(\eta, \tau) \propto \eta^{-1} (t\delta_2)^{\delta_2} \tau^{\delta_2-1} e^{-t\delta_2\tau} / \Gamma(\delta_2)$$

となる。 $\delta_1 = \delta_2 = 0$ の場合、 $\pi(\eta, \tau) \propto \eta^{-1} \tau^{-1}$ となる。

$\delta_1 = 0$ のとき、事後密度は次のように書ける。

$$(2.3) \quad \pi(\eta, \tau|\mathbf{x}) = \eta^{n-1} \tilde{x}^{n\tau} \tau^{n+\delta_2-1} e^{-n\eta\hat{\mu}_m^\tau - t\delta_2\tau} / k(\mathbf{x}, \delta_2),$$

ここで、 $k(\mathbf{x}, \delta_2)$ は正規化定数である。

3. 提案推定量

本節では、前節で定義した $\delta_1 = 0$ のときの事後密度 $\pi(\eta, \tau|\mathbf{x})$ のもとで、ワイブル分布の母数 (μ, τ) に対するベイズ推定量を提案する。ベイズ推定理論では、適当な関数 g_1, g_2 について $(g_1(\mu, \tau), g_2(\mu, \tau))$ を求め、その事後平均により推定する。前述したように、指數型分布族においては自然母数が g_1, g_2 の候補になる。自然母数の事後平均による推定量をプラグインした予測子は、二乗損失、およびカルバック・ライブラー損失の事後平均をそれぞれ最小にするという 2 つの最適性を満たす。これらの 2 つの最適性から、自然母数の事後平均の良好な性能を期待できる。

そこで、まず τ を固定したときの自然母数 η の事後平均 $\check{\eta}(\tau)$ を考えると、式(2.3)より、 $k(\mathbf{x}, \delta_2) = \tilde{x}^{n\tau} \tau^{n+\delta_2-1} e^{-t\delta_2\tau} \int \eta^{n-1} e^{-n\eta\hat{\mu}_m^\tau} d\eta$ であるから、

$$\check{\eta}(\tau) = \int \eta^n e^{-n\eta\hat{\mu}_m^\tau} d\eta / \int \eta^{n-1} e^{-n\eta\hat{\mu}_m^\tau} d\eta$$

を得る。ここで、 $n\eta\hat{\mu}_m^\tau = r$ とおいて、ガンマ関数を利用することで、右辺の分子は $(n\hat{\mu}_m^\tau)^{-n-1}\Gamma(n+1)$ となる。同様に、右辺の分母は $(n\hat{\mu}_m^\tau)^{-(n-1)-1}\Gamma(n)$ となるから、

$$\check{\eta}(\tau) = 1/\hat{\mu}_m^\tau$$

が得られる。この推定量は好ましい性質を持つことが期待できる。一方、 τ の推定量 $\hat{\tau}$ は $\pi(\eta, \tau)$ のもとでの τ の事後平均として求める。

$\delta_1 = 0$ のとき、周辺事後密度 $\int \pi(\eta, \tau | x) d\eta$ は、 $\tau^{n+\delta_2-1} \tilde{x}^{n\tau} \hat{\mu}_m^{-n\tau} e^{-t\delta_2\tau}$ に比例するから、 τ の事後平均は次式で得られる。

$$(3.1) \quad \hat{\tau} = \int_0^\infty \left(\frac{\tilde{x}^\tau}{\hat{\mu}_m^\tau} \tau \right)^n \tau^{\delta_2} e^{-t\delta_2\tau} d\tau / \int_0^\infty \left(\frac{\tilde{x}^\tau}{\hat{\mu}_m^\tau} \tau \right)^n \tau^{\delta_2-1} e^{-t\delta_2\tau} d\tau.$$

この推定量を $\check{\eta}(\tau)$ にプラグインして $\hat{\eta} = \check{\eta}(\hat{\tau})$ を得る。対応する μ の推定量は

$$(3.2) \quad \hat{\mu} = 1/\hat{\eta}^{1/\hat{\tau}}$$

として求める。本研究では、 $(\hat{\mu}, \hat{\tau})$ を母数 (μ, τ) の提案推定量とする。推定量 $(\hat{\mu}, \hat{\tau})$ による予測分布と $(\hat{\eta}, \hat{\tau})$ による予測分布は共通になる。

4. リスク比較

推定量 $\check{\theta}$ は推定目標 θ に対する損失 $L(\check{\theta}, \theta)$ を用いて、そのリスク $E[L(\check{\theta}, \theta)]$ で比較される。比較対象の推定量としては MLE を採用する。 (μ, τ) を母数の真値、 $(\check{\mu}, \check{\tau})$ をその任意の推定量とする。また、母数 (μ, τ) の各要素 μ と τ についても個別に比較する。真値 (μ, τ) のもとで、ワイブル分布から標本サイズ n のデータを生成し、推定値を計算する。すべての数値実験において、繰り返し数は 10,000 とし、損失関数の算術平均によってリスクの推定値とする。また、 $\mu = 1$ と設定しても μ に関して的一般性を失わない。事前分布(2.2)における想定値 t については、一般性を失うが $t = 1$ とする場合を扱う。この設定は $\tau = 1$ の場合は適切な仮定となる。

4.1 二乗損失

まず、 τ についての二乗損失 $L_s(\check{\tau}, \tau) = (\check{\tau} - \tau)^2$ を考える。数値実験により推定されたリスクが表 1 である。 δ_2 の値が少しでも大きくなると、リスクが小さくなっていることが観察される。特に、標本サイズが小さいときにその傾向は顕著である。 $\delta_2 = 0$ の提案推定量のケースに注目すると、選択されたすべての n と τ に対して、その推定されたリスクは MLE のそれよりも小さい。標本サイズが大きくなると、すべてのリスクは近づく傾向にある。特に、 $\delta_2 = 0$ の提案推定量と MLE のリスクは $n = 20$ のときに近い。事前分布において $\delta_2 \neq 0$ のときに $t = 1$ と設定しているにも関わらず、 $\tau \neq 1$ であっても性能が良くなることは好ましい結果である。これはゲンベル分布の母数推定にも直接的に関連している。

次に、 μ についての二乗損失 $L_s(\check{\mu}, \mu) = (\check{\mu} - \mu)^2$ を考える。推定されたリスクの結果を表 2 に示す。 τ が小さいところでは、MLE よりも提案推定量の推定リスクが小さい。 τ が大きくなると、両者の推定リスクに大きな差はなくなる。これはどの標本サイズ n でも同じ傾向である。 $\tau = 0.5, 5, 10$ において、 δ_2 が大きくなると $\delta_2 = 0$ のときに比べて提案推定量の性能が悪くなる。これは、 δ_2 が大きくなるにつれて τ は $\tau = 1$ の近傍に集まるため、性能が低下していると考えられる。ただし、その推定リスクに大きな差は見受けられない。 δ_2 は $t = 1$ に対する信念の度合いであり、 μ の推定には直接的な影響が少ないことに依る。

次に、 $\log \mu$ の二乗損失の場合を表 3 に示す。MLE に比べて、提案推定量の推定リスクが大きいが、その差は小さい。また、 μ の場合と同様に、標本サイズに依らず τ が大きくなると差

表 1. 二乗損失のもとでの τ の推定リスク.

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	11.87	1.828	1.311	18.67
	1	47.47	4.936	3.195	74.67
	2	189.88	11.97	6.473	298.68
	5	1186.74	28.34	9.644	1866.73
	10	4746.96	35.16	8.491	7466.93
5	0.5	0.1874	0.1814	0.1761	0.1970
	1	0.7497	0.6742	0.6124	0.7879
	2	2.999	2.342	1.883	3.152
	5	18.74	9.890	5.788	19.70
	10	74.97	22.01	9.249	78.79
10	0.5	0.03580	0.03585	0.03590	0.03756
	1	0.1432	0.1398	0.1365	0.1502
	2	0.5729	0.5313	0.4934	0.6009
	5	3.580	2.857	2.309	3.756
	10	14.32	9.012	6.190	15.02
20	0.5	0.01114	0.01118	0.01122	0.01153
	1	0.04456	0.04419	0.04383	0.04611
	2	0.1783	0.1726	0.1672	0.1844
	5	1.114	1.006	0.914	1.153
	10	4.456	3.612	3.059	4.611

表 2. 二乗損失のもとでの μ の推定リスク.

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	3.491	3.508	3.525	3.568
	1	0.377	0.374	0.371	0.381
	2	0.0865	0.0857	0.0850	0.0867
	5	0.01520	0.01531	0.01548	0.01516
	10	0.00403	0.00420	0.00438	0.00401
5	0.5	1.554	1.563	1.572	1.593
	1	0.223	0.222	0.221	0.225
	2	0.0531	0.0528	0.0525	0.0532
	5	0.00902	0.00906	0.00913	0.00899
	10	0.00234	0.00240	0.00249	0.00233
10	0.5	0.627	0.629	0.632	0.639
	1	0.1131	0.1129	0.1128	0.1140
	2	0.02737	0.02728	0.02721	0.02741
	5	0.004499	0.004504	0.004519	0.004486
	10	0.001144	0.001159	0.001183	0.001139
20	0.5	0.2537	0.2543	0.2549	0.2569
	1	0.05442	0.05438	0.05435	0.05469
	2	0.01347	0.01345	0.01343	0.01348
	5	0.002192	0.002193	0.002196	0.002188
	10	0.000553	0.000557	0.000563	0.000552

表 3. 二乗損失のもとでの $\log \mu$ の推定リスク.

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	1.731	1.724	1.717	1.721
	1	0.4328	0.4324	0.4320	0.4303
	2	0.1082	0.1088	0.1093	0.1076
	5	0.0173	0.0177	0.0182	0.0172
	10	0.00433	0.00458	0.00483	0.00430
5	0.5	0.9777	0.9744	0.9713	0.9707
	1	0.2444	0.2441	0.2438	0.2427
	2	0.0611	0.0613	0.0614	0.0607
	5	0.0098	0.0099	0.0101	0.0097
	10	0.00244	0.00254	0.00266	0.00243
10	0.5	0.4673	0.4664	0.4656	0.4648
	1	0.1168	0.1167	0.1166	0.1162
	2	0.02921	0.02923	0.02926	0.02905
	5	0.00467	0.00471	0.00475	0.00465
	10	0.00117	0.00119	0.00122	0.00116
20	0.5	0.2238	0.2236	0.2234	0.2231
	1	0.05596	0.05593	0.05590	0.05577
	2	0.013989	0.013995	0.014002	0.013944
	5	0.00224	0.00225	0.00226	0.00223
	10	0.000560	0.000565	0.000573	0.000558

が小さくなり, δ_2 の値による影響が小さい. このリスクはゲンベル分布に関係する. $\log \mu$ の性能は MLE よりもわずかに劣るがほぼ同等であることと, 先の τ の二乗損失の結果を踏まえれば, ゲンベル分布においても本アプローチは有効であることが期待される.

ここで, μ と τ を含む $\eta = \mu^{-\tau}$ の二乗損失の場合を検討する. 表 4 に結果を示す. 標本サイズが小さいとき, MLE と $\delta_2 = 0$ のときの提案推定量が大きく, $\delta_2 > 0$ では τ が大きくなると急激に推定リスクが小さくなる. $n = 10$ ではそれぞれの推定リスクの差は小さくなるが, 提案推定量はすべて MLE よりも小さい. $n = 20$ でその差はほぼ無くなる. その逆数 η^{-1} の二乗損失は表 5 に示される. 計算不能のケースが $n = 3$ で含まれている. これは τ の推定値が極端に大きい場合が含まれていたためである. 全体的な傾向は表 4 と類似しており, 提案推定量は MLE よりも推定リスクが小さいことが観察される.

4.2 カルバック・ライブラー損失

プラグイン予測子 $p(x|\tilde{\mu}, \tilde{\tau})$ を \check{p} とする. 真の分布を p とする. カルバック・ライブラー損失を $L_e[(\tilde{\mu}, \tilde{\tau}), (\mu, \tau)]$ と書く. ここで, $L_e[(\tilde{\mu}, \tilde{\tau}), (\mu, \tau)] = D(\check{p}, p)$ である. すると, $L_e[(\tilde{\mu}, \tilde{\tau}), (\mu, \tau)]$ は

$$n \left[\log \frac{\tilde{\tau}}{\tau} - \tau \log \frac{\tilde{\mu}}{\mu} + \left(1 - \frac{\tau}{\tilde{\tau}}\right) \psi(1) - 1 + \frac{\tilde{\mu}^\tau}{\mu^\tau} \Gamma\left(1 + \frac{\tau}{\tilde{\tau}}\right) \right].$$

となる. $L_e[(\tilde{\mu}, \tilde{\tau}), (\mu, \tau)]$ における推定リスクを表 6 にまとめる. ここでも, MLE のリスクよりも提案推定量のリスクのほうが全体的に小さい傾向が見受けられる. δ_2 を大きくするとリスクは小さくなる. 特に, τ が大きくなるとその傾向は顕著になる. 逆に, $\tau = 0.5$ のときの δ_2 の大きさによる影響は小さい. また, 標本サイズの違いによるリスクの変化は, 二乗損失の場合と比べると小さい.

表 4. 二乗損失のもとでの $\eta = \mu^{-\tau}$ の推定リスク.

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	1.01×10^{32}	1.39×10^{22}	4.11×10^{17}	6.71×10^{31}
	1	1.01×10^{32}	1.22×10^{17}	6.91×10^{11}	6.71×10^{31}
	2	1.01×10^{32}	2.68×10^{11}	2.34×10^6	6.71×10^{31}
	5	1.01×10^{32}	5.07×10^4	57.79	6.71×10^{31}
	10	1.01×10^{32}	44.30	2.54	6.71×10^{31}
5	0.5	1.06×10^4	5.96×10^3	3.54×10^3	1.40×10^4
	1	1.06×10^4	2.81×10^3	936.56	1.40×10^4
	2	1.06×10^4	756.66	116.66	1.40×10^4
	5	1.06×10^4	43.23	3.91	1.40×10^4
	10	1.06×10^4	3.51	0.83	1.40×10^4
10	0.5	0.2620	0.2625	0.2630	0.2673
	1	0.2620	0.2597	0.2575	0.2673
	2	0.2620	0.2543	0.2474	0.2673
	5	0.2620	0.2398	0.2223	0.2673
	10	0.2620	0.2198	0.1918	0.2673
20	0.5	0.07348	0.07357	0.07365	0.07394
	1	0.07348	0.07343	0.07338	0.07394
	2	0.07348	0.07315	0.07284	0.07394
	5	0.07348	0.07235	0.07129	0.07394
	10	0.07348	0.07109	0.06897	0.07394

表 5. 二乗損失のもとでの $\eta^{-1} = \mu^\tau$ の推定リスク.

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	—	3.69×10^{36}	1.17×10^{17}	—
	1	—	2.58×10^{16}	7.99×10^6	—
	2	—	3.97×10^6	687.78	—
	5	—	52.90	1.59	—
	10	—	1.39	0.33	—
5	0.5	3.40×10^3	1.60×10^3	807.85	5.09×10^3
	1	3.40×10^3	639.74	163.08	5.09×10^3
	2	3.40×10^3	131.72	14.75	5.09×10^3
	5	3.40×10^3	5.27	0.70	5.09×10^3
	10	3.40×10^3	0.65	0.25	5.09×10^3
10	0.5	0.2493	0.2498	0.2504	0.2620
	1	0.2493	0.2440	0.2391	0.2620
	2	0.2493	0.2332	0.2193	0.2620
	5	0.2493	0.2054	0.1748	0.2620
	10	0.2493	0.1708	0.1296	0.2620
20	0.5	0.0716	0.0719	0.0722	0.0733
	1	0.0716	0.0715	0.0713	0.0733
	2	0.0716	0.0705	0.0694	0.0733
	5	0.0716	0.0679	0.0645	0.0733
	10	0.0716	0.0638	0.0576	0.0733

‘—’は計算不能のケースが含まれていたことを表す。

表 6. カルバッック・ライブラー損失 $L_e[(\check{\mu}, \check{\tau}), (\mu, \tau)]$ のもとでの推定リスク.

n	τ	Proposed			
		$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	1.494	1.464	1.439	1.510
	1	1.494	1.417	1.353	1.510
	2	1.494	1.334	1.213	1.510
	5	1.494	1.142	0.958	1.510
	10	1.494	0.958	0.904	1.510
5	0.5	1.236	1.229	1.223	1.251
	1	1.236	1.212	1.189	1.251
	2	1.236	1.178	1.127	1.251
	5	1.236	1.094	1.000	1.251
	10	1.236	1.001	0.952	1.251
10	0.5	1.107	1.105	1.103	1.116
	1	1.107	1.098	1.090	1.116
	2	1.107	1.085	1.066	1.116
	5	1.107	1.052	1.013	1.116
	10	1.107	1.014	0.989	1.116
20	0.5	1.035	1.034	1.033	1.040
	1	1.035	1.031	1.028	1.040
	2	1.035	1.025	1.017	1.040
	5	1.035	1.011	0.993	1.040
	10	1.035	0.994	0.981	1.040

また、双対なカルバッック・ライブラー損失 $L_m[(\check{\mu}, \check{\tau}), (\mu, \tau)] = L_e[(\mu, \tau), (\check{\mu}, \check{\tau})]$ を考えることもできる。損失 $L_m[(\check{\mu}, \check{\tau}), (\mu, \tau)]$ の場合の結果を表 7 に示す。得られたリスクは、総じて提案推定量のほうが MLE よりも小さい。また、 δ_2 が大きくなると、リスクは小さくなる傾向であり、標本サイズが小さいほどその傾向は顕著である。標本サイズが $n = 3$ のとき、計算不能になるケースが含まれている。これは推定された τ の値が大きくなるケースであり、損失関数のガンマ関数が計算できなくなる。一方、 $\delta_2 > 0$ の場合は $n = 3$ においても計算値が得られている。

4.3 偏り

次に、推定量の期待値と真値の差である偏りを調べる。まずは、 τ の推定値の平均を表 8 に示す。 n が小さいとき、いずれも真の τ から離れている。MLE の偏りの絶対値が最も大きい。 $\delta_2 = 0.2$ のときは比較的偏りの絶対値が小さい。 n が大きくなると提案推定量の偏りの絶対値は小さくなる。また、選択した n と τ のいずれにおいても MLE より提案推定量の偏りの絶対値のほうが小さい。 δ_2 を大きくすれば、偏りの絶対値もまた小さくなる。

次に、 μ の推定値の平均を表 9 に与える。 n と τ がともに小さい場合を除けば、真の $\mu = 1$ に対して、明らかに偏りの絶対値が小さいことが見受けられる。特に $n = 3$ であっても $\tau \geq 1$ においてその偏りの絶対値は小さい。 $\tau = 0.5$ のとき、相対的に偏りの絶対値が大きくなるが、提案推定量は MLE よりも偏りの絶対値が小さい。一方で、 τ が大きくなると、 $\delta_2 > 0$ の提案推定量が MLE よりも偏りの絶対値が大きくなる。これは $t = 1$ と設定していることに起因する。

表 10 に $\log \mu$ の推定結果を示す。同様に、 τ が大きい場合には標本サイズが小さくても偏りの絶対値が小さいことが見受けられる。また、ほとんどのケースで負になる。 $\tau = 0.5$ のケースを除けば、いずれの推定法においても顕著な差はない。 $\tau = 0.5$ の場合は MLE の偏りの絶対

表 7. 双対なカルバック・ライブラー損失 $L_m[(\check{\mu}, \check{\tau}), (\mu, \tau)]$ のもとでの推定リスク.

Proposed					
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	—	5.62×10^{69}	1.96×10^{32}	—
	1	—	7.12×10^{30}	2.13×10^{12}	—
	2	—	6.09×10^{11}	1.85×10^5	—
	5	—	2.79×10^3	9.37	—
	10	—	7.89	1.30	—
5	0.5	9.67×10^{10}	4.39×10^9	3.18×10^8	2.75×10^{11}
	1	9.67×10^{10}	1.94×10^8	2.03×10^6	2.75×10^{11}
	2	9.67×10^{10}	1.32×10^6	2.48×10^3	2.75×10^{11}
	5	9.67×10^{10}	216.71	4.96	2.75×10^{11}
	10	9.67×10^{10}	4.54	1.42	2.75×10^{11}
10	0.5	2.399	2.396	2.394	2.501
	1	2.399	2.339	2.285	2.501
	2	2.399	2.233	2.091	2.501
	5	2.399	1.963	1.661	2.501
	10	2.399	1.629	1.244	2.501
20	0.5	1.368	1.372	1.376	1.398
	1	1.368	1.362	1.355	1.398
	2	1.368	1.342	1.317	1.398
	5	1.368	1.287	1.216	1.398
	10	1.368	1.206	1.092	1.398

‘-’は計算不能のケースが含まれていたことを表す。

表 8. τ の推定値の平均.

Proposed					
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	1.146	1.070	1.038	1.175
	1	2.291	2.028	1.908	2.350
	2	4.583	3.724	3.345	4.699
	5	11.457	7.718	6.333	11.748
	10	22.913	12.345	9.238	23.497
5	0.5	0.703	0.704	0.705	0.715
	1	1.405	1.390	1.377	1.430
	2	2.811	2.716	2.633	2.859
	5	7.027	6.353	5.832	7.148
	10	14.053	11.499	9.830	14.296
10	0.5	0.575	0.577	0.578	0.582
	1	1.150	1.148	1.147	1.164
	2	2.300	2.278	2.257	2.328
	5	5.751	5.562	5.387	5.819
	10	11.502	10.702	10.009	11.638
20	0.5	0.533	0.534	0.535	0.537
	1	1.066	1.066	1.065	1.073
	2	2.132	2.124	2.116	2.147
	5	5.331	5.256	5.184	5.367
	10	10.661	10.334	10.026	10.734

表 9. μ の推定値の平均.

Proposed					
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	1.4560	1.4609	1.4657	1.4712
	1	1.0395	1.0369	1.0346	1.0452
	2	0.9765	0.9711	0.9664	0.9792
	5	0.9794	0.9725	0.9672	0.9805
	10	0.9877	0.9808	0.9762	0.9882
5	0.5	1.2676	1.2718	1.2759	1.2825
	1	1.0224	1.0215	1.0207	1.0286
	2	0.9846	0.9817	0.9789	0.9877
	5	0.9872	0.9830	0.9793	0.9884
	10	0.9924	0.9880	0.9844	0.9930
10	0.5	1.1425	1.1449	1.1473	1.1526
	1	1.0147	1.0145	1.0143	1.0193
	2	0.9937	0.9923	0.9910	0.9960
	5	0.9941	0.9920	0.9901	0.9950
	10	0.9965	0.9942	0.9921	0.9969
20	0.5	1.0689	1.0702	1.0714	1.0748
	1	1.0072	1.0072	1.0072	1.0100
	2	0.9969	0.9962	0.9956	0.9983
	5	0.9971	0.9961	0.9951	0.9977
	10	0.9983	0.9971	0.9960	0.9986

表 10. $\log \mu$ の推定値の平均.

Proposed					
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	-0.2888	-0.2839	-0.2791	-0.2772
	1	-0.1444	-0.1465	-0.1483	-0.1386
	2	-0.0722	-0.0776	-0.0824	-0.0693
	5	-0.0289	-0.0359	-0.0414	-0.0277
	10	-0.0144	-0.0214	-0.0261	-0.0139
5	0.5	-0.1759	-0.1719	-0.1680	-0.1632
	1	-0.0880	-0.0886	-0.0892	-0.0816
	2	-0.0440	-0.0469	-0.0497	-0.0408
	5	-0.0176	-0.0218	-0.0256	-0.0163
	10	-0.0088	-0.0132	-0.0169	-0.0082
10	0.5	-0.0820	-0.0796	-0.0773	-0.0726
	1	-0.0410	-0.0411	-0.0413	-0.0363
	2	-0.0205	-0.0219	-0.0232	-0.0182
	5	-0.0082	-0.0103	-0.0123	-0.0073
	10	-0.0041	-0.0064	-0.0085	-0.0036
20	0.5	-0.0399	-0.0387	-0.0375	-0.0343
	1	-0.0200	-0.0200	-0.0200	-0.0172
	2	-0.0100	-0.0106	-0.0113	-0.0086
	5	-0.0040	-0.0050	-0.0060	-0.0034
	10	-0.0020	-0.0031	-0.0043	-0.0017

表 11. η の推定値の平均.

n	τ	Proposed			
		$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	1.01×10^{14}	1.58×10^9	8.39×10^6	8.28×10^{13}
	1	1.01×10^{14}	4.56×10^6	1.19×10^4	8.28×10^{13}
	2	1.01×10^{14}	7.47×10^3	37.93	8.28×10^{13}
	5	1.01×10^{14}	8.98	2.22	8.28×10^{13}
	10	1.01×10^{14}	2.13	1.53	8.28×10^{13}
5	0.5	3.1510	2.7170	2.4266	3.3979
	1	3.1510	2.3217	1.9500	3.3979
	2	3.1510	1.8950	1.5957	3.3979
	5	3.1510	1.5051	1.3772	3.3979
	10	3.1510	1.3715	1.3040	3.3979
10	0.5	1.1059	1.1051	1.1044	1.1028
	1	1.1059	1.1060	1.1061	1.1028
	2	1.1059	1.1077	1.1094	1.1028
	5	1.1059	1.1126	1.1186	1.1028
	10	1.1059	1.1200	1.1315	1.1028
20	0.5	1.0414	1.0409	1.0404	1.0389
	1	1.0414	1.0415	1.0416	1.0389
	2	1.0414	1.0427	1.0439	1.0389
	5	1.0414	1.0462	1.0508	1.0389
	10	1.0414	1.0518	1.0613	1.0389

値が小さい傾向にあるが、その差はわずかである。

次に、 η, η^{-1} の推定値の平均を表 11、表 12 に示す。表 11 から、 $n = 3$ のとき、提案推定量の $\delta_2 = 0$ と MLE はともに近い値をとっているが、どちらも真値から大きく離れている。 $n = 5$ では提案推定量の偏りの絶対値が小さい。 $n = 10$ では MLE の偏りの絶対値が小さいが、提案推定量との差はわずかである。 n が大きくなると、どの推定値も真値に近い。 $\delta_2 > 0$ において $t = 1$ とおいた影響が $n = 10, 20$ における $\tau = 5, 10$ のケースで見受けられる。表 12 では $n = 10$ で提案推定量における $\delta_2 = 0$ と MLE が同等で、それ以外のすべてのケースで提案推定量のほうが偏りの絶対値が小さい。

5. 実データへの適用

本節では、推定量の選択の必要性を確認するために、実データに適用し推定値の違いを比較する。ある材料の 10 個の軸受の時間単位の疲労寿命データ 152.7, 172.0, 172.5, 173.3, 193.0, 204.7, 216.5, 234.9, 262.6, 422.6 を利用する。このデータは McCool (1974) が提供しており、Cohen and Whitten (1988) でも分析されている。提案推定量の設定において、第 4 節の数値実験と同様に $t = 1$ とする。推定した結果を表 13 に記載する。まず、提案推定量は、 $\delta_2 = 0$ から $\delta_2 = 0.2$ まで連続的に推定値が変化している。次に、提案推定量と MLE を比較すると、 $\delta_2 = 0$ のときでも両者の推定値に違いがある。 δ_2 が大きくなると、 τ の推定値が小さくなり、MLE との違いが大きくなる。これは $t = 1$ の設定に起因すると考えられる。事前情報が無ければ、 $\delta_2 = 0$ とすることが推奨される。

この推定値の違いは、適切な推定量の選択の必要性を示すものである。第 4 節の数値実験の結果は、提案推定量の選択を支持する。この計算についての R コードはウェブ上の補足資料に

表 12. η^{-1} の推定値の平均.

		Proposed			
n	τ	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$	MLE
3	0.5	2.17×10^{189}	1.92×10^{16}	3.43×10^6	8.14×10^{253}
	1	2.19×10^{189}	1.61×10^6	52.85	8.14×10^{253}
	2	2.18×10^{189}	38.46	2.23	8.14×10^{253}
	5	2.17×10^{189}	1.58	1.14	8.14×10^{253}
	10	2.17×10^{189}	1.12	0.96	8.14×10^{253}
5	0.5	1.9994	1.7574	1.6063	2.1626
	1	1.9994	1.5518	1.3736	2.1626
	2	1.9994	1.3432	1.2050	2.1626
	5	1.9994	1.1515	1.0638	2.1626
	10	1.9994	1.0553	0.9775	2.1626
10	0.5	1.0653	1.0666	1.0679	1.0725
	1	1.0653	1.0640	1.0628	1.0725
	2	1.0653	1.0591	1.0532	1.0725
	5	1.0653	1.0451	1.0278	1.0725
	10	1.0653	1.0246	0.9945	1.0725
20	0.5	1.0236	1.0243	1.0250	1.0270
	1	1.0236	1.0234	1.0233	1.0270
	2	1.0236	1.0218	1.0200	1.0270
	5	1.0236	1.0169	1.0105	1.0270
	10	1.0236	1.0091	0.9962	1.0270

表 13. McCool (1974) のデータに対する推定値.

Proposed			
	$\delta_2 = 0$	$\delta_2 = 0.1$	$\delta_2 = 0.2$
μ	245.446	245.049	244.661
τ	2.874	2.848	2.823
			246.409
			2.936

記載されている。

6. 考察

ワイブル分布の母数推定に対して提案したベイズ推定量の性能は、リスク比較の結果、既存推定量としての MLE の性能を上回る。特に、母数 (μ, τ) の同時推定の性能は有望な結果である。今回の比較では $\delta_1 = 0$ という限定的なものを扱い、 δ_2 を変化させて検証を行った。 δ_2 は τ の推定に影響を与えるものであるため、 μ の推定には大きな影響がなかった。また、 $\hat{\mu}$ や $\log \hat{\mu}$ の偏りは小さいことが観察された。自然母数の事後平均による推定量の偏りが小さくなることを Sakumura and Yanagimoto (2024) が報告している。本研究の提案推定量も自然母数に着想を得たものであり、これらの設定でも近似的な不偏性が観察されたことは興味深い。

打切りや切断が含まれるデータに対して、Kundu and Mitra (2016) は共役な事前分布を仮定した例を紹介している。一方で、無情報事前分布については特に言及はない。このような打切りや切断を含むデータにおいても、本研究における事前分布の設定については特段の変更はなく、形式的には提案手法を適用できる。ただし、その性能については検証が必要である。

参考文献

- Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference, *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**(2), 113–128, <https://doi.org/10.1111/j.2517-6161.1979.tb01066.x>.
- Cohen, A. C. and Whitten, B. J. (1988). *Parameter Estimation in Reliability and Life Span Models*, 1st ed., CRC Press, Boca Raton, Florida, <https://doi.org/10.1201/9781003066064>.
- Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families, *The Annals of Statistics*, **7**(2), 269–281, <http://www.jstor.org/stable/2958808>.
- Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distribution*, Volume 1, 2nd ed., Wiley & Sons, New York.
- Kundu, D. and Mitra, D. (2016). Bayesian inference of Weibull distribution based on left truncated and right censored data, *Computational Statistics & Data Analysis*, **99**, 38–50, <https://doi.org/10.1016/j.csda.2016.01.001>.
- McCool, J. I. (1974). *Inferential Techniques for Weibull Populations*, Aerospace Research Laboratories Report ARL, TR 74-0180, Wright-Patterson Air Force Base, Dayton, Ohio.
- Nelson, W. (1990). *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses*, John Wiley & Sons, New York, <https://doi.org/10.1002/9780470316795>.
- Sakumura, T. and Yanagimoto, T. (2024). Very small bias observed in Bayesian estimators even for small sample sizes, *Communications in Statistics — Simulation and Computation*, **53**(11), 5594–5604, <https://doi.org/10.1080/03610918.2023.2196380>.
- Sun, D. (1997). A note on noninformative priors for Weibull distributions, *Journal of Statistical Planning and Inference*, **61**(2), 319–338, [https://doi.org/10.1016/S0378-3758\(96\)00155-3](https://doi.org/10.1016/S0378-3758(96)00155-3).
- Xu, A., Fu, J., Tang, Y. and Guan, Q. (2015). Bayesian analysis of constant-stress accelerated life test for the Weibull distribution using noninformative priors, *Applied Mathematical Modelling*, **39**(20), 6183–6195, <https://doi.org/10.1016/j.apm.2015.01.066>.
- Yanagimoto, T. and Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e -mixture, *Journal of Statistical Planning and Inference*, **139**(9), 3064–3075, <https://doi.org/10.1016/j.jspi.2009.02.005>.

Bayesian Estimation under the Conjugate-gamma Prior for the Weibull Distribution

Takenori Sakumura¹ and Takemi Yanagimoto²

¹Faculty of Science and Engineering, Hosei University

²The Institute of Statistical Mathematics

Bayesian inference contains two important issues; the choice of the parameters to be estimated and that of prior density. In the former issue, a canonical parameter is a candidate in an exponential family. In the latter issue, when only weak information is available, a conjugate prior is often assumed. A non-informative prior is assumed, when an objective assumption is preferred. The latter prior is regarded as the limit of the former one. Applying the optimality properties of Bayesian estimators in the exponential family, we construct a Bayesian estimator when the sampling density follows the Weibull distribution. A conjugate-gamma prior is employed, which is an extension of a non-informative prior. The performance of the proposed estimator is evaluated favorably in terms of the squared and the Kullback-Leibler losses. Risk comparison studies are conducted, which support the desirable performance of the proposed estimator. The approximate unbiasedness of the proposed estimator is also observed. We also apply the proposed estimator to a real dataset, which indicates the need for selecting a suitable estimator. We conclude that the proposed estimator is promising.

信頼性解析とツールの活用

岡村 寛之[†]・鄭 俊俊[†]・土肥 正[†]

(受付 2024 年 11 月 29 日；改訂 2025 年 2 月 28 日；採択 2 月 28 日)

要 旨

本稿では、信頼性分野で使用されるツールについて紹介する。信頼性評価は、故障木やマルコフモデルに代表される確率モデルを用いたアプローチと、信頼性試験などから得られるデータをもとにモデルの同定を行う統計的手法に大別される。これらの手法はいずれも、信頼性を評価するために信頼性解析ツールが必要となる。本稿では、信頼性に関する基礎知識を紹介し、その後、具体的な信頼性解析ツールについて説明する。

キーワード：信頼性解析ツール、確率モデル、統計的手法、信頼性評価。

1. はじめに

近年、IoT や AI を用いたシステムがあらゆる分野に応用されつつある。特に自動運転など我々の命に関わるようなセーフティクリティカルシステムが複雑化の一途を辿り、その信頼性確保が急務な課題となっている。このような状況下、システムの信頼性を定量的に評価することは、製品開発のあらゆる段階において必要不可欠である。

信頼性評価は、大きく分けて二つのアプローチが存在する。一つは、故障木やマルコフモデルに代表される確率モデルを用いたアプローチである。このアプローチでは、システムの構造や部品の故障率をもとに、システム全体の信頼度や故障率を算出する。もう一つは、信頼性試験などから得られるデータをもとに、システムの信頼性モデルを同定する統計的手法である。このアプローチでは、過去のデータから将来の信頼性を予測することができる。これらの手法はいずれも、信頼性を評価するために信頼性解析ツールが必要となる。信頼性解析ツールは、複雑なシステムの信頼性モデルを構築し、様々な指標を算出することで、信頼性評価を支援する。近年では、商用ソフトウェアからオープンソースソフトウェアまで、様々なツールが開発・提供されている。

本稿では、信頼性分野で使用されるツールとその事例について紹介する。まず、信頼性に関する基礎知識として、確率モデルを用いた信頼性評価手法と統計的手法について概説する。その後、具体的な信頼性解析ツールについて、商用ソフトウェア、オープンソースソフトウェア、学術分野で開発されたツールなどを例に挙げ、それぞれの特徴や機能について説明する。

本稿の構成は以下の通りである。2 節では、信頼性に関する基礎知識について述べる。3 節では、確率モデルを用いた信頼性解析の基礎のうち、故障木を用いた静的モデルとその解析について言及する。特に、ツール化に必要なアルゴリズムを中心に概説する。4 節は、マルコフ連鎖を用いた信頼性解析の基礎と、数値的な計算アルゴリズムについて紹介する。5 節では、

[†] 広島大学 大学院先進理工系科学研究科：〒 739-8527 広島県東広島市鏡山 1-4-1; okamu@hiroshima-u.ac.jp, jzheng@hiroshima-u.ac.jp, dohi@hiroshima-u.ac.jp

信頼性における統計解析の基礎と、相型分布と呼ばれる分布のパラメータ推定手続きを中心に紹介する。6節では、信頼性に関するツールの紹介を行い、一つの信頼性評価事例を紹介する。最後に7節において、信頼性解析ツールの限界および今後の課題について言及する。

2. 信頼性の基礎知識

信頼性とは「ある期間において、所定の条件の下で、対象が意図した機能を遂行する能力」として定義されている(Barlow and Proschan, 1965)。ここで「ある期間」および「所定の条件」は、信頼性が時間の経過ならびに使用環境によって変化するものであり、信頼性を評価する際にはこれらの条件を明確に定義する必要がある。また、「意図した機能」とは、製品やシステムが期待される性能を発揮することが前提となり、一般的な故障を含むより広い概念として規定されている。信頼性は、製品やシステムが安全に運用されるために不可欠な要素であり、製品が備えるべき一つの重要な要素となっている。

信頼性工学では、信頼性を故障確率と関連付けて定量化することから始まる。故障確率とは、ある期間内に対象が故障する確率を表し、信頼性が高いほど故障確率は低くなる。確率変数 X を対象の寿命とし、 $F(x)$ を寿命に対する累積分布関数(故障分布関数)とする。このとき、対象の信頼度は時刻 t において故障しない(寿命を迎えていない)確率として定義され、次のように表される。

$$(2.1) \quad R(t) = P(X > t) = 1 - F(t).$$

ここで、 $R(t)$ は信頼度関数と呼ばれ、信頼度が時間の経過に伴って単調に減少することを表している。信頼度の時間変化を表す信頼度関数 $R(t)$ は、信頼性評価の基本的な指標として用いられる。また、故障分布関数の導関数 $f(t) = dF(t)/dt$ を用いて、故障率関数 $\lambda(t)$ が次のように定義される。

$$(2.2) \quad \lambda(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)}.$$

故障率は、時間の経過に伴う故障の発生頻度を示し、信頼性評価において重要な指標の一つである。例えば、故障率が一定である場合、時間経過による故障率の変化が見られないことを意味し、これは故障がランダムに発生する状況(偶発故障)を表す。この場合、故障分布は指数分布に従う。一方、対象が経年劣化する場合には、故障率は時間とともに増加する。また、初期不良のように時間経過とともに故障率が減少する場合もある。このように、故障率が時間に応じて増減する場合、故障分布は指数分布よりも一般的なワイブル分布などに従う。故障率と信頼度関数には以下の関係が成立する。

$$(2.3) \quad R(t) = e^{-\int_0^t \lambda(s)ds}.$$

上式から $\Lambda(t) = \int_0^t \lambda(s)ds = -\log R(t)$ となり、 $\Lambda(t)$ を累積ハザード関数と呼び、これも信頼度評価によく用いられる。

一方で、信頼性工学では上記の時間要素を除いた信頼度あるいは故障確率(不信頼度)を用いることもよくある。その場合、信頼度は対象が特定の時点で機能する確率を表し、故障確率は文字通りその時点で既に故障している確率である。時間要素を除いた信頼度および故障確率は時刻 t を省略した記号で信頼度 R のように記述されることが多い。

信頼性工学における各種の分析では、信頼度関数 $R(t)$ 、信頼度 R および故障確率を求めることが主な目的となる。代表的な信頼性解析として故障木(FT: Fault Tree)解析や信頼性ブロック図(RBD: Reliability Block Diagram)解析がある(Barlow and Proschan, 1965; Birnbaum

et al., 1961). 故障木解析は、システムが故障に至る原因を木構造で表現し、各部品の故障や何らかのイベントがシステム全体の信頼性に与える影響を分析する。信頼性ブロック図解析もシステム故障と部品の故障の関係をブロック図で表現する手法であり、主に直列・並列と呼ばれるシステムの冗長性構造を視覚的に与える手法である。これらの手法は、システムを構成する部品とシステムの関係をモデル化することで、部品の信頼度からシステムの信頼度を同定する手法である。一般に、これらのモデルは静的モデルであり、部品の信頼度 R からシステムの信頼度 R を算出することを主な目的としている。ただし、Dynamic Fault Tree のように「故障の順番」の時間要素が含まれるものもある (Dugan et al., 1992).

静的な解析と同様に、故障に至る過程をモデル化するアプローチであるが、時間による変化を考慮した手法として、マルコフモデルを用いた信頼性解析がある。マルコフモデルでは、システムを離散状態に分類し状態遷移をマルコフ過程で表現する。特に、信頼性解析では正常稼働状態と故障状態に分類し、故障状態に到達するまでの時間を考慮することで信頼度関数 $R(t)$ を求める。また、マルコフモデルと故障木を複合させることで、部品の信頼度関数 $R(t)$ からシステムの信頼度関数 $R(t)$ を求めるハイブリッド手法も存在する (Trivedi and Bobbio, 2017).

他方、データから直接信頼度あるいは信頼度関数を同定するアプローチとして統計解析による手法がある。信頼性分野における統計解析では、主に信頼性試験や運用データから得られるデータをもとに故障分布を同定する。信頼性試験とは、部品やシステムを一定の条件下で運用し、故障や劣化の減少を観測する手法である。一般に、システムの信頼性試験や運用データを収集することは莫大なコストを必要とするため、信頼性試験では部品を対象に行われることが多い。実際に、信頼性試験から得られたデータから部品の故障分布を最尤推定法やベイズ推定で推定した後に、先に紹介した確率モデルを組み合わせることで、最終的にシステムの信頼度を評価することができる。

3. FT 解析

3.1 モデル表現

故障木(以下、FT)はシステム故障とその原因(部品の故障など)の関係を表すモデルとして利用され、システム故障に至る原因解析などを事前に抽出するために利用される。基本的なFTでは原因と事象を AND ゲート(すべての原因が発生したら該当する事象が発生)、OR ゲート(少なくとも一つの原因が発生したら該当する事象が発生)あるいは k -out-of- n ゲート(n 個の原因のうち k 個が発生したら該当する事象が発生)でモデル化していく。FT は何らかの事象を表すノードと AND/OR/ k -out-of- n ゲートから構成される木構造となっており、親ノードに対応する事象の発生条件をゲートと子ノードで表現する。例えば、図 1(a)は AND ゲートを用いて Event 1 と Event 2 が発生した時に Event S が発生する関係を表しており、図 1(b)は OR ゲートを用いて Event 1 と Event 2 の少なくともどちらか一方が発生した時に Event S が発生する関係を表している。実際の FT では複数の階層で構成されトップ事象(木の根ノードに対応する事象)の発生条件を原因事象(木の葉ノードに対応する事象)を用いて記述する。

FT の解析を行うために構造関数(structure function)を定義する。いま、FT における原因事象を表す変数 $x_i, i = 1, \dots, n$ を

$$(3.1) \quad x_i = \begin{cases} 1 & (\text{原因 } i \text{ が発生している}) \\ 0 & (\text{原因 } i \text{ が発生していない}) \end{cases}$$

と定義する。この時、トップ事象の発生を表す変数 x_S が以下の式で与えられるものとする。

$$(3.2) \quad x_S = \varphi(x_1, x_2, \dots, x_n).$$

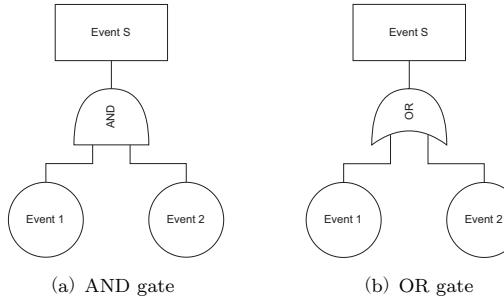


図 1. FT の例.

ここで, $\varphi(x_1, x_2, \dots, x_n)$ は構造関数と呼ばれる論理関数であり, 各原因事象の状態 x_i に応じてトップ事象の状態(0 または 1)を出力する. 例えば, 図 1(a)で, Event 1, Event 2 の状態(0 または 1)を表す変数 x_1, x_2 が与えられた時, Event S の状態 x_S は

$$(3.3) \quad x_S = x_1 \wedge x_2$$

として与えられる. ここで \wedge は論理積を表す. また, 図 1(b)の場合は

$$(3.4) \quad x_S = x_1 \vee x_2$$

となる. ここで, \vee は論理和を表す. 一般に AND ゲートは直列構造と呼ばれ, どちらか一方の部品が壊れるとシステム全体が壊れるような構造を表す. 一方, OR ゲートは並列構造と呼ばれ, 複数の部品が同時に壊れない限りシステム全体が壊れないような構造を表す(冗長設計). k -out-of- n ゲートは AND ゲート, OR ゲートの中間に位置するゲートであり, n 個の部品のうち k 個以上が壊れた場合にシステム全体が壊れるような構造を表す. 例えば, $k = 2, n = 3$ の場合, その構造関数は

$$(3.5) \quad x_S = (x_1 \wedge x_2) \vee (x_1 \wedge x_3) \vee (x_2 \wedge x_3)$$

となる. 構造関数の論理演算による表記は簡単ではあるが実際のトップ事象の発生確率を求めるためには構造関数の多項式表現あるいは, 後述する BDD(Binary Decision Diagram)を用いた手法が必要となる.

3.2 BDD による定量解析

FT では, 各原因事象の発生確率からトップ事象の発生確率を算出することが必要となる. トップ事象の計算には構造関数の多項式表現が必要となる. これは, 積事象を $x \wedge y \rightarrow xy$, 和事象を $x \vee y \rightarrow 1 - (1 - x)(1 - y)$ と置き換えることで得られる. いま, 構造関数 $\varphi(x_1, \dots, x_n)$ から対応する多項式関数 $f_\varphi(x_1, \dots, x_n)$ が得られたとする. X_1, \dots, X_n を原因事象の発生を表す指標確率変数, X_S をトップ事象の発生を表す指標確率変数とすると, トップ事象の発生確率は

$$(3.6) \quad P(X_S = 1) = f_\varphi(p_1, p_2, \dots, p_n)$$

となる. ただし, ここでは原因事象が互いに独立であることを仮定している. ここで, $p_i = P(X_i = 1)$ は原因事象 X_i の発生確率である. しかしながら, 繰り返し事象や k -out-of- n ゲートを含む FT では上記の単純な置き換えでは計算できない. 繰り返し事象とは, 同じ原因事象が葉ノードに複数回現れることを表す. 図 2 は繰り返し事象を含む FT の例を示してい

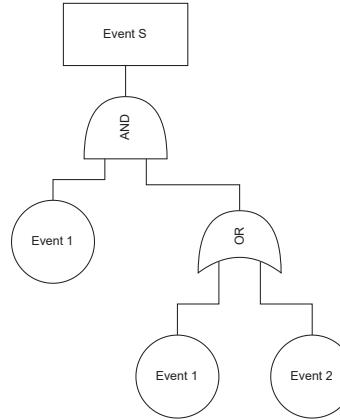


図 2. 繰り返し事象を含む FT の例.

る. この場合, Event 1 が葉ノードに複数回現れている. この FT に対する構造関数は

$$(3.7) \quad \varphi(x_1, x_2) = (x_1 \vee x_2) \wedge x_1$$

となる. 論理積, 論理和を単純に $x \wedge y \rightarrow xy$, $x \vee y \rightarrow 1 - (1 - x)(1 - y)$ で置き換えた多項式は

$$(3.8) \quad f_\varphi(x_1, x_2) = x_1^2 + x_1 x_2 - x_1^2 x_2$$

となる. $x_1^2 = x_1$ を考慮して単純化すると, 式(3.8)は $\varphi(x_1, x_2) = x_1$ となる. つまり, 求めたい多項式は $f_\varphi(x_1, x_2) = x_1$ であるが, 繰返し事象を含む FT に対しては, 直並列の公式と多項式演算だけでなく, 適切な単純化処理を行う必要がある. このような問題から, FT では構造関数を BDD (Binary Decision Diagram) で表現する手法が行われる (Rauzy, 1993). BDD とは論理関数を非循環グラフにより表す手法であり, 同じ論理値を出力するサブグラフを共有することで論理関数をコンパクトに表現することができる (Bryant, 1986).

論理関数 $\varphi(x_1, \dots, x_n)$ が与えられたとき, 原因事象 1 の状態に着目するとシャノン分解から以下のように記述できる.

$$(3.9) \quad \varphi(x_1, \dots, x_n) = (x_1 \wedge \varphi(1, x_2, \dots, x_n)) \vee (\bar{x}_1 \wedge \varphi(0, x_2, \dots, x_n)).$$

ここで, $\bar{x}_1 = 1 - x_1$ である. 同様に $\varphi(1, x_2, \dots, x_n)$, $\varphi(0, x_2, \dots, x_n)$ の x_2, \dots, x_n についてもさらにシャノン分解を行うことができ, 最終的には $\varphi(1, 0, \dots, 1)$ のように具体的な値が与えられた時の論理値まで分解できる. BDD では, 上記の部分的に値が与えられた $\varphi(1, x_2, \dots, x_n)$ や $\varphi(0, x_2, \dots, x_n)$ をノードとする二分木を構成することで論理関数を表現する. 図 3(a), 図 3(b), 図 3(c)はそれぞれ式(3.3), 式(3.4), 式(3.5)の構造関数に対応する BDD 表現を示している. 一番上にあるノードから各変数が取る値の枝を辿ることで最終的な構造関数の出力として 0 または 1 が得られることがわかる.

また, 構造関数の多項式に対するシャノン分解は以下の再帰式となる.

$$(3.10) \quad f_\varphi(x_1, \dots, x_n) = x_1 f_\varphi(1, x_2, \dots, x_n) + (1 - x_1) f_\varphi(0, x_2, \dots, x_n).$$

これと, 式(3.6)の性質を利用すると BDD 表現からトップ事象の発生確率(故障確率や信頼度)を計算するアルゴリズムを構築することができる. 以下の Algorithm 1 は BDD で表現された構造関数をつかってトップ事象の発生確率を計算する手続きを示している.

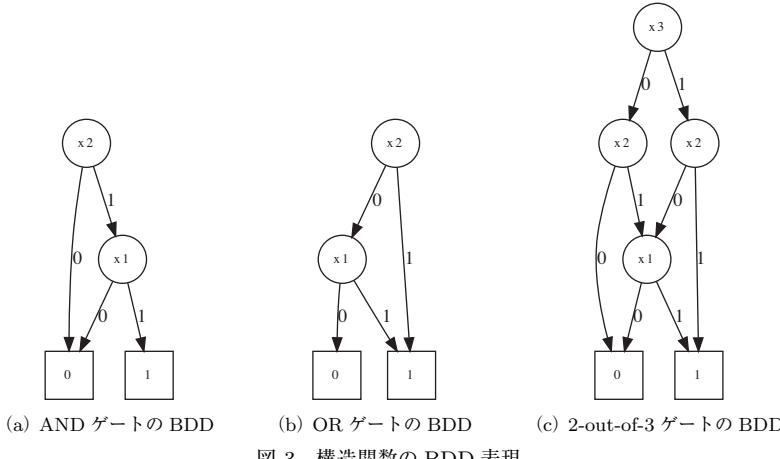


図 3. 構造関数の BDD 表現.

Algorithm 1 Computation of Top Event Probability.

```

function prob(node)
    return visited[node] if haskey(visited, node)
    return 0 if is_zero(node)
    return 1 if is_one(node)
    p = get_prob(node)
    result = p * prob(get_one(node)) + (1-p) * prob(get_zero(node))
    visited[node] = result
    return result
end

```

ここで、`visited`, `haskey` はノードをキーとしたハッシュテーブルとハッシュテーブルに特定のキーがあるかどうか確認をする関数を表している。これは、一度計算した確率を何回も計算しないようにするためのキャッシュとして機能する。関数 `get_prob` は BDD ノードに対応する変数の発生確率を取得するための関数。`is_zero`, `is_one` は、それぞれ 0-終端, 1-終端ノードかどうかを判別する関数、`get_one`, `get_zero` はそれぞれ、1-枝, 0-枝の子ノードを取得するための関数である。

3.3 BDD による定性解析

静的モデルの構造関数の分析では、極小カットベクトル (Minimal Cut Vector) の導出が重要な役割を果たす。構造関数 $\varphi(x_1, \dots, x_n)$ の入力値 $x = (x_1, \dots, x_n)$ を状態ベクトルと呼ぶことにすると、極小カットベクトルとは、 $\varphi(x) = 0$ となる状態ベクトルのうち、任意の x_i の値が 0 から 1 に変化した新しい状態ベクトル x' が $\varphi(x') = 1$ となるような状態ベクトル x を示す。また、極小カットベクトルの集合を極小カット集合 (MCS: Minimal Cut Set) と呼ぶ。これは、現在正常に稼働している部品の一つでも故障すると、システムが故障するような部品の状態を意味している。FT が AND, OR, k -out-of- n ゲートで構成されるとき、このシステムは単調、つまり、部品の故障によってシステム状態がより故障しにくい状態には移行しない性質を持つ。システムが単調であるとき、MCS と構造関数は 1 対 1 に対応するため、通常の FT に対して MCS は一意に決定される。例えば、式(3.3)のシステムの場合、MCS は $\{(0, 1), (1, 0)\}$ となり、式(3.4)の場合は $\{(0, 0)\}$ となる。また、式(3.5)の場合は、 $\{(0, 0, 1), (0, 1, 0), (1, 0, 0)\}$ となる。このような、MCS の導出は信頼性の定性的な分析において重要な役割を果たす。

FT が繰り返し事象を含む場合、MCS の導出は容易ではない。ここでは、BDD を用いた MCS の導出アルゴリズムの紹介を行う (Rauzy, 1993)。基本的なアイデアは構造関数を表す BDD から、MCS の集合を表す BDD を再構成する。シャノン分解によって構造関数が

$$(3.11) \quad \varphi(x_1, \dots, x_n) = (x_1 \wedge \varphi(1, x_2, \dots, x_n)) \vee (\overline{x_1} \wedge \varphi(0, x_2, \dots, x_n))$$

のように分解されることを思い出す。システムが単調であることを仮定すると、 $\varphi(0, x_2, \dots, x_n)$ の (x_2, \dots, x_n) に対する MCS は $x_1 = 0$ とすることで $\varphi(x_1, \dots, x_n)$ の MCS になっていることが容易にわかる。一方、 $\varphi(1, x_2, \dots, x_n)$ の (x_2, \dots, x_n) に対する MCS を求めたとき、 $\varphi(0, x_2, \dots, x_n)$ の (x_2, \dots, x_n) に対する MCS と同じ集合があった場合、明らかに MCS ではない。そのため、 $\varphi(1, x_2, \dots, x_n)$ の MCS から $\varphi(0, x_2, \dots, x_n)$ の MCS を除いた集合を求め、その集合要素に対して $x_1 = 1$ を付け加えることで $\varphi(1, \dots, x_n)$ の MCS を求めることができる。

以下の Algorithm 2 は、上記の考えに基づいて構造関数の BDD 表現から MCS を導出するアルゴリズムである (Rauzy, 1993)。

Algorithm 2 MCS.

```

function minsol(node)
    return node if is_zero(node) || is_one(node)
    x = minsol(get_zero(node))
    z = get_one(node) - x # set difference
    y = minsol(z)
    return create_node(x, y)
end

```

関数 `minsol` は再帰的に BDD の 0-枝と 1-枝を辿るようにして実行される。まず 0-枝のノードに対して `minsol` を実行し、MCS を得る。一方、1-枝については 0-枝に含まれる MCS を除いたノードに対して `minsol` を実行して MCS を取得し、最終的に、それぞれ 0,1 の値を決定した際に得られた MCS を子ノードとするノードを `create_node` 関数を使って作成している。アルゴリズム内の `is_zero`, `is_one` は、それぞれ 0-終端, 1-終端ノードかどうかを判別する関数。関数 `get_one` は 1-枝の子ノードを取得する。また演算子-は BDD 上の集合の差の演算を行う。

4. マルコフ解析

4.1 モデル表現

FT では原因と事象の関係を静的にモデル化している。そのため、時系列的な関係(例えば、Event 1 → Event 2 の順で発生したときだけ Event S が発生する)を表現することができない。そこで、より一般的なシステムの動特性を表現するためにマルコフ連鎖モデルがよく用いられる (Trivedi and Bobbio, 2017)。マルコフ連鎖とは離散状態の確率過程であり、将来の確率法則が現在の状態のみに依存して決定する性質をもつ。

マルコフ連鎖は離散時間あるいは連続時間の確率過程として定義することができるが、本稿では連続時間マルコフ連鎖を考える。いま、システムの状態が $\mathcal{S} = \{1, 2, \dots, n\}$ の離散状態で表現されているものとする。時刻 t におけるシステム状態を表すため、状態空間 \mathcal{S} 上の連続時間マルコフ連鎖 $\{X(t); t \geq 0\}$ を考える。時刻 t における状態確率ベクトル(行ベクトル)を $x(t) = [P(X(t) = i)]_{i=1, \dots, n}$ として定義する。時刻 0 における状態確率ベクトル $x(0)$ が与えられているもとで、任意時刻 t の状態確率ベクトル $x(t)$ は次の微分方程式を解くことによって得られる。

$$(4.1) \quad \frac{d}{dt} \mathbf{x}(t) = \mathbf{x}(t) \mathbf{Q}$$

ここで \mathbf{Q} は無限小生成行列と呼ばれ, (i, j) 要素が状態 i から状態 j への推移率で構成される正方行列である.

システムの動特性に基づいて信頼性尺度を算出する場合, システムの各状態に対する報酬ベクトルを与えることが多い. 列ベクトル \mathbf{r} の i 番目の要素 $[\mathbf{r}]_i$ に対して次の定義を行う.

$$(4.2) \quad [\mathbf{r}]_i = \begin{cases} 1 & \text{システムが故障している} \\ 0 & \text{システムが稼働している} \end{cases}$$

この時, 時刻 t でシステムが故障している確率 $F_S(t)$ は

$$(4.3) \quad F_S(t) = \mathbf{x}(t) \mathbf{r}$$

となる. 一方, 無限時間システムを稼働させた場合の状態確率ベクトルは, ある一定の条件のもと, 次の方程式を満足する定常分布 π で与えられる.

$$(4.4) \quad \pi \mathbf{Q} = \mathbf{0}, \quad \pi \mathbf{1} = 1.$$

ここで $\mathbf{1}$ は全ての要素が 1 の列ベクトル, $\mathbf{0}$ は全ての要素が 0 の列ベクトルである. 故障確率と同様に, 報酬ベクトル \mathbf{r} を用いると, システムが定常状態において故障している確率は

$$(4.5) \quad F_S = \pi \mathbf{r}$$

で与えられる.

4.2 計算アルゴリズム

マルコフ連鎖モデルでは, 状態確率ベクトルの算出が必要となる. 一般に状態確率ベクトルの計算は $\mathbf{x}(t)$ を求める計算(過渡解)と π を求める計算(定常解)で区別される. 過渡解を求める計算は式(4.1)の常微分方程式を数値的に解く方法と, 式(4.1)から得られる以下の行列指数表現から算出する手法がある.

$$(4.6) \quad \mathbf{x}(t) = \mathbf{x}(0) \exp(\mathbf{Q}t).$$

ここで \exp は行列指数関数であり, 任意の正方行列 \mathbf{A} に対して

$$(4.7) \quad \exp(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2!} + \cdots = \sum_{u=0}^{\infty} \frac{\mathbf{A}^u}{u!}$$

と定義される. ここで \mathbf{I} は単位行列である.

行列指数関数の計算は一般に数値的に困難であるが, マルコフ連鎖の場合は一様化と呼ばれる効率的な計算手法が用いられる (Stewart, 1994). 一様化では, $\mathbf{P} = \mathbf{I} + \mathbf{Q}/q$ で得られる離散時間マルコフ連鎖の推移確率行列 \mathbf{P} を用いる. ここで q は \mathbf{Q} の対角要素の絶対値の最大値である. このとき式(4.7)を使うと, 式(4.6)は

$$(4.8) \quad \mathbf{x}(t) = \mathbf{x}(0) \sum_{u=0}^{\infty} \frac{(qt)^u}{u!} e^{-qt} \mathbf{P}^u$$

となる. 許容誤差 ε に対して以下を満たす r_{\max} を定義する.

$$(4.9) \quad \sum_{u=0}^{r_{\max}} \frac{(qt)^u}{u!} e^{-qt} \geq 1 - \varepsilon.$$

ここで, r_{\max} はパラメータ qt のポアソン分布の分位点となっていることに注意する. このとき, 過渡解は

$$(4.10) \quad \mathbf{x}(t) \approx \tilde{\mathbf{x}}(t) = \mathbf{x}(0) \sum_{u=0}^{r_{\max}} \frac{(qt)^u}{u!} e^{-qt} \mathbf{P}^u$$

として計算され, 確率の性質から $\|\mathbf{x}(t) - \tilde{\mathbf{x}}(t)\|_1 \leq \varepsilon$ が成り立つ. ここで $\|\cdot\|_1$ はベクトルの 1 ノルムを表す.

一方, 定常解は式(4.4)の連立方程式を解くことによって得られる. 無限小生成行列 \mathbf{Q} は疎行列であることが多いため, Gauss-Seidel 法のような繰返し手法が使われる. 具体的に無限小生成行列が上三角行列 \mathbf{U} , 下三角行列 \mathbf{L} , 対角行列 \mathbf{D} によって次のように分解されるものとする.

$$(4.11) \quad \mathbf{Q} = \mathbf{U} + \mathbf{L} - \mathbf{D}.$$

これらの準備のもと, π が収束するまで次の更新を繰り返す.

$$(4.12) \quad \pi \leftarrow \frac{\pi \mathbf{U}(\mathbf{D} - \mathbf{L})^{-1}}{\pi \mathbf{U}(\mathbf{D} - \mathbf{L})^{-1} \mathbf{1}}.$$

通常の Gauss-Seidel 法とは異なり, $\pi \mathbf{1} = 1$ を保証するため正規化定数で更新値を割っていることに注意する (Bolch et al., 2006).

5. 統計解析

5.1 故障分布

信頼性分野における統計解析は故障分布(寿命分布)の推定を行うことが主な目的になる. 一般的に統計解析における推定はパラメトリックな方法とノンパラメトリックな方法がある. 寿命分布の推定においても双方用いられるが, 信頼性分野ではパラメトリックな方法がよく用いられる. 寿命分布の推定は, 信頼性試験やフィールドデータから得られた寿命データを用いて行われる. つまり, 故障時間のデータが必要となるが, 故障の観測がそもそも希な事象であり, 十分なサンプルが得られないことが多い, そのため, 十分なサンプル数を必要とするノンパラメトリックな方法は信頼性分野ではあまり用いられない. さらに, 指標として平均故障時間などの要約統計量を求めることが多いため, パラメトリックな方法が適している. 典型的に信頼性工学において故障分布として用いられるものを表 1 に示す (Meeker and Escobar, 1998).

5.2 相型分布

ここでは表 1 に示した故障分布とは異なる枠組みの連続時間マルコフ連鎖に基づく相型分布について説明する. 相型分布(Phase-Type Distribution, PH 分布)は, マルコフ過程の吸収時間として定義される確率分布である (Neuts, 1981). この分布は, 指数分布やその一般化として, さまざまな確率分布を近似する柔軟性を持ち, 信頼性工学をはじめ多くの分野で利用されている. いま, m 個の過渡状態と 1 つの吸収状態を持つ連続時間マルコフ連鎖を考え, その無限小生成行列を \mathbf{Q} とする. このとき, \mathbf{Q} は次のように表される.

$$(5.1) \quad \mathbf{Q} = \begin{pmatrix} \mathbf{T} & \boldsymbol{\tau} \\ \mathbf{0} & 0 \end{pmatrix}.$$

ここで, \mathbf{T} は $m \times m$ の行列であり, $\boldsymbol{\tau}$ は m 次元の列ベクトルである. このとき, 相型分布の確率密度関数および分布関数は次のように表される.

表 1. 信頼性工学における主な故障分布.

分布名	特徴・用途
指数分布(Exponential)	故障率が一定. 偶発故障のモデル化に適用.
ガンマ分布(Gamma)	段階的に進む劣化による故障をモデル化.
切断正規分布(Truncated Normal)	非負の寿命データをモデル化.
対数正規分布(Log-Normal)	故障時間が非対称分布を示す場合に適用.
切断ロジスティック分布(Truncated Logistic)	寿命データが非負かつ裾が重い場合に利用.
対数ロジスティック分布(Log-Logistic)	故障率が増減する場合や裾が重い分布に適用.
極値分布(ワイブル分布など)(Extreme Value/Weibull)	初期不良, 偶発故障, 経年劣化など広範な故障現象をモデル化可能.

$$(5.2) \quad f_{PH}(t) = \alpha \exp(\mathbf{T}t)\tau, \quad t \geq 0.$$

$$(5.3) \quad F_{PH}(t) = 1 - \alpha \exp(\mathbf{T}t)\mathbf{1}, \quad t \geq 0.$$

ここで, α は時刻 0 における状態を定義する確率ベクトル(行ベクトル)である. 相型分布のパラメータは $(\alpha, \mathbf{T}, \tau)$ であり, これらのパラメータを適切に選ぶことで, さまざまな確率分布を表現することができる. 実際には $\tau = -\mathbf{T}\mathbf{1}$ であるため, (α, \mathbf{T}) で十分であるが, ここでは記号の便宜上 $(\alpha, \mathbf{T}, \tau)$ とする. 特に, 一つの過渡状態 $m = 1$ とした場合, 相型分布は指数分布と一致する.

相型分布の重要な特性として閉包性が挙げられる (Maier and O'Cinneide, 1992). これは, 署み込み, 混合, 最小値, 最大値といった操作を施しても相型分布としての性質が保たれることを意味する. PH 分布の閉包性を具体例で示す. 独立な相型分布に従う確率変数 X_A と X_B が, それぞれパラメータ $(\alpha_A, \mathbf{T}_A, \tau_A)$ および $(\alpha_B, \mathbf{T}_B, \tau_B)$ を持つとする. このとき, 次の操作によって生成される確率変数 X_C も相型分布に従う. ただし, 以下では記号の便宜上 $\mathbf{0}$, \mathbf{O} は適切な次元の零ベクトルおよび零行列を表す.

(i) 署み込み (Convolution) : $X_C = X_A + X_B$.

$$(5.4) \quad \alpha_C = \begin{pmatrix} \alpha_A & \mathbf{0} \end{pmatrix}, \quad \mathbf{T}_C = \begin{pmatrix} \mathbf{T}_A & \tau_A \alpha_B \\ \mathbf{O} & \mathbf{T}_B \end{pmatrix}, \quad \tau_C = \begin{pmatrix} \mathbf{0} \\ \tau_B \end{pmatrix}.$$

(ii) 混合 (Mixture) : $X_C = \chi X_A + (1 - \chi) X_B$. ただし, χ は確率 p で X_A を選ぶ指標確率変数.

$$(5.5) \quad \alpha_C = \begin{pmatrix} p\alpha_A & (1 - p)\alpha_B \end{pmatrix}, \quad \mathbf{T}_C = \begin{pmatrix} \mathbf{T}_A & \mathbf{O} \\ \mathbf{O} & \mathbf{T}_B \end{pmatrix}, \quad \tau_C = \begin{pmatrix} \tau_A \\ \tau_B \end{pmatrix}.$$

(iii) 最小値 (Minimum) : $X_C = \min(X_A, X_B)$.

$$(5.6) \quad \alpha_C = \alpha_A \otimes \alpha_B, \quad \mathbf{T}_C = \mathbf{T}_A \oplus \mathbf{T}_B, \quad \tau_C = \tau_A \otimes \mathbf{1} + \mathbf{1} \otimes \tau_B.$$

ただし, \otimes はクロネッカーノルム積, \oplus はクロネッカーノルム和を表す.

(iv) 最大値 (Maximum) : $X_C = \max(X_A, X_B)$.

$$(5.7) \quad \alpha_C = \begin{pmatrix} \alpha_A \otimes \alpha_B & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{T}_C = \begin{pmatrix} \mathbf{T}_A \oplus \mathbf{T}_B & \tau_A \otimes \mathbf{I} & \mathbf{I} \otimes \tau_B \\ \mathbf{O} & \mathbf{T}_B & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{T}_A \end{pmatrix}, \quad \tau_C = \begin{pmatrix} \mathbf{0} \\ \tau_B \\ \tau_A \end{pmatrix}.$$

FT を使ったモデリングにおいて、AND ゲートで上位イベントが発生するためには、すべての下位イベントが発生する必要がある。つまり、上位イベントの発生時刻は下位イベントの最大値で表される。逆に、OR ゲートで上位イベントが発生するためには、いずれかの下位イベントが発生すればよいため、OR ゲートにおいて上位イベントが発生する時刻は下位イベントの最小値で表される。換言すると、FT において各原因事象の発生時間(各部品の故障時間)が相型分布に従う場合、上位イベントの発生時間も相型分布に従う。このような性質から、システム故障時間を相型分布で表現することの妥当性が示唆される。

相型分布は基礎となるマルコフ連鎖の構造により循環型と非循環型に分類される。マルコフ連鎖の状態遷移で、すべての過渡状態に高々 1 回しか到達しない場合を循環型、2 回以上到達する可能性がある場合を非循環型と呼ぶ。明らかに、循環型は非循環型の特別な場合であるため、分布の表現能力としては非循環型の方が高い。一方で、位相数(マルコフ連鎖の過渡状態数)を十分大きく取ることで、循環型は非循環型で近似できる。この性質から非循環型の相型分布は、理論的に任意の確率分布を任意の精度で近似できることが知られている(Asmussen and Koole, 1993)。さらに、先の畳み込み、混合、最小値、最大値の操作において、 X_A, X_B が非循環型の相型分布であるならば明らかに X_C も非循環型の相型分布となることがわかる。

非循環型の相型分布に関する性質として、あらゆる非循環型の相型分布は標準形に変換することができるところが知られている(Cumani, 1982)。以下は標準形の一つである CF1(Canonical Form 1)と呼ばれる形式である。

$$(5.8) \quad \alpha = (\alpha_1, \alpha_2, \dots, \alpha_m),$$

$$(5.9) \quad T = \begin{pmatrix} -\lambda_1 & \lambda_1 & & & O \\ & -\lambda_2 & \lambda_2 & & \\ & & \ddots & \ddots & \\ & & & -\lambda_{m-1} & \lambda_{m-1} \\ O & & & & -\lambda_m \end{pmatrix}, \quad \tau = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \lambda_m \end{pmatrix}.$$

ここで、 α_i は状態 i に初期状態がある確率、 λ_i は状態 i から状態 $i+1$ への遷移率を表し、 $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$ である。CF1 以外にも、Cox 分布型の標準形もある。これらの標準形は、同じ位相数を持つあらゆる非循環型の相型分布に対してパラメータ数が最小であることが示されている。具体的に CF1 の場合は、 $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)$ とすると、 (α, λ) の $2m$ 個のパラメータで表現され、これが位相数 m の非循環型の相型分布の最小パラメータ数となっている。ここで、 $\alpha_1 = 1$ であるためパラメータ自由度は $2m - 1$ であることに注意する。

このような性質から、FT モデルの故障分布が非循環型の相型分布で表現され、あらゆる非循環型が標準形に変換されることを考えると、故障データに基づいて $2m$ 個の標準形パラメータを推定することは、確率モデリングの観点からも妥当な故障分布推定が実施できていることがわかる。

5.3 パラメータ推定

統計解析による信頼性評価では、データから故障分布のパラメータを推定する必要がある。故障分布推定では、信頼性試験やフィールドデータから得られた故障データを用いるが、故障そのものが希な事象であるため信頼性試験やフィールドデータでは有限な時間で観測を打ち切った右打ち切りデータを扱うことが多い。

次のデータ構造を持つ打ち切りデータを考える。

$$(5.10) \quad \mathcal{D} = \{(t_1, \delta_1), (t_2, \delta_2), \dots, (t_N, \delta_N)\}.$$

ここで t_i は観測された故障時刻, δ_i は打ち切り指示を表す指標である. $\delta_i = 0$ の場合, 時刻 t_i で故障が観測されたことを意味し, $\delta_i = 1$ の場合, 時刻 t_i で故障が観測されないまま試験が終わったことを意味する. このデータに対する故障分布の最尤推定を考える. 以下は上記の \mathcal{D} に対する尤度関数である.

$$(5.11) \quad L(\boldsymbol{\theta}) = \prod_{i=1}^N f(t_i; \boldsymbol{\theta})^{\delta_i} R(t_i; \boldsymbol{\theta})^{1-\delta_i}.$$

ここで, $R(t; \boldsymbol{\theta})$ は信頼度関数に対応し, $f(t; \boldsymbol{\theta}) = -dR(t; \boldsymbol{\theta})/dt$ は密度関数である. $\boldsymbol{\theta}$ はパラメータベクトルである. 最尤推定法では上記の尤度を最大にするパラメータ $\boldsymbol{\theta}$ を求める.

一方, 相型分布を故障分布として信頼性解析を行う場合, 分布パラメータが非常に多く上記の尤度あるいは対数尤度を直接最大にするパラメータを見つけることが難しい. そのため EM アルゴリズムを用いたアルゴリズムが提案されている (Asmussen et al., 1996; Olsson, 1996). EM アルゴリズムは, 不完全データに対する最尤推定手法である. 不完全データとは, 観測されたデータと観測されなかったデータの両方を含むデータであり, EM アルゴリズムは, 観測されたデータからパラメータベクトルを推定するための反復的な手法である (Dempster et al., 1977; Wu, 1983). ここで, \mathcal{D} と \mathcal{U} を観測されたデータと観測されなかったデータとする. EM アルゴリズムは E ステップと M ステップからなる. E ステップでは, 事後分布 $p(\mathcal{U}|\mathcal{D}; \boldsymbol{\theta}')$ に基づく期待対数尤度 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ を計算する. M ステップでは, 期待対数尤度 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ を最大化するようなパラメータベクトル $\boldsymbol{\theta}$ を求める. これらのステップをパラメータベクトルが収束するまで繰り返す.

Olsson (1996) で与えられた打ち切りデータを用いた相型分布に対する EM アルゴリズムの式を示す. 以下の確率変数を定義する.

- $T^{(k)}$: k 番目のデータの故障時刻.

実際のデータは打ち切りデータであるため, すべてのサンプルに対して故障時刻が観測されているとは限らないことに注意する. 一方, 相型分布を支配するマルコフ連鎖の状態が観測されていないため, 次の未観測データ(確率変数)を導入する.

- $B_i^{(k)}$: $T^{(k)}$ のマルコフ連鎖が状態 i で開始したかどうかを表す指標確率変数.
- $Z_i^{(k)}$: $T^{(k)}$ のマルコフ連鎖が吸収状態に到達するまでに状態 i に滞在する時間.
- $Y_i^{(k)}$: $T^{(k)}$ のマルコフ連鎖が吸収状態に到達する直前に状態 i に滞在するかどうかを表す指標確率変数.
- $N_{i,j}^{(k)}$: $T^{(k)}$ のマルコフ連鎖が状態 i から状態 j に遷移する回数.

このとき, EM アルゴリズムの M ステップは以下のように表される.

$$(5.12) \quad \alpha_i \leftarrow \frac{\sum_{k=1}^N \mathbb{E}[B_i^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}]}{N}, \quad i = 1, \dots, m,$$

$$(5.13) \quad \tau_i \leftarrow \frac{\sum_{k=1}^N \mathbb{E}[Z_i^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}]}{\sum_{k=1}^N \mathbb{E}[Z_i^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}]}, \quad i = 1, \dots, m,$$

$$(5.14) \quad \lambda_{i,j} \leftarrow \frac{\sum_{k=1}^N \mathbb{E}[N_{i,j}^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}]}{\sum_{k=1}^N \mathbb{E}[Z_i^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}]}, \quad i = 1, \dots, m, \quad j = 1, \dots, m, \quad i \neq j.$$

ここで $\mathcal{D}_k = (t_k, \delta_k)$ である.

E ステップにおける期待値を計算するため以下のベクトルならびに行列を定義する.

$$(5.15) \quad \mathbf{f}(t) = \boldsymbol{\alpha} \exp(\mathbf{T}t),$$

$$(5.16) \quad \overline{\mathbf{f}}(t) = \boldsymbol{\alpha}(-\mathbf{T})^{-1} \exp(\mathbf{T}t),$$

$$(5.17) \quad \mathbf{b}(t) = \exp(\mathbf{T}t)\boldsymbol{\tau},$$

$$(5.18) \quad \overline{\mathbf{b}}(t) = \exp(\mathbf{T}t)\mathbf{1},$$

$$(5.19) \quad \mathbf{H}(t) = \int_0^t \exp(\mathbf{T}s)\boldsymbol{\tau}\boldsymbol{\alpha} \exp(\mathbf{T}(t-s))du,$$

$$(5.20) \quad \overline{\mathbf{H}}(t) = \int_0^t \exp(\mathbf{T}s)\boldsymbol{\tau}\boldsymbol{\alpha}(-\mathbf{T})^{-1} \exp(\mathbf{T}(t-s))ds.$$

このとき、未観測データに対する各期待値は以下のように計算される。

$$(5.21) \quad \mathbb{E}[B_i^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}] = \delta_k \frac{\alpha_i [\mathbf{b}(t_k)]_i}{\alpha \mathbf{b}(t_k)} + (1 - \delta_k) \frac{\alpha_i [\overline{\mathbf{b}}(t_k)]_i}{\alpha \overline{\mathbf{b}}(t_k)},$$

$$(5.22) \quad \mathbb{E}[Y_i^{(k,l)} | \mathcal{D}_k; \boldsymbol{\theta}] = \delta_k \frac{[\mathbf{f}(t_k)]_i \tau_i}{\alpha \mathbf{b}(t_k)} + (1 - \delta_k) \frac{[\overline{\mathbf{f}}(t_k)]_i \tau_i}{\alpha \overline{\mathbf{b}}(t_k)},$$

$$(5.23) \quad \mathbb{E}[Z_i^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}] = \delta_k \frac{[\mathbf{H}(t_k)]_{i,i}}{\alpha \mathbf{b}(t_k)} + (1 - \delta_k) \frac{[\overline{\mathbf{b}}(t_k) \overline{\mathbf{f}}(0) + \overline{\mathbf{H}}(t_k)]_{i,i}}{\alpha \overline{\mathbf{b}}(t_k)},$$

$$(5.24) \quad \mathbb{E}[N_{i,j}^{(k)} | \mathcal{D}_k; \boldsymbol{\theta}] = \delta_k \frac{\lambda_{i,j} [\mathbf{H}(t_k)]_{j,i}}{\alpha \mathbf{b}(t_k)} + (1 - \delta_k) \frac{\lambda_{i,j} [\overline{\mathbf{b}}(t_k) \overline{\mathbf{f}}(0) + \overline{\mathbf{H}}(t_k)]_{j,i}}{\alpha \overline{\mathbf{b}}(t_k)}.$$

上記の $\mathbf{H}(t)$ ならびに $\overline{\mathbf{H}}(t)$ は畳み込み積分が必要となるため計算コストが高い。Okamura et al. (2011, 2013) では一様化を適用した効率的な計算法を提案している。そのアルゴリズムでは、非循環型の相型分布に対して EM ステップの計算量が $O(mK)$ となっており、より大きな位相数の分布推定が可能となっている。

6. 信頼性解析ツール

6.1 FT ツール

FT は信頼性解析や安全性評価で頻繁に使用されており、商用およびフリーのソフトウェアがいくつか存在する。特に、機能安全規格(例: ISO 26262, IEC 61508)では、FTA(Fault Tree Analysis)を安全性分析の手法として推奨しており、一部の商用ソフトウェアはこれらの規格に対応している。商用ソフトウェアとしては、Reliability Workbench や RiskSpectrum によるソフトウェア群などが挙げられる。これらは製品開発における利用を想定しているため高い性能を持つ代わりにライセンス料が高い。

FT の研究においては、オープンソースやフリーソフトウェアの利用が想定される。フリーソフトウェアとしては、OpenFTA (<https://github.com/luyangshang/OpenFTA>) が広く知られている。OpenFTA は、静的 FT の構築と解析を支援するオープンソースのツールであり、定性的および定量的な解析を提供している。広島大学のグループは Julia 言語をベースとしたオープンソースの FT 解析ツール FaultTree.jl (<https://github.com/JuliaReliab/FaultTree.jl>) を開発しており、OpenFTA と同様に定量・定性解析を提供し、Jupyter Notebook 上の IJulia カーネルで利用することができるため、非常に柔軟な利用方法が可能である。

FT による解析例として、61 個の部品と 84 個のゲートからなるモデルを用いる (Rauzy, 1993)。実際に Julia 言語による記述例を付録 A に示す。表 2 は、FaultTree.jl により、BDD を構築する時間(BDD creation), MCS を導出する時間(MCS extraction), 信頼度を算出する時間(Reliability)を計測した結果である。なお数値例に利用した環境は Apple M2, 16GB memory

表 2. FT 解析の計算時間.

計算時間(秒)	
BDD creation	1.07
MCS extraction	3.08
Reliability	0.43

である。

BDD の構築について、対象モデルは 546,399 個のノードと 1,092,794 個の枝から構成される BDD を構築した。モデルの規模としては非常に大きいが、BDD の作成については、1 秒程度となっており、実用的に十分な速度を有している。FT から BDD の構築には BDD に対する Apply 演算と呼ばれるアルゴリズムを利用するが、これが有効に機能したものと考えられる。次に MCS の算出については、3 秒程度の時間がかかった。実際に MCS のための BDD は 11,318 個のノードと 22,632 個の枝から構成され、MCS の要素は 46,188 個となった。こちらについてもかなり大規模であるが、実用的な時間での分析が行えていることがわかる。定量的な指標である信頼度については 0.5 秒程度とかなり速く計算が行えている。

6.2 マルコフ解析ツール

マルコフ連鎖モデルに対する解析は行列操作であるため MATLAB や Python の Numpy などの汎用的な数値計算ライブラリが利用されることが多い。一方で、マルコフ連鎖解析でよく使われる一様化手法や定常分布の計算に特化したツールやライブラリも存在する。これらのツールは、特にマルコフ連鎖に関連する大規模な行列操作や、状態遷移に伴う複雑な確率分布の計算を効率的に実行できるように設計されている。例えば、SHARPE (<https://sharpe.pratt.duke.edu>) はマルコフ連鎖モデルだけでなく、FT や信頼性ブロック図も含む包括的な信頼性解析ツールである。また、PRISM (<https://www.prismmodelchecker.org>) は確率モデル検査のためのツールで、マルコフ連鎖や確率的モデルの解析に広く使用されている。Julia 言語を用いたものでは MarkovChain.jl (<https://github.com/wangnangg/MarkovChains.jl.git>) や、広島大学グループで開発した NMarkov.jl (<https://github.com/JuliaReliab/NMarkov.jl.git>) がある、これはパラメータ感度解析を行うこともできる。

マルコフ連鎖モデルに基づく信頼性解析の例として、図 4 に示すような生産システムを考える。システムは 4 つのステーションからなり、各ステーションでは与えられた部品を加工して次のステーションに送る。各ステーションの容量(処理中 + バッファで加工待ちの部品数)は 5 とし、送り先のステーションの容量がいっぱいの場合は処理を中断して送り先の容量が空くのを待つこととする。ステーション 1 で加工された部品は、ステーション 2 とステーション 3 へ分解されて送られ、ステーション 2 とステーション 3 で加工された部品はステーション 4 に送られる。ステーション 4 ではステーション 2 とステーション 3 で加工された部品を組み立て出荷する。一方で、ステーション 2 とステーション 3 で加工された部品は検査され、不良品がある場合は廃棄される。ステーション 2 とステーション 3 で加工される部品はペアで利用されるため、一方で不良品がある場合、部品のバランスがおかしくなる。一方で大量の不良品が発生すると、ステーション 2 あるいはステーション 3 の容量がいっぱいになり、生産が中断される(デッドロックが起こる)。このシステムの信頼性としてデッドロックが発生する確率をマルコフ連鎖モデルにより解析する。

デッドロックの解析は、特定のイベントの順序が発生要因となる場合がある。つまり、時間にともなう状態変化を正確に捉える必要があるため、単純な FT による分析では対応できず、

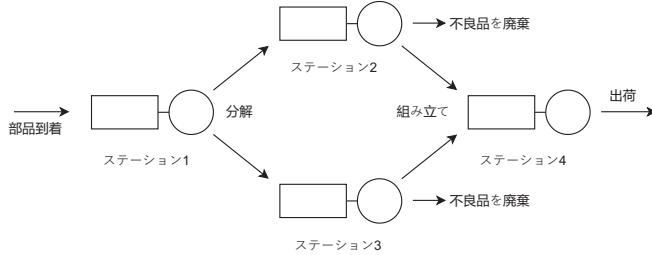


図 4. Fork-Join 型生産システム。

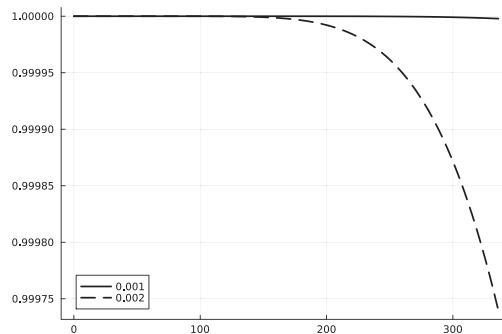


図 5. マルコフ解析による信頼度関数。

マルコフ連鎖によるモデル化および解析を必要とする。このような例としては、そのほかにも通信における輻輳解析や、ソフトウェアのバグ解析などが挙げられる。一方、FT は航空システムや電力システムなど、主に個々の部品の故障が独立してシステム故障に影響するシステムの解析を、より少ない計算量で実施できるため適している。

図 5 はステーション 1 への部品の到着が $\lambda = 5.0$ [1/hour]、全ステーションの処理率がすべて一律で $\mu = 10.0$ [1/hour]、ステーション 2 とステーション 3 の不良率が $f = 1/1000$ あるいは $f = 2/1000$ とした時の信頼度関数、つまり 366 時間(24 時間 \times 14 日間)稼働させたときにデッドロックの発生する確率を NMarkov.jl パッケージを使って計算した結果を示している。このような評価を行うことで、不良率 $f = 1/1000$ 程度であれば、デッドロックが発生するのは希であり、実用上ほとんど問題ないが $f = 2/1000$ であれば、デッドロックの発生に対する対策が必要であることがわかる。

生産システムに対するマルコフ連鎖モデルの状態数は 7,776 で、NMarkov.jl によって生産システムの信頼度関数を計算するのに要した時間は $f = 1/1000$ の場合で 1.56 秒、 $f = 2/1000$ の場合で 1.57 秒であった。ここで、生産システムに対するマルコフ連鎖モデルの無限小生成行列は確率ペトリネットツール gospn (<https://github.com/JuliaReliab/gospn>) によって生成された。数千状態という中規模モデルであるが、いずれの場合も計算時間が 2 秒以下と言う実用的な時間で解析が行えている。

6.3 統計解析ツール

データから分布パラメータを推定するための統計解析ツールとしては、R や Python の Scipy などの汎用的な統計解析ライブラリが利用されることが多い。特に、R の survival パッケージ

表 3. 相型分布のパラメータ推定の計算時間.

位相数	計算時間(秒)
10	0.523
20	0.954
30	1.458
40	2.842
50	3.680
60	4.821
70	5.969
80	7.209
90	8.494
100	9.855

ジを利用すると、打ち切りデータから典型的な指数分布、ガンマ分布、ワイブル分布などのパラメータ推定を行うことができる。一方で、相型分布に特化した統計解析ツールも存在する。広島大学グループが開発した R の mapfit パッケージ(<https://github.com/okamumu/mapfit>)は、EM アルゴリズムに基づいた相型分布のパラメータ推定を実行することができる。現在、打ち切りデータに対する推定は censored ブランチにのみ実装されているが、今後のバージョンでメインブランチにマージされる予定である。

Olsson (1996) で使われている右打ち切りデータを用いた相型分布に対する EM アルゴリズムの実行例を示す。R での利用方法は以下のコマンドで利用できる。

```
# mapfit の censored ブランチのインストール
devtools::install_github("okamumu/mapfit", ref="censored")

# mapfit パッケージの読み込み
library(mapfit)

# x: 故障時刻, delta: 打ち切りかどうか
# 打ち切りデータ ... は適切なデータを入力する
data <- data.frame(x=c(...), delta=c(...))

# 100 フェーズの CF1 を指定して推定
result <- phfit.surv(ph = cf1(100),
                      x = data$x,
                      delta = data$delta)

# 推定されたモデル
result$model
```

表 3 は mapfit パッケージを用いて、右打ち切りデータ(サンプル数 100)から相型分布のパラメータを推定するのに要した時間を示している。なお、数値例に利用した環境は Apple M2, 16GB memory である。位相数が増加するにつれて計算時間が増加しているが、位相数が 100 の場合でも 10 秒程度で推定が完了しており、十分な速度で推定が行えることがわかる。

7. まとめと今後の課題

本稿では、信頼性解析として、FT 解析、マルコフ解析、相型分布を用いた統計解析の三つの手法を紹介し、それらを支援するツールの紹介を行った。大規模で複雑なシステムの信頼性解析ではこれらのツールによる支援が必要不可欠である。一方で、現在のツールでは超大規模なシステムに対する解析には限界がある。FT 解析においては、BDD の構築や MCS の抽出に

比較的多くの計算コストがかかるため近似手法も含めた解析手法とそれを実装したツールが求められる。また、マルコフ解析においては、対象となるシステムの状態爆発の問題がある。先の例で見たように比較的簡単なモデルであっても指数的に状態数が増加するため、マルコフモデルについては構築と計算の双方における効率化が求められる。相型分布に対する統計解析においては、大きな状態数を持つ相型分布に対しても十分な速度でパラメータ推定が行える。しかしながら一方で、過適合の問題があり、罰則項を導入するなどの対策が求められる。また、実務的な観点からは、これらのツールを統合化し、同じプラットフォームを通じて信頼性解析が行える環境の構築が求められる。

付録 A : FaultTree.jl による FT の記述例

```

using FaultTree
ft = FTree()

@repeated ft c[1:61]

g62 = c[1] & c[2]; g63 = c[1] & c[3]; g64 = c[1] & c[4]; g65 = c[1] & c[5];
g66 = c[1] & c[6]; g67 = c[1] & c[7]; g68 = c[1] & c[8]; g69 = c[1] & c[9];
g70 = g62 | c[10]; g71 = g63 | c[11]; g72 = g64 | c[12]; g73 = g65 | c[13];
g74 = g62 | c[14]; g75 = g63 | c[15]; g76 = g64 | c[16]; g77 = g65 | c[17];
g78 = g62 | c[18]; g79 = g63 | c[19]; g80 = g64 | c[20]; g81 = g65 | c[21];
g82 = g62 | c[22]; g83 = g63 | c[23]; g84 = g64 | c[24]; g85 = g65 | c[25];
g86 = g62 | c[26]; g87 = g63 | c[27]; g88 = g64 | c[28]; g89 = g65 | c[29];
g90 = g66 | c[30]; g91 = g68 | c[31]; g92 = g67 | c[32]; g93 = g69 | c[33];
g94 = g66 | c[34]; g95 = g68 | c[35]; g96 = g67 | c[36]; g97 = g69 | c[37];
g98 = g66 | c[38]; g99 = g68 | c[39]; g100 = g67 | c[40]; g101 = g69 | c[41];
g102 = g66 | c[42]; g103 = g68 | c[43]; g104 = g67 | c[44]; g105 = g69 | c[45];
g106 = ftkofn(ft, 3, g70, g71, g72, g73); g107 = ftkofn(ft, 3, g74, g75, g76, g77);
g108 = ftkofn(ft, 3, g78, g79, g80, g81); g109 = ftkofn(ft, 3, g82, g83, g84, g85);
g110 = ftkofn(ft, 3, g86, g87, g88, g89); g111 = ftkofn(ft, 3, g94, g95, g96, g97);
g112 = ftkofn(ft, 3, g98, g99, g100, g101); g113 = g90 & g92; g114 = g91 & g93;
g115 = g102 & g104; g116 = g103 & g105; g117 = g113 | c[46]; g118 = g114 | c[47];
g119 = g107 | g108 | c[52]; g120 = g109 | g110; g121 = g66 | g117 | c[48];
g122 = g68 | g118 | c[49]; g123 = g67 | g117 | c[50]; g124 = g69 | g118 | c[51];
g125 = ftkofn(ft, 2, g121, g123, g122, g124); g126 = g111 | g112 | g125 | c[53];
g127 = g115 & g120; g128 = g116 & g120; g129 = g62 | g127 | c[54];
g130 = g63 | g128 | c[55];
g131 = g64 | g127 | c[56]; g132 = g65 | g128 | c[57]; g133 = g62 | g129 | c[58];
g134 = g63 | g130 | c[59]; g135 = g64 | g131 | c[60]; g136 = g65 | g132 | c[61];
g137 = ftkofn(ft, 3, g133, g134, g135, g136); g138 = g106 | g119 | g137;
g139 = g62 | g66 | g117 | g129 | c[48]; g140 = g63 | g68 | g118 | g130 | c[49];
g141 = g64 | g67 | g117 | g131 | c[50]; g142 = g65 | g69 | g118 | g132 | c[51];
g143 = g139 & g140 & g141 & g142; g144 = g111 | g112 | g143 | c[53];
top = g126 & g138 & g144;

@parameters ft begin
    // setting parameters for c[1] to c[61]
    c[1] = 0.01
    ...
    省略 ...
end

@time bdd = ftbdd!(ft, top)
@time begin
    result = mcs(ft, bdd)
    mincuts = extractpath(result)
end
@time fresult = prob(ft, bdd)

```

参考文献

- Asmussen, S. and Koole, G. (1993). Marked point processes as limits of Markovian arrival streams, *Journal of Applied Probability*, **30**, 365–372.
- Asmussen, S., Nerman, O. and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm, *Scandinavian Journal of Statistics*, **23**(4), 419–441.
- Barlow, R. E. and Proschan, F. (1965). *Mathematical Theory of Reliability*, John Wiley & Sons, New York.
- Birnbaum, Z. W., Esary, J. D. and Saunders, S. C. (1961). Multi-component systems and their reliability, *Technometrics*, **3**(1), 55–77.
- Bolch, G., Greiner, S., de Meer, H. and Trivedi, K. S. (2006). *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, 2nd ed., John Wiley & Sons, New York.
- Bryant, R. (1986). Graph-based algorithms for Boolean function manipulation, *IEEE Transactions on Computers*, **C-35**(8), 677–691.
- Cumani, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions, *Microelectronics and Reliability*, **22**(3), 583–602.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **B-39**, 1–38.
- Dugan, J. B., Bavuso, S. J. and Boyd, M. A. (1992). Dynamic fault-tree models for fault-tolerant computer systems, *IEEE Transactions on Reliability*, **41**(3), 363–377.
- Maier, R. and O'Cinneide, C. (1992). A closure characterisation of phase-type distributions, *Journal of Applied Probability*, **29**(1), 92–103.
- Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*, John Wiley & Sons, New York.
- Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Dover Publications, Mineola, New York.
- Okamura, H., Dohi, T. and Trivedi, K. S. (2011). A refined EM algorithm for PH distributions, *Performance Evaluation*, **68**(10), 938–954.
- Okamura, H., Dohi, T. and Trivedi, K. S. (2013). Improvement of EM algorithm for phase-type distributions with grouped and truncated data, *Applied Stochastic Models in Business and Industry*, **29**(2), 141–156.
- Olsson, M. (1996). Estimation of phase-type distributions from censored data, *Scandinavian Journal of Statistics*, **23**, 443–460.
- Rauzy, A. (1993). New algorithms for fault trees analysis, *Reliability Engineering & System Safety*, **40**(3), 203–211.
- Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, New Jersey.
- Trivedi, K. S. and Bobbio, A. (2017). *Reliability and Availability Engineering: Modeling, Analysis, and Applications*, Cambridge University Press, Cambridge, UK.
- Wu, C. F. J. (1983). On the convergence properties of the EM algorithm, *Annals of Statistics*, **11**(1), 95–103.

Reliability Analysis and Utilization of Tools

Hiroyuki Okamura, Junjun Zheng and Tadashi Dohi

Graduate School of Advanced Science and Engineering, Hiroshima University

In this paper, we introduce tools commonly used in the field of reliability. Reliability evaluation can be broadly categorized into two approaches: one involves probabilistic models, such as fault trees and Markov models, while the other uses statistical methods to identify models based on data obtained from reliability tests. Both approaches rely on reliability analysis tools for evaluation. This paper provides fundamental knowledge about reliability and explains specific reliability analysis tools in detail.

極値統計論に基づくモデリング

吉田 拓真¹・北野 利一²

(受付 2024 年 11 月 29 日；改訂 2025 年 5 月 13 日；採択 5 月 26 日)

要 旨

豪雨や地震などに起因する自然災害、ファイナンスにおける金融リスク、そして製品寿命など様々な現象に関するデータを扱う応用分野において、データ全体の中で極めて大きい、または小さい値の発生確率の見積りは信頼性評価やリスク管理の観点から重要な課題である。この課題に対する統計学的なテーマはデータの最大値や最小値、あるいはそれらに近い分位点の予測となる。そのためには、裾の挙動にのみ焦点を当てた確率モデルを構築することが不可欠である。極値統計学はそのための方法論を示すものであり、その稀な事象が生起する分布の裾のモデリングに数理的に重要な枠組みを与える。本稿では極値統計学の基本的な考え方、統計理論、モデリングについて総説する。特に、極値統計手法で実データ分析を行ったときに、その解析結果を現象の頻度特性の把握へとフィードバックする際に重要な役割を担う概念である再現期間と再現レベルは少し掘り下げた解説を行う。また、気象データや株価データなど極値統計学が活躍するデータはクラスターデータの形式で得られることが多い。そのため、クラスターデータに対する極値統計モデリングの方法論についても議論する。本稿では統計ソフト R のパッケージを紹介しながら降雨量データへ極値統計モデリングの適用例を示す。

キーワード：一般化極値分布、一般化パレート分布、極値統計学、クラスターデータ、空間データ、再現期間。

1. はじめに

実用科学の様々な分野で得られるデータには時に極端な値が含まれる。その極端な値が起きる原因の特定は難しいことが多い、外れ値として扱われ、解析から取り除かれることも多い。しかし、自然災害対策や製品開発現場では、そういった極端な値が起きる確率を正しく見積もることが重要となることもある。極端なデータとはデータ全体を見ると分布の裾に位置するようなものを指す。実際に、データ全体は正規分布で表現できても裾の部分にブレがあるような例は多い。そもそもデータのボリュームゾーンと裾を同一の分布(モデル)で表現することは難しい。よって、極端な値に相当する分布の裾の予測にはそれに特化したモデリングが重要となる。極値統計学はデータ全体の分布の裾の挙動にのみ焦点を当て、その部分を取り出して新たな確率分布を再構築するための手法である。

極値統計学を用いた裾の予測モデルとして大きく分けて二つの確率分布が用いられる。一つは一般化極値分布、もう一つは一般化パレート分布である。一般化極値分布はいわゆるデータ

¹ 鹿児島大学 大学院理工学研究科：〒890-0065 鹿児島県鹿児島市郡元 1-21-35; yoshida@sci.kagoshima-u.ac.jp

² 名古屋工業大学 大学院工学研究科：〒466-8555 愛知県名古屋市昭和区御器所町; kitano@nitech.ac.jp

の最大値の確率分布を記述し, 一般化パレート分布はデータ全体のうちの閾値超過データの確率分布を表現する。2つのモデルで共通する特徴は分布の裾の挙動は大きく三つのタイプに分類されることである。一つは正規分布や指数分布のように指数関数の速度で裾が減衰する指数テール, もう一つはベータ分布や一様分布のように有限のエンドポイントが存在するようなショートテール, 最後の一つは t 分布やパレート分布のように極端な値が発生しやすいヘビーテールである。興味ある現象のリスク評価を行う際に, 分布の裾が上の三つのどのタイプに属するかを知ることは重要であろう。多くの統計モデリングはデータに対して特定の分布モデルを想定するところからスタートするが, それはこの裾のタイプを一つに絞っていることになる。極端な値の発生メカニズムをモデル化するにあたり, どの裾タイプを持つか? から調べることは極値統計論を適用する応用上の観点からは重要となる。一般化極値分布, 一般化パレート分布はこれらの裾のタイプを一つのパラメータで統一的に表現しているので, このパラメータを推定することで裾タイプの判定から予測モデル構築までを一連に行うことができる。

本稿では, 分布の裾挙動をモデリングするための極値統計学の基礎理論, 予測モデリングの構築, 結果の解釈の仕方について先行研究のサーベイ, 実データ適用を交えて総説する。特に, 結果の解釈として再現レベル, 再現期間と呼ばれる重要な概念があるが, これらの概念は極値統計モデルを現象の頻度特性の把握にフィードバックする際に重要な役割を担うため, 少し掘り下げた解説を行う。第2章ではデータの最大値が持つ分布として一般化極値分布を用いたモデルリングを解説する。閾値超過部分のモデリングとして有用な一般化パレート分布は第3章で議論する。第4章では一般化極値分布, 一般化パレート分布の基礎モデルの発展についてサーベイする。さらに, 気象データ, 金融データなど極値統計学分野が活躍する分野はデータはクラスターデータの形で与えられることが多い。第5章ではクラスターデータに対する極値統計モデリングについても議論する。第6章でまとめと, 本稿で述べることが出来なかった多変量極値分布について簡単に触れる。

2. 一般化極値分布

2.1 一般化極値分布(Generalized Extreme Value distribution, GEV)

確率変数 X_1, \dots, X_n は独立同一に分布関数 F に従って発生されたとする。つまり, $F(x) = P(X_i \leq x)$ である。このとき, 最大値 $\max_{1 \leq i \leq n} X_i$ の分布は

$$\begin{aligned} P\left(\max_i X_i \leq x\right) &= P(X_1 \leq x, \dots, X_n \leq x) \\ &= \prod_{i=1}^n P(X_i \leq x) \\ &= F^n(x) \end{aligned}$$

となり, 一般に $F(x) < 1$ である x に対して $F^n(x) \rightarrow 0(n \rightarrow \infty)$ と 0 に退化してしまう。一方で, 最大値に対して位置と尺度を n に応じてうまく調整した場合は有限の確率分布を持つことが知られている。Fisher and Tippet (1928), Gnedenko (1943) は適当な分布 F に対して, 数列 $\mu_n = \mu_n(F)$, $\sigma_n = \sigma_n(F)$, 分布関数 G が存在し,

$$P\left(\frac{\max_i X_i - \mu_n}{\sigma_n} \leq x\right) = F^n(\sigma_n x + \mu_n) \rightarrow G(x)$$

が成立することを示し, さらに G が F の条件によって三つのタイプに分類されることを示した。Jenkinson (1955) はその三タイプの確率分布を統一的に表現した。すなわち, G は分布関数 F に依存するパラメータ $\gamma = \gamma(F) \in \mathbb{R}$ を用いて,

$$(2.1) \quad G(x) = G_\gamma(x) = \exp[-(1 + \gamma x)_+^{-1/\gamma}]$$

と表される。ただし、 $(x)_+ = \max\{x, 0\}$ である。また、 $\gamma = 0$ のとき、 $G_0(x) = \lim_{\gamma \rightarrow 0} G_\gamma(x) = \exp[-e^{-x}]$ となる。実際に、(2.1) は γ の符号に応じて

$$G_\gamma(x) = \begin{cases} \exp[-(1 + \gamma x)^{-1/\gamma}], & \gamma > 0 \quad (\text{F}) \\ \exp[-(1 + \gamma x)_+^{-1/\gamma}], & \gamma < 0 \quad (\text{W}) \\ \exp[-e^{-x}], & \gamma = 0 \quad (\text{G}) \end{cases}$$

と三つのタイプに分類される。タイプ(F)はフレッシュタイプ(Fréchet type)、(W)は逆ワイブルタイプ(reversed Weibull type)、(G)はガンベルタイプ(Gumbel type)と呼ばれる。一般に、 $\gamma > 0$ の(F)はヘビーテール、 $\gamma < 0$ に対応する(W)はショートテール、 $\gamma = 0$ の(G)は指数テールである。特に、タイプ(W)は $1 + \gamma x > 0$ を満たす必要があるため、有限エンドポイント $x < -1/\gamma$ が存在する。この統一表現されたモデル(2.1)は一般化極値分布(Generalized Extreme Value distribution, GEV)と呼ばれている。実データ分析では F が未知のため、当然 γ とその符号も未知である。データ分析では、基本的に極値分布のタイプを事前に指定することができないため、三タイプの極値分布の統一表現はデータ分析を行う上で大変有用であろう。一般に、分布関数 F がパラメータ γ の極値分布 G_γ を持つとき、 $F \in \mathcal{D}(G_\gamma)$ と表される。

ここで、(2.1)を書き直すと、

$$P\left(\max_i X_i \leq x\right) = F^n(x) \approx G_\gamma\left(\frac{x - \mu_n}{\sigma_n}\right)$$

と近似できることから、 (μ_n, σ_n) もパラメータ (μ, σ) と見ると

$$P\left(\max_i X_i \leq x\right) \sim G_\gamma\left(\frac{x - \mu}{\sigma}\right) = \exp\left[-\left(1 + \gamma \frac{x - \mu}{\sigma}\right)_+^{-1/\gamma}\right]$$

で近似できることになる。右辺の分布を $GEV(\mu, \sigma, \gamma)$ と書き、 μ を位置パラメータ、 σ を尺度パラメータ、 γ を形状パラメータと呼ぶ。GEV はデータの最大値の確率分布をモデリングすることで、元の分布 F の裾を表現するモデルとみなせる。

次に一般化極値分布を理解する上で、特に重要な性質を挙げる。自然数 h について、 hn 個の確率変数 X_1, \dots, X_{hn} は独立同一に分布関数 F に従うとする。また、 $F \in \mathcal{D}(G_\gamma)$ とする。このとき、 n が十分に大きいとすると、ある (μ_{hn}, σ_{hn}) が存在して

$$P\left(\max_{1 \leq i \leq hn} X_i \leq x\right) = F^{hn}(x) \sim G_\gamma\left(\frac{x - \mu_{hn}}{\sigma_{hn}}\right)$$

が成立する。一方で(2.1)より、

$$P\left(\max_{1 \leq i \leq hn} X_i \leq x\right) = \{F^n(x)\}^h \sim \left\{G_\gamma\left(\frac{x - \mu_n}{\sigma_n}\right)\right\}^h$$

と書けるので、

$$G_\gamma\left(\frac{x - \mu_{hn}}{\sigma_{hn}}\right) \approx \left\{G_\gamma\left(\frac{x - \mu_n}{\sigma_n}\right)\right\}^h$$

が成立することになる。この近似から

$$(2.2) \quad \mu_{hn} = \mu_n + \sigma_n \frac{h^\gamma - 1}{\gamma}, \quad \sigma_{hn} = h^\gamma \sigma_n$$

という関係式が直接計算より得られる。この(2.2)は極値に対する相似性を表現していると言える。さらに、 h は自然数から実数に拡張でき、その場合は確率過程におけるブロックサイズ

の増加率に相当する。

一般化極値分布の応用のための事項をいくつか紹介する。 (2.1)の確率変数を $-X_i$ に置き換えると, $P(\max_i(-X_i) \leq x) = P(\min_i X_i > -x)$ であり, (2.1)において, $y = -x$, パラメータ $\alpha, \lambda > 0$ について $\gamma = -1/\alpha, \mu_n \rightarrow -\lambda, \sigma_n \rightarrow \lambda/\alpha$ とすると

$$P\left(\min_i X_i > y\right) \approx \exp\left[-\left(\frac{y}{\lambda}\right)^\alpha\right]$$

となり, ワイブル分布に帰着する。このとき, $\gamma < 0$ なので, (W) である。例えば, X_i を i 番目の製品の故障までの時間とすると, $P(\min_i X_i > y)$ は“製品は少なくとも y (時間) までは故障しないという事象の確率”に相当する。このように, 信頼性工学において故障寿命予測によく用いられるワイブル分布 (Weibull, 1939, 1951) は極値統計学に深い関連がある。同様に, 極値統計論は最弱リンク理論 (weakest link theory) の基礎にもなっている (Meeker et al., 2021)。

また, F を正規分布 $\mathcal{N}(0, 1)$ とする。このとき, $\mu_n = \sqrt{2 \log n - \log(\log n) - \log(4\pi)}$, $\sigma_n = 1/\mu_n$ と取ることで $\gamma = 0$ としたときの (2.1) が成立する (de Haan and Ferreira, 2006)。これは正規分布に従う確率変数の最大値がガンベル分布に収束することを意味するが, その収束が遅いことは有名である。そのため, penultimate 近似の議論がある (Fisher and Tippett, 1928)。一方で, F が自由度 ν の t 分布であれば, $\gamma = 1/\nu$ となることがすぐにわかる。つまり, $\gamma > 0$ となるわけであるが, t 分布が代表的なヘビーテールとなる分布であることは周知の事実であろう。 t 分布の性質から $\gamma = 1/2$ で分散が発散, $\gamma = 1$ で期待値も発散することがわかる。

このように, 各分布について (2.1) が成立するかどうかを確認するのは可能なものも多いが, 一般にそれなりに手間であり容易ではない。ただし, ある条件下では簡単に示すことができる。いま, 元の分布関数 F は 2 階微分が連続とする。このとき, von-Mises (1936) は以下を示した: (2.1) が成立するための十分条件は

$$\lim_{y \rightarrow y^*} \frac{d}{dy} \frac{1 - F(y)}{F'(y)} = \gamma$$

または

$$(2.3) \quad \lim_{y \rightarrow y^*} \frac{\{1 - F(y)\}F''(y)}{\{F'(y)\}^2} = -\gamma - 1$$

が成立することである。ここで, $y^* = \sup\{y : F(y) < 1\}$, F' , F'' は F の 1 階, 2 階微分である。この性質は一般に von-Mises condition と呼ばれており, この性質を使うと指定した F について (2.1) が成立するかどうか簡単に確認できる。なお, 式 (2.3) はその上の式の左辺の微分を実行しただけであるが, GEV への収束 (特に, ガンベルタイプ $\gamma = 0$) に対して, より具体的なイメージが可能である。例えば, $\gamma = 0$ に収束する分布関数 F に対して, ハザードレート $F'(y)/\{1 - F(y)\}$ が定数に収束する場合 (例えば指数分布が該当する), 十分大きな値である y について

$$-\frac{F''(y)}{F'(y)} = \frac{F'(y)}{1 - F(y)} \approx \text{const}$$

が成立する。この性質はガンベルタイプを特徴づける性質を表しており, ガンベルタイプの別称を指数タイプということもあるのはこの性質に由来する (北野 他, 2025)。なお, 正規分布はガンベルタイプの極値分布を持つが, そのハザードレートは単調増加し, 定数に収束しない。また, フレシェタイプや逆ワイブルタイプもハザードレートは定数に収束しないが, それらの分布については (2.3) により,

$$\frac{F'(y)}{1 - F(y)} \propto \frac{F''(y)}{F'(y)}$$

が成立することを示唆している。

2.2 GEV のパラメータ推定

前章で触れたように GEV は最大値の分布であるから、データからある意味での最大値を計算しないといけない。しかし、データ全体の最大値はひとつだけであり、そこからパラメータ推定をするのは難しい。そこで、以下に述べるブロック最大化法(block maxima)を用いるのが通例である。まず、データ X_1, \dots, X_n を k 個のブロックに分割する： $\{X_1, \dots, X_{n_1}\}$, $\{X_{n_1+1}, \dots, X_{n_2}\}, \dots, \{X_{n_{k-1}+1}, \dots, X_n\}$ 。便宜上、 $n_0 = 1, n_k = n$ と書く。このとき、各ブロックの最大値を $Y_i = \max_{n_{i-1} < j \leq n_i} X_j$ とすると、 $\{Y_1, \dots, Y_k\}$ が集まる。この k 個のデータに対して、 $\text{GEV}(\mu, \sigma, \gamma)$ を想定し、位置、尺度、形状パラメータ (μ, σ, γ) を推定する。GEV の密度関数は

$$f_{\text{GEV}}(x|\mu, \sigma, \gamma) = \frac{1}{\sigma} \left(1 + \gamma \frac{x - \mu}{\sigma}\right)^{-\frac{1}{\gamma} - 1} \exp \left[-\left(1 + \gamma \frac{x - \mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right]$$

なので、最尤推定量は

$$(\hat{\mu}, \hat{\sigma}, \hat{\gamma}) = \underset{\mu, \sigma, \gamma}{\operatorname{argmax}} \sum_{i=1}^k \log f_{\text{GEV}}(Y_i|\mu, \sigma, \gamma)$$

と定義される。

このように、ブロック最大化法はデータをブロックに分割して疑似的に最大値を複数個構成して、それを元にパラメータ推定をする。ここで議論になるのがデータのブロック分割方法であろう。例えば、 X_i が期間内で i 日目に観測されたデータとしよう。すると、 $\{X_i : i = 1, \dots, 365\}$ は最初の 1 年間で得られたデータになる。よって、数十年分の観測データがある場合、各年単位をブロックとすると年最大値データを考えることになり、それを元に作られた GEV は年最大値分布と解釈される。同様に、年最大値の代わりに月最大値をはじめとする任意の期間最大値分布を GEV でモデリングすることが可能である。ただし、注意点もある。例えば、2.5 節、3.5 節、5.3 節でも扱っている日降雨量データに対して月最大値分布を考えたいとしよう。しかし、日本では基本的に豪雨は台風シーズンの夏季に発生しやすく、冬季では発生しにくいため、各月の降雨量が同分布に従うという仮定が非現実的となる。そのため、一般に時系列構造だけでなく、解析対象のデータの領域知識を考慮したブロック分割を行う必要がある。また、時系列構造がない独立に得られたデータにおいても、ブロックの決め方は難しい。データをランダムに等分割することは可能であるが、上の例と同じく各ブロックのデータを同分布と見なしてよいか、さらには、“何の最大値を考えているのか?”といった解釈性が問題となる。GEV によるモデリングでは基本的に意味のあるブロック分割を考える必要があり、適切な分割数 k は解析の目的に応じて定まると言える。

2.3 最尤推定量の漸近理論

パラメータベクトル $\theta = (\mu, \sigma, \gamma)^\top$ について、GEV における対数尤度関数を

$$\ell_{k, \text{GEV}}(\theta) = \ell_{k, \text{GEV}}(\mu, \sigma, \gamma) = \frac{1}{k} \sum_{i=1}^k \log f_{\text{GEV}}(Y_i|\mu, \sigma, \gamma)$$

と書く。また、フィッシャー情報行列を

$$I_{\text{GEV}}(\theta) = E \left[\frac{\partial}{\partial \theta} \ell_{k, \text{GEV}}(\theta) \frac{\partial}{\partial \theta^\top} \ell_{k, \text{GEV}}(\theta) \right]$$

とする。このとき, Bücher and Segers (2017) は, Y_1, \dots, Y_k が独立同一に $\text{GEV}(\mu, \sigma, \gamma), \gamma > -1/2$ に従うとき, 最尤推定量 $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{\gamma})$ について

$$(2.4) \quad \sqrt{k}(\hat{\theta} - \theta) \xrightarrow{D} \mathcal{N}_3(\mathbf{0}_3, I(\theta)^{-1})$$

が成立することを示した (Smith, 1985 も参照されたい)。ここで, $\mathbf{0}_3 = (0, 0, 0)^\top$, \mathcal{N}_p は p 変量正規分布を表す記号である。

この結果で注意したいのは, 推定量の収束レートが元々のデータ数 n ではなく, 元のデータから構成されたブロック数 k で決まるところにある。一般に極値理論によると, Y_i が GEV に従うのは n が大きいときの極限での話なので厳密には $Y_i \sim \text{GEV}(\mu, \sigma, \gamma)$ という仮定は少々強いものとなっている。したがって, 上の結果では漸近分布の期待値は 0 となっているが, 本来はブロックの作り方に応じたバイアスが生じるはずである。そうでないと, 収束レートが k に依存するので, 結局 $k = n$ のときが最も推定量の精度が良いことになる。しかし, $k = n$ はすべてのデータを使うことを意味し, “GEV が最大値の分布である” ことに反してしまう。また, 上でも述べたように, GEV による解析ではブロックはデータの領域知識から定まることが多い。その意味で, 一般には(2.4)のように, $\text{GEV}(\mu, \sigma, \gamma)$ をブロックとその数 k を固定としたパラメトリックモデルとして考えることになる。

2.4 GEV の再現期間 (Return Period) と再現レベル (Return Level)

極値統計モデリングを実データに適用し, その結果を検討する際に特に重要な概念となるのが再現期間 (Return Period), 再現レベル (Return Level) である。いま, $Y \sim \text{GEV}(\mu, \sigma, \gamma)$ とすると,

$$P(Y > y) = 1 - \exp \left[- \left(1 + \gamma \frac{y - \mu}{\sigma} \right)^{-1/\gamma} \right]$$

より, 与えられた確率 $p \in (0, 1)$ に対して, $p = P(Y > z)$ を満たす $z = z(p)$ は

$$z(p) = \begin{cases} \mu + \frac{\sigma}{\gamma} \{ [-\log(1-p)]^{-\gamma} - 1 \}, & \gamma \neq 0, \\ \mu - \sigma \log(-\log(1-p)), & \gamma = 0 \end{cases}$$

で与えられる。例えば, 想定規模に対するブロックサイズを T とし, $p = 1/T$ とすると, ブロック内で生成される T 個のサンプルのうち 1 つが $z(1/T)$ を超過すると解釈できる。これを書き直すと

$$(2.5) \quad T = \frac{1}{P(Y > z(1/T))}$$

である。さて, ここで Y がどのような確率変数であったかを思い出してほしい。GEV の構成によると, Y はブロック最大値であった。そして, 多くの実用例では GEV は時系列で得られたデータに利用される。例えば, 1 日単位で得られるデータに対してブロックサイズを 365 とすると Y は年最大値である。そのとき, T 個のサンプルのうち 1 つが $z(1/T)$ を超過するというのは, T 年間のうち 1 年間は得られるデータの最大値が $z(1/T)$ を超過すると解釈できる。言い換えると, $z(1/T)$ を超えるのは T 年間のうち(平均的に)1 年程度であるというわけである。このように, GEV が降雨量のような時系列データの極端な事象の確率モデルに利用されていることから T は“期間”と解釈されることが多く, 固定した確率 p の逆数 T とその分位点 $z(p) = z(1/T)$ はそれぞれ再現期間, 再現レベルと呼ばれる。

例えば, 建造物や製品の耐久性試験に関わる衝撃力データ(波の高さ, 地震の大きさ, 自動車の衝突時の速度など含む)では, 今後 50, 100, 200 年間など長期間を想定した T に基づく $z(1/T)$ を計算し, その値に耐えうる製品を造るような設計がなされる。その意味で, T は実際

の観測期間よりも長い期間を想定されることが多い。この再現期間、再現レベルという名称は極値統計学以外ではあまり見かけないが、実際には再現レベルは分位点そのものであるので、その点に注意すると理解しやすい。

再現期間については次のような解釈もできる。ここで設定しているように、ブロックサイズを1年とすると、サンプルは1年に1回しか得られない。よって、想定サンプルサイズ T が $T = 1, 2, \dots$ と増えていく状況では同じように1年、2年…と時間が経過していく。つまり、 T は何らかの時間である。ここで T_0 を生起確率 p の幾何分布に従う確率変数とみると $E[T_0] = 1/p$ となることから、 $T = E[T_0]$ とみるとこれは事象 $z(p)$ が起こるまでの平均的な待ち時間と解釈できる。このように再現期間 T を幾何分布の期待値であると解釈をすれば、極端な事象 $z(p)$ が T 年に1度は必ず起こるわけではないことを強く認識できる。

さて、再現期間・レベルは、パラメータ推定値の診断図としても機能する。まず期間 $T > 1$ に対して、 $(T, z(1/T))$ を再現レベル図(Return Level Plot)と呼ぶ。データ Y_1, \dots, Y_k がGEVにうまく当てはまっているならば、 Y_1, \dots, Y_k の順序統計量 $Y_{(1)} \geq \dots \geq Y_{(k)}$ と、再現期間点(確率点の逆数) $T_i = 1/p_i, p_i = i/(k+1), i = 1, \dots, k$ として、 $\{(T_i, Y_{(i)}): i = 1, \dots, k\}$ の多くの点が曲線 $(T, z(1/T))$ の周りに分布することが期待できる。これはいわゆるQQ-plotの考え方と同じである。なお、 $z(1/T)$ は未知パラメータ (μ, σ, γ) に依存するので、実際には推定値をプラグインする。

再現期間・レベルについて、上では元となるデータ Y_i は年最大値、つまりブロックは1年間であるとしていた。それでは、一般にブロックが $N(> 1)$ 年、つまり、元となるデータが N 年最大値である場合はどのように考えるのだろうか。いま、 N 年最大値の確率変数を $Y^N \sim GEV(\mu_N, \sigma_N, \gamma)$ と書く。 (Y^1, μ_1, σ_1) は先ほどまでの1年最大値 (Y, μ, σ) に対応する。このとき、GEVの分布関数を $G_{N,\gamma}((x - \mu_N)/\sigma_N)$ とし、 $p^N = 1 - G_{N,\gamma}((x - \mu_N)/\sigma_N)$ を固定したときの x についての逆数を $z_N(p^N)$ とする。このとき、再現期間を T^N と書くと、

$$T^N = \frac{N}{p^N}$$

で表される。ここで、 p^N, T^N はそれぞれ p, T の N 乗という意味ではなく、新しい記号として導入している。これを用いて再現レベル図を作成することは当然できるのだが、すると当然ブロックの作り方によって結果が変わってくる。このことは数学的には自然なことなのだが、データ解釈の立場に立つと理解をしづらい。ちなみに、(2.2)より再現レベルはブロックの作り方に依存しない。これを踏まえて、もう少し再現期間について考察してみる。Coles (2001)は別の再現レベル図として $(1/\{-\log(1 - p^N)\}, z(p^N))$ を用いている。これは再現期間を改めて

$$(2.6) \quad T^N = \frac{N}{-\log(1 - p^N)}$$

と定義していることと同値である。実はこの設定の下で、(2.2)より(2.6)は任意の N で

$$T = T^N = \left(1 + \frac{\gamma(x - \mu)}{\sigma}\right)^{1/\gamma}$$

となり、ブロックの作り方に依存しなくなることがわかる。したがって、この新しい T を用いて $(T, z(1/T))$ を再現レベル図として定義する方がデータ解釈の意味では理解しやすい。これを踏まえると再現期間の定義は2通りあることになり、Tawn (1988)もこの点に言及している。また、北野 (2020)はこのことについて、稀な事象の生起をポアソン分布で表現した場合のモデリングとの関連も示している。

2.5 GEV の適用例

1976 年 1 月 1 日～2023 年 12 月 31 日までに東京都八王子市の降雨量観測所で観測された日降雨量データに一般化極値分布を適用する。降雨量を始めとする気象関連データは気象庁の HP からダウンロードできる。八王子市観測所は 1976 年 1 月 1 日から降雨量観測を始めており、1975 年以前はデータの蓄積がない。さて、48 年間でデータの蓄積は 17521 日分あるが、そのうち降雨がない日(降雨量 0)は 12097 日で 69% に相当し、実際に降雨があった日は 5424 日であった。降雨日は年平均 113 日(標準偏差 12.2)であった。図 1 に降雨量をプロットしている。図中の赤点は各年の最大降雨量である。

さて、GEV は独立同分布に従う確率変数に対して適用される。よって、まずは独立同分布性に言及する。降雨量データが各日で同分布に従うことを調べるのは困難であるが、興味あるのは分布の裾部分なので全体を表現する分布には興味がない。しかし、例えばブロックサイズを 1 年とした年間の最大降雨量に同分布を仮定するのは自然であろう。次に独立性に言及する。降雨量データは通常、データ全体が独立に生起しているとは考えにくい。しかし、今回は豪雨のみに関心があり、晴れや小雨には関心がない。よって、極値事象に限定した自己相関のような従属性の有無を調べる。ここでは、Coles et al. (1999) が提案した(上側)裾従属係数(裾依存係数とも言う)を利用した

$$\chi(\ell) = \lim_{q \rightarrow 1} P(F(X_{i+\ell}) > q | F(X_i) > q) = \lim_{q \rightarrow 1} \frac{P(F(X_{i+\ell}) > q, F(X_i) > q)}{1 - q}, \quad \ell \geq 1$$

を用いる。Spearing et al. (2023) もこの基準を用いて極値事象の時系列の定常性の有無を調べている。裾従属係数は本来 2 変量の閾値超過確率であり、次章で述べる单変量の閾値超過確率である一般化パレート分布と密接に関連している。簡単にその性質を述べると、 $0 \leq \chi(\ell) \leq 1$ であり、 $\chi(\ell) = 0$ のときは裾独立(asymptotic independence)、 $\chi(\ell) > 0$ のとき、裾従属(asymptotic dependence)と言われる。また、 $\chi(\ell) = 1$ のときは完全従属(full dependence)である。裾従属係数の計算は統計ソフト R のパッケージ extRemes が便利である (Gilleland and Katz, 2016)。ここでの裾従属係数 $\chi(\ell)$ は、時系列に得られたデータについて i 日目の(一様分布に変換された)データが大きな値を持ったとき、続けて ℓ 日後にもまた大きな値を持つ確率と解釈される。特に、降雨量データに対して $\ell = 1$ とすると、ある程度の大暴雨が降った日の翌日も大雨であった確率を見ていることになる。これは“2 日連続で大雨が降る確率”とは異なるので注意して解釈されたい。実際には $\chi(\ell)$ は 1 に近い q で、 F や P は経験分布関数で近似する。八王子市の日降雨量では、全体の 98% 分位点が 46mm に相当した。よって、 $q = 0.98$ で計算すると、 $\chi(\ell) = 0.131$ であった。つまり、観測期間全体を通してある程度の降雨があった日の翌日も同程度かそれ以上の降雨であるケースは稀だったので、降雨量に関する極値事象は独立に生起されると仮定して GEV を当てはめる。ちなみに $\ell > 1$ とすると裾従属係数はより小さい値に

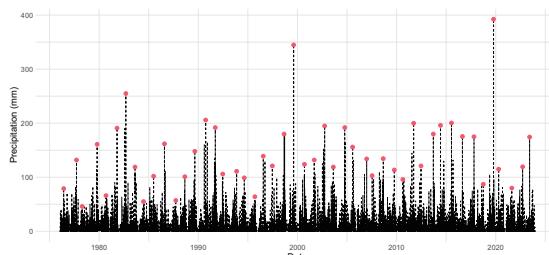


図 1. 1976 年 1 月 1 日～2023 年 12 月 31 日における八王子市での日降雨量(mm)。赤点は各年の(48 年分)の最大降雨量。

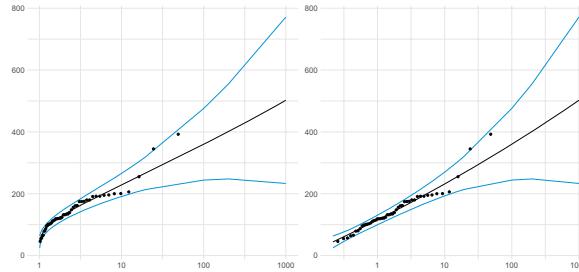


図 2. 八王子市の降雨量データの再現期間(x 軸)と再現レベル(y 軸). 左は再現期間を $T = 1/p$, 右は再現期間を $T = 1/(-\log(1-p))$ としたもの.

なったが, これは直感的にも明らかであろう. 気象データに関する推測では季節依存なども重要な要素となる. 特に降雨量は台風シーズンは大きな値を取る傾向がある. しかし, 今回はそのような解析は省略する.

さて, 八王子市の降雨量データの極値解析に移る. ここではブロックサイズは 1 年とする. つまり, 年最大値データ(図 1 の赤点)を GEV に適用する. なお, 解析には R パッケージ `ismev` を用いた (Stephenson, 2018). すると, パラメータの最尤推定値は $(\hat{\mu}, \hat{\sigma}, \hat{\gamma}) = (115, 48.1, 0.04)$ となった. 形状パラメータ γ の 95% 信頼区間は $[-0.156, 0.236]$ となった. 形状パラメータは符号の正負どちらも取り得るため, それほどヘビーテールしていないことがわかる. 言い換えると, 年最大値データはヘビーテールする可能性はあるもののそれなりに安定した裾挙動をしていることがわかる. 続いて再現期間・レベルを確認する. 図 2 には 2 つの再現期間に対する再現レベルをプロットしている. 図中の曲線(黒)が再現期間に対する再現レベル, 点が年最大値データの再現期間に対する QQ-plot, 青線が漸近理論から求められる再現レベルの 95% 信頼区間である. 左は再現期間が $T = 1/p = 1/P(X > z(1/T))$ なので $T > 1$ である. 右図は再現期間 $T = 1/\{-\log(1 - P(X > z(1/T)))\}$ なので, $T > 0$ が定義域となる. このブロック最大によって得られた結果は, 1 年最大値データであることを強く意識するのであれば左図が理解しやすい. しかし, それゆえにブロックサイズの柔軟性はない. 他方, 右図は(2.2)に示した相似性の図示表現となっているため, ブロックサイズによらない解釈ができる. このモデリングは観測期間 48 年としたため, $T > 48$ のデータが存在しない. しかし, 再現レベル図から $T > 48$ の将来の豪雨時の降雨量を見積もることができる.

GEV を用いた波浪や潮位などに関する気象データ解析は Haigh et al. (2010), Caires (2011a, 2011b), Wahl et al. (2013) なども参照されたい.

2.6 上位順序統計量の極値分布

GEV は観測データの最大値分布として定義されていた. パラメータ推定の際はブロック最大値を利用したが, それでも推定に使うデータ数は少ない. また, あるブロックの 2 番目に大きいデータは他のブロックの最大値よりも大きな値を持つこともあるだろう. そこで, 各ブロックについて上位 1 番目から r 番目($r > 1$)までのデータを使ってパラメータ推定を行う方法もある (Smith, 1986; Tawn, 1988). いま, $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F$ の順序統計量を $X_{(1)} \geq \dots \geq X_{(n)}$ とする. このとき, F が(2.1)を満たすとすると, $(X_{(1)}, \dots, X_{(r)})$ の同時分布は $n \rightarrow \infty$ の下で密度関数

$$(2.7) \quad g(x_{(1)}, \dots, x_{(r)}) = \exp \left[- \left[1 + \gamma \left(\frac{x_{(r)} - \mu}{\sigma} \right) \right]^{-1/\gamma} \right] \prod_{i=1}^r \frac{1}{\sigma} \left[1 + \gamma \left(\frac{x_{(i)} - \mu}{\sigma} \right) \right]^{-1/\gamma-1}$$

を持つ極値分布で近似される。ここで, $x_{(1)} \geq \cdots \geq x_{(r)}$ であり, $1 + \gamma(x_{(i)} - \mu)/\sigma > 0 (i = 1, \dots, r)$ である。 $\gamma = 0$ のときの密度関数は(2.7)を $\gamma \rightarrow 0$ として得られる。ここで $r = 1$ のときは(2.1)に帰着する。(2.7)の証明は高橋・志村(2016)の2.3節が詳しい。

パラメータ (μ, σ, γ) の推定の際は、各ブロックの $1 \sim r$ 番目のデータが独立に(2.7)に従うと見なして最尤推定を行えばよい。ブロック数を k とするとパラメータ推定に使えるデータ数が kr 個となり推定値が安定する。Guades Soares and Scotto(2004)は検定に基づく r の決め方を議論している。

3. 一般化パレート分布

3.1 一般化パレート分布(Generalized Pareto distribution)

さて、(2.1)が成立するとき、元の分布関数 F は閾値 w について、 w に依存する尺度 σ_w が存在して

$$\begin{aligned} P(X_i > w + \sigma_w x | X_i > w) &= \frac{P(X_i > w + \sigma_w x)}{P(X_i > w)} \\ (3.1) \quad &\rightarrow (1 + \gamma x)_+^{-1/\gamma} \quad \text{as } w \rightarrow \infty \end{aligned}$$

が成立する(証明は付録Aを参照)。これを書き直すと、改めて $\sigma = \sigma_w$ とおいて

$$P(X_i - w > x | X_i > w) \approx \left(1 + \frac{\gamma x}{\sigma}\right)_+^{-1/\gamma}$$

を得る。右辺の分布は一般化パレート分布(Generalized Pareto distribution, GP)と呼ばれ、尺度パラメータ σ 、形状パラメータ γ を用いて $GP(\sigma, \gamma)$ で表現される。 $\gamma = 0$ のとき、右辺は $1 - e^{-x/\sigma}$ と指数分布になり、 $\gamma < 0$ で有限のエンドポイント $x < -\sigma/\gamma$ が得られる。GPはデータ全体のうち閾値 w より大きい部分が持つ確率分布であると言える。また、 γ の符号に関する解釈はGEVと同じである。形状パラメータは閾値に依存せず、元の分布関数 F にのみ依存するが、尺度パラメータは閾値 w と F に依存する。しかし、結局 F が未知なので、 σ_w 自身をパラメータだと思ってこれも推定するのが基本方針となる。この方法の枠組みは、閾値を超える部分のモデリングすることから、閾値超過法(peak over threshold method)と呼ばれている(Davison and Smith, 1990)。

3.2 GP の最尤推定

GPでは閾値より大きいデータのみで解析する。いま、閾値 w に対して、 $X_i > w$ となるデータについて $Y_i = X_i - w$ とし、 $\{Y_1, \dots, Y_k\}$ と再ラベリングする。この k 個のデータに GP を当てはめる。GPの密度関数は

$$f_{GP}(y | \sigma, \gamma) = \frac{1}{\sigma} \left(1 + \frac{\gamma y}{\sigma}\right)_+^{-1/\gamma-1}, \quad y > 0$$

であるので、最尤推定量は

$$(\hat{\sigma}, \hat{\gamma}) = \underset{\sigma, \gamma}{\operatorname{argmax}} \sum_{i=1}^k \log f_{GP}(Y_i | \sigma, \gamma)$$

である。

統計モデル $GP(\hat{\sigma}, \hat{\gamma})$ はその分位点を用いて診断できる。いま、固定した $p \in (0, 1)$ に対して、 $p = P(Y_i > y)$ を満たす $y = y(p)$ は

$$y(p) = \frac{\sigma}{\gamma}(p^{-\gamma} - 1)$$

となる。よって、 $p_i = i/(k+1), i = 1, \dots, k, \{Y_1, \dots, Y_k\}$ の順序統計量を $Y_{(1)} \geq \dots \geq Y_{(k)}$ とすると、モデルの当てはまりが良ければ $\{(y(p_i), Y_{(i)}) : i = 1, \dots, k\}$ は 45 度線 $y = x$ に沿うように分布するはずである。実際には、 (σ, γ) にはその推定量 $(\hat{\sigma}, \hat{\gamma})$ を代入する。これは実質的に GP に対する QQ-plot である。

GP は多くのデータに適用可能である。しかし、GP は GEV とは違いデータ全体に対して閾値を設定するので、年最大などの閾値の取り方で分布の解釈を与えるのは難しい。したがって、閾値 w 、または k の選択が重要となる。GEV のようにデータ解釈ベースに設定されることもあるが、基本的には調整パラメータと思って何らかの意味で最適化する必要がある。閾値の選択法は様々なものが提案されているが、ここでは標準的な方法である平均残差生存図 (mean residual life plot, Coles, 2001) を紹介する。平均残差生存図は平均超過量図 (mean excess plot) とも呼ばれる。ここでは、 $\gamma < 1$ を仮定する。これは実用上は決して強い仮定ではない。実際、 $\gamma = 1$ はコーシー分布の裾と同程度であり、かなり重い裾である。閾値 w_0 と確率変数 $X - w_0 \sim \text{GP}(\sigma, \gamma), (X > w_0)$ について

$$E[X - w_0 | X > w_0] = \frac{\sigma w_0}{1 - \gamma}$$

が成立する ($\gamma \geq 1$ のとき、期待値は発散)。ここで、 $w > w_0$ となる別の閾値について

$$E[X - w | X > w] = \frac{\sigma w}{1 - \gamma} = \frac{\sigma w_0 + \gamma w}{1 - \gamma}$$

となり (付録 A を参照)、条件付き期待値は w_0 より大きい閾値 w に関して傾きが固定の線形構造を持つ。よって、 $k = k_w = \sum_{i=1}^n I(X_i > w)$ とするとき、

$$E[X - w | X > w] \approx \frac{1}{k} \sum_{i=1}^n (X_i - w) I(X_i > w)$$

となるので、 $w > w_0$ の部分で

$$(3.2) \quad \left(w, \frac{1}{k} \sum_{i=1}^n (X_i - w) I(X_i > w) \right)$$

は線形構造を持つはずである。このように、(3.2)を図示して、安定的に線形構造を有する部分として w_0 を決定すればよいことがわかる。ただし、極値統計論としては $E[X - w | X > w]$ の w に関する線形構造に着目するのは妥当であるが、実データに対する (3.2) は簡単に解釈できないことが多い (明確な線形構造を見出すことが難しい)。特に、 w が大きくなりすぎると、 $X_i > w$ となるデータ数が少なくなり、不安定な挙動をすることが多くなる。実際には、(3.2) により判断できるのは、適切な閾値が存在すると考えられる区間 (これを (w_L, w_U) と表す) にとどまり、その区間内で実際にパラメータ推定を行い、推定値の挙動が安定している箇所を最適値として用いる。なお、区間 (w_L, w_U) の設定についても、(3.2) のグラフの形状やデータ知識などをもとに解析者の判断により決定される。パラメータの推定値の挙動であるが、極値統計論によると γ は閾値に対して不变であるが、尺度パラメータは $w > w_L$ について、 $\sigma_w = \sigma_{w_L} + \gamma(w - w_L)$ となり、閾値に対して線形となる。よって、

$$\sigma^* = \sigma_w - \gamma w \equiv \text{const}$$

とすれば σ^* は理論的には閾値不变となる。このように、閾値 (w_L, w_U) の範囲で $(w, \hat{\sigma}^*), (w, \hat{\gamma})$ の安定性を評価し (定数的挙動をしていることを確認)、最終的に適切と思われる閾値をひとつ

決定する。なお、閾値 w が決まれば、閾値より大きいデータ数 $k = \sum_{i=1}^n I(X_i > w)$ も決まる。

上記の方法は閾値をオートマティックに決定するものではない。伝統的なオートマティックな閾値選択方法はレビュー論文 Scarrott and MacDonald (2012) を参照されたい。近年では Wadsworth (2016), Northrop et al. (2017), Murphy et al. (2024) などが新規手法を開発している。

3.3 GP の漸近理論

GP では閾値を調整パラメータと見ることが多いため、その影響を考慮した理論構築がなされる。そこで(3.1)を改めて

$$(3.3) \quad P(X_i > w + \sigma_w x | X_i > w) \approx (1 + \gamma x)_+^{-1/\gamma} + A(w)B(x) \quad \text{as } w \rightarrow \infty$$

と書く。ここで、 A は $\lim_{w \rightarrow \infty} A(w) = 0$ を満たす関数であり、 $B(x)$ は x の関数である。 A や B の具体的な形は de Haan and Ferreira (2006) の Theorem 2.3.8 で与えられている。要は $A(w)$ が閾値の選択による GP の近似バイアスであり、これが推定量 $(\hat{\sigma}, \hat{\gamma})$ のバイアスに影響するのである。

GP の対数尤度関数を改めて

$$\ell_{k, \text{GP}}(\sigma, \gamma | w) = \frac{1}{k} \sum_{i=1}^k \log f_{\text{GP}}(X_i - w | \sigma, \gamma) I(X_i > w)$$

と書く。GP における最尤推定量 $(\hat{\sigma}, \hat{\gamma})$ の漸近理論は 2 通りのものが知られている。ひとつは Smith (1987) によるもので、この論文では閾値 $w = w_n$ は n に応じて変動する実数列と考えて、 $n \rightarrow \infty$ のとき、 $w_n \rightarrow \infty$, $P(Y_i > w_n) \rightarrow 0$, $n P(Y_i > w_n) \rightarrow \infty$ と仮定している。つまり、閾値より大きいデータの個数 $k = \sum_{i=1}^n I(X_i > w_n)$ 自体が確率変数で $k/n - P(Y_i > w_n) \xrightarrow{P} 0$ である。すると、 $P(Y_i > w_n) \rightarrow 0$ という仮定は実質的に $k/n \rightarrow 0$ と同じであり、元のデータ数 n が増える度に閾値より大きいデータ数 k も増えるが n と比較するとその数は少ないという状況を表現していることになる。さて、このような状況で、また適切な仮定の下で

$$\sqrt{k} \begin{bmatrix} \frac{\hat{\sigma}}{\sigma} - 1 \\ \hat{\gamma} - \gamma \end{bmatrix} \xrightarrow{D} \mathcal{N}_2(\boldsymbol{\lambda}_1, \Sigma_1)$$

が成り立つ。ここで、 $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \lambda_{1,2})$ は $A(w_n)$ や $B(x)$ に依存して決まるバイアス項であり、 $A(w_n)$ と k によって決まる関数 $\boldsymbol{\lambda}(k)$ に対して、 $\sqrt{k}\boldsymbol{\lambda}(k) \rightarrow \boldsymbol{\lambda}_1$ と仮定されていると解釈できる。このバイアスの詳細な式は Smith (1987) を参照されたい。また、 Σ_1 は

$$\Sigma_1 = \begin{bmatrix} 2(\gamma + 1) & -(1 + \gamma) \\ -2(1 + \gamma) & (1 + \gamma)^2 \end{bmatrix}$$

であり、これは $\ell_{k, \text{GP}}(\sigma, \gamma | w)$ のフィッシャー情報行列から導かれる。一般的に、 w_n が大きいとデータは GP によく当てはまるので(3.3)の近似精度が上がる($A(w_n)$ が小さくなる)。しかし、その分 k が小さくなり、パラメータ推定に用いるデータ数が少なくなることから推定量の分散は増大する。逆に、 w_n が小さいと k は大きくなり推定量の挙動は安定する(分散が減少)が、(3.3)の精度が悪くなる($A(w_n)$ の値が大きくなる)。そのため、収束レートである $k^{-1/2}$ はバイアスと分散のトレードオフのバランスを取る役割を果たしていることになる。

Drees et al. (2004) は Smith (1987) とは別のシナリオを考えている。元のデータ X_1, \dots, X_n の順序統計量を $X_{(1)} \geq \dots \geq X_{(n)}$ とするとき、彼らは閾値を上から $k+1$ 番目のデータ: $w = X_{(k+1)}$ と仮定している。つまり、パラメータ推定に使うデータは $\{Y_i = X_{(i)} - X_{(k+1)}, i = 1, \dots, k\}$ で

あるとみなせる。その上で、Smith (1987) と同じ解釈が可能な設定： $n \rightarrow \infty$ の下で $k \rightarrow \infty$, $k/n \rightarrow 0$ を考えている。大まかにいうと、Smith (1987) の設定では閾値 w が実数、閾値を超えるデータ数 k が確率変数であり、Drees et al. (2004) では閾値が確率変数、閾値を超えるデータ数 k は実数となっている。この設定は、2.6 節で述べた上位順序統計量の極値分布においてブロック数を 1 とした場合に関連している。Drees et al. (2004) は以上のような設定といいくつかの適切な条件の下で

$$(3.4) \quad \sqrt{k} \begin{bmatrix} \frac{\hat{\sigma}}{\sigma} - 1 \\ \hat{\gamma} - \gamma \end{bmatrix} \xrightarrow{D} \mathcal{N}_2(\boldsymbol{\lambda}_2, \Sigma_2)$$

が成立することを示した。ただし、 $\boldsymbol{\lambda}_2 = (\lambda_{21}, \lambda_{22})$ は Smith (1987) の理論と同じく分布の近似バイアス $A(w)$ に依存する項である。しかし、 $A(w)$ の設定が Smith (1987) とは異なるので $\boldsymbol{\lambda}_1$ との直接の比較は難しい。また、分散共分散行列は

$$\Sigma_2 = \begin{bmatrix} 2(1 + \gamma) + \gamma^2 & -(1 + \gamma) \\ -2(1 + \gamma) & (1 + \gamma)^2 \end{bmatrix}$$

で与えられる。ここで注目するのは Σ_2 の $(1, 1)$ 成分が Σ_1 のそれとは異なっていることである。この違いは閾値の取り方の違いによるものである（本質的な意味合いはほとんど等しいとしても）。結果から閾値を実数とした方が尺度パラメータ $\hat{\sigma}/\sigma$ の推定量の分散が小さいことがわかる。ただし、推定量の分散が小さいからと言って閾値を固定で考える方がよいと結論付けるのは早計である。なぜならば、両者のバイアスは一概にどちらが小さいか判断できるものではないためである。またバイアスは、要は GP の近似バイアスから発生しているが、その分布関数の近似誤差の推定は一般に困難である。分布関数の近似誤差、ひいてはバイアスを推定する研究も Gomes and Martins (2002), Gomes and Pestana (2007), Caeiro and Gomes (2011) などで議論されているが、このバイアスを補正した推定量を考える場合の最適な閾値 w （または閾値より大きいデータ数 k ）は補正しない場合とは異なるのでやはり直接の比較は難しい。

このような背景から、 (σ, γ) の信頼区間や仮説検定を行う場合はバイアスを 0 と仮定し、 Σ_1, Σ_2 に γ の推定量を代入するいわゆるプラグイン法が主流となっている。近年では、de Haan and Zhou (2024) が極値解析におけるブートストラップ法の理論的正当性を示しているため、ブートストラップ信頼区間などは今後広く利用されると思われる。

なお、Drees et al. (2004) の理論は推定に用いるデータが $Y_i = X_{(i)} - X_{(k+1)}$ なので Y_1, \dots, Y_k が独立な確率変数ではない。したがって、(3.4) は通常の最尤推定の証明が通用せず、順序統計量、極値統計理論、確率過程を融合させた証明を元に構築されている。

3.4 GP の再現レベル

いま、 $X \sim F$ として、閾値 w を超えた部分が $GP(\sigma, \gamma)$ に従うとすると、 $x > w$ に対して、

$$\begin{aligned} P(X > x) &= P(X > w)P(X > w + x - w | X > w) \\ &= P(X > w) \left(1 + \frac{\gamma(x - w)}{\sigma} \right)^{-1/\gamma} \end{aligned}$$

と表すことができる。したがって、固定した $p \in (0, 1)$ に対して、 $p = P(X > z)$ を満たす $z = z(p)$ は

$$z(p) = w + \frac{\sigma}{\gamma} \left\{ \left(\frac{p}{P(X > w)} \right)^{-\gamma} - 1 \right\}$$

となる。 X と同じ分布を持つ元のデータ X_1, \dots, X_n に対して閾値 w を超えるデータ数を k と

すると, $P(X > w) \approx k/n$ と推定できるので, これに置き換えて

$$z(p) \approx w + \frac{\sigma}{\gamma} \left\{ \left(\frac{np}{k} \right)^{-\gamma} - 1 \right\}$$

と近似される. 実際には, (σ, γ) にはその推定量 $(\hat{\sigma}, \hat{\gamma})$ が用いられる.

GP に対しても再現レベルという言葉は使われる. 想定サンプルサイズ m に対して, $p = 1/m$ とすると, m 回に 1 回程度 $z(p) = z(1/m)$ を超過するイベントが発生すると解釈できる. また, もしデータが 1 年で R 個で得られたとすると ($R = 365$ とすると 1 日単位で得られるデータに相当), $T = m/R$ として $z(1/m) = z(1/(RT))$ は T 年に 1 回超過する値, つまり再現期間 T 年の再現レベルであると解釈できる. 想定サンプルサイズの考え方方が GEV と若干異なるので注意されたい. GP ではこの T が(年数の)再現期間と定義される. これをもう少し紐解くと, $1/p = m = RT$ より,

$$T = \frac{1}{Rp} = \frac{1}{RP(X > z(1/T))} = \frac{1}{RP(X > w)} \left(1 + \frac{\gamma(z(1/T) - w)}{\sigma} \right)^{1/\gamma}$$

となる.

再現レベルの定義は GEV に対応させて

$$\frac{1}{m} = 1 - \exp[-P(X > x)] \approx 1 - \exp \left[- \left(1 + \frac{\gamma(x - w)}{\sigma} \right)^{-1/\gamma} \frac{k}{n} \right]$$

を満たす x として定義されることもある. この場合, x は $P(X > x) = -\log(1 - 1/m)$ の逆数として計算される. m が大きいときは $\log(1 - 1/m) \approx -1/m$ なので上のケースとほぼ等しい. 3.5 節の実例では年数再現期間 T について $m = 365 \times T$ を想定したが, どちらも違いはほとんどなかった.

2.4 節では GEV の再現期間の定義は 2 通りあると述べたが, そのうち GP の再現期間に対応するのは(2.6)であることがその定義からわかる.

3.5 GP の適用例

2.5 節と同じ八王子市の降雨量データに今度は GP を当てはめる. GEV で利用した R パッケージ `ismev` でも GP を実行できるが, ここではもうひとつの極値解析の標準的なパッケージである `eva` を使う (Barder and Yan, 2020). まずは閾値の選択を行う. 図 3 には 3.2 節で議論した平均残差生存図(MRL, 左上)とパラメータ $(w, \hat{\gamma}), (w, \hat{\sigma}^*)$ の挙動(右上, 左下)を示している. これらは挙動が安定している箇所の確認が目的なので信頼区間はあえて載せていない. これらの 3 つのグラフでは縦線で (w_L, w_U) を示している. この範囲内で MRL とパラメータは比較的安定して定数挙動していることがわかる. ここでは, 図中の赤丸である $w = 92$ 用いて GP を適用する. 右下図にはデータと閾値を図示している. このとき, 閾値より大きいデータ数は $k = 79$ で $k/n \approx 0.005$ であった. この最適な閾値を用いたときの GP のパラメータの最尤推定値は $(\hat{\sigma}, \hat{\gamma}) = (45.9, 0.05)$ となった. バイアスを 0 とした場合の形状パラメータ γ の 95% 信頼区間は $[-0.183, 0.287]$ であった. 2.5 節で示した GEV の形状パラメータの推定値とほぼ等しくなっている. γ はほぼ 0 でありガンベルタイプに近いが, GEV 解析と同様に若干ヘビーテールとなっている.

次に, 当てはまったモデル $GP(\hat{\sigma}, \hat{\gamma})$ の適合度を視覚的に確認する. 図 4 の左図は 3.2 節で解説した $\{(y(p_i), Y_{(i)}) : i = 1, \dots, k\}$ と $y = x$ の直線, また, $(\hat{\sigma}, \hat{\gamma})$ の漸近理論とデルタ法から導かれる $y(p)$ の 95% 信頼区間である. すべてのデータが信頼区間に収まっていることが確認できる. 図 4 の右図は 3.4 節において $R = 365$ と置いた再現期間(年数) T (対数スケール)と再

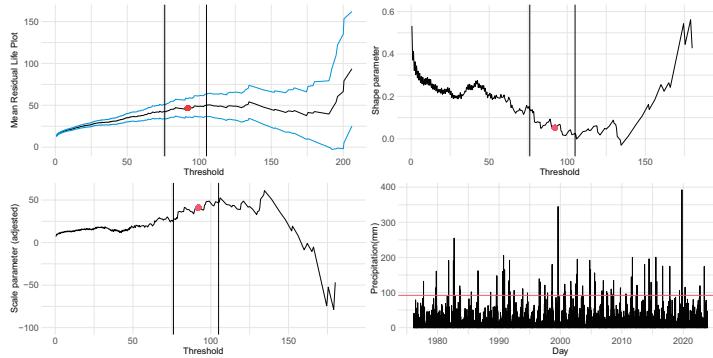


図 3. 八王子市の降雨量データの閾値選択. 左上図は平均残差生存図(青線は信頼区間). 右上は $(w, \hat{\gamma})$. 左下は $(w, \hat{\sigma}^*)$. 赤点は $w_{opt} = 92$. 縦実線は (w_L, w_U) . 右下図は降雨量データと閾値(赤線).

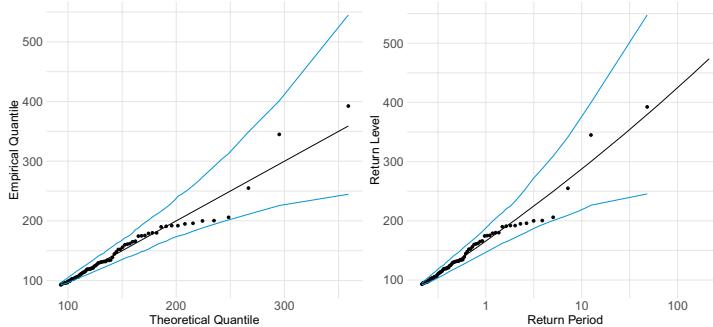


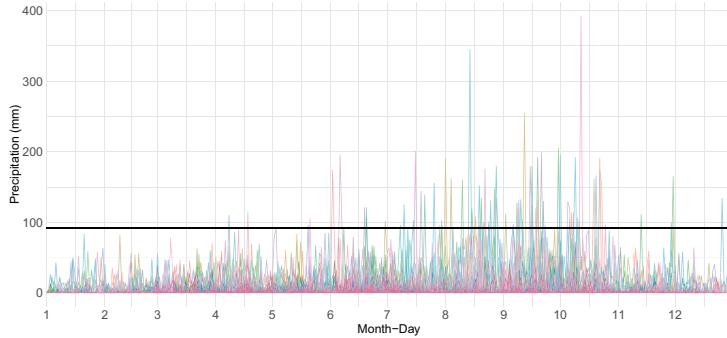
図 4. 八王子市の降雨量データに対する GP の QQ-plot(左図)と再現レベル図(右図).

現レベル $z(1/T)$ を示している. 青線は 95% 信頼区間である. 1976 年以降, 観測期間 48 年としたときの八王子観測所の降雨量の最大値は 392.5mm であったが, これは再現期間 50 年の再現レベルより少し大きな値となっていることが確認できる.

今回の GP を用いた解析では観測期間全体を通して日降雨量を独立な事象として扱った. しかし, 日本特有の事情として, まず豪雨は台風も多い夏季に起こりやすい. よって, 季節変動を考慮した深化研究は今後の解析として興味深い.

図 5 に各年の月日に対する降雨量を重ねて描いている. この図より, 閾値超過する豪雨イベントは多くが夏季に集中していることがわかる. 実際に閾値 w (横線)を超過しているのはほとんどが 7 月から 10 月である. このように, 日本の降雨量データは本来季節性があるので, 定常な GP を適用しにくく, 本来は季節性を共変量として考慮するなどの工夫をするのが自然である. 一方で, 降雨の極値解析に伝統的に GEV, すなわち年最大値を用いるのは季節性を暗に消すためであるとも考えられる.

先に挙げた Caires (2011a, 2011b), Wahl et al. (2013) は GEV だけでなく GP による実データ解析も行っており, このことは両方のモデルの振る舞いから総合的にデータ解釈を行うこと重要性を物語っていると言えるだろう.

図 5. 各年の各月日の降雨量. 横線は閾値 $w_{opt} = 92$.

3.6 パレート型分布

一般化パレート分布において γ の符号が正であるとわかっているとき, (3.1)において, $\sigma_w \approx w\gamma$ ととることができます. このとき, 任意の $x \geq 1$ について,

$$P\left(\frac{X_i}{w} > x \mid X_i > w\right) = \frac{P\left(\frac{X_i}{w} > x\right)}{P(X_i > w)} \rightarrow x^{-1/\gamma} \text{ as } w \rightarrow \infty$$

を得る. (3.5)の右辺の分布はパレート型分布 (Pareto-type distribution) と呼ばれる.

一般にデータ解析をするとき, γ の符号そのものも未知であることが多いが, “データはヘビーテールする”, または“ヘビーテールであると想定しておく方が無難である”という事前情報があるならば, 最初から $\gamma > 0$ と仮定し, パレート型分布を当てはめることもある. 他にも例えば尖度が正規分布の尖度である 3 に比べて大きい場合はヘビーテールするとみなすこともある. 特に株価のデータや(地域と種類によるが)様々な気象データは $\gamma > 0$ と仮定できる場合が多い.

GP と同様に, PT でも閾値より大きいデータのみを用いて解析する. 閾値 w に対して, $X_i > w$ となるデータについて $Y_i = X_i/w$ とし, $\{Y_1, \dots, Y_k\}$ と再ラベリングする. すると, $Y_i, i = 1, \dots, k$ はパレート型分布 (PT) に従うとみなせる. PT の密度関数は

$$f_{\text{PT}}(x|\gamma) = \frac{1}{\gamma} x^{-1/\gamma-1}, \quad x \geq 1$$

となるので, γ の最尤推定量は

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmax}} \sum_{i=1}^k \log f_{\text{PT}}(Y_i|\gamma) = \frac{1}{k} \sum_{i=1}^k \log Y_i$$

と単純な対数変換後の標本平均の形で与えられる. ちなみに, 元のデータ X_1, \dots, X_n の順序統計量を $X_{(1)} \geq \dots \geq X_{(n)}$ とし, $w = X_{(k+1)}$ (大きい方から数えて $k+1$ 番目のデータ) とすると

$$\hat{\gamma} = \frac{1}{k} \sum_{i=1}^k \log \frac{X_{(i)}}{X_{(k+1)}}$$

となり, 最尤法とは別のアプローチで得られる Hill 推定量 (Hill, 1975) と呼ばれる推定量と一致する.

GP では閾値を固定点と見るか確率変数と見るかで $\hat{\sigma}$ の分散が異なったが $\hat{\gamma}$ は同じであった. PT ではパラメータは形状パラメータ γ のみなので, 閾値の考え方の違いは漸近理論には

表れない。この場合、 $\hat{\gamma}$ の漸近的性質は Hill (1975) が示しており、GP とほぼ同様の条件下で

$$\sqrt{k}(\hat{\gamma} - \gamma) \xrightarrow{D} N(\lambda, \gamma^2)$$

が成立する。ここで、 λ はやはり (3.3) の近似から発生するバイアスである。

PT を用いる場合の再現レベルの解釈は GP と同様である。固定した $p \in (0, 1)$ に対して、 $p = P(X > z(p)) = P(X/w > z(p)/w | X > w)P(X > w)$ を満たす $z(p)$ は

$$z(p) \approx w \left\{ \frac{p}{P(X > w)} \right\}^{-\gamma}$$

で与えられる。よって、 $P(X > w) \approx k/n$ で置き換えると $z(p) \approx w\{(np)/k\}^{-\gamma}$ となる。3.4 節と同様に、想定サンプルサイズ T に対する再現レベル $z(1/T)$ が導かれる。閾値より大きいデータ数を k とすると、形式的に $w = z(k/n)$ となるので

$$(3.5) \quad z(p) = z(k/n) \left(\frac{p}{k/n} \right)^{-\gamma}$$

を得る。特にこの場合、より高位の分位点 $z(p)$ はそれなりに良い推定が可能な分位点 $z(k/n)$ の定数倍であるという解釈ができる。これは 4.2 節で述べる極値分位点の基礎理論にもなっている。

4. 発展

第 3 章まで極値統計学の基本的な分布である GEV, GP を用いた分析方法を述べた。本章では、それらのさらなる展開を 2 つにしぶって紹介する。ただしそれらは別々の問題というわけではなく、密接に繋がりがある。

4.1 回帰分析

興味ある変数の極値挙動を GEV, GP, PT でモデリングする方法を述べたが、そのひとつの応用が共変量を取り込んだ回帰分析である。すなわち、目的変数 Y 、共変量 X のペアを考え、共変量 $X = x$ と固定した下での Y の分布関数 $P(Y \leq y | X = x) = F(y|x)$ が (2.1) を満たすと仮定する。回帰分析においては、(2.3) に対応する

$$\lim_{y \rightarrow y^*(x)} \frac{\{1 - F(y|x)\}F''(y|x)}{\{F'(y|x)\}^2} = -\gamma(x) - 1$$

を仮定する場合が多い。ここで、 $y^*(x) = \sup\{y | F(y|x) < 1\}$ である。さらに $\gamma(x)$ を x に関する関数として推定する場合は γ が x について連続微分可能である仮定を追加する。GEV の場合は 3 つのパラメータが x に依存した GEV($\mu(x), \sigma(x), \gamma(x)$) を考え、GP では GP($\sigma(x), \gamma(x)$)、PT では PT($\gamma(x)$) とし、それぞれの関数の推定問題に帰着する。GEV 回帰の主な手法は、Yee and Stephenson (2007), Castro-Camino et al. (2022), Zhong et al. (2022), GP 回帰は Beirlant and Goegebeur (2004), Chavez-Demoulin and Davison (2005), Youngman (2019), PT 回帰は Wang and Tsai (2009), Goegebeur et al. (2015), Lin et al. (2022) などが挙げられる。しかし、GEV や GP では形状パラメータの推定量の精度が悪くなる例も多いことが知られており、この部分だけは定数: $\gamma(x) \equiv \gamma$ を仮定するケースも多い。また、非線形関数 $\gamma_0(x)$ に対して $\gamma(x) = \exp[\gamma_0(x)]$ と仮定し、 $\gamma(x) > 0$ に限定して議論させている場合も多い。同様に $\gamma(x) < 0$ に限定して議論することも可能だが、それならば $\gamma(x) \equiv 0$ の単に指数分布を考えれば conservative なモデルとなるので実用例は少ない。実際に、GEV や GP は形状パラメータの符号によって分布関数が(不連続に)異なるので、 $\gamma(x)$ と x に依存させて関数推定した $\hat{\gamma}(x)$ が x

によって符号が変わると解釈が難しくなる。よって、理論研究でも $\gamma(x)$ の符号は固定で語られる場合がほとんどである。共変量 x によって $\gamma(x)$ の符号が変わるモデルに関する統計理論研究は未だに未開発であり、重要な課題である。極値回帰の R パッケージとしては Youngman (2022) が `evgam` を開発している。このパッケージは非線形回帰の有名なパッケージ `mgcv` を極値分布に適用したものであり、多くのユーザーにとって扱いやすい。

最後に、回帰分析とは背景が異なるが関連研究として、時間 $t \in [0, 1]$ や空間 $s \in \mathbb{R}^2$ に依存する確率過程と極値モデルの融合研究も盛んである。極値確率過程は、Smith (1989), Einmahl et al. (2016), de Haan and Zhou (2020) などが知られている。de Haan and Ferreira (2006) にも確率過程の極値理論がまとめられている。空間極値統計学(spatial extremes)についてはレビュー論文である Davison et al. (2012), Huser and Wadsworth (2022) に丁寧にまとめられている。Einmahl et al. (2022) は時空間極値モデルを提案し、統計理論の構築にも成功している。

4.2 極値分位点推定

ここでは先行研究が多い GP を考える。3.4 節の設定で再現レベル $z(p)$ は実質的に X の $100(1-p)\%$ 分位点であった。また、閾値 w もデータの最小値と最大値の間にあるのだから X の分位点であると解釈できる。実際に、 $P(X > w) \approx k/n$ であるとき、 $w \approx z(k/n)$ となる。このとき、再現レベルは

$$(4.1) \quad z(p) \approx z(k/n) + \frac{\sigma}{\gamma} \left\{ \left(\frac{np}{k} \right)^{-\gamma} - 1 \right\}$$

と表すことができる。確率点 k/n の周りは極値エリアであるもののそれなりにデータ数を確保できる。一方で、確率点 p に相当するエリアにはデータがほとんど存在しない。通常はそのような点の分位点 $z(p)$ は直接推定できないが、(4.1) はそれなりに良い推定が可能な分位点 $z(k/n)$ に極値理論から導かれる定数 $(\sigma/\gamma)\{((np)/k)^{-\gamma} - 1\}$ を足すいわば“外挿(extrapolation)”の形で構成されていることがわかる。このように外挿を駆使して推定されるような分位点 $z(p)$ は極値分位点(extreme quantile)と呼ばれ、GEV や GP など極値分布の議論を介さず、直接分位点を推定する文脈で語られることも多い。もちろん極値分位点は再現レベルと本質的に同定義である。

極値分位点は Weissman (1978) を皮切りに様々なモデルの応用がなされている。特に、目的変数 Y の極値分位点を説明変数 X に関連させて回帰モデルの形式での発展がなされ、極値分位点回帰(extreme quantile regression, EQR) と呼ばれている。その場合は $X = x$ を与えた下での Y の分位点を $z_Y(\cdot|x)$ と書くと、上の結果から直ちに

$$(4.2) \quad z_Y(p|x) \approx z_Y(k/n|x) + \frac{\sigma(x)}{\gamma(x)} \left\{ \left(\frac{np}{k} \right)^{-\gamma(x)} - 1 \right\}$$

が導かれる。ただし、尺度パラメータと形状パラメータも基本的には説明変数に依存するものとして考えられる。4.1 節でも述べた形状パラメータ $\gamma(x)$ の連続性と符号の問題は当然極値分位点推定でも起きる。そのため、多くの場合で $\gamma(x) > 0$ が仮定される。大まかな方法としては、まず $z_Y(k/n|x)$ を通常の分位点回帰 (Koenker, 2005 等) で推定し、 $\gamma(x), \sigma(x)$ を GP 回帰で推定し、(4.2) よりより高位な分位点の推定を行う。最初から $\gamma(x) > 0$ と仮定し、PT を考えるときは条件付き EQR は

$$z_Y(p|x) \approx z_Y(k/n|x) \left(\frac{p}{k/n} \right)^{-\gamma(x)}$$

となる。極値分位点回帰は先行研究が豊富である。Chernozhukov (2005) は外挿は行っていないものの極値統計学の文脈で分位点回帰の裾部分を理論評価している。Daouia et al. (2013)

はEQRのノンパラメトリック推定を確立している。近年ではGnecco et al. (2024)がランダムフォレスト, Allouche et al. (2024)がニューラルネットワーク, Richards and Huser (2024)が深層学習を用いたEQRを開発しており、目覚ましい発展がなされている。

5. クラスターデータに対する極値モデリング

極値統計モデリングは様々な分野で利用されているが、多くの場合、データ構造としてクラスターデータの形式で得られる。代表的なものは気象データで、例えば降雨量データは日本全国で約1300地点で観測・集計されている。このようなデータは距離が近いところは関連があり、その情報を取り入れた解析をすることで自然現象の理解を深めること、また予測精度の向上に繋がることもある。単にクラスター間で予測値を平滑化するだけでもわかることは多い。

いま、 $X_{ij}, i = 1, \dots, n_j, j = 1, \dots, J$ を J 個のクラスターで i 番目に観測されたデータとする。

クラスターが地理情報を含む場合は、これを空間データととらえることもできる。その場合は、主にコンパクトな空間 $\mathcal{S} \subset \mathbb{R}^2$ における位置 $s \in \mathcal{S}$ に対して、データを $X_i(s)$ と確率過程表現を行うこともある。ただし、降雨量データのように観測地点が固定の場合 (\mathcal{S} が有限の離散点) で観測時点も等しい場合は単に多変量データ $\{(X_{i1}, \dots, X_{iJ}) : i = 1, \dots, n\}$ として扱うことも多い。この場合は、各観測所に付与するダミー変数を説明変数とした場合の回帰分析の文脈でも記述できる。この章ではクラスターデータに対する極値統計モデリングの例を述べる。以降はモデルとしてGPを想定するが、GEVでも同様の解析が可能である。

5.1 クラスター統合

ここでは多変量データ $\{(X_{i1}, \dots, X_{iJ}) : i = 1, \dots, n\}$ の各 $j = 1, \dots, J$ について、 $F_j(x) = P(X_{ij} \leq x)$ の裾のモデリングを行う。パラメータ推定にあたり J 個のクラスター情報を統合できれば単純にデータ数が増え、パラメータ推定が安定する。いま、 j 番目のクラスターに対して、閾値 w_j としたとき、 X_{1j}, \dots, X_{nj} の中で $Y_{ij} = X_{ij} - w$ が 0 より大きいものを集め、 $\{Y_{1,j}, \dots, Y_{k_j,j}\}$ と再ラベリングする。これらのデータが $\text{GP}(\sigma_j, \gamma_j)$ に従うとする：

$$P(Y_{i,j} > y) = \left(1 + \frac{\gamma_j y}{\sigma_j}\right)_+^{-1/\gamma_j}, \quad j = 1, \dots, J.$$

ここで、Yee and Stephenson (2007) などで扱われている代表的なクラスター統合では

$$\gamma = \gamma_1 = \dots = \gamma_J$$

が仮定される。つまり、形状パラメータは同一であり、尺度パラメータはクラスター固有のものとしている。一般に、 σ_j は閾値 w_j に依存するので、 σ_j のみの統合は有効な効果が得られない。一方で、形状パラメータは位置尺度不变なので、閾値と尺度パラメータによらず統合を検討できる。また、これは $F_1, \dots, F_J \in \mathcal{D}(G_\gamma)$ という仮定を置いていくことになり、 J 個のクラスターの元の分布関数そのものは異なるが、極値分布は共通であることを意味する。このような仮定は極値分布論と整合性がとれているため考えやすい。この仮定の下でパラメータ $(\gamma, \sigma_1, \dots, \sigma_J)$ を推定する。その最尤推定量は

$$(5.1) \quad (\hat{\gamma}, \hat{\sigma}_1, \dots, \hat{\sigma}_J) = \underset{\gamma, \sigma_1, \dots, \sigma_J}{\text{argmax}} \sum_{j=1}^J \sum_{i=1}^{k_j} \log f(Y_{i,j} | \gamma, \sigma_j)$$

で与えられる。これにより、各 j について

$$P(X_{ij} > y) = P(X_{ij} > w) \left(1 + \frac{\hat{\gamma}(y - w_j)}{\hat{\sigma}_j} \right)^{-1/\hat{\gamma}}$$

から再現レベルを計算できる。ただし、 $P(X_{ij} > w_j) \approx k_j/n$ である。

モデルとして GEV を扱う場合は、 j 番目のクラスターが持つパラメータを $(\mu_j, \sigma_j, \gamma_j)$ とするとき、GP と同様に γ_j のみクラスターで共通として、 (μ_j, σ_j) は統合せず、クラスター固有のパラメータとする。

5.2 統合するクラスターの選定

クラスターの統合は有効な場合はよいが、むやみに統合してよいわけではない。統合するクラスターの選定を何らかの意味で正当化したい。例えば、降雨量データであれば観測地点が近い地域では同じ日に豪雨が観測されるだろう。このように、同じタイミングで極値データが発生するクラスター同士は統合を検討してよさそうである。異なるクラスター間の極値データの同時発生を測る指標として 2.5 節でも用いた裾従属係数を改めて考える。極値データの同時発生を測るための裾従属係数は異なる 2 つのクラスター番号 j, k について、

$$\chi_{jk} = \lim_{q \rightarrow 1} P(F_j(X_{ij}) > q | F_k(X_{ik}) > q) = \lim_{q \rightarrow 1} \frac{P(F_j(X_{ij}) > q, F_k(X_{ik}) > q)}{1 - q}$$

で定義される。裾従属係数 χ_{jk} が比較的大きい値を持つとき、 j 番目と k 番目のクラスターは同時に極値データが発生する確率が高いと判断できる。

いま、主に解析したいクラスター番号を 1 とする。このとき、残りの $J - 1$ 個のクラスターのうち、ある $\delta \in (0, 1)$ に対して

$$\{j \in \{2, \dots, J\} | \chi_{1j} > \delta\}$$

となるクラスターのみを集め、改めて $\{(X_{i1}, \dots, X_{iK}) : i = 1, \dots, n\}$ と書く。ここで、 $K \leq J$ は J 個のうち、1 番目のクラスターとの裾従属係数が高いクラスターの個数である。この K 個のクラスターは統合を検討できるだろう。

本手法について一点補足をしておく。本章では $\{(X_{i1}, \dots, X_{iJ}) : i = 1, \dots, n\}$ の多変量確率変数を考えている。したがって、本来は J 変量の多変量分布を検討すべきであろう。実際に、裾従属係数の値が大きいクラスター同士は独立とは考えにくい。さて、多変量分布関数はコピュラ C と X_{1j}, \dots, X_{nj} の周辺分布関数 F_j を用いて

$$P(X_{i1} \leq x_1, \dots, X_{iJ} \leq x_J) = C(F_1(x_1), \dots, F_J(x_J))$$

と表せる。すると対応する密度関数は

$$(5.2) \quad f(x_1, \dots, x_J) = c(F_1(x_1), \dots, F_J(x_J)) \prod_{j=1}^J f_j(x_j)$$

と表せる。ここで、 $c(u_1, \dots, u_J) = \partial^J C(u_1, \dots, u_J) / \partial u_1 \cdots \partial u_J$ であり、 f_j は F_j の周辺密度関数である。さらに f_j に GP を仮定する場合は

$$f_j(x_j) = \begin{cases} P(X_{ij} < w_j) f_{u,j}(x_j) & x_j < w_j \\ P(X_{ij} \geq w_j) \frac{1}{\sigma} \left(1 + \frac{\gamma(x_j - w_j)}{\sigma} \right)^{-1/\gamma_j - 1} & x_j \geq w_j \end{cases}$$

を考えることになる。ただし、 $f_{u,j}$ は $X_{ij} < w_j$ の部分の密度関数であるが、極値解析には用いない形式的なものである。すると、(5.2) の対数尤度は

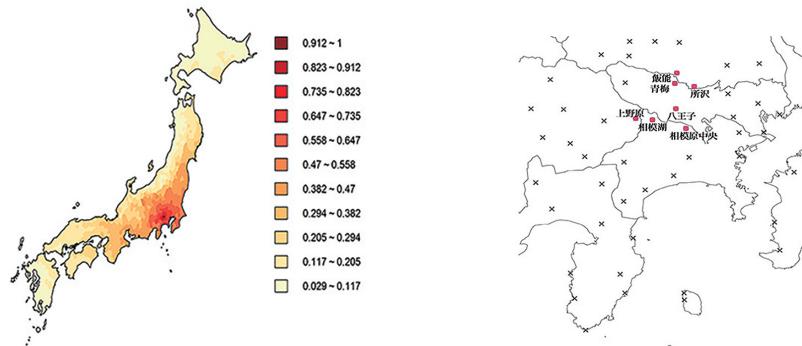


図 6. 左：八王子市と他の 1127 箇所の観測所の裾従属係数. 右：八王子市と裾従属係数が 0.75 以上となった観測所 (赤点). \times は周辺の観測所.

$$\sum_{i=1}^n \log f(X_{i1}, \dots, X_{iJ}) = \sum_{i=1}^n \log c(F_1(X_{i1}), \dots, F_J(X_{iJ})) + \sum_{i=1}^n \sum_{j=1}^J \log f_j(X_{ij})$$

となる. 従属関数 $c(F_1(X_{i1}), \dots, F_J(X_{iJ}))$ に含まれる F_j を経験分布関数で置き換えると c の推定と周辺尤度の部分の推定は互いに影響しなくなる. このテクニックは極値解析に関わらず, コピュラを用いた推測ではよく議論されている (Nelsen, 2006). これより, 5.1 節ではクラスター間の従属性を多変量分布としてモデリングせず単に周辺分布の統合を行っているが, これは従属関数 $c(F_1(X_{i1}), \dots, F_J(X_{iJ}))$ の推定と周辺分布の推定を別で考えていることに相当する. なお, 極値解析において従属関数と周辺分布の推定を別で考えることの理論的な正当性は Genest et al. (1995) と Genest and Segers (2009) によって議論されている.

5.3 データ適用

第 2, 3 章と同様の日降雨量データを扱う. 全国には約 1300 箇所の雨量観測所があるが, 最近開設された場所を除いた $J = 1128$ 箇所の観測所について, 東京都八王子市を基準にして全国の各観測所との裾従属係数を調べる. ただし, 観測所によって降雨量観測を開始した時点が異なるため, 裾従属係数の計算のために用いるデータの観測期間を 2000 年 1 月 1 日～2023 年 12 月 31 日とした. 結果を図 6 に示している. 左図から, 八王子とその周辺は従属性が高く, 離れるほど従属性が低いことがわかる. これは直感的にも明らかで, 八王子市で豪雨が降ると, 近隣も雨量は多いが遠方地域では同日に晴れであることもあるだろう. 統合クラスターの選定は単純に観測地点間の距離をベースに考えても似たような結果を得るだろうが, 裾従属係数はきちんと地理情報を反映した結果を返しているし, 株価など地理情報を含まないデータに対しても有効なので汎用性が高い.

八王子との裾従属係数が 0.75 以上であった観測所は埼玉県所沢, 飯能, 東京都青梅, 山梨県上野原, 神奈川県相模原中央, 相模湖の 6 箇所であった (図 6 右図を参照). 八王子を含めた 7 箇所 ($K = 7$) に対して, クラスターの統合解析を行う. それぞれの地域を $j = 1, \dots, 7$ とし, 極値分布として共通の極値分布を持つ, つまり, $F_j \in \mathcal{D}(G_\gamma)$ と仮定する. 特にここでは GP を極値モデルとして想定する. まずは各クラスターで閾値を設定する. 八王子の閾値は $w = 92.5$ であり, これは上位 79 個のデータを使うことに相当した. 八王子以外の観測所も上位 79 個のデータを GP に当てはめる. これに相当する閾値を図 7 に平均残差生存図とともに示している (縦線). 結果からおよそすべての観測所でこの値周辺で比較的安定して線形構造を有している

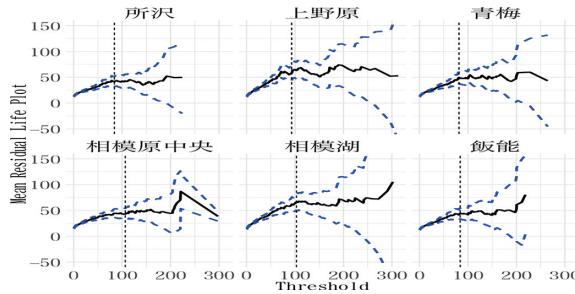


図 7. 飯能, 所沢, 青梅, 上野原, 相模原中央, 相模湖の降雨量データの平均残差生存図. 青線は 95% 信頼区間.

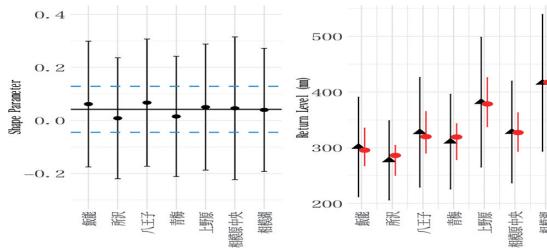


図 8. 左: クラスター統合前後の形状パラメータの推定値(左図)と 50 年再現レベル(右図).
右図では黒が統合前, 赤が統合後の結果.

ことがわかった.

図 8(左)は 7 地点毎とクラスター統合後の形状パラメータの推定値である. まず, 各地点における黒点がクラスター統合前の形状パラメータ推定値と 95% 信頼区間(エラーバー)であり, 青線がクラスター統合後の推定値(95% 信頼区間)である. クラスター統合後の γ の信頼区間は尤度(5.1)のヘシアンに基づいて構成した. 推定値は $\hat{\gamma} = 0.047$ となり, 95% 信頼区間幅は $[-0.04, 0.134]$ となりその幅は 0.17 となった. 3.5 節で八王子観測所のみで解析した信頼区間幅は約 0.47 であったため, かなり幅が縮小された(各観測所のエラーバーの長さと破線間の長さを比較しても一目瞭然である). このように, 推定の安定性が向上したと言える.

結果から, 所沢, 青梅, 相模湖は形状パラメータの推定値が若干ではあるが上方修正されていることがわかる. 他の観測所は下方修正されているがあまり変化は見られなかった. また, クラスター統合後は信頼区間幅が縮小されたことも見て取れる. これにより, より安定した推測が可能になったと言えるだろう.

図 8(右)には, クラスター統合前(黒)と統合後(赤)の 7 地点での 50 年再現レベル降雨量と 95% 信頼区間を示している. クラスター統合後の再現レベルの信頼区間は(5.1)のヘシアンとデルタ法により構成した. 形状パラメータが大きく上方修正された所沢ではクラスター統合後の再現レベルは大きく見積もられた. それ以外の観測所では大きな変化が見られなかったもののクラスター統合により各観測所の予測値が平滑化され, また信頼区間幅も縮小されたことによって, 安定した解釈性の高いモデルとなった.

今回は簡単な極値モデリングのクラスター統合を行ったが, 重要な留意点がある. 例えば, 3 つのクラスター(番号を 1, 2, 3 とする)を考えたとき, $\mathcal{X}_{12}, \mathcal{X}_{23} > \delta$ でも $\mathcal{X}_{13} < \delta$ となる可能性がある. この場合, (X_{i1}, X_{i2}, X_{i3}) を統合できるかどうかは判断が難しい. したがって, この方法のみで統合するクラスターを決定する場合には, 基点となるクラスターを 1 つ設定しな

ければならない。ただし、基点を変えると統合できるクラスターが異なる、いわば一意性が担保されないのが問題となる。近年では基点を用いずにクラスター選定とモデリングを同時に使う方法が Rohrbeck and Tawn (2021), Dupuis et al. (2023) によって開発されている。しかし、いずれもクラスター統合の前後でモデルの当てはまりを比較する発見的な手法となっている。そのため、統合後のクラスター数をある程度決定しておく必要があり、クラスター数が大きい場合の解析には向かない。このように、極値データのクラスター統合は重要な研究に位置付けられており、さらなる展開が望まれる手法である。

また、気象データの地域統合に関する研究として、Hosking and Wallis (1993, 1997) に代表される地域頻度解析 (Regional Frequency Analysis, RFA) は有名である。RFA では気象学的に均質とみなせる地域のデータをプールする方法であるが、この方法はいわば複数地域の分布関数そのものを同一視する。一方で今回我々が用いた統合モデルは地域間の極値分布は同じであるが分布関数そのものには差異があるというもので、RFA とは若干異なる。これにより、図 8 右図では再現レベルの地域の微小な差異を捉えることができている。ただし、再現レベルがほぼ等しい地域は完全に統合する RFA ベースの解析も検討できるだろう。

また、降雨量データ解析について述べると、事前の解析から標高と降水量は共変関係にあることがわかっている。今回はクラスター以外の情報は考慮していないが、その他の有効な説明変数を用いるとより踏み込んだ解析が期待できるだろう。

6.まとめ

本総説では極値統計学の基礎的な方法論を実データ適用を交えながら議論した。特に、ターゲット変数が 1 次元の場合に焦点を絞り、GEV と GP の 2 つのモデリングを扱った。本総説では触れなかつたが、GEV と GP どちらを用いるべきか? と疑問を持つこともあるだろう。一概にどちらが優れていることはないが、Bücher and Zou (2021) はそれぞれのモデルの特徴を踏まえながら比較している。

発展研究としては回帰分析に言及したが、そこでも極値データとして興味ある目的変数はあくまで 1 次元であった。一方で、例えば降雨量データについて、日本では台風時に大きな値を持つことから風速との関連は興味深い。その場合は降雨量、風速が共に大きい値を持つ 2 変量極値データの分析が興味の対象となる。他にも、豪雨時に 2 つの河川が同時に氾濫するリスク、洪水(外水氾濫)と内水氾濫の同時発生(日雨量と時間雨量に起因)や豪雨と高潮の同時発生するリスク、複数の株が同時に極端に高い/低い値を持つリスク予測のモデリングなど従属関係のモデリングも重要である。その場合は多変量極値統計モデル (Multivariate extreme value model) の利用が考えられる。多変量極値分布の先駆けとしては Tiago de Oliveira (1958), Geffroy (1958/59), Gumbel (1960), Sibuya (1960) が挙げられる。また、2 変量正規分布から導かれる極値分布である Hüsler-Reiss model (Hüsler and Reiss, 1989) は最も代表的な多変量極値モデルのひとつである。以降、これまでに多くの多変量極値モデルが提案された (Beirlant et al., 2004)。現在では高次元極値モデルを Engelke and Hitz (2020) がグラフィカルモデリングの枠組みで提案し、Engelke and Ivanovs (2021) で高次元極値モデルのスパース化によるアプローチについて概観している。彼らは次元が膨大な中で極値従属が特徴的な少数の変数を抜き出し、モデルを簡略化する方法をまとめている。特に、条件付き独立を利用した方法は Engelke and Hitz (2020) と深い関連がある。2 変量極値モデルの基本的な性質は北野 (2021) にまとめられている。

本稿では GEV と GP をあくまで独立な確率変数の極値論として紹介した。一方で、この極値統計論は点過程論 (Point Process) として記述することもできる (付録 B を参照)。極値統計

に関する点過程については 1 変量の場合は Leadbetter et al. (1983), Resnick (1987), Coles (2001), 多変量極値の場合は Resnick (2007), その訳本の国友・栗栖 (2021) を参照されたい。

最後に, 本稿で扱った実データ分析は統計ソフト R の代表的な極値統計分析パッケージ `extRemes`, `ismev`, `eva` を用いた。他にも多くの極値統計のためのパッケージが存在し, Stephenson and Gilleland (2006), Gilleland et al. (2013), Belzile et al. (2023) にまとめられている。特に, Belzile et al. (2023) には Python のパッケージについても言及している。和書としては西郷・有本 (2020) が R を用いた極値統計解析について詳しく解説している。

謝 辞

本総説の執筆にあたり, 「統計数理」の編集委員, ならびに三名の査読者の方々には数々の貴重なアドバイス, 的確なコメントをいただきました。心より御礼申し上げます。本研究は日本学術振興会科学研究費助成事業基盤研究(B)(課題番号: 23K28043), 基盤研究(C)(課題番号: 22K11935)の助成を受けております。

付録 A: GEV と GP の同値性

(2.1) と (3.1) の同値性を述べる。ここでは, (2.1) を仮定した下で (3.1) が成立することのみ示すが, 逆も同様である。また, $\gamma \neq 0$ のときのみ証明するが, $\gamma = 0$ のときの証明はより簡単である。

(2.1) より, $P(\max_i X_i \leq x) = P(X_i \leq x)^n = F^n(x)$ について, 十分に大きな x, n に対して

$$F^n(x) \approx \exp \left[- \left(1 + \frac{\gamma(x - \mu_n)}{\sigma_n} \right)^{-1/\gamma} \right]$$

であった。ただし, $1 + \gamma(x - \mu_n)/\sigma_n > 0$ となる x を選んでいる。ここで, $\log F^n(x) = n \log F(x) \approx n(1 - F(x))$ より,

$$1 - F(x) \approx -\frac{1}{n} \left(1 + \frac{\gamma(x - \mu_n)}{\sigma_n} \right)^{-1/\gamma}$$

を得る。すると, $\sigma_w = \sigma_n + \gamma(w - \mu_n)$ と置くと以下を得る:

$$\begin{aligned} P(X_i > w + x | X > w) &= \frac{1 - F(w + x)}{1 - F(w)} \\ &\approx \frac{(1 + \gamma(w + x - \mu_n)/\sigma_n)^{-1/\gamma}}{(1 + \gamma(w - \mu_n)/\sigma_n)^{-1/\gamma}} \\ &= (1 + \gamma \frac{x - w}{\sigma_w})^{-1/\gamma}. \end{aligned}$$

また, 別の閾値 $w^* > w$ を用いたとき, $\sigma_{w^*} = \sigma_n + \gamma(w^* - w - \mu_n)$ を用いて

$$P(X_i > w^* + x | X > w^*) \approx \left(1 + \gamma \frac{x - w^*}{\sigma_{w^*}} \right)^{-1/\gamma}$$

が成立することが簡単にわかる。つまり, 閾値を少し変化させたとき, GP の形状パラメータは不变であるが, 尺度パラメータは線形に少しシフトする。

付録 B: GEV の点過程表現

確率変数 $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} F$ は (2.1) を満たすとする。このとき, 十分に大きい x について

$$P\left(\max_i X_i \leq x\right) = F^n(x) \approx G_\gamma\left(\frac{x-\mu}{\sigma}\right)$$

となる. このとき, $-\log F^n(x) = -n \log F(x) \approx n(1 - F(x))$ より,

$$P(X_i > x) = 1 - F(x) \approx -\frac{1}{n} \log G_\gamma\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{n} \left(1 + \frac{\gamma(u-\mu)}{\sigma}\right)^{-1/\gamma}$$

となる. したがって, 十分に大きい値 u について

$$p = P(X_i > u) \approx \frac{1}{n} \left(1 + \frac{\gamma(u-\mu)}{\sigma}\right)^{-1/\gamma}$$

とすると, X_1, \dots, X_n の内 u を超過する個数は二項分布 $B(n, p)$ に従う. 言い換えると, 領域 $A \subset \mathbb{R}$ における点過程 $N_n(A) = |\{i \in \{1, \dots, n\} : X_i \in A\}|$ と $I_u = [u, \infty)$ について $N_n(I_u) \sim B(n, p)$ が成立する. すると, 二項分布がポアソン分布に収束することを利用すると, $n \rightarrow \infty$ の下で

$$N_n(I_u) \xrightarrow{D} N(I_u) \sim Po(\lambda)$$

を得る. ただし,

$$\lambda = \left(1 + \frac{\gamma(u-\mu)}{\sigma}\right)^{-1/\gamma}$$

である. ここで, $P(\max_i X_i < x) = P(N_n(I_x) = 0)$ なので, 確かに

$$P\left(\max_i X_i < x\right) = P(N_n(I_x) = 0) \rightarrow P(N(I_x) = 0) = \exp[-\lambda] = G_\gamma\left(\frac{x-\mu}{\sigma}\right)$$

を満たす. これがGEVのポアソン点過程表現である. 同様に, GPも点過程表現できる (Coles, 2001).

参考文献

- Allouche, M., Girard, S. and Gobet, E. (2024). Estimation of extreme quantiles from heavy-tailed distributions with neural networks, *Statistics and Computing*, **34**, <https://doi.org/10.1007/s11222-023-10331-2>.
- Barde, B. and Yan, J. (2020). eva: Extreme Value Analysis with Goodness-of-Fit Testing, R package version 0.2.6., <https://CRAN.R-project.org/package=eva> (最終アクセス日 2025年6月4日).
- Beirlant, J. and Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for Pareto-type distributions, *Journal of Multivariate Analysis*, **89**, 97–118.
- Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*, John Wiley & Sons, Chichester.
- Belzile, L., Dutang, C., Northrop, P. and Opitz, T. (2023). A modeler's guide to extreme value software, *Extremes*, **26**, 595–638.
- Bücher, A. and Segers, J. (2017). On the maximum likelihood estimator for the generalized extreme-value distribution, *Extremes*, **20**, 839–872.
- Bücher, A. and Zhou, C. (2021). A horse race between the block maxima method and the peak-over-threshold approach, *Statistical Science*, **36**, 360–378.
- Caeiro, F. and Gomes, M. I. (2011). Asymptotic comparison at optimal levels of reduced-bias extreme value index estimators, *Statistica Neerlandica*, **65**, 462–488.
- Caires, S. (2011a). Extreme value analysis: Wave data, JCOMM Technical Report, **57**, World Meteorological Organization, Geneva.

- Caires, S. (2011b). Extreme value analysis: Still water level, JCOMM Technical Report, **58**, World Meteorological Organization, Geneva.
- Castro-Camilo, D., Huser, R. and Rue, H. (2022). Practical strategies for generalized extreme value models for extremes, *Environmetrics*, **33**, <https://doi.org/10.1002/env.2742>.
- Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modeling of sample extremes, *Journal of the Royal Statistical Society Series C*, **54**, 207–222.
- Chernozhukov, V. (2005). Extremal quantile regression, *Annals of Statistics*, **33**, 806–839.
- Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, New York.
- Coles, S. G., Heffernan, J. and Tawn, J. (1999). Dependence measures for extreme value analyses, *Extremes*, **2**, 339–365.
- Daouia, A., Gardes, L. and Girard, S. (2013). On kernel smoothing for extremal quantile regression, *Bernoulli*, **19**, 2557–2589.
- Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds, *Journal of the Royal Statistical Society Series B*, **52**, 393–442.
- Davison, A. C., Padoan, S. and Ribatet, M. (2012). Statistical modelling of spatial extremes (with discussion), *Statistical Science*, **27**, 161–186.
- de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory — An Introduction*, Springer, New York.
- de Haan, L. and Zhou, C. (2020). Trends in extreme value indices, *Journal of the American Statistical Association*, **116**, 1265–1279.
- de Haan, L. and Zhou, C. (2024). Bootstrapping extreme value estimators, *Journal of American Statistical Association*, **119**, 382–393.
- Drees, H., Ferreira, A. and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index, *Annals of Applied Probability*, **14**, 1179–1201.
- Dupuis, D. J., Engelke, S. and Trapin, L. (2023). Modeling panels of extremes, *Annals of Applied Statistics*, **17**, 498–517.
- Einmahl, J. H., de Haan, L. and Zhou, C. (2016). Statistics of heteroscedastic extremes, *Journal of the Royal Statistical Society Series B*, **78**, 31–51.
- Einmahl, J. H., Ferreira, A., de Haan, L., Neves, C. and Zhou, C. (2022). Spatial dependence and space-time trend in extreme events, *Annals of Statistics*, **50**, 30–52.
- Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes, *Journal of the Royal Statistical Society Series B*, **82**, 871–932.
- Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes, *Annual Review of Statistics and Its Application*, **8**, 241–270.
- Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest members of a sample, *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.
- Geffroy, J. (1958/59). Contribution à la théorie des valeurs extrêmes, *Publications de l'Institut de Statistique de l'Université de Paris*, **7**, 37–121, and **8**, 123–184.
- Genest, C. and Segers, J. (2009). Rank-based inference for bivariate extreme-value copulas, *Annals of Statistics*, **37**, 2990–3022.
- Genest, C., Ghoudi, K. and Rivest, L. P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika*, **82**, 543–552.
- Gilleland, E. and Katz, R. W. (2016). extRemes 2.0: An extreme value analysis package in R, *Journal of Statistical Software*, **72**, 1–39.
- Gilleland, E., Ribatet, M. and Stephenson, A. (2013). A software review for extreme value analysis, *Extremes*, **16**, 103–119.

- Gnecco, N., Terefe, E.M. and Engelke, S. (2024). Extremal random forests, *Journal of American Statistical Association*, **119**, 3059–3072.
- Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum of d'unesérie Aléatoire, *Annals of Mathematics*, **44**, 423–453.
- Goegebeur, Y., Guillou, A. and Stupler, G. (2015). Uniform asymptotic properties of a nonparametric regression estimator of conditional tails, *Annales de l'Institut Henri Poincaré-Probabilités et Statistiques*, **51**, 1190–1213.
- Gomes, M. I. and Martins, M. J. (2002). “Asymptotically unbiased” estimators of the extreme value index based on external estimation of the second order parameter, *Extremes*, **5**, 5–31.
- Gomes, M. I. and Pestana, D. D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator, *Journal of American Statistical Association*, **102**, 280–292.
- Guedes Soares, C. and Scotto, M. G. (2004). Application of the r largest-order statistics for long-term predictions of significant wave height, *Coastal Engineering*, **51**, 387–394.
- Gumbel, E. J. (1960). Multivariate extremal distributions, *Bulletin of International Statistical Institute*, **39**, 471–475.
- Haigh, I. D., Nicholls, R. and Wells, N. (2010). A comparison of the main methods for estimating probabilities of extreme still water levels, *Coastal Engineering*, **57**, 838–849.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **13**, 331–341.
- Hosking, J. R. M. and Wallis, J. R. (1993). Some statistics useful in regional frequency analysis, *Water Resources Research*, **29**, 271–281.
- Hosking, J. R. M. and Wallis, J. R. (1997). *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge University Press, Cambridge.
- Huser, R. and Wadsworth, J. L. (2022). Advances in statistical modeling of spatial extremes, *WIREs Computational Statistics*, **14**, <https://doi.org/10.1002/wics.1537>.
- Hüsler, J. and Reiss, R. D. (1989). Maxima of normal random vectors: Between independence and complete dependence, *Statistics and Probability Letters*, **7**, 283–286.
- Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–171.
- 北野利一 (2020). 再現期間再探訪、極値理論の工学への応用 (18), 統計数理研究所共同研究リポート, No.445, 63–70.
- 北野利一 (2021). 成分毎の最大値と閾値を超過する多変量極値：それらの相互関係、それらの単純極値分布と乱数生成法, 日本統計学会誌, **51**, 123–156.
- 北野利一, 加藤紗也, 平松健太郎 (2025). 年最大日降水量とピーク流量の従属性と治水計画におけるカバー率, 土木学会論文集, **81**, <https://doi.org/10.2208/jscej.24-16135>.
- Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, Cambridge.
- 国友直人, 栗栖大輔 (訳) (2021). 『極値現象の統計分析：裾の重い分布のモデリング』, 朝倉書店, 東京.
- Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.
- Lin, R., Leng, C. and You, J. (2022). Semiparametric tail index regression, *Journal of Business & Economic Statistics*, **40**, 82–95.
- Meeker, W. Q., Escobar, L. A. and Pascual, F. G. (2021). *Statistical Methods for Reliability Data*, 2nd ed., Wiley, New Jersey.
- Murphy, C., Tawn, J. A. and Varty, Z. (2024). Automated threshold selection and addoxiated inference uncertainty for univariate extremes, *Technometrics*, **67**, 215–224.
- Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer, New York.
- Northrop, P. J., Attalides, N. and Jonathan, P. (2017). Cross-validatory extreme value threshold selec-

- tion and uncertainty with application to ocean storm severity, *Journal of the Royal Statistical Society Series C: Applied Statistics*, **66**, 93–120.
- Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*, Springer, New York.
- Resnick, S. I. (2007). *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*, Springer, New York.
- Richards, J. and Huser, R. (2024). Extreme quantile regression with deep learning, arXiv, <https://doi.org/10.48550/arXiv.2404.09154>.
- Rohrbeck, C. and Tawn, J. A. (2021). Bayesian spatial clustering of extremal behavior for hydrological variables, *Journal of Computational and Graphical Statistics*, **30**, 91–105.
- 西郷達彦, 有本彰雄 (2020). 『Rによる極値統計学』, オーム社, 東京.
- Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification, *REVSTAT - Statistical Journal*, **10**, 33–60.
- Sibuya, M. (1960). Bivariate extreme statistics, I, *Annals of Institute of Statistical Mathematics*, **11**, 195–210.
- Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, **72**, 67–90.
- Smith, R. L. (1986). Extreme value theory based on the r largest annual events, *Journal of Hydrology*, **86**, 27–43.
- Smith, R. L. (1987). Estimating tails of probability distributions, *The Annals of Statistics*, **15**, 1174–1207.
- Smith, R. L. (1989). Extremevalue analysis of environmental time series: An application to trend detection in ground-level ozone, *Statistical Science*, **4**, 367–377.
- Spearing, J. Tawn, J. Irons, D. and Paulden, T. (2023). A framework for statistical modelling of the extremes of longitudinal data, applied to elite swimming, arXiv, <https://doi.org/10.48550/arXiv.2306.12419>.
- Stephenson, A. G. and Gilleland, E. (2006). Software for the analysis of extreme events: The current state and future directions, *Extremes*, **8**, 87–109.
- Stephenson, R. A. (2018). ismev: An Introduction to Statistical Modeling of Extreme Values, R package version 1.42, <https://CRAN.R-project.org/package=ismev> (最終アクセス日 2025年6月4日).
- 高橋倫也, 志村隆彰 (2016). 『極値統計学(ISMシリーズ: 進化する統計数理)』, 近代科学社, 東京.
- Tawn, J. A. (1988). An extreme-value theory model for dependent observations, *Jornal of Hydrology*, **101**, 227–250.
- Tiago de Oliveira, J. (1958). Extremal distributions, *Faculdade de Ciencias de Lisboa. Série A: Matemática*, **7**, 219–227.
- von Mises, R. (1936). La distribution de la plus grande de n valeurs, *Revue de Mathématiques de l'Union Interbalkanique*, **1**, 141–160.
- Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection, *Technometrics*, **58**, 116–126.
- Wahl, T., Haigh, I. D., Jensen, J. and Pattiarchi, C. (2013). Estimating extreme water level probabilities: A comparison of the direct methods and recommendations for best practise, *Coastal Engineering*, **81**, 51–66.
- Wang, H. and Tsai, C. L. (2009). Tail index regression, *Journal of the American Statistical Association*, **104**, 1233–1240.
- Weibull, W. (1939). A statistical theory of the strength of materials, *Ingeniörsvetenskapsakademiens Handlingar (Royal Swedish Academy of Engineering Sciences)*, **151**, 1–45.
- Weibull, W. (1951). A statistical distribution function of wide applicability, *Journal of Applied Me-*

- chanics, **18**, 293–297.
- Weissman, I. (1978). Estimation of parameters and larger quantiles based on the k largest observations, *Journal of American Statistical Association*, **73**, 812–815.
- Yee, T. W. and Stephenson, A. G. (2007). Vector generalized linear and additive extreme value models, *Extremes*, **10**, 1–19.
- Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U. S. wing gusts, *Journal of the American Statistical Association*, **114**, 1865–1879.
- Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models, *Journal of Statistical Software*, **103**, 1–26.
- Zhong, P., Huser, R. and Opitz, T. (2022). Modeling nonstationary temperature maxima based on extremal dependence changing with event magnitude, *Annals of Applied Statistics*, **16**, 272–299.

Statistical Modeling Using Extreme Value Theory

Takuma Yoshida¹ and Toshikazu Kitano²

¹Graduate School of Science and Engineering, Kagoshima University

²Department of Architecture, Nagoya Institute of Technology

In applied fields dealing with various real-world phenomena such as natural disasters caused by heavy rainfall or earthquakes, financial risks in finance, and product lifespans, estimating the probabilities of extremely large or small values within the overall data is a critical issue from the perspective of reliability assessment and risk management. The statistical topic that addresses this issue is the prediction of the maximum or minimum values of the data, or quantiles close to these extremes. For this purpose, it is essential to construct probabilistic models that focus solely on the tail behavior. Extreme value statistics provide the methodological framework for such analyses, offering a mathematically rigorous approach to modeling the tails of distributions where rare events occur. Naturally, however, modeling for real-world data requires tailored approaches that consider the unique characteristics of the data. This paper provides an overview of the fundamental concepts, statistical theories, and modeling techniques in extreme value statistics. Special attention is given to the concept of return periods, which play a critical role in understanding the frequency characteristics of phenomena based on the analysis results derived from applying extreme value statistical methods to real data. In addition, data sets where extreme value statistics are applied, such as meteorological and financial market data, are often obtained in the form of clustered data. Therefore, this paper also discusses methodological approaches to extreme value statistical modeling for clustered data. Finally, we illustrate the application of extreme value statistical modeling to rainfall data and present relevant R packages to demonstrate the practical implementation.

スポーツ外傷予測への信頼性解析の応用： 大相撲力士に対する実証分析

太田 修平¹・木村 光宏²

(受付 2024 年 11 月 30 日；改訂 2025 年 3 月 24 日；採択 3 月 25 日)

要　　旨

スポーツ外傷予測とは、対象者に将来起こりうるスポーツ外傷の潜在的なリスクを評価することである。近年では、スポーツ選手が過去に負った外傷のイベントヒストリーデータに基づいて、時系列的に選手の外傷リスクを評価・予測する手法がいくつか提案されている。本稿では、点過程を用いた外傷発生を表す確率モデルを詳説し、そのモデルを用いた外傷予測の方法について述べる。また、大相撲力士に対する外傷予測の実証分析を通して、外傷の連鎖性を考慮できる自己励起型の点過程が、外傷予測に有効であることを示す。さらに本稿は、力士の外傷予測に関する先行研究では明らかにされなかった、予測モデルのパラメータ推定に使用されていない未知のデータに対する、モデルの予測精度の検証を新たに行い、モデルの妥当性を補強する。最後に、予測結果を外傷予防に応用する観点から、外傷予測の今後の課題を述べる。

キーワード：点過程、確率予測、スポーツ外傷、信頼性解析。

1. はじめに

スポーツ選手にとって、スポーツ外傷予防は、自身のパフォーマンスを維持し、キャリアを長く保つために重要である。ここで、スポーツ外傷は、外から加わる 1 回の大きな力(衝撃)によって起こる怪我を指し、例えば捻挫、脱臼、骨折などを含む。一方で、外傷予防には、選手の日常的な心身のケアが必要であり、そのためには時間的・経済的なコストがかかる。そしてそのコストの削減のために、選手に対する効率的な外傷予防の介入が求められている。とくに、選手ごとの外傷リスクを評価できれば、個別最適化された効率的な外傷予防が可能となる。このとき、スポーツ外傷予測は、外傷リスクの評価に有用な手法である。

スポーツ外傷予測は、対象者の健康状態や実績データから、将来起こりうる外傷の潜在的なリスクを評価することである (Bahr and Holme, 2003; Pfirrmann et al., 2016)。予測によって、対象者はリスクを事前に把握することができ、適切な予防措置を講じることで、自身の健康と安全を守ることができる。近年では、スポーツ選手らの日常的な活動量や健康状態のモニタリングデータの記録が進んでおり、データの利活用の観点から、外傷予測は挑戦的な研究分野となっている (Fiscutean, 2021; Rossi et al., 2021)。とくに、外傷予測では、外傷に関するモニタリングデータに対して、機械学習の手法が主に用いられている (Rossi et al., 2018; Rommers et al., 2020; 小笠原, 2021)。

¹ 神奈川大学 工学部：〒221-8686 神奈川県横浜市神奈川区六角橋 3-27-1; ota@kanagawa-u.ac.jp

² 法政大学 理工学部：〒184-8584 東京都小金井市梶野町 3-7-2; kim@hosei.ac.jp

スポーツ外傷は対象者に繰り返し発生すると考えると、いつ外傷が発生したかを記録した外傷履歴は、外傷予測の重要な情報である (Shrier et al., 2009; Parekh et al., 2012)。信頼性解析や生存時間解析の観点から、外傷履歴をイベントヒストリーデータとして扱うことで、不確実に生じる外傷を統計的にモデル化することができ、そのモデルを用いることで将来に起こりうる、潜在的な外傷の発生確率を予測することができる。例えば、Parekh et al. (2012) は定常ポアソン過程によって、学生ラグビー選手を対象とした外傷予測モデルを提案している。しかし、このモデルは、定常ポアソン過程の無記憶性により、外傷の再発など、過去に生じた外傷が将来の外傷の起こりやすさに与える影響を考慮できないという課題がある (Parekh et al., 2012; Rizou et al., 2017)。すなわち、長期的に対象者の外傷予測を行う場合には、定常ポアソン過程は適していない。

スポーツ外傷予測において、定常ポアソン過程モデルの無記憶性の問題を解決するために、Ota and Kimura (2023) は大相撲力士を対象として、外傷の長期的な影響を考慮した外傷予測モデルを提案した。とくに提案手法は信頼性工学におけるシステムのモデリングと信頼性解析の手法を応用し、力士の休場をシステムの故障と見なして、その故障の発生を自己励起過程でモデル化した。このモデルは、自己励起過程の自己励起性によって、過去の外傷履歴が将来に起こりうる外傷に与える影響を考慮できる。その結果、予測精度の向上に成功している。しかし、提案されたモデルの予測精度は、モデルの学習データに対する適合度でのみ評価されており、学習に使用していない、未知のデータに対するモデルの予測精度は評価されていない。実用の観点からは、予測モデルの妥当性が解明されていないという課題がある。

本稿では、こうした近年のスポーツ外傷予測の背景を踏まえて、外傷予測のための点過程モデルを詳述する。2章では、点過程モデルに用いる外傷の分類と計測単位について述べる。さらに大相撲力士の外傷履歴データを例に、外傷発生の統計的な特徴を概観する。3章では Ota and Kimura (2023) を再訪し、点過程を用いた外傷予測手法を解説する。予測手法として、定常ポアソン過程と自己励起過程をそれぞれ用いた、2つの予測モデルを導入する。そして4章では、力士の外傷履歴データを例にとり、予測モデルの適合度の評価と、予測の数値例を示す。また、Ota and Kimura (2023) では明らかにされなかった、モデルのパラメータ推定に使用されていない未知のデータに対する、モデルの予測精度の検証を新たに行い、予測モデルの妥当性を補強する。

2. スポーツ外傷データ

2.1 準備

スポーツ現場で発生するスポーツ外傷には、擦り傷のように傷の処置を適切に実施すれば、対象者がすぐに競技に復帰できるものから、前十字靭帯損傷のように、競技への復帰に半年以上を要するものまでさまざまである (砂川, 2015)。そこで、まず、スポーツ外傷の分類について述べる。

スポーツ外傷の分類として、医療系国家資格や専門資格保有者に対応された外傷は、medical attention injury と定義される (砂川 他, 2022)。さらに、medical attention injury のうち、練習や試合に参加できないスポーツ外傷はタイムロス外傷、参加できるものはノンタイムロス外傷として定義される。

スポーツ外傷の調査では、対象者の活動単位を暴露(exposure)と呼ぶ。暴露の代表的な数え方には、athlete-exposure(AE) と athlete-hour(AH) の 2 つが挙げられる。AE とは、対象者が 1 回の練習または試合に参加する単位を指し、比較的に集計が容易である。一般的には、1,000AEあたりのスポーツ外傷件数が調査される。AH とは、対象者が 1 時間の練習または試合に参加

する単位を指すため、スポーツ外傷のリスクが総合時間の影響を受けやすいと考える場合は有用であるが、AEよりも集計が複雑である。

以上の定義を踏まえて、次節では、AEを活動単位としたタイムロス外傷の外傷履歴の例を示す。

2.2 大相撲力士のスポーツ外傷データ

本節では、大相撲力士を例にとり、スポーツ外傷履歴データの可視化と、外傷発生の統計的特徴を紹介する。なお、力士は平均体重が160kgであり、他のスポーツ選手と比べて外傷を負いやすい（清水他, 2021）。そのため、力士の外傷履歴データは多く存在するという特徴がある。これが、本研究が力士を例にとる理由である。

まず、力士の外傷履歴を、力士が休場するまでに要した試合数（取り組み数）と定義する。休場とは、力士が規定の休場届を日本相撲協会に提出して、本場所の取り組みに出場しないことである（金指, 2002）。力士は謹慎処分などの例外を除き、負傷または病気のため本場所に出場できない場合のみ休場できる。休場の多くはタイムロス外傷が原因であることを踏まえて、本研究は休場に関するデータで外傷履歴を定義する。

本研究は、力士の外傷履歴データとして、一般公開されているデータベース（相撲レファレンス, 2024）で公開されている、力士の生涯戦歴数[単位：1,000 AE]および休場時刻[単位：1,000 AE]を利用する。生涯戦歴数は、力士が初土俵を踏んでから数えた本場所での勝ち、負け、休場の試合数の総和である。休場時刻は力士が休場をした時点の生涯戦歴数である。このデータからData-Aと、Data-Aには含まれないData-Bの2つのデータセットを作成する。なお力士の番付は階級があり、下から順に、前相撲、序ノ口、序二段、三段目、幕下、十両、幕内と呼ばれる。幕内の定員は42名である。また、大相撲では1年に6回開催される本場所において力士同士の対戦が行われる（伊藤, 2019）。

Data-Aは、初土俵を1973年から2003年に踏み、幕内の番付を1場所以上維持したことがある力士（該当者数=209）のデータである。このデータはモデルのパラメータを推定するため使用する。Data-Bは、2019年の1年間における幕内力士のデータである。このデータは推定されたスポーツ外傷予測モデルを用いた、外傷予測の数値計算例のために使用する。また、未知のデータに対する予測精度の検証のためにData-Bを使用する。なお、幕内経験者以外のデータを除外する理由は、データの層別化のためである。

図1はData-Aに含まれる209人の力士の休場データを示す。図において、力士によっては×印が立て続けに並んだ箇所が見られる。このことから、多くの休場の発生が連鎖的に発生していることが分かる。これは、スポーツ外傷の長期的な影響によって、外傷を負った力士は、その外傷を再発しやすいや、外傷部位をかばうことで他の部位に負担がかかり外傷を負いやすいという、実際の力士の特徴が表れていると考えられる。実際に、力士の前十字靭帯損傷の再発頻度は、他のスポーツ選手よりも高いことが知られている（Shimizu et al., 2020）。

次に、力士のスポーツ外傷の起こりやすさの時間変化を可視化するために、力士の平均的な休場の発生頻度を示す。時刻 t (≥ 0) を力士の生涯戦歴数として、力士が一定期間内で休場する頻度を示す指標として、時刻 t における平均休場率を次のように定義する。

$$(2.1) \quad \text{平均休場率} = \frac{\text{時刻 } t \text{ で出場かつ } (t, t + \Delta t] \text{ で休場した力士の人数}}{\text{時刻 } t \text{ で出場した力士の人数}}.$$

ただし、 Δt は正の実数とする。平均休場率は信頼性解析における故障率とも解釈できる（Chioldo and Lauria, 2015）。図2はData-Aに含まれる力士の平均休場率の振る舞いを表す。平均休場率はおよそ1回目の休場時刻の平均値 \bar{t}_1 まで一定であり、その後は t が大きくなるに

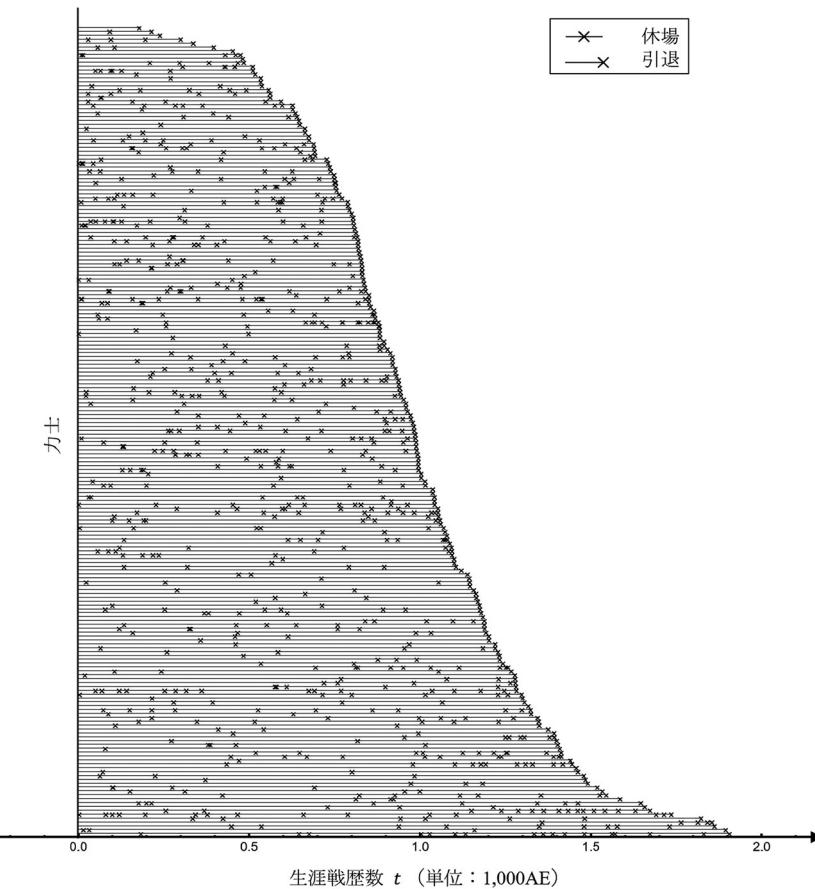


図 1. Data-A に含まれる 209 人の力士の休場データ。各線は各力士の休場データを意味する。力士の並びは引退までの生涯戦歴数についての昇順である。

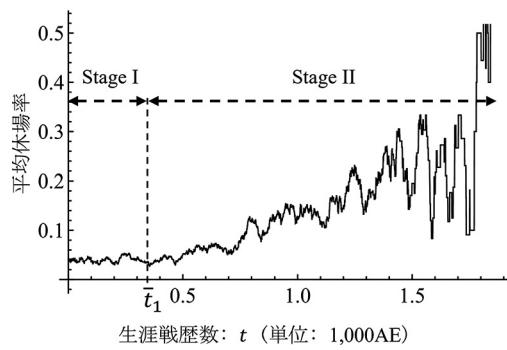


図 2. 平均休場率の振る舞い。

つれて増加する。このような平均休場率の振る舞いは、信頼性工学で良く知られたバスタブ曲線の偶発故障期と摩耗故障期として捉えることができる(真壁 他, 2002; Trivedi, 2016)。本研究は以上の点に着目し、3章で休場予測モデルを導入する。

3. 予測モデル

本章では、2.2節の議論を踏まえて、スポーツ外傷予測モデルとして定常ポアソン過程、および自己励起過程をそれぞれ用いたモデルを導入する。それぞれのモデルの入力は、ある時点までの対象者の外傷履歴データであり、出力は、 t 時間以内に新たに外傷を発症する確率である。 T_{ij} を対象者 i が j 回目の外傷を発症した時刻を表す確率変数とする($i = 1, 2, \dots, n$, $j = 1, 2, \dots$)。ただし、外傷の発生の順序を考慮して、 $T_{i1} < T_{i2} < \dots$ とする。また、外傷の発生時刻の集合を $H_t = \{T_{ij} | T_{ij} < t\}$ とする。そして、 T_{ij} はある強度関数 $\lambda(t|H_t)$ に従っているとする。ここで、直感的には、強度関数とは時刻 t における瞬間的な外傷発生率を意味する(Rizoou et al., 2017)。厳密には、時刻 t までに発生した外傷の発生回数 N_t に対して、強度関数は次式で与えられる(山中他, 2017)。

$$(3.1) \quad \lambda(t|H_t) = \lim_{\Delta t \rightarrow 0} \frac{E[N_{t+\Delta t} - N_t|H_t]}{\Delta t}.$$

また強度関数は、ハザードレートとも呼ばれ信頼性工学では故障発生率の評価に用いられている(Ota and Kimura, 2017; Wu, 2019)。さらに、予測モデルには以下のことを仮定する。

仮定1：対象者らのすべての外傷発生は独立で同一な確率過程に従う。

仮定2：対象者が外傷を負っている時間は、調査期間に対して短く無視できるとする。すなわち、対象者の外傷は直ちに治癒され、競技を再開できるとする。

以上の定義と仮定に基づいて、次節から、2種類の外傷予測モデルを導入する。

3.1 定常ポアソン過程モデル

本節では、定常ポアソン過程を用いたスポーツ外傷予測モデルを導入する。強度関数が時刻 t によらず定数 λ_0 としたとき、すなわち

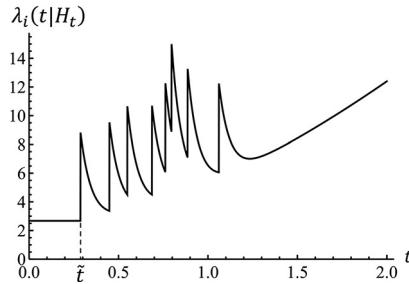
$$(3.2) \quad \lambda_p(t) = \lambda_0,$$

ならば、 T_{ij} はパラメータ λ_0 の指数分布に従い、外傷の発生回数 N_t は定常ポアソン過程に従うと言う。定常ポアソン過程によるスポーツ外傷予測では、将来起こりうる外傷が過去の外傷履歴によらないという仮定が必要である。この性質は定常ポアソン過程の無記憶性とも呼ばれる。従って、もし T_{ij} が $j = 1, 2, \dots$ に関して互いに独立であり、かつ時刻 t に依存しないならば、定常ポアソン過程は、1つのパラメータのみのシンプルな記述で外傷予測を行えるモデルである。例えば、Parekh et al. (2012)は定常ポアソン過程を用いて、ラグビーにおける男子生徒の外傷発生をモデル化した。しかし、先述のとおり、長い調査期間においては、対象者の外傷の起こりやすさは変化するため、定常ポアソン過程は外傷予測に適さない。

3.2 自己励起過程モデル

本節では、自己励起過程を用いたスポーツ外傷予測モデルを導入する。自己励起過程とは、比較的短い時間内に、複数のイベントが集中して起こりやすいという規則性を持った点過程であり、イベントの発生が過去のすべてのイベントの発生に依存する現象を表せる(Hawkes, 1971; Hawkes, 2018)。外傷予測においては、外傷発生の連鎖性を考慮するのに有用な点過程である。Ota and Kimura (2023)は、次の強度関数 $\lambda_i(t|H_t)$ で T_{ij} をモデル化した。

$$(3.3) \quad \lambda_i(t|H_t) = \begin{cases} \lambda_0 & (t \leq T_{i1}) \\ \lambda_0 + ab(t - T_{i1})^{b-1} + \sum_{T_{ij} < t} \alpha e^{-\beta(t - T_{ij})} & (t > T_{i1}). \end{cases}$$

図 3. 自己励起過程による予測モデルの強度関数の振る舞い例 ($b > 1$).

ただし, $\lambda_0, a, b, \alpha, \beta > 0$ であり, さらに $H_t = \{T_{ij} | T_{ij} < t\}$ は対象者 i において時刻 $[0, t)$ に発生したすべての外傷発生時刻の集合であり, 外傷履歴を意味する.

式(3.3)は, 外傷発生率の 2段階の振る舞いを表現するために場合分けされている. すなわち, 対象者の 1回目の外傷が発生するまでの $t \leq T_{i1}$ に対して, 式(3.3)は外傷発生率が一定であることを表している. 一方で $t > T_{i1}$ の場合, 式(3.3)の第 2 項は外傷発生率が時間とともに増減する振る舞いを表しており, 第 3 項は過去に発生したすべての外傷が, 将來の外傷発生に与える長期的な影響を表している. このように外傷発生率を場合分けすることで, 本モデルでは休場するほどの外傷を一度負った力士は, その後その外傷をかばうことで, 体の他の部位に負担がかかり, 戦歴とともに外傷の可能性が高まることを表現している. 一方で, 一度も休場していない力士は引退するまで外傷発生率の変化がしにくいことも, 本モデルは表現している.

このモデルにおいて, 時刻 T_{i1} は強度関数の変化点になっている. 変化点を含む強度関数のモデリングに関しては, 例えば, ソフトウェア信頼性の分野で研究されている (Inoue and Yamada, 2015). また, $a = \alpha = 0$ の場合, $\lambda_i(t|H_t)$ は定常ポアソン過程の強度関数に一致するため, このモデルは定常ポアソン過程モデルの拡張モデルである. モデルのパラメータは, 対象者らの外傷履歴データから統計的に推定することが可能である. 具体的な推定方法は, 次節で述べる.

図 3 は, $b > 1$ における $\lambda_i(t|H_t)$ の振る舞いの例を表している. 強度は $t = 0$ から 1回目のスポーツ外傷が発生した \tilde{t} まで一定であるが, \tilde{t} の直後に強度は α だけジャンプする. このジャンプによって, 対象者は外傷直後に再び外傷を負いやすいことを表現している. その後, 強度は指数関数的に減少し, 再び外傷が発生した時点でジャンプし, このことを時間とともに繰り返している. そして, 外傷が発生せず十分に時間が経過したおよそ $t = 1.2$ 以降は, 強度が単調増加している. このように, 自己励起過程による外傷予測モデルは, 時刻と外傷履歴に応じた外傷発生率の変化を表現可能である.

一方で, $0 < b < 1$ の場合, 強度は時間に対して減少して一定の値に収束する. これは, 対象者が時間とともに体の使い方を覚えていくことで, 対象者のスポーツ外傷のリスクが低下する現象を, 自己励起過程による外傷予測モデルは表現できると言える. このような現象は, 分析する期間を限定すれば観測できると考えられる. 例えば, 齊藤 他 (2020) および大垣 他 (2022) によれば, サッカー選手の外傷発生数がシーズン後期に近づくにつれて, 徐々に減少していることが報告されている. この場合, 分析期間を 1 シーズンに限定してスポーツ外傷の予測モデルを構築する場合は, $0 < b < 1$ となる可能性がある.

対象者の将来時点におけるスポーツ外傷の発生確率は, 式(3.3)から求められる. $p_{ij}(w|t)$ を, 時刻 t において対象者 i に次の j 回目の外傷がその後 w (> 0) 時間以内に発生する確率とする. このとき, $p_{ij}(w|t)$ は次式で与えられる.

$$(3.4) \quad p_{ij}(w|t) = \begin{cases} 1 - e^{-\lambda_0 w} & (j = 1) \\ 1 - e^{-\int_t^{t+w} \lambda_i(s|H_t) ds} & (j \geq 2). \end{cases}$$

対象者が時刻 t までに一度も外傷を負っていない場合, 式(3.4)は t によらず定数で求められ, それ以外の場合, w, t, H_t に依存する. 例えば, 対象者が大相撲力士の場合, 15 日間(= 0.015AE)開催される本場所中の外傷の発生確率は $p_{ij}(0.015|t)$ で求められる.

また, 個々の対象者の外傷の発生確率から, 対象集団全体での外傷発生数の評価ができる. 対象者の人数を N , 確率変数 X を本場所中に外傷が発生する対象者数, および X の確率質量関数を $q_x \equiv \Pr[X = x]$ とする. そして, 時刻 t_i を対象者 i の AE とすると, その後 w 時間以内に各対象者に外傷が発生する確率は $p_{ij}(w|t_i)$ となる. このとき, X はポアソン二項分布(Chen and Liu, 1997)に従うことが知られており, q_x は次式で与えられる.

$$(3.5) \quad q_x = \sum_{S' \in S_x} \left\{ \prod_{i \in S'} p_{ij}(w|t_i) \prod_{i \in (I \setminus S')} (1 - p_{ij}(w|t_i)) \right\}.$$

ここで, $I = \{1, 2, \dots, N\}$ であり, S_x は集合 I から選択可能な x 個の整数からなる, すべての部分集合である. さらに, X の期待値と分散は以下のようになる.

$$(3.6) \quad E[X] = \sum_{i=1}^N p_{ij}(w|t_i),$$

$$(3.7) \quad V[X] = \sum_{i=1}^N (1 - p_{ij}(w|t_i)) p_{ij}(w|t_i).$$

このように, 特定の対象者の集団に潜在する外傷リスクを, $q_x, E[X], V[X]$ で要約することが可能である.

3.3 パラメータの最尤推定

本節では, スポーツ外傷予測モデル $\lambda_i(t|H_t)$ のパラメータ $\lambda_0, a, b, \alpha, \beta$ の推定方法として, 最尤推定を紹介する.

変数 ν_i を対象者 i が引退するまでのスポーツ外傷の発生回数とする. このとき, 式(3.3)で与えられる自己励起過程モデルに対する対数尤度関数 $\log \mathcal{L}$ は, 以下のように表される.

$$(3.8) \quad \log \mathcal{L}(\lambda_0, a, b, \alpha, \beta|H_t) = \sum_{i=1}^n \left\{ \sum_{j=1}^{\nu_i} \log \lambda_i(t_{ij}|H_t) - \int_0^{t_{i\nu_i}} \lambda_i(s|H_t) ds \right\}.$$

さらに, 式(3.8)は以下のように再帰的に計算できる(Ogata, 1988).

$$(3.9) \quad \log \mathcal{L}(\lambda_0, a, b, \alpha, \beta|H_t) = \sum_{i=1}^n \left\{ \sum_{j=1}^{\nu_i} \log(h_i(t_{ij}|H_t) + \alpha R(j)) - \int_0^{t_{i\nu_i}} h_i(s|H_t) ds - \sum_{j=1}^{\nu_i} \frac{\alpha}{\beta} (1 - e^{-\beta(t_{i\nu_i} - t_{ij})}) \right\}.$$

ただし,

$$(3.10) \quad h_i(t|H_t) = \begin{cases} \lambda_0 & (t \leq t_{i1}) \\ \lambda_0 + ab(t - t_{i1})^{b-1} & (t > t_{i1}), \end{cases}$$

$$(3.11) \quad R(j) = \begin{cases} 0 & (j = 1) \\ e^{-\beta(t_{ij} - t_{ij-1})}(1 + R(j-1)) & (j \geq 2). \end{cases}$$

そして、式(3.9)を最大化することで、パラメータ $\lambda_0, a, b, \alpha, \beta$ の最尤推定量として $\hat{\lambda}_0, \hat{a}, \hat{b}, \hat{\alpha}, \hat{\beta}$ が得られる。また、推定量の標準誤差は対数尤度関数のヘッセ行列から求められる (Serfling, 1980)。

4. 実証分析

本章では、大相撲力士を対象としたスポーツ外傷予測の結果を示す。大相撲において、力士の外傷をいかに減らすかということは長年の課題である。例えば、2019年において幕内力士42人の内、本番所当たり平均して5.2人が休場している。しかし、依然として各力士がどれだけ外傷を負いやすいかは明らかになっておらず、外傷予防の導入が進んでいない。そのため、力士の外傷リスクを把握するために、外傷予測は重要である。

本章の流れを図4に示す。図4において、四角は工程を表し、矢印は各工程の入力と出力を表している。まず、過去に引退した力士のスポーツ外傷の履歴データから、予測モデルのパラメータを推定し、モデルの適合度の評価を行う。そして、推定されたモデルを用いて外傷予測の数値例を示す。

4.1 パラメータ推定結果

大相撲力士のスポーツ外傷の発生をモデル化するために、2.2節で定義したData-Aを用いて、式(3.3)のパラメータを最尤推定 (Ogata, 1981; Lehmann and Casella, 1998)する。Data-Aは力士の休場と引退を、外傷の発生とみなしたデータである。パラメータの最尤推定量は、式(3.9)で与えられる対数尤度関数を最大化することで得られる。ここで $\hat{\lambda}_0, \hat{a}, \hat{b}, \hat{\alpha}, \hat{\beta}$ を、それぞれパラメータ $\lambda_0, a, b, \alpha, \beta$ の最尤推定量とする。そして、パラメータの最尤推定量を $\lambda_i(t|H_t)$, $p_{ij}(w|t)$, q_x に代入したものを、それぞれ $\hat{\lambda}_i(t|H_t)$, $\hat{p}_{ij}(w|t)$, \hat{q}_x とする。

なお、休場は引退により打ち切られるが、この打ち切り (Blanche et al., 2013; Emura and Chen, 2018)が推定結果に与える影響は小さく、無視できるものと仮定する。ここでは打ち切りの影響を無視するが、推定結果の妥当性は、次節以降のモデル検証および予測精度の検証で保証される。

表1にパラメータの推定値を示す。この表は、推定値、標準誤差、およびAIC (Akaike Information Criterion; 赤池情報量規準)をまとめている。AICはモデルのパラメータ数 k および最大対数尤度 $\log \mathcal{L}$ に対して、 $AIC = 2k - 2\log \mathcal{L}$ で与えられる。AICの第一項はパラメータ数に対する罰則項であり、第二項はデータに対するモデルに当てはまりの良さを表している。このことから AIC が小さいモデルほど、データに適合しているといえる。よって、より小さい値の AIC をもつ自己励起過程によるモデルは、定常ポアソン過程によるモデルよりも Data-A に適

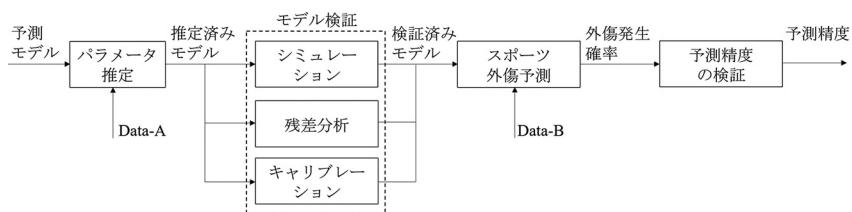


図4. スポーツ外傷予測モデルの検証と予測の流れ。

表 1. パラメータの推定結果と各モデルの AIC(括弧の中身は標準誤差である)。

モデル	$\hat{\lambda}_0$	\hat{a}	\hat{b}	$\hat{\alpha}$	$\hat{\beta}$	AIC
定常ポアソン過程	4.671 (0.149)	—	—	—	—	-1058.03
自己励起過程	2.685 (0.170)	1.399 (0.298)	2.689 (0.300)	2.866 (0.630)	7.626 (2.663)	-1267.93

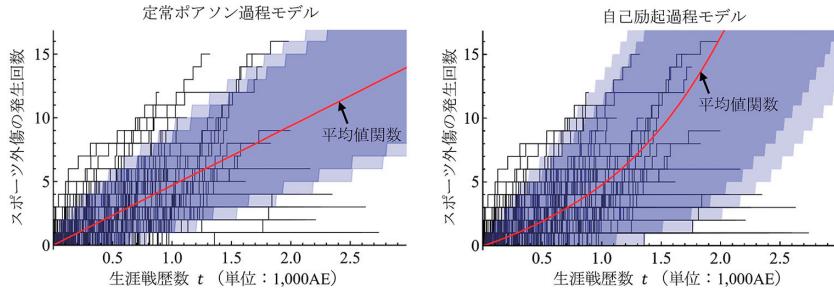


図 5. 予測モデルの信頼区間。黒線は Data-A に含まれる 209 名の力士のスポーツ外傷の発生のパス、薄いおよび濃い領域はシミュレーションにもとづく平均値関数の 99% および 95% 信頼区間をそれぞれ表している。

合していることが分かる。

推定結果について考察する。自己励起過程による予測モデルにおいて、 $\hat{\lambda}_0 = 2.685$ であるから、1 回目の外傷が発生するまでの時間の期待値は $0.372(=1/\hat{\lambda}_0)$ ということが分かる。初土俵を踏んで間もない力士が、この期待値までに休場する確率は以下のように求められる。

$$(4.1) \quad \hat{p}_{i1}(0.372|t) = 1 - e^{-\hat{\lambda}_0 0.372} = 1 - e^{-1} = 0.632.$$

また、1 度も外傷が発生していない力士が、15 日間開催される本場所において、休場する確率は $\hat{p}_{i1}(0.015|t) = 0.0395$ と推定できる。そして、 $\hat{b} > 1$ より、力士は 1 回目の休場後に図 2 の Stage II の摩耗故障期 (Shooman, 2003) に入り、 $\hat{\alpha} > 0$ より、外傷発生の長期的な影響があることが分かる。

4.2 適合度評価

本節では、前節で推定されたモデルの妥当性を検証する。推定された強度関数 $\hat{\lambda}_i(t|H_t)$ に基づくシミュレーション (Ogata, 1981) により、時間に対するスポーツ外傷の発生回数の期待値および信頼区間を求める。なおシミュレーションでは有意水準 γ に対して $100(1 - \gamma)\%$ 信頼区間 (Dohi et al., 2020) を、シミュレーション結果の $\gamma/2$ および $1 - \gamma/2$ パーセント点で導出する。図 5 は推定された定常ポアソン過程モデルと自己励起過程モデルに基づくシミュレーション結果、および Data-A の経験的な結果の重ね合わせを、それぞれ示している。例えば、自己励起過程モデルによる予測モデルによれば、 $t = 1.0$ までの外傷の発生回数の期待値(すなわち平均値関数の値)は 4.8 となる。図 5 より、定常ポアソン過程による予測モデルでは、信頼区間に含まれないデータがグラフ左上および右下で見受けられるが、自己励起過程による予測モデルでは信頼区間に含まれないデータが減っていることが分かる。

次に予測モデルの残差 (Ogata, 1988) を調べることで、Data-A に対する予測モデルのそれぞれの適合度を評価する。 $i = 1, 2, \dots, n$ に対して、 $\{t_{i1}, t_{i2}, \dots, t_{i\nu_i}\}$ を、強度関数が $\lambda_i(t|H_t)$ で

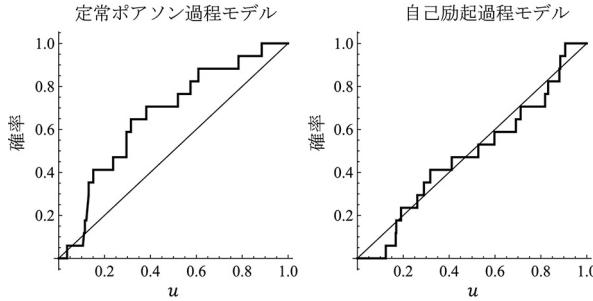


図 6. Data-A に含まれる力士 1 名の休場時刻データを予測モデルで変換した $\{u_{i1}, u_{i2}, \dots, u_{i\nu_i}\}$ の経験分布関数.

ある点過程の実現値とする. 連続した t_{ij-1} と t_{ij} に対して, 以下の $\lambda_i(t|H_t)$ の積分を考える.

$$(4.2) \quad \Lambda_i(t_{ij-1}, t_{ij}) = \int_{t_{ij-1}}^{t_{ij}} \lambda_i(s|H_t) ds.$$

点過程の時間変換により, $\{\tau_{i1}, \tau_{i2}, \dots, \tau_{i\nu_i}\}$ を以下のように定義する.

$$(4.3) \quad \tau_{i1} = \int_0^{t_{i1}} \lambda_i(s|H_t) ds,$$

$$(4.4) \quad \tau_{ij} = \tau_{ij-1} + \Lambda_i(t_{ij-1}, t_{ij}).$$

ここで, $\tau_{ij} - \tau_{ij-1} = \Lambda_i(t_{ij-1}, t_{ij})$ は, 期待値が 1 の指数分布に従う. また, $\{\tau_{i1}, \tau_{i2}, \dots, \tau_{i\nu_i}\}$ は残差と呼ばれる. さらに, $j = 1, 2, \dots, \nu_i$ に対して, $u_{ij} = 1 - e^{-\tau_{ij}}$ は区間 $[0, 1]$ 上の独立で同一な一様分布に従う. 従って, もし推定された強度関数 $\hat{\lambda}_i(t|H_t)$ が真の強度関数 $\lambda_i(t|H_t)$ の精度の良い近似であれば, $\{u_{i1}, u_{i2}, \dots, u_{i\nu_i}\}$ は区間 $[0, 1]$ 上の独立で同一な一様分布に従うことが分かる.

図 6 は *u-plot* (Abdel-Ghaly et al., 1986) として知られ, 推定された定常ポアソン過程と自己励起過程による, それぞれの予測モデルの強度関数から得られた $\{u_{i1}, u_{i2}, \dots, u_{i\nu_i}\}$ の経験分布関数を表している. これら経験分布が理論分布である一様分布に適合しているかを, コルモゴロフ・スマルノフの適合度検定(有意水準 5%)で調べる. 結果として, Data-A に含まれる 209 人の力士に対して, 定常ポアソン過程による予測モデルでは 31 人の力士データが適合しなかったが, 自己励起過程による予測モデルでは 8 人の力士のみが適合しなかった. このことから, 自己励起過程による予測モデルの方が定常ポアソン過程による予測モデルよりも Data-A に適合しているといえる. 以上の議論を踏まえて, 以降では自己励起過程による予測モデルのみを分析する.

最後にキャリブレーション (Gerds and Kattan, 2021) によって, 予測モデルの予測精度を検証する. キャリブレーションとは, 予測モデルが输出した確率がどの程度, 実際のイベントの発生率を近似しているかを評価し, さらに, 必要に応じて出力を補正する手法である. キャリブレーションの手順は以下のとおりである.

(1) Data-A に含まれる各力士が, 初土俵から引退するまでの各本場所で休場する確率の推定値を求める. このとき, 力士は幕下以下の場合に本場所で 7 日間対戦し, 幕下よりも上の場合は 15 日間対戦することを考慮して, 対象力士が休場する確率の推定値を, それぞれ $\hat{p}_{ij}(0.007|t)$ と $\hat{p}_{ij}(0.015|t)$ で求める.

(2)区間 $[0,0.3]$ を 0.05 刻みで分割し、推定された確率をそれらが含まれる区間ごとにグループ化する。そして、各グループに含まれる推定された確率の平均値を求め、各グループの代表値とする。

(3)各グループにおける実際の休場の発生率を求める。すなわち、各グループが含むデータに対して、データ数で実際に休場した力士の人数を割った値を求める。

(4)手順(2)で求めた推定された確率と、手順(3)で求めた実際の休場の発生率の 2 变量データを散布図にする。

図 7 は Data-A に対するキャリブレーションの結果を示す。ここで各区間 $[0,0.05]$, $[0.05,0.1]$, $[0.1,0.15]$, $[0.15,0.2]$, $[0.2,0.25]$, $[0.25,0.3]$ のサンプルサイズは、それぞれ 9438, 7075, 1485, 217, 39, 5 となった。例えば、区間 $[0.1,0.15]$ に含まれるデータでは、休場する確率の推定値の平均が 0.118 で、実際の休場の発生率は 0.122 であった。この 2 つの値が近いほど、予測が実際の結果を再現できたと言える。従って、図 7 において各点が対角線付近にあることから、予測モデルは平均的に精度よく力士のスポーツ外傷の発生を予測できているといえる。

ただし、区間 $[0.25,0.3]$ における点は、休場する確率の過小評価により、対角線から最も離れるという評価結果となった。この過小評価の主な要因として、(i)サンプルサイズが 5 と小さいこと、(ii)引退による外傷発生の打ち切りを無視する仮定の影響が考えられる。まず、サンプルサイズが小さいことに関しては、休場する確率が 0.25 以上となる力士の数が少ないと影響している。そのような力士は、生涯戦歴数が多く、休場が多くても力士を続けられるほどの高位の番付を維持できる実力を有している必要がある。これらの条件を満たす力士として、唯一休場しても番付が下がらない特例が適用される横綱が該当する可能性が高い。実際に、図 7 において区間 $[0.25,0.3]$ における点は、すべて同一の横綱のサンプルのみから計算された。このように、区間 $[0.25,0.3]$ における点は限られたデータのみから導出されたため、過小評価が生じたと考えられる。次に、引退による打ち切りの影響を無視する仮定の影響について考える。生涯戦歴数が大きくなるにつれて、引退によるデータの打ち切りの影響は強くなるため、休場する確率が過小評価されやすい。以上より、自己励起過程による予測モデルは、とくに横綱のような生涯戦歴数の多い力士に対して、スポーツ外傷の発生確率を過小評価する可能性があると言える。

4.3 予測結果例

本節では、前節のパラメータ推定と検証の結果を踏まえて、Data-B に含まれるすべての力士

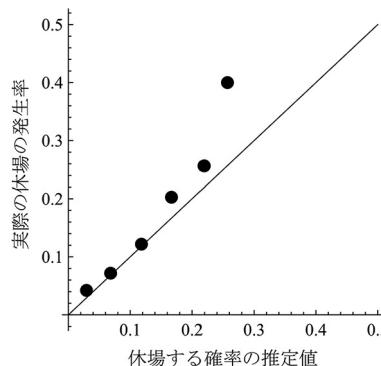
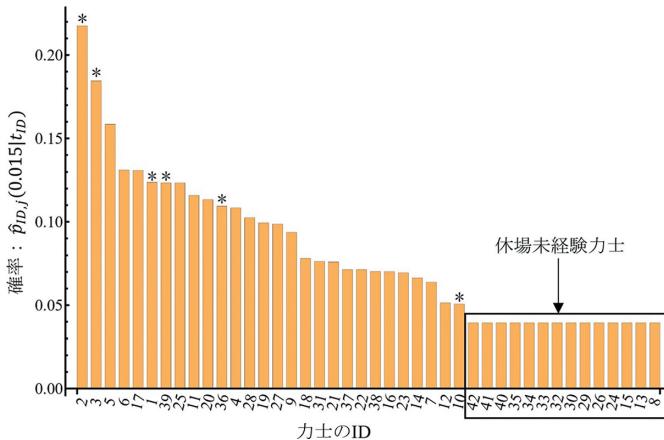


図 7. 自己励起過程による予測モデルのキャリブレーション。

表 2. 2019 年 1 月場所における幕内力士 42 名のデータ.

ID	しこ名	ID	しこ名	ID	しこ名
1	稀勢の里	15	正代	29	宝富士
2	白鵬	16	松鳳山	30	阿炎
3	鶴竜	17	琴奨菊	31	佐田の海
4	高安	18	隠岐の海	32	勢
5	豪栄道	19	碧山	33	輝
6	栃ノ心	20	嘉風	34	明生
7	貴景勝	21	千代大龍	35	矢後
8	玉鷲	22	阿武咲	36	琴勇輝
9	妙義龍	23	竜電	37	豊山
10	御嶽海	24	大栄翔	38	千代翔馬
11	栃煌山	25	魁聖	39	千代の国
12	逸ノ城	26	朝乃山	40	琴恵光
13	錦木	27	貴ノ岩	41	大奄美
14	北勝富士	28	遠藤	42	大翔丸

図 8. 2019 年 1 月場所における幕内力士 42 名の休場する確率の推定値 $\hat{p}_{ID,j}(0.015|t_{ID})$. ID と四股名の対応は表 2 に記載されている. *は実際の休場力士を表す.

に対するスポーツ外傷の予測例を示す. 本例題では, 自己励起過程モデルで外傷予測を行い, モデルのパラメータには, 4.1 節で得られたパラメータの推定値を用いる. 予測モデルの入力は, $i = 1, 2, \dots, 42$ および $j = 1, 2, \dots$ に対して, 第 i 力士の生涯戦歴数 t_i と j 回目までの休場時刻であり, 出力は第 i 力士が休場する確率の推定値 $\hat{p}_{ij}(0.015|t_i)$ である. 表 2 は Data-B に含まれる力士の名前と便宜的に割り付けた ID を表している. なお, 6 人の力士 (ID = 1, 2, 3, 10, 36, 39) が, 実際に本場所で休場した.

図 8 は, 各力士が本場所において休場する確率の推定値 $\hat{p}_{ID,j}(0.015|t_{ID})$ を表している. この図から, 各力士が他の力士と比べてどれだけスポーツ外傷を負いやすい状態にあるかが分かる. 例えば, 42 人の力士の内, 白鵬 (ID = 2, $j = 10$) は休場する確率が 21.7% で最も高く, 次いで鶴竜 (ID = 3, $j = 9$) が 18.5% で豪栄道 (ID = 5, $j = 8$) が 15.9% である. 14 人の力士は過去に休場したことが一度もないため, 休場する確率は 3.95% であった. とくに, 玉鷲 (ID = 8, $j = 1$) は休場したことが一度もない力士のなかで最も長い生涯戦歴数 1.151 を有しており, こ

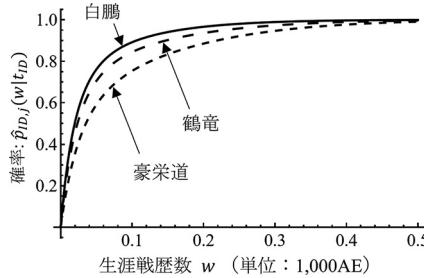


図 9. 白鵬 (ID = 2, $j = 10$), 鶴竜 (ID = 3, $j = 9$), 豪栄道 (ID = 5, $j = 8$) の次のスポーツ外傷発生時間の累積分布関数。原点 $w = 0$ は 2019 年 1 月の本場所初日に該当する。

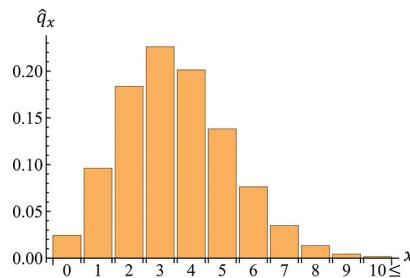


図 10. 2019 年 1 月の本場所における休場者数の推定値の分布。

れは $e^{-1.151\hat{\lambda}_0} = 0.0455$ の確率で起こりうる稀な事象であることが分かる。

図 9 は、白鵬、鶴竜、豪栄道の将来における休場発生確率 $\hat{p}_{ID,j}(w|t_{ID})$ の変化を示している。この図から、それぞれの力士が生涯戦歴数 0.05 以内、すなわち 50 回の取り組み以内、に休場する確率は 50% 以上あることが分かる。また、白鵬は分析対象の本場所において、もっとも生涯戦歴数が長く休場回数も多いため、この図においても将来における休場の確率が最も大きいことが分かる。

2019 年 1 月の本場所において、潜在的にスポーツ外傷で休場する人数 X を推定することも可能である。この人数 X は、大相撲を運営する日本相撲協会が、本場所においてどれくらいの力士が休場するリスクがあるのかを把握することができ、日々の取り組みの組み合わせを考える際に有用である。図 10 は X の推定された確率分布を表している。 X の期待値は 3.401、標準偏差は 1.746 である。また、この本場所において、実際には ID = 1, 2, 3, 10, 36, 39 の計 6 人の力士が休場したが、その発生確率は式(3.5)より $\hat{q}_6 = 0.070$ であったことが分かる。

4.4 予測精度の評価

本節では、Ota and Kimura (2023) では検証されなかった、パラメータの推定に用いていない未知のデータに対する、予測モデルの予測精度を検証する。具体的には、パラメータの推定に使用していない Data-B を用いて、2019 年の各本場所における全幕内力士の実際の休場または出場の同時確率を、定常ポアソン過程モデルと自己励起過程モデルでそれぞれ求め、比較・評価する。

ここで M を、ある本場所で休場した力士の ID の集合とする。そして以下の \tilde{p} のように、その本場所で実際に休場した力士らの休場する確率の推定値と、実際に出場した力士らの出場する確率 (= 1 - 休場する確率) の推定値の積を考える。

表 3. 2019 年の各本場所における予測精度 \tilde{p} の評価結果. 記号*は値が大きいモデルを表す.

モデル	2019 年 1 月	2019 年 3 月	2019 年 5 月	2019 年 7 月	2019 年 9 月	2019 年 11 月
定常ポアソン過程	7.71×10^{-9}	3.83×10^{-3}	1.46×10^{-6}	7.71×10^{-9}	7.71×10^{-9}	5.59×10^{-10}
自己励起過程	$2.25 \times 10^{-7}*$	$4.89 \times 10^{-3}*$	$4.37 \times 10^{-6}*$	$2.30 \times 10^{-7}*$	$4.37 \times 10^{-7}*$	$8.91 \times 10^{-9}*$

$$(4.5) \quad \tilde{p} = \underbrace{\left(\prod_{m \in M} \hat{p}_{mj_m}(0.015|t_m) \right)}_{\text{休場した力士の休場する確率の推定値の積}} \underbrace{\left(\prod_{m \in \{1, 2, \dots, 42\} \setminus M} 1 - \hat{p}_{mj_m}(0.015|t_m) \right)}_{\text{実際に出場した力士の出場する確率の推定値の積}}.$$

ただし, j_m は m 番目の力士が, 対象とする本場所で休場した場合の入門してからの累積休場回数を表す. 式(4.5)は, $m = 1, 2, \dots, n$ に対して, 成功確率が p_{mj_m} で失敗確率が $1 - p_{mj_m}$ の n 回の独立なベルヌーイ試行の同時確率と言える (浜田, 2022). この確率が大きいモデルほど, 実際に起きたイベントの発生確率を大きく見積もったことを意味するので, 予測精度が高いと言える.

Data-B に含まれる 2019 年の各本場所における予測精度 \tilde{p} の評価結果を表 3 に示す. 例えば, 2019 年 1 月の本場所では, $M = \{1, 2, 3, 10, 36, 39\}$ であり, 定常ポアソン過程モデルの場合, $\tilde{p} = 7.71 \times 10^{-9}$, 自己励起過程モデルの場合, $\tilde{p} = 2.25 \times 10^{-7}$ となり, 自己励起過程モデルの方が値が大きく予測精度が高いと言える. とくに, 図 8 において, 自己励起過程モデルは ID = 2, 3 の力士のように実際に休場した力士ほど休場する確率を高く評価し, 出場している力士ほど休場する確率を低く評価している. このことから, 自己励起過程モデルは実際のデータとの整合性があると言える. 一方で, 定常ポアソン過程モデルの場合, 生涯戦歴数や外傷履歴に関わらず, どの力士でも同一の確率で休場すると予測するため, 実際のデータとの整合性の評価が難しい. 以上の理由から, 自己励起過程モデルの \tilde{p} は定常ポアソン過程モデルのものよりも高くなっている. 同様に, すべての本場所において自己励起過程モデルの方が \tilde{p} が大きく, 予測精度が高いことが分かる. したがって, 4.2 節の適合度評価の結果と同様に, スポーツ外傷予測において, 過去のスポーツ外傷の影響を考慮した自己励起過程によるモデルのほうが, それを考慮していない定常ポアソン過程によるモデルよりも, 予測精度が高いことが分かった.

4.5 予測モデルの限界

本節では, 本稿で紹介したスポーツ外傷の予測モデルの限界について述べる. 限界の 1 つとして, 本稿が紹介した外傷予測手法を, 力士以外のスポーツ選手へ適用した場合に, その予測精度が保証されない点がある. 自己励起過程モデルは, 大相撲力士の外傷履歴の特徴から考案されたモデルであり, 他のスポーツ選手の外傷予測に適用できるかは確認されていない. 今後は, 相撲以外のスポーツ選手の外傷履歴データから, モデルのパラメータ推定および外傷予測の精度検証が求められる.

また, スポーツ外傷予防への予測結果の応用の限界がある. 外傷予測に利用できるデータは, Modifiable なデータと, Non-modifiable なデータに分類される (Bahr and Holme, 2003; Joyce and Lewindon, 2016). Modifiable なデータとは, BMI 指数, 筋力, 柔軟性といった, トレーニングや生活習慣で修正可能なデータであり, Non-modifiable なデータとは, 年齢, 性別, 過去の外傷履歴といった, 修正不可能なデータを意味する (Hägglund et al., 2006). 本稿で紹介した外傷予測は, Non-modifiable なデータである外傷履歴と AE のみを使用しているため, 外傷の発生確率を低減させるような外傷予防をするために, 対象者が何をすべきかは明らかにできない. 今後はこの問題を解決するために, Modifiable なデータを共変量として考慮できる外傷

予測モデルの開発が必要である (Therneau and Grambsch, 2000; Meira-Machado et al., 2009)。例えば、力士に関する Modifiable なデータとして、小川 他 (1973) が調査した力士の筋力や運動能力が共変量として利用可能である。また、対象者の心理的要因が外傷のリスクと関係すること (Bahr and Krosshaug, 2005; Kimotsuki et al., 2024) を踏まえて、モデルに心理的要因を導入することも今後の研究に期待される。

5. おわりに

本稿では、対象者の外傷履歴を用いた統計的なスポーツ外傷予測手法について詳説した。外傷予測では、繰り返し発生するスポーツ外傷をイベントヒストリーデータと見なして、信頼性解析・生存時間解析の観点から、スポーツ外傷の発生を点過程でモデル化した。モデルでは自己励起過程によって外傷発生の連鎖性や、時間に対する発生率の単調増加性が考慮された。そして、大相撲力士を例に、自己励起過程モデルによる外傷予測の妥当性を示した。その結果、外傷発生の連鎖性を考慮できない定常ポアソン過程モデルよりも、自己励起過程モデルの方が、より正確に外傷予測できることが確認できた。

今後の課題としては、4.5 節で述べたように、他のスポーツ選手の外傷予測への自己励起過程モデルの適用や、スポーツ外傷予防に効果的な変数を予測モデルに追加した新しいモデルの開発が挙げられる。

謝 辞

本稿の作成にあたって、編集委員および匿名の査読者 2 名から内容を改善する貴重なコメントを多く頂きました。ここに記して感謝いたします。本研究の一部は、JSPS 科研費(課題番号 19K04892 および 21K14373) および神奈川大学工学研究所共同研究(A)の助成を受けたことを付記します。

参 考 文 献

- Abdel-Ghaly, A. A., Chan, P. Y. and Littlewood, B. (1986). Evaluation of competing software reliability predictions, *IEEE Transactions on Software Engineering*, **SE-12**(9), 950–967.
- Bahr, R. and Holme, I. (2003). Risk factors for sports injuries—A methodological approach, *British Journal of Sports Medicine*, **37**(5), 384–392.
- Bahr, R. and Krosshaug, T. (2005). Understanding injury mechanisms: A key component of preventing injuries in sport, *British Journal of Sports Medicine*, **39**(6), 324–329.
- Blanche, P., Dartigues, J. F. and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks, *Statistics in Medicine*, **32**(30), 5381–5397.
- Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions, *Statistica Sinica*, **7**(4), 875–892.
- Chiodo, E. and Lauria, D. (2015). Some basic properties of the failure rate of redundant reliability systems in industrial electronics applications, *IEEE Transactions on Industrial Electronics*, **62**(8), 5055–5062.
- Dohi, T., Zheng, J. and Okamura, H. (2020). Data-driven software reliability evaluation under incomplete knowledge on fault count distribution, *Quality Engineering*, **32**(3), 421–433.

- Emura, T. and Chen, Y. H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*, Springer, Singapore.
- Fiscutean, A. (2021). Data scientists are predicting sports injuries with an algorithm, *Nature*, **592**(7852), 10–11.
- Gerds, T. A. and Kattan, M. W. (2021). *Medical Risk Prediction Models: With Ties to Machine Learning*, Chapman and Hall/CRC, Boca Raton.
- Hägglund, M., Waldén, M. and Ekstrand, J. (2006). Previous injury as a risk factor for injury in elite football: A prospective study over two consecutive seasons, *British Journal of Sports Medicine*, **40**(9), 767–772.
- 浜田宏 (2022). ベイズで広がる数理社会学の世界, 理論と方法, **37**(1), 124–137.
- Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes, *Biometrika*, **58**(1), 83–90.
- Hawkes, A. G. (2018). Hawkes processes and their applications to finance: A review, *Quantitative Finance*, **18**(2), 193–198.
- Inoue, S. and Yamada, S. (2015). Software reliability assessment with multiple changes of testing-environment, *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, **E98-A**(10), 2031–2041.
- 伊藤勝治 (2019). 『相撲見物』, 青幻舎, 京都.
- Joyce, D. and Lewindon, D. (eds.) (2016). *Sports Injury Prevention and Rehabilitation*, Routledge, New York.
- 金指基 (2002). 『相撲大事典』, 現代書館, 東京.
- Kimotsuki, Y., Ogasawara, I., Iwasaki, S., Nagai, K., Nishida, K., Kanzaki, N., Hoshino, Y., Matsushita, T., Kuroda, R., Uchida, R., Take, Y. and Nakata, K. (2024). Goal orientation and desire for approval were associated with sports injuries among young Japanese athletes, *AJPM Focus*, **3**(4), <https://doi.org/10.1016/j.focus.2024.100236>.
- Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.
- 真壁肇, 鈴木和幸, 益田昭彦 (2002). 『品質保証のための信頼性入門』, 日科技連, 東京.
- Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data, *Statistical Methods in Medical Research*, **18**(2), 195–222.
- 大垣亮, 陣内峻, 八田倫子 (2022). スポーツ外傷・障害・疾病における疫学データの活用方法, 日本アスレティックトレーニング学会誌, **8**(1), 33–41.
- 小笠原一生 (2021). 機械学習によるスポーツ外傷予測, 予防の可能性, 体育の科学, **71**(11), 778–784.
- Ogata, Y. (1981). On Lewis' simulation method for point processes, *IEEE Transactions on Information Theory*, **27**(1), 23–31.
- Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *Journal of the American Statistical Association*, **83**(401), 9–27.
- 小川新吉, 古田善伯, 山本恵三, 永井信雄 (1973). 相撲力士の体力科学的研究 (その2) (関取の体力と発達), 体力科学, **22**(2), 45–55.
- Ota, S. and Kimura, M. (2017). A statistical dependent failure detection method for n-component parallel systems, *Reliability Engineering & System Safety*, **167**, 376–382.
- Ota, S. and Kimura, M. (2023). Statistical injury prediction for professional sumo wrestlers: Modeling and perspectives, *PLoS ONE*, **18**(3), <https://doi.org/10.1371/journal.pone.0283242>.
- Parekh, N., Hodges, S. D., Pollock, A. M. and Kirkwood, G. (2012). Communicating the risk of injury in schoolboy rugby: Using Poisson probability as an alternative presentation of the epidemiology, *British Journal of Sports Medicine*, **46**(8), 611–613.
- Pfirrmann, D., Herbst, M., Ingelfinger, P., Simon, P. and Tug, S. (2016). Analysis of injury incidences in

- male professional adult and elite youth soccer players: A systematic review, *Journal of Athletic Training*, **51**(5), 410–424.
- Rizou, M. A., Lee, Y., Mishra, S. and Xie, L. (2017). Hawkes processes for events in social media, *Frontiers of Multimedia Research* (ed. S. F. Chang), 191–218, Association for Computing Machinery and Morgan & Claypool, San Rafael.
- Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., Lenoir, M., D'Hondt, E. and Witvrouw, E. (2020). A machine learning approach to assess injury risk in elite youth football players, *Medicine & Science in Sports & Exercise*, **52**(8), 1745–1751.
- Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J. and Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning, *PLoS ONE*, **13**(7), <https://doi.org/10.1371/journal.pone.0201264>.
- Rossi, A., Pappalardo, L. and Cintia, P. (2021). A narrative review for a machine learning application in sports: an example based on injury forecasting in soccer, *Sports*, **10**(1), <https://doi.org/10.3390/sports10010005>.
- 齊藤和快, 安斎健太郎, 岡林務, 今城栄祐, 五十子圭佑, 竹内真太, 西田裕介 (2020). Jリーグ・ディビジョン2に所属するチームにおける2年間の傷害調査, 理学療法科学, **35**(1), 33–39.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, Toronto.
- Shimizu, S., Nagase, T., Tateishi, T., Nakagawa, T. and Tsuchiya, M. (2020). Second anterior cruciate ligament injuries after anterior cruciate ligament reconstruction in professional sumo wrestlers: A case series, *Orthopaedic Journal of Sports Medicine*, **8**(2), 1–5.
- 清水禎則, 長瀬寅, 立石智彦, 佐藤哲也, 中川照彦, 土屋正光 (2021). 平成時代の大相撲力士の傷害統計, 日本整形外科スポーツ医学会雑誌, **41**(3), 201–208.
- Shooman, M. L. (2003). *Reliability of Computer Systems and Networks: Fault Tolerance, Analysis, and Design*, John Wiley & Sons, New York.
- Shrier, I., Steele, R. J., Hanley, J. and Rich, B. (2009). Analyses of injury count data: Some do's and don'ts, *American Journal of Epidemiology*, **170**(10), 1307–1315.
- 相撲レファレンス (2024). 力士, <https://sumodb.sumogames.de/Rikishi.aspx?l=j> (最終アクセス日 2025年3月26日).
- 砂川憲彦 (2015). 『スポーツ現場の傷害調査』, ブックハウス・エイチディ, 東京.
- 砂川憲彦, 真鍋知宏, 半谷美夏, 細川由梨, 奥脇透, 広瀬統一, 増島篤 (2022). スポーツ外傷・障害および疾病調査に関する提言書: 日本臨床スポーツ医学会・日本アスレティックトレーニング学会共同声明, 日本アスレティックトレーニング学会誌, **7**(2), 155–171.
- Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.
- Trivedi, K. S. (2016). *Probability & Statistics with Reliability, Queuing and Computer Science Applications*, 2nd ed., John Wiley & Sons, Hoboken.
- Wu, S. (2019). A failure process model with the exponential smoothing of intensity functions, *European Journal of Operational Research*, **275**(2), 502–513.
- 山中卓, 中川秀敏, 杉原正顯 (2017). Hawkes過程による信用リスク伝播のモデリングとその応用, 応用数理, **27**(1), 5–12.

Application of Reliability Analysis to Sports Injury Prediction: A Case Study for Sumo Wrestlers

Shuhei Ota¹ and Mitsuhiro Kimura²

¹Faculty of Engineering, Kanagawa University

²Faculty of Science and Engineering, Hosei University

Sports injury prediction is a way to evaluate the risks of injuries that may occur in athletes in the future. In recent years, based on event history data, several methods have been proposed to assess and predict injury risk in athletes over time. In this paper, we explain stochastic models using point processes to model sports injury occurrences and how to apply the model to injury prediction. Besides, via a case study of injury prediction for professional sumo wrestlers, a self-exciting point process that can consider sequential injury occurrences is useful for precise injury prediction. Furthermore, to reinforce the validity of the model, we newly investigate the predictive accuracy of the model using data that is not used to estimate the parameters of the predictive model, which was not revealed in previous studies on predicting injuries to sumo wrestlers. Finally, we discuss some future works of injury prediction from the viewpoint of its application to injury prevention.

生存解析における樹木法と最近の発展

杉本 知之¹・丸尾 和司²・下川 敏雄³

(受付 2024 年 12 月 2 日；改訂 2025 年 5 月 6 日；採択 5 月 14 日)

要 旨

決定木やランダムフォレストなどの樹木法は、データサイエンス分野をはじめとして、他の実質科学領域やビジネスの実務で利用されている統計・機械学習ツールである。過去 10 年において、樹木法やフォレストからの変数重要度指標の推測、フォレストの一致性和漸近正規性に基づいた因果ツリーや因果フォレストなどの因果推論ツールとしての方法論や理論の発展も目覚ましい。本稿では、このような樹木法における方法や理論の最近の発展を、生存時間解析での利用の観点で各手法とその特徴を調査し総合的に報告する。

キーワード：CART, Cox 回帰, 生存木, ランダムフォレスト, 変数重要度, 因果フォレスト。

1. はじめに

関心のあるイベントに対するリスク要因分析、将来起こるイベントからのリスク予測のため、生存解析に基づく様々な回帰モデリング手法が適用できる。統計的性質、解釈性、計算可能性といった多くの利便性の観点から、対数ハザードの観点で線形な Cox 回帰モデリングをデフォルトの選択肢として使用することも多い。樹木法は、より柔軟な仮定のもとで、高次の交互作用や変数の非線形要因に自動的に対応する分析手法として人気が高く、とくに、解釈性の高い決定木(CART) (Breiman et al., 1984; 下川 他, 2013), 予測性能の高いランダムフォレスト (Breiman, 2001) は、科学研究での利用だけでなく、ビジネスにおける実務でも広く応用されるようになっている。そして、樹木法の構成方法の対案 (例えば, Loh, 2002; Hothorn et al., 2006b), 経時データや生存データへの方法の拡張 (例えば, Segal, 1992; LeBlanc and Crowley, 1992; Arano et al., 2010; 江村, 2023) も数多く提案してきた。生存解析におけるランダムフォレスト法 (Ishwaran et al., 2008) は、高性能なリスク予測モデルを構築する Cox 回帰の代替手法を提供する (例えば, Morvan et al., 2020; Liao et al., 2024 など)。また、このような生存フォレスト法に基づく変数重要度は、予後因子の同定のため、Cox 回帰のような全所的線形モデリングの代替もしくは補完として有用であることも多く報告されるようになっている (例えば, Hsich et al., 2011; He et al., 2020; Liu et al., 2021 など)。

樹木法、ランダムフォレスト法は広く利用されているにもかかわらず、複数の要素を巧妙に組み合わせたブラックボックスの性質をもっているため、数理統計的な性質の解明の難度は高い。ランダムフォレストの一貫性の早期の研究では、Breiman (2004) はランダムフォレストの

¹ 大阪大学 基礎工学研究科：〒560-8531 大阪府豊中市待兼山町 1-3; sugimoto.tomoyuki.es@osaka-u.ac.jp

² 筑波大学 医学医療系：〒305-8575 茨城県つくば市天王台 1-1-1; kazushi.maruo@gmail.com

³ 和歌山県立医科大学 医学部：〒641-8509 和歌山市紀三井寺 811-1; toshibow2000@gmail.com

簡略化バージョンに対して技術ノートを与えている。その後, Scornet et al. (2015)は, ブートストラップが非復元抽出サンプリングにおき換えられている点を除き, 元のランダムフォレストの現実的な設定のもとで, 共変量次元を固定し, 2乗期待値の意味での一致性を与え, さらに, Wager (2016) [Theorem 4.3] は, 共変量次元を固定しない高次元の状況において, ランダムフォレストの各点一致性を示している。Wager and Athey (2018)は, 正直性(Honesty)などの概念とダブルサンプリング手法を導入して, ランダムフォレストによる応答関数の各点推定において漸近正規性を確立することに成功し, フォレスト法に基づく条件付きの処理効果の(因果)推定法の因果フォレストを提案している(中村, 2020)。応答関数の各点推定値に関する分散の一致推定値を, ランダムフォレストなどのバギング学習器に対する分散推定の無限小ジャックナイフ法(Wager et al., 2014; Efron, 2014)を適合させて計算可能としている点も実用的なツールとして著しく有用である。因果フォレスト法(Wager and Athey, 2018)の着想は, ランダムフォレスト法の推定方程式による定式化として一般化され(Athey et al., 2019), その生存解析への応用として Cui et al. (2023) は右側打ち切りデータに利用可能な理論とツールへと発展させていている。また, Athey and Wager (2019)では, クラスター化された連続観測値に対して, 因果フォレストを用いてノンパラメトリックなランダム効果あてはめの変化形を与えている。本論文では, このような樹木法の発展を体系的に概説する。

2. 生存樹木法とその例示

2.1 生存データ

生存解析における基本的データ集合の設定を与える。個人 i の真のイベント時間 T_i^* , 真の右側中途打ち切り時間 C_i に対して, 観測イベント時間は $T_i = \min(T_i^*, C_i)$ である。 $\Delta_i = \mathbb{1}(T_i^* \leq C_i)$ (1: 完全データ, 0: 打ち切りデータ) は右側中途打ち切り指標であり, 個人 i の共変量は p 個の共変量のベクトル $Z_i = (Z_{i1}, \dots, Z_{ip})$ で構成される。また, 観測機構は独立打ち切りの仮定を満たすとする, すなわち, T_i^* と C_i は, 共変量 Z_i を所与として条件付きに独立であると仮定する。データは独立した N 人の被験者 $\mathcal{L}_N = \{(T_i, \Delta_i, Z_i), i = 1, \dots, N\}$ に対して利用可能である。ここでは, 一貫して, 共変量 Z_i は, 時点 0 で測定され, 時間変化しないとする。

例示として, R パッケージ `randomForestSRC` 内のデータ `peakV02` を用いる。これは心肺運動負荷試験を受けた 2231 名の収縮期心不全患者の予後におけるリスク因子を同定するための研究(Hsich et al., 2011)からのデータ集合であり, 我々の例示では, 全データのうち男性 $N = 1629$ 名を用いる。このとき, 死亡までの観測時間(T)と打ち切り指標(Δ)に関する共変量(Z)として, 糖尿病や冠動脈疾患などの 8 項目の過去の病歴・治療歴, β 遮断薬やカルシウム拮抗薬など 15 項目の薬剤使用状況, トレッドミル運動時間や最高酸素摂取量などの 6 つの心機能指数, BUN などの 5 つの血液検査値, および年齢, 人種, BMI, 喫煙からなる $p = 38$ 個の変数を含む。図 1 は, Cox 回帰を適用し, 情報量規準 AIC によって変数選択されたハザード比に関する結果であり, 変数の順番はハザード比に関する統計的有意性の順に並び替えていく。Cox 回帰では, 予後因子の同定とその大きさの評価をハザード比の観点で解釈できる点で優れている。また Cox 回帰のあてはめから, 有意義な予後予測モデルの作成を行うこともできる。この予測性能については, 2.3 節においてランダムフォレストなどの樹木法と比較する。

2.2 生存木

生存木の開発は, CART 法(Breiman et al., 1984)や自動交互作用検出(Auto Interaction Detection; AID)法(Morgan and Sonquist, 1963)などの既存の樹木構造接近法を生存データに拡張することを目的として展開された(例えば, Gordon and Olshen, 1985; Ciampi et al., 1986;

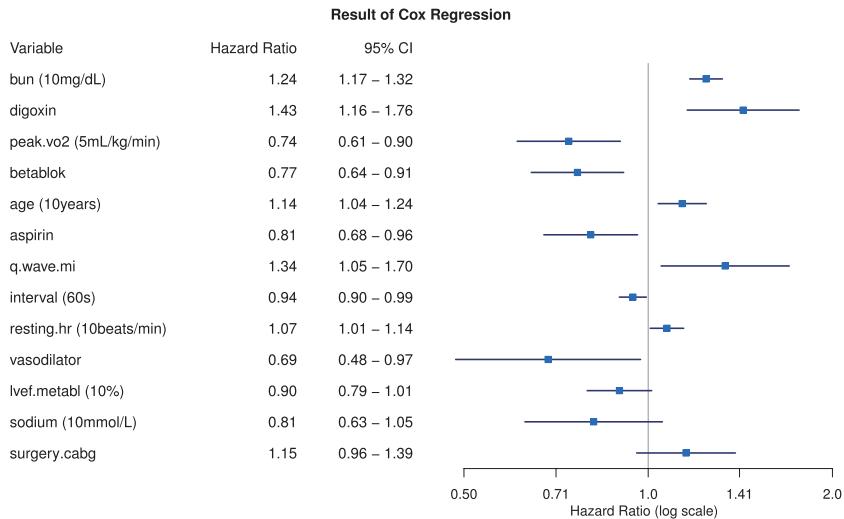


図 1. データ peakV02(男性)に対する AIC に基づいて選択された 13 個の共変量の Cox 回帰の結果。

```
tree<-rpart(Surv(ttodead, died)~., data=peakV02m, cp=0)
tree0<- prune(tree, cp=tree$cptable[which.min(tree$cptable[, "xerror"]),"CP"])
plot(as.party(tree0))
```

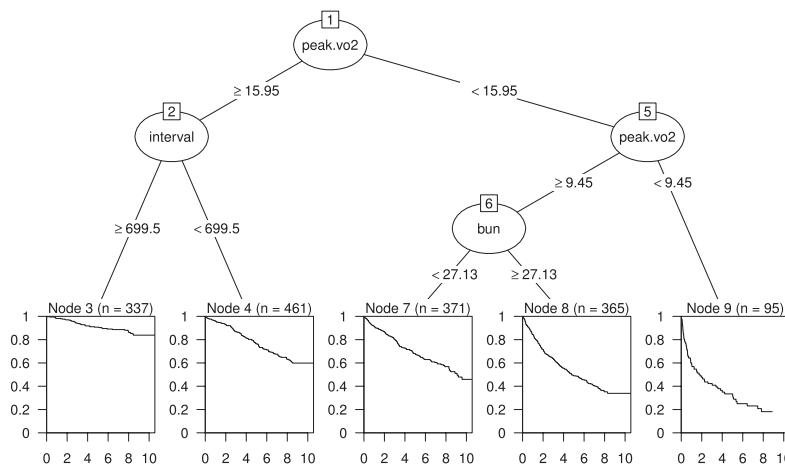


図 2. データ peakV02(男性)に対する最小交差検証生存木。

Segal, 1988; Davis and Anderson, 1989; LeBlanc and Crowley, 1992). 基本的な生存木の方法論が確立された後は、生存木のアンサンブル拡張を含め、多変量生存データ、相関のある生存データ、時間依存共変量、離散尺度イベント時間変数などを扱う問題などの研究も進められた。生存木に関するレビューは LeBlanc and Crowley (1995), Bou-Hamad et al. (2011), Zhou and McArdle (2015) などが詳しく、Zhou and McArdle (2015)の図 2 は、公表された生存木の分岐ルール、刈り込み、実装の分類を明確に与えている。

生存木を含む樹木構造接近法では、トレーニング・データとして、全データ集合 \mathcal{L}_N のすべ

て、もしくは \mathcal{L}_N から復元抽出もしくは非復元抽出した部分集合を用いて木を構築する。図 2 は、データ peakV02(男性)を用いて、R の rpart 関数により計算された生存木を partykit を通して描画したものである。 \mathcal{T} を構築された生存木とする。図 2 の生存木 \mathcal{T} は 9 個のふし (node) h_ℓ , $\ell = 1, \dots, 9$ により構成されている。各ふしは対応するルールに該当する個人ラベル i の集合を表すものである。とくに、最も上位にあるふし h_1 はルート (root) と呼ばれる。終端ふし (terminal node) の集まり $\tilde{\mathcal{T}} = \{h_3, h_4, h_6, h_8, h_9\}$ は、葉 (leaf) とも呼ばれ、通常、終端ふし内のデータを用いて(生存)予測子を構築する。図 2 の終端ふしの要約には、Kaplan-Meier 推定値が描かれており、終端ふし 3 から 9 にかけて予後が悪くなる様子が見て取れる。最高酸素摂取量 (peak.vo2) は Cox 回帰の結果と同様に予後に重要な因子となったが、トレッドミル運動時間 (interval) や尿素窒素値 (bun) との組み合わせ (交互作用) に基づいたより解釈しやすい予後予測が得られている。

R の rpart 関数で作成される現在の生存木は、CART 法で用いる残差を生存時間データに適応させるためデビアンス残差を利用した方法 (LeBlanc and Crowley, 1992) の指數分布版が採用されている。 h_L と h_R を、ふし h から生成される娘ふしとする。説明変数 Z_{ij} が数値型であれば、分岐点 c_j を用いて、 $\{i \in h_L : Z_{ij} < c_j\}$ と $\{i \in h_R : Z_{ij} \geq c_j\}$ の 2 分割を行う。カテゴリ型であれば、組み合わせに基づく 2 分割となる。 $R(h)$ をふし内残差とする。親ふしから娘ふし h_L, h_R への 2 分岐ルールの選択は、ふし h 内の不純度指標の $R(h)$ を最も減少させる h_L, h_R のペア、すなわち、減少量 $\Delta R_h(s_h) = R(h) - \{R(h_L) + R(h_R)\}$ を最も大きくする 2 分岐ルール $s_h \in \mathcal{S}_h$ が選択される。ここに、 \mathcal{S}_h は、各ふし h すべての説明変数の対応するすべての 2 分岐ルールの候補集合である。適切な樹木の大きさを決定するために、CART 法 (Breiman et al., 1984) では、複雑度コスト (cost-complexity) 削り込み法と交差検証法を提案している。一方、生存木では、ログランク検定統計量に基づく分岐ルール (Segal, 1988) のニーズも高い。LeBlanc and Crowley (1993) は、2 分岐ルールに検定統計量の最大化を採用する場合に、複雑度分岐 (split-complexity) 法を提案している。複雑度パラメータの統計的解釈は、前者の場合は予測指標の罰則項、後者の場合は有意性指標の自由度調整項に該当する。

決定木はこのような形で作成されるが、データの見方についての一つの提案であって、実地では、生存木を含め、決定木の公式的な作り方に忠実に拘る必要はない。例えば、分岐候補値は、解釈のしやすいカテゴリ分けにしてもよいし、分岐基準もデータの特性から関心のある統計的指標に変更してよい。また、決定木は、例えば、観測データの 90% で学習させた場合に得られる木構造が大きく異なるなど、データのばらつきにより異なる木が生成される。安定した生存木を利用したい場合には、例えば 90% のデータでランダムサンプリングを行い推定された木構造の安定性を確かめることが推奨される。決定木は人が解釈のしやすい思考構造を提供するため、例えば、ブートストラップ標本を用いて類似する性能をもつ生存木を複数作り出して、分析者の仮説に合致する木を探し出すことも考えられる。

2.3 生存フォレスト

ランダムフォレスト (RF) (Breiman, 2001) は、大きく成長させて過剰あてはめを引き起こした樹木予測子を平均化することで過学習を回避するという他の機械学習法 (多くは縮小により過学習を回避する) の中では見られない独自性をもち、高い予測性能をもつノンパラメトリックな機械学習手法を提供するにも関わらず、並列計算も可能という点が顕著な特徴である。そのような RF を生存データに適応させる研究が Breiman (2003), Hothorn et al. (2006a), Ishwaran et al. (2008) などにより進められた。生存フォレストの構成方法は標準的な RF と同様であるが、生存解析に特有の分岐方法や予測指標が必要になる。生存フォレストの一般的な計算方法は以下の通りである：

ステップ 1. B 個のブートストラップ標本 $\mathcal{L}_N^{\text{IB}(b)}$ $b = 1, \dots, B$ を抽出する。

ステップ 2. 各ブートストラップ標本 $\mathcal{L}_N^{\text{IB}(b)}$ に基づいて生存木を成長させる(刈り込み無し):

(i) 各生存木のふしにおいて、ある説明変数(予測変数、特徴量)の集まり(単一、線形結合)を選択する。

(ii) (i)で選択した説明変数によって分割されるすべての 2 群の中から、右側打ち切りデータに適した分割基準(例えば、ログランク検定)に基づいて、最良の二分割(娘ふしへの分割)を見つける。

(iii) 各娘ノードに対して、(i) - (iii) の手順を停止基準に達するまで再帰的に繰り返す。

ステップ 3. B 個の生存木の終端ふしからの生存情報を集約し、生存予測モデルのアンサンブルを得る。

生存フォレストによるリスク予測方法を記述するため、 $\mathcal{T}^{(b)}$ を b 番目のブートストラップ標本に対して生成された生存木とし、共変量 z をもつ個体が位置する終端ふしを $\mathcal{T}^{(b)}(z)$ とする。ブートストラップ標本 $\mathcal{L}_N^{\text{IB}(b)}$ 内には、元のデータ集合 \mathcal{L}_N の個体が複数含まれ得る。個体 i が b 番目のブートストラップ標本 $\mathcal{L}_N^{\text{IB}(b)}$ に出現する回数を $\aleph_i^{(b)}$ とする。 $\mathcal{L}_N^{\text{IB}(b)}$ に個体 i が含まれていない場合、 $\aleph_i^{(b)} = 0$ である。共変量 z に対応する生存木の終端ふし $\mathcal{T}^{(b)}(z)$ において、 $\bar{\mathcal{N}}^{(b)}(t|z)$ を時刻 t までの打ち切りされていないイベント数、 $\bar{\mathcal{Y}}^{(b)}(t|z)$ を時刻 t におけるアトリスク数とすると、これらは、計数過程の記法 (Fleming and Harrington, 1991) を用いて

$$\bar{\mathcal{N}}^{(b)}(t|z) = \sum_{i=1}^N \aleph_i^{(b)} \mathbb{1}(\mathcal{Z}_i \in \mathcal{T}^{(b)}(z)) \mathcal{N}_i(t), \quad \bar{\mathcal{Y}}^{(b)}(t|z) = \sum_{i=1}^N \aleph_i^{(b)} \mathbb{1}(\mathcal{Z}_i \in \mathcal{T}^{(b)}(z)) \mathcal{Y}_i(t).$$

と書ける。ここに、 $\mathcal{N}_i(t) = \mathbb{1}(T_i \leq t, \Delta_i = 1)$ 、 $\mathcal{Y}_i(t) = \mathbb{1}(T_i > t)$ である。Ishwaran et al. (2008) の生存フォレストでは、各生存木に基づく Nelson-Aalen 推定量 $\hat{H}^{(b)}(t|z) = \int_0^t d\bar{\mathcal{N}}^{(b)}(u|z) / \bar{\mathcal{Y}}^{(b)}(u|z)$ を集約することで予測子を構築する。すなわち、ブートストラップ標本のインパック (in-bag) データを用いて、各木の終端ふしに対する条件付き累積ハザード関数を $\hat{H}^{(b)}(t|z)$ により推定することで、生存フォレストの予測生存関数を

$$(2.1) \quad \hat{S}^{\text{RSF}}(t|z) = \exp \left(-\frac{1}{B} \sum_{b=1}^B \hat{H}^{(b)}(t|z) \right)$$

として構成する。一方、Hothorn et al. (2004) では、予測生存関数の構成に

$$(2.2) \quad \hat{S}^{\text{CSF}}(t|z) = \prod_{u \leq t} \left(1 - \frac{\sum_{b=1}^B d\bar{\mathcal{N}}^{(b)}(u|z)}{\sum_{b=1}^B \bar{\mathcal{Y}}^{(b)}(u|z)} \right) \approx \exp \left(- \int_0^t \frac{\sum_{b=1}^B d\bar{\mathcal{N}}^{(b)}(u|z)}{\sum_{b=1}^B \bar{\mathcal{Y}}^{(b)}(u|z)} \right)$$

を用いることを提案している。予測生存関数の(2.1) と (2.2) の違いに関して、Mogensen et al. (2012) は関係式

$$\frac{\sum_{b=1}^B d\bar{\mathcal{N}}^{(b)}(u|z)}{\sum_{b=1}^B \bar{\mathcal{Y}}^{(b)}(u|z)} = \frac{1}{B} \sum_{b=1}^B \left[\frac{\bar{\mathcal{Y}}^{(b)}(u|z)}{\frac{1}{B} \sum_{b=1}^B \bar{\mathcal{Y}}^{(b)}(u|z)} \right] \frac{d\bar{\mathcal{N}}^{(b)}(u|z)}{\bar{\mathcal{Y}}^{(b)}(u|z)}$$

から、累積ハザード関数の集約において、(2.2) ではリスク数の多い終端ふしほどより多くの重みを割り当てる一方、 $\hat{S}^{\text{RSF}}(t|z)$ はすべての終端ふしに等しい重みを割り付ける違いがあることを指摘している。

良い性能をもつ RF を構築するための本質は、樹木あてはめの残差の相関が樹木間で小さくなるようにランダム量を上手く注入することである。そのような調整 (turning) パラメータとして代表的なものは、各分岐でランダムに選択される候補変数の数 (`mtry`) と最小ふしサイズ (`nodesize`) である (例えば、Segal, 2004; Probst et al., 2019)。変数の線形結合を分岐に用いる

RF(Forest-RC) (Breiman, 2001) は計算上のコストから R, Python では実装されていないが, 調整パラメータの選択肢が増え, 予測性能のさらなる向上が見込める (杉本 他, 2005). 各調整パラメータの目安となる(デフォルト)値はあるが, 実際には, データの特徴によって最適な調整パラメータは異なるため, 計算コストに配慮しつつ, 調整パラメータのいくつかの組み合わせを事前に試行した上で, 良い予測性能を与えるものを選択することが望ましい. このことは 3.2 節以降で紹介する因果フォレストにおいても同様である.

予測性能 あてはめたモデルの予測性能の評価では, 生存木やフォレストに尤度基準の導入が難しいことから, Harrell et al. (1982) の C-指標 (Heagerty and Zheng, 2005) や Brier スコアが用いられる. また, 生存予測では C-指標が 2 分類予測の ROC 曲線 AUC ほど理解しやすい指標ではなく, 期待 Brier スコア $BS^*(t; \hat{S}) = E[\{1(T_i^* > t) - \hat{S}_i(t)\}^2]$ には, 個人生存関数の推定値が真の生存関数と等しい場合に最小になるという絶対的基準があるため, 生存時間データの予測性能評価では, Brier スコアが好まれることも多い. ここに, $\hat{S}_i(\cdot)$ は, 個人 i の生存関数の推定値である. 右側中途打ち切りの生存データにおける Brier スコアの推定値として

$$(2.3) \quad \widehat{BS}(t) = \frac{1}{N} \sum_{i=1}^N [\{1 - \hat{S}_i(t)\}^2 \mathbb{1}(T_i > t) \hat{S}_i^C(t)^{-1} + \hat{S}_i(t)^2 \mathbb{1}(T_i \leq t, \Delta_i = 1) \hat{S}_i^C(T_i)^{-1}]$$

が提案されている (Graf et al., 1999). ここに, $\hat{S}_i^C(t)$ は個人 i の右側中途打切り時間の生存関数の推定値であり, Kaplan-Meier 推定法を用いることが基本的であるが, Cox 回帰など共変量を用いたモデル推定も有用で, 性能を向上させることができることが報告されている. 期待 Brier スコアは小さいほど良いという性質をもつが, 実際には(2.3) のように推定する必要ある. Brier スコアによる予測性能の評価は, 交差検証法などを用いてモデル構築と期待値に関する推定のために起こるバイアスを防ぐことが望ましい. 全体的な時間の総合指標として, $\widehat{BS}(t)$ を対象となる時間の上(ここでは $(0, T_{\max}]$ とする)で合計した累積 Brier スコア

$$\widehat{IBS} = \frac{1}{T_{\max}} \int_0^{T_{\max}} \widehat{BS}(t) dt \approx \frac{1}{T_{\max}} \sum_m \widehat{BS}(t_m) (t_m - t_{m-1})$$

が利用できる. 図 3 は, 二つの Cox 回帰(Cox.full, Cox.AIC), 生存木(SCART), Ishwaran et al.

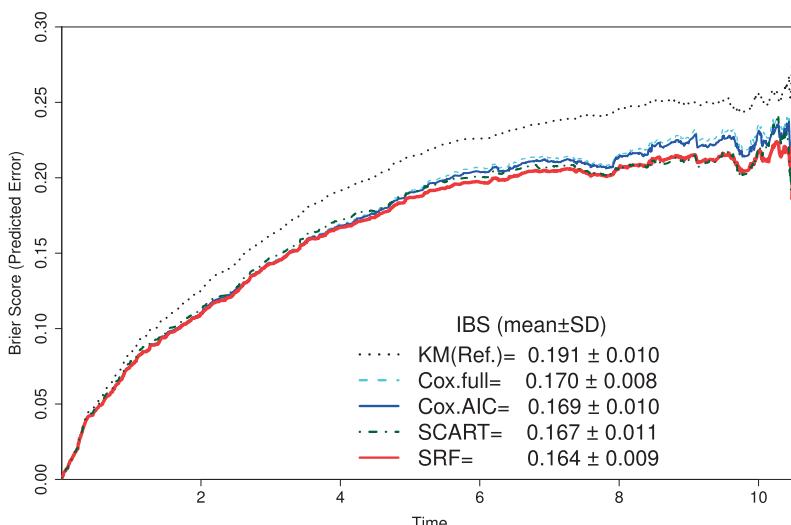


図 3. データ peakV02(男性)における Cox 回帰, 生存木(SCART), 生存フォレスト(RSF)の Brier スコアの比較.

(2008) の生存フォレスト (RSF) (ライブラリ `randomForestSRC` の `rfsrc` 関数) をデータ `peakV02` (男性) に適用した結果を示す。予測性能は、20 回の内的妥当性検証 (ホールドアウト法: 訓練 80%, 検証 20%) を用いた Brier スコアの平均で要約している。このデータでは、AIC 選択の Cox 回帰 (`Cox.AIC`) よりも生存フォレストの方が僅かではあるが累積 Brier スコアの性能はよく、とくに RSF と SCART は時間的に後半の予測でより優位である。生存木は、一般に、AIC 選択 Cox 回帰よりも累積 Brier スコアの性能が落ちることが多い (今回のデータの傾向とは異なる)。なお、Brier スコア算出のライブラリ `pec` は現時点で `rpart` 関数からの生存木の結果に対応していない。実装において、生存木の予測生存関数に、終端ふしの部分集団の Kaplan-Meier 推定値を用いると標本サイズが小さくなり Brier スコアが不安定となり得るため、図 3 では、終端ふしをダミー変数とする Cox 回帰の Breslow 推定値に基づく推定生存関数を用いて予測 Brier スコアを安定化させている。なお、生存フォレストの調整パラメータは、ここでは、予測性能と計算時間の観点で、`nodesize` は 15 (`rfsrc` 関数のデフォルト値)、`mtry` は 5 (デフォルト値は $\lceil \sqrt{p} \rceil = 7$) に設定した。調整パラメータの設定、 N が小さいときの挙動に関するさらなる議論は、Web Supplementary Materials A に与える。

3. 変数効果の推測

本節では、生存フォレストによる変数重要度指標、因果生存フォレストによる変数効果の統計的推測の最近の発展を通して、フォレストからの変数効果の推測方法を概説する。

3.1 変数重要度指標

樹木法では、各手法に応じて変数重要度 (VIMP) の推定ツールが提案されている (杉本 他, 2005)。手法間で多少の違いはあるものの、概念的には、変数の並び替えやマスキングを活用することにあり、それらの操作のもとで、各変数に起因する予測誤差を評価し、各変数の重要度を評価する。フォレスト法では、重要度の評価において計算コストが高い交差検証法を避けることができる (Breiman, 1996, 2001)。すなわち、オリジナルな CART 法とは異なり、ブートストラップ標本 $\mathcal{L}_N^{\text{IB}^{(b)}} := \mathcal{L}_N^{\text{IB}}(\Theta^{(b)})$ を訓練データに用いて各樹木 $\mathcal{T}^{(b)}$ を構成するため、訓練データであるブートストラップ標本で用いられなかった、いわゆるアウトオブバック (OOB) データ $\mathcal{L}_N^{\text{OOB}}(\Theta^{(b)}) := \mathcal{L}_N \setminus \mathcal{L}_N^{\text{IB}}(\Theta^{(b)})$ を評価データに用いて予測誤差を推定することができる。ここに、 $\Theta^{(b)}$ は木を成長させることと OOB データの並び替えに関するランダム性を含む要因とする。並び替え (Breiman-Cutler) 重要度の計算では、OOB データの j 番目の共変量のすべての値 $Z_{ij}, i \in \mathcal{L}_N^{\text{OOB}}(\Theta^{(b)})$ をランダムに並び替える。このとき、元の Z_{ij} から並び替えられた共変量値を \tilde{Z}_{ij} と書くと、これらの新たなデータの共変量

$$\tilde{Z}_i^{(j)} = (Z_{i1}, \dots, Z_{ij-1}, \tilde{Z}_{ij}, Z_{ij+1}, \dots, Z_{ip}), i \in \mathcal{L}_N^{\text{OOB}}(\Theta^{(b)})$$

を木にあてはめて誤差率を計算する。この新しい誤差率が元の OOB 誤差率を上回る量が、木における j 番目の変数の重要度

$$\text{Imp}(Z^{(j)}, \Theta^{(b)}, \mathcal{L}_N) = \frac{\sum_{i \in \mathcal{L}_N^{\text{IB}}(\Theta^{(b)})} \{\ell(S_i, \hat{S}_i(\tilde{Z}_i^{(j)}, \Theta^{(b)}, \mathcal{L}_N)) - \ell(S_i, \hat{S}_i(Z_i, \Theta^{(b)}, \mathcal{L}_N))\}}{|\mathcal{L}_N^{\text{IB}}(\Theta^{(b)})|}$$

として定義して、フォレスト全体の木に対する平均

$$\text{Imp}(Z_j, \Theta^{(1)}, \dots, \Theta^{(B)}, \mathcal{L}_N) = \frac{1}{B} \sum_{b=1}^B \text{Imp}(Z^{(j)}, \Theta^{(b)}, \mathcal{L}_N)$$

を取ることで VIMP を算出する。ここに、 S_i と \hat{S}_i は、それぞれ、個人 i の生存関数の真値と

推定値, $\ell(S_i, \hat{S}_i(\mathbf{Z}_i, \Theta^{(b)}, \mathcal{L}_N))$ は生存木に関する損失関数であり, 典型的なものとして Brier スコアもしくは C-指標を想定する. 無限に生成した木に対する VIMP 推定量は, $B \rightarrow \infty$ の極限を取ることで次のように定義される:

$$Imp(Z_j, \mathcal{L}_N) = \mathbb{E}_{\Theta}[Imp(Z^{(j)}, \Theta, \mathcal{L}_N)].$$

上記の定式化では, いずれもフォレスト予測子を明示的に使用せず, 各々が単一の木の予測子のみで書かれることに注意したい.

フォレストからの VIMP の分散推定における深刻な問題として, ブートストラップの直接的な適用が機能しないことが挙げられる. フォレストの構築において, 既にブートストラップ標本を使用しており, 分散推定のために改めてブートストラップを適用すれば, 二重ブートストラップ標本を生成する必要があり, 計算上の問題に加えて, 通常の OOB 利用における妥当性に問題が生じる. Ishwaran and Lu (2018) は, 二重ブートストラップ標本において, 妥当な OOB の生成確率は 0.164 であることを導き, VIMP の分散推定のための.164 ブートストラップ法, およびこの方法から派生して, サブサンプリング法と d 例削除ジャックナイフ法を提案している. Williamson et al. (2021) は生存フォレストへの直接的な適用ではないものの, 変数のマスキングに基づく VIMP 指標に対して, ノンパラメトリックな影響関数を導出し, VIMP 指標の補正推定量を提案し, 提案の推定量の漸近ガウス性と漸近分散の一致推定量を与えている. ただし, VIMP 指標がゼロである帰無仮説では, パラメータ空間の境界上にある設定のため, 帰無仮説のもとでは不向きな補正方法であり, さらなる工夫を要する課題としている.

3.2 因果フォレスト

VIMP 指標の性質の理論的研究が徐々に進行する中で, Wager and Athey (2018) は, 樹木法に基づく条件付きの処理効果の(因果)推定に関する画期的な方法として因果フォレスト法を提案した. 因果木や因果フォレストに対する主要な着想は, 利用可能なデータ集合 \mathcal{L}_N を, 樹木を構成するデータ集合 ($\mathcal{L}_\bullet^{\mathcal{I}}$) と, 処理効果を推定するためのデータ集合 ($\mathcal{L}_\bullet^{\mathcal{J}}$) に分割して各プロセスを実施することで彼らの用語「正直さ(honesty)」を実現することである. つまり, この正直さの着想によって, 樹木の分岐の決定(予測モデルの構造推定)と処理効果の推定において, 同じデータを用いないことで各推定の独立性を与え, 因果推定の偏りを減らすことができる. このとき, 樹木構成でトレーニングデータの半分を捨てるため, 関数推定に関する情報損失が生じるが, $\mathcal{L}_\bullet^{\mathcal{I}}$ と $\mathcal{L}_\bullet^{\mathcal{J}}$ のデータ分割を再度ランダムサンプリングし直すため, この意味においてデータを無駄にすることがないと主張している. このような樹木の生成方式は, ダブルサンプル木(double-sample tree)と呼ばれ, 以下にそのアルゴリズムに纏めたものを与える:

ステップ 0. データ集合 \mathcal{L}_N の個人 i は応答観測値 $Y_i (= y(T_i, \Delta_i))$ と特徴量 \mathbf{Z}_i をもつ. 因果木では, 特徴量から関心のある処理変数を分離し, その処理変数を W_i と書く. 葉の最小サイズを k とする.

ステップ 1. 個体ラベル $\{1, \dots, N\}$ から M_N 個の要素をランダムに非復元抽出して得たデータ集合を $\mathcal{L}_{M_N}^{\text{sub}}$ とする. $\mathcal{L}_{M_N}^{\text{sub}}$ を二つの互いに素なデータ集合 $\mathcal{L}_{\lfloor M_N/2 \rfloor}^{\mathcal{I}(b)}, \mathcal{L}_{\lceil M_N/2 \rceil}^{\mathcal{J}(b)}$ にランダム分割する.

ステップ 2. 再帰分割により木を成長させる. 樹木成長では, $\mathcal{L}_{\lfloor M_N/2 \rfloor}^{\mathcal{J}(b)}$ からの任意のデータと, $\mathcal{L}_{\lfloor M_N/2 \rfloor}^{\mathcal{I}(b)}$ からの特徴量のみを用いて分岐を選択する(因果木の場合では処理変数 W は葉の例数管理のため用いる). ただし, データ集合 $\mathcal{L}_{\lfloor M_N/2 \rfloor}^{\mathcal{I}(b)}$ の応答観測値は用いない. また各分岐の構成において, 回帰木では, $\mathcal{L}_{\lfloor M_N/2 \rfloor}^{\mathcal{I}(b)}$ の観測値を k 個以上含むように, 因果木では, $\mathcal{L}_{\lfloor M_N/2 \rfloor}^{\mathcal{I}(b)}$ の観測値を各処理群ごとに k 個以上含むように管理する必要がある.

ステップ3. ステップ2で得た木 $T^{(b)}$ に, データ集合 $\mathcal{L}_{[M_N/2]}^{\mathcal{I}(b)}$ をあてはめて, 葉ごとの応答観測値の要約を求める.

ダブルサンプル木は, 因果推論のためだけでなく, 通常の樹木法にも適用可能であり, 関数近似の区間推定等に利用可能である. 因果ツリーの文脈では, 因果推論におけるSUTVA, 無交絡性(Unconfoundedness), 重複あり(overlap)の基本条件は成り立つとする. 十分に小さい葉内においては, 個人 $i \in T^{(b)}(z)$ に対応する応答と処理変数のペアがランダム化実験で生じたように見えるため潜在的な交絡調整効果が期待できる. このとき, 条件付き処理効果(CATE) $\tau(z) = E[Y_i^{(1)} - Y_i^{(0)} | Z_i = z]$ を, 終端ふし $T^{(b)}(z)$ を用いて

$$(3.1) \quad \hat{\tau}_{\text{tree}}^{(b)}(z) = \frac{\sum_{\{i \in T^{(b)}(z) : W_i = 1\}} Y_i}{|\{i \in T^{(b)}(z) : W_i = 1\}|} - \frac{\sum_{\{i \in T^{(b)}(z) : W_i = 0\}} Y_i}{|\{i \in T^{(b)}(z) : W_i = 0\}|}$$

により推定する. ここに, $Y_i^{(1)}$ と $Y_i^{(0)}$ はそれぞれ処理 $W_i = 1, 0$ を与えたときの Y_i の潜在的な結果である. 因果フォレストは, B 個の単一のダブルサンプル木からの結果 $\hat{\tau}_{\text{tree}}^{(b)}(z)$ を集約して処理効果のより強固な推定値 $\hat{\tau}(z) = B^{-1} \sum_{b=1}^B \hat{\tau}_{\text{tree}}^{(b)}(z)$ をもたらす. Wager and Athey (2018) はこの推定値 $\hat{\tau}(z)$ が, いくつかの正則条件(サブ標本サイズ M_N の条件, 正直性, ランダム分岐, α -正則性, 対称性, Lipschitz 連続)のもとで, $N \rightarrow \infty$ として

$$(3.2) \quad (\hat{\tau}(z) - \tau(z)) / \sqrt{V[\hat{\tau}(z)]} \xrightarrow{D} N(0, 1)$$

となる漸近正規性を確立し, 分散 $V[\hat{\tau}(z)]$ の一致推定値

$$(3.3) \quad \hat{V}_{IJ}(z) = \frac{N-1}{N} \left(\frac{N}{N-M_N} \right)^2 \sum_{i=1}^N \widehat{\text{Cov}}[\hat{\tau}_{\text{tree}}(z), \mathbb{N}_i]^2$$

を与えていた. ここに, $\widehat{\text{Cov}}[\hat{\tau}_{\text{tree}}(z), \mathbb{N}_i] = \sum_b (\hat{\tau}_{\text{tree}}^{(b)}(z) - \hat{\tau}(z)) (\mathbb{N}_i^{(b)} - \bar{\mathbb{N}}_i) / B$ であり, $\bar{\mathbb{N}}_i = B^{-1} \sum_b \mathbb{N}_i^{(b)}$ であり, $\mathbb{N}_i^{(b)}$ は第 i サンプルが b 番目の木 $T^{(b)}$ に使用されれば 1, そうでなければ 0 を示す(ダブルサンプル木では, 第 i サンプルが, データ集合 $\mathcal{L}_{[M_N/2]}^{\mathcal{I}(b)}$ または $\mathcal{L}_{[M_N/2]}^{\mathcal{J}(b)}$ のいずれかに含まれることを意味する). 正則条件についての詳細, 定理の証明は Wager (2016), Wager and Athey (2018) を参照されたい.

3.3 因果生存フォレスト

Wager and Athey (2018) の因果フォレストは, 連続な応答観測値 Y_i に対する方法であるため, 生存時間データに適用するためには, 更なる工夫が必要である. 因果フォレストそのものを生存解析用に開発し直すか, 生存時間情報 (T_i, Δ_i) のある変換 y を与えて, $Y_i = y(T_i, \Delta_i)$ とおくことが考えられる. 例えば, 変換 y として, 通常のCART樹木を生存データに適用するために帰無モデルのマルチングル残差を利用する方法 (Therneau et al., 1990) は直截的である(ただしこの場合の結果の解釈は難しい). また, 中途打ち切りがなければ, ある定数 $x > 0$ (horizon) に対して, RMST(制限付き平均生存時間)の推測では, $y(T_i, 1) = T_i \wedge x$, 生存率の推測では, $y(T_i, 1) = \mathbb{1}(T_i > x)$ と定めることができる. これに先立って, Athey et al. (2019) は, 因果フォレストに基づく処理効果の推測において, ダブルサンプルフォレストと傾向フォレストを統合する動機から, 観測値を直接的に用いる形式から推定方程式

$$(3.4) \quad \sum_{i=1}^N \alpha_i(z) \psi_{\hat{\tau}(z), \hat{\eta}(z)}(Y_i) = 0$$

に基づく方法へと拡張し, 因果フォレストを局所推定の枠組みで一般化した. ここに,

$\psi_{\hat{\tau}(z), \hat{\eta}(z)}(\cdot)$ は関心のある推定量 $\hat{\tau}(z)$ と局外推定量 $\hat{\eta}(z)$ を含む式であり, 重み $\alpha_i(z)$ は $B^{-1} \sum_{b=1}^B \alpha_i^{(b)}(z)$ と構成され, それぞれの $\alpha_i^{(b)}(z) = \mathbb{1}(i \in \mathcal{T}^{(b)}(z))$ は, 生成した b 番目の單一の木 $\mathcal{T}^{(b)}(z)$ から算出される. このような Athey et al. (2019) の一般化 RF の枠組みでの因果木の生成でも, 正直性を保証するため, 互いに素な二つのデータ集合に分け, 一方のデータ集合を樹木構成に用い, もう一方のデータ集合を $\alpha_i^{(b)}(z)$ の計算に用いる. ただし, 樹木構成における分割基準は, 通常の CART 法の分割基準とは異なり, 親ふしあにおいて

$$\widetilde{\Delta R}(\hbar_L, \hbar_R) = \frac{1}{|\{i \in \hbar_L\}|} \left(\sum_{i \in \hbar_L} \rho_i \right)^2 + \frac{1}{|\{i \in \hbar_R\}|} \left(\sum_{i \in \hbar_R} \rho_i \right)^2$$

を最大にする娘ふしあ \hbar_L, \hbar_R を選択する. ここに, ρ_i は疑似アウトカムと呼ばれ, 親ふしあにおける $\tau(z)$ の推定値の計算に対する第 i 観測値の影響関数に負符号を付けたものに対応する. 疑似アウトカム ρ_i は, 最小二乗回帰のときには回帰残差であり, 因果推論の文脈では処理効果の個人推定値に対応するため, 疑似アウトカムを目的変数にすることで標準的な機械学習手法をそのまま適用できる(さらなる詳細は Athey et al., 2019 を参照). 一般化 RF の因果木生成のアルゴリズムを以下に纏める:

ステップ 0-1. 因果フォレストのステップ 0-1 と同様に行う(全データから部分標本を非復元抽出し, 互いに素なデータ集合 $\mathcal{L}_{[M_N/2]}^{\mathcal{J}(b)}, \mathcal{L}_{[M_N/2]}^{\mathcal{I}(b)}$ にランダム分割する).

ステップ 2. $\mathcal{L}_{[M_N/2]}^{\mathcal{J}(b)}$ のデータ集合を用い, 再帰的分割により樹木を成長させる. なお, 樹木成長では, 各ふしあにおいて $\widetilde{\Delta R}(\hbar_L, \hbar_R)$ を最大にする娘ふしあ \hbar_L, \hbar_R を選択する分割基準を用いる. このような方式で十分成長させた樹木を $\mathcal{T}^{\mathcal{J}(b)}$ とする.

ステップ 3. ステップ 2 で作成した樹木 $\mathcal{T}^{\mathcal{J}(b)}$ に, $\mathcal{L}_{[M_N/2]}^{\mathcal{I}(b)}$ のデータ集合をあてはめて, $\alpha_i^{\mathcal{J}(b)}(z), i \in \mathcal{L}_{[M_N/2]}^{\mathcal{I}(b)}$ を計算する.

ステップ 4. ステップ 1-3 を B 回繰り返して, 集計された $\{\alpha_i(z)\}$ を用いて, (3.4) の解を求める.

Cui et al. (2023) の因果生存フォレストの方法は, この一般化 RF の局所推定の方法に沿って, Robins et al. (1994) の推定方程式を組み合わせて生存解析に応用している. Robinson (1988) の推定量を構成する式を

$$\psi_{\tau}(T_i^*, \mathbf{Z}_i, W_i; \hat{e}, \hat{m}) = \{W_i - \hat{e}(\mathbf{Z}_i)\} \{y(T_i^*) - \hat{m}(\mathbf{Z}_i) - \tau(W_i - \hat{e}(\mathbf{Z}_i))\}$$

とする. ここに, 真のイベント時間 T_i^* は必ずしも観測可能ではなく, $\hat{e}(\mathbf{Z}_i)$ と $\hat{m}(\mathbf{Z}_i)$ は $e(z) = P(W_i = 1 | \mathbf{Z}_i = z)$, $m(z) = E[y(T_i^*) | \mathbf{Z}_i = z]$ の交差あてはめ(cross-fitting, 例えは, Schick, 1986; Chernozhukov et al., 2018) によって得られた推定値とし, $y(T)$ は $y(T) = T \wedge x$ または $y(T) = \mathbb{1}(T > x)$ のいずれかである.

Cui et al. (2023) は, 中途打切りデータに対して, 処理効果を推定するためのある推定方程式 $E[\tilde{\psi}_{\tau}(T_i, \Delta_i, \mathbf{Z}_i, W_i, x)] = 0$ を得るために, Robins et al. (1994) を一般化して, 次のスコア関数

$$(3.5) \quad \begin{aligned} \tilde{\psi}_{\tau}(T_i, \Delta_i, \mathbf{Z}_i, W_i, x) &= \Delta_i^x \psi_{\tau}(T_i, \mathbf{Z}_i, W_i) / S_{W_i}^C(T_i \wedge x | \mathbf{Z}_i) \\ &+ (1 - \Delta_i^x) E[\psi_{\tau}(T_i^*, \mathbf{Z}_i, W_i) | T_i^* \wedge x > T_i \wedge x, \mathbf{Z}_i, W_i] / S_{W_i}^C(T_i \wedge x | \mathbf{Z}_i) \\ &- \int_0^{T_i \wedge x} \frac{\lambda_{W_i}^C(t | \mathbf{Z}_i)}{S_{W_i}^C(t | \mathbf{Z}_i)} E[\psi_{\tau}(y(T_i^*), \mathbf{Z}_i, W_i) | T_i^* \wedge x > t, \mathbf{Z}_i, W_i] dt \end{aligned}$$

を用いることを提案している. ここに, $\Delta_i^x = \Delta_i \mathbb{1}(T_i \geq x)$, $S_w^C(t | \mathbf{z}) = P(C_i \geq t | W_i = w, \mathbf{Z}_i = \mathbf{z})$ は中途打切り時間の条件付き生存関数, $\lambda_w^C(t | x) = -\frac{d}{dt} \log S_w^C(t | x)$ は対応するハザード関数

である。 (3.5) に、完全データの場合のスコア関数 $\psi_\tau(T_i^*, \mathbf{Z}_i, W_i; \hat{e}, \hat{m})$ をあてはめて、実際の推定計算で用いる 2 重ロバスト・スコア関数

$$(3.6) \quad \begin{aligned} & \tilde{\psi}_\tau(T_i, \Delta_i, \mathbf{Z}_i, W_i, x; \hat{e}, \hat{m}, \hat{\lambda}_w^C, \hat{S}_w^C, \hat{Q}_w) \\ &= \left(\frac{\hat{Q}_{W_i}(T_i \wedge x | \mathbf{Z}_i) + \Delta_i^x [y(T_i) - \hat{Q}_{W_i}(T_i \wedge x | \mathbf{Z}_i)] - \hat{m}(\mathbf{Z}_i) - \tau(W_i - \hat{e}(\mathbf{Z}_i))}{\hat{S}_{W_i}^C(T_i \wedge x | \mathbf{Z}_i)} \right. \\ & \quad \left. - \int_0^{T_i \wedge x} \frac{\hat{\lambda}_{W_i}^C(t | \mathbf{Z}_i)}{\hat{S}_{W_i}^C(t | \mathbf{Z}_i)} [\hat{Q}_{W_i}(t | \mathbf{Z}_i) - \hat{m}(\mathbf{Z}_i) - \tau(W_i - \hat{e}(\mathbf{Z}_i))] dt \right) (W_i - \hat{e}(\mathbf{Z}_i)) \end{aligned}$$

を得る。ここに、 $\hat{Q}_w(t|x) = E[y(T_i^*) | \mathbf{Z}_i = z, W_i = w, T_i \wedge x > t]$ は変換された生存情報 $y(T_i^*)$ の条件付き期待値であり、 $\hat{Q}_w(t|x)$ 、 $\hat{S}_w^C(t|x)$ 、 $\hat{\lambda}_w^C(t|x)$ は交差あてはめされた局外パラメータの推定値である。 $\hat{Q}_w(t|x)$ は推定生存関数の数値積分によって、 $\hat{\lambda}_w^C(t|x)$ は $-\log(\hat{S}_w^C(t|x))$ の前向き差分によって推定される。生存木成長の分割ルール「 $\widetilde{\Delta R}(\hbar_L, \hbar_R)$ の最大化」で用いる疑似アウトカム ρ_i ($i \in \hbar$) は、属するふしあにおいて、 $\rho_i = \tilde{\psi}_{i, \hat{\tau}_h} / (|\hbar|^{-1} \sum_{j \in \hbar} U_j)$ を用いる (Cui et al., 2023)。ここに、 $\tilde{\psi}_{i, \tau}$ は $\tilde{\psi}_\tau(T_i, \Delta_i, \mathbf{Z}_i, W_i, x; \hat{e}, \hat{m}, \hat{\lambda}_w^C, \hat{S}_w^C, \hat{Q}_w)$ の簡略表記であり、 U_j は

$$U_j = U(T_j, \mathbf{Z}_j, W_j, x; \hat{e}, \hat{\lambda}_w^C, \hat{S}_w^C) = (W_j - \hat{e}(\mathbf{Z}_j))^2 \left(\frac{1}{\hat{S}_{W_j}^C(T_j \wedge x | \mathbf{Z}_j)} - \int_0^{T_j \wedge x} \frac{\hat{\lambda}_{W_j}^C(t | \mathbf{Z}_j)}{\hat{S}_{W_j}^C(t | \mathbf{Z}_j)} dt \right)$$

であり、 $\hat{\tau}_h$ は $\sum_{i \in \hbar} \tilde{\psi}_{i, \tau} = 0$ を解くことで得られるふしあ内での τ の推定値である。

一般化 RF のアルゴリズムのステップ 0-4 に従って、因果木を生成し、各重みを評価することで、フォレスト全体としての重み $\{\alpha_i(z)\}$ が算出される。このステップ 4 における処理効果のフォレスト型の推定を行う際には、スコア関数 (3.6) の中に必要とされる局外成分 \hat{e} 、 \hat{m} 、 $\hat{\lambda}_w^C / \hat{S}_w^C$ 、 \hat{S}_w^C 、 \hat{Q}_w を推定し、それらをフォレストの局所推定方式 (3.4) に組み合わせることで得られる推定方程式

$$(3.7) \quad \sum_{i=1}^N \alpha_i(z) \tilde{\psi}_{i, \tau(z)}(T_i, \Delta_i, \mathbf{Z}_i, W_i, x; \hat{e}, \hat{m}, \hat{\lambda}_w^C, \hat{S}_w^C, \hat{Q}_w) = 0$$

を用いる。処理効果のフォレスト推定値 $\hat{\tau}(z)$ は (3.7) を満たす解として得られる。

Cui et al. (2023) は Wager and Athey (2018) の因果フォレストに類似する正則条件と局外成分の一致性のもとで推定量 $\hat{\tau}(z)$ が真値 $\tau(z)$ の周りで漸近的正規であることを与えている。漸近分散 σ_N^2 は、関数的デルタ法により、 $\sigma_N^2(z) = \text{Var}(\sum_i \alpha_i(z) \tilde{\psi}_{i, \tau(z)}) / E[U(T, \mathbf{Z}, W, x; e, \lambda_w^C, S_w^C | \mathbf{Z} = z)]^2$ の形式で与えられる。 $E[U(T, \mathbf{Z}, W, x; e, \lambda_w^C, S_w^C) | \mathbf{Z} = z]$ はフォレストに基づく経験推定値で推定でき、 $\text{Var}(\sum_i \alpha_i(z) \tilde{\psi}_{i, \tau(z)})$ はスマールバックのブートストラップ法 (Athey et al., 2019) により推定可能である。なお、スマールバックのブートストラップ法のサンプリング計画や一致性についての詳細は、R パッケージ `grf` の実装コードや Athey et al. (2019) を参照されたい。現時点では因果生存フォレストを応用した研究 (例えば、Desai et al. (2024) では駆出率保持心不全に対するスピロノラクトンの有効性評価のために実施された RCT データに因果生存フォレストを適用し、治療効果の異質性に影響を与える要因を検討している) は限られているが、因果生存フォレストは伝統的なロジスティック回帰や Cox 回帰よりも柔軟に集団レベルの曝露効果を個人レベルに分解し、理論的妥当性をもって治療効果の個人不均一性の特定もしくは仮説を生み出すツールとして有望である。

4. データ適用例

ここでは、データ `peakv02`(男性) を用いて生存フォレストや因果生存フォレストの適用例を

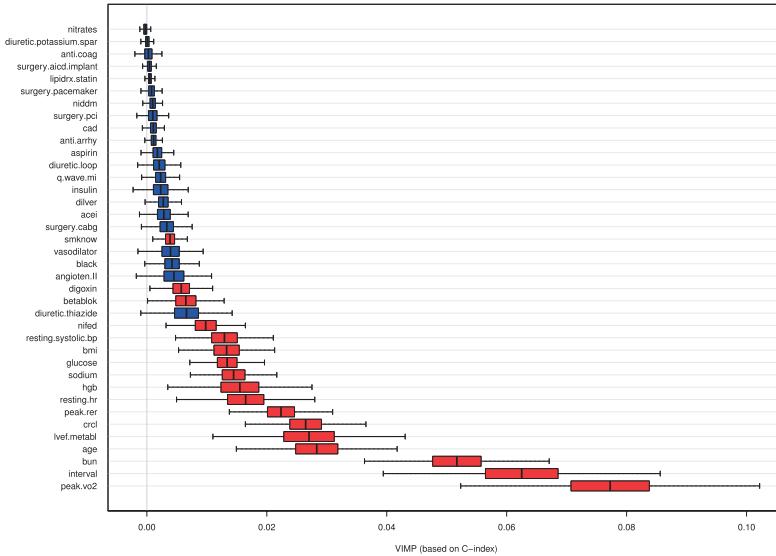


図4. データ peakV02(男性)における生存フォレストの変数重要度(VIMP)プロット.

示す。まず、3.1節で説明した生存フォレストによるVIMPを図4に与える。例えば心筋梗塞二次予防において推奨されてきたが近年議論になっている β 遮断薬の投与(betablock)は、Cox回帰と同様に有意な傾向にある。各VIMPの近似的95%信頼区間は d 例削除ジャックナイフ法(Ishwaran and Lu, 2018)により作成され、この方法の漸近理論の確立は依然として課題であるがシミュレーション研究では良好な性能が示されている。なお、図4はRライブラリのrandomForestSRC(Ishwaran et al., 2021)の次のコマンドにより描画した：

```
set.seed(124)
rf <- rfsrc(Surv(ttodead, died)~., data=peakV02m)
rf.smp.o <- subsample(rf, B=25, subratio=0.5)
plot.subsample(rf.smp.o, xlab="VIMP", sort=T)
```

Hsieh et al. (2011)で指摘されたように、選ばれる重要な予後因子は、Cox回帰と生存フォレストで類似するが、重要度の順位にはいくらかの違いが生じる。そのため、VIMPは、実際の研究では、変数選択の目的や重要な予後因子の同定のためにCox回帰の代替もしくは補完として有用である(例えば、He et al., 2020; Liu et al., 2021など)一方、ハザード比のような直接的な解釈ができないため、集団疫学的な視点からは活用しにくい側面があることには留意したい。

次に、処理変数ではないが例示のため、最高酸素摂取量(peak.vo2)を興味ある変数と仮定して生存フォレストと因果生存フォレストの結果を比較する。図5の左図は生存フォレストから予測された生存関数に対して、最高酸素摂取量を12から30まで変化させた部分従属プロットである。この部分従属プロットは(2.1)の推定生存関数ではなく、個々の生存木 $T^{(b)}$ から推定された生存関数 $\hat{S}^{(b)}(t|z)$ に、特定の変数 Z_j の値(ここではpeak.vo2)を c におき換えた $T^{(b)}$ のOOBデータをあてはめて、フォレストの上で平均化した

$$\hat{S}_j^{\text{OB}}(t|c) = \frac{1}{N} \sum_{i \in \mathcal{L}_N} \frac{1}{|O_i|} \sum_{b \in O_i} \hat{S}^{(b)}(t|Z_i, Z_j = c)$$

により与えられる (Ishwaran and Lu, 2019). ここに, O_i は個人 i が OOB であることを記録した生存木の番号である. 図 5 の中央は生存関数の部分従属プロットを 5 年生存率で表現したもので peak.vo2 が 20 までは緩やかに増加し, 20 以降ではあまり変化がないことがわかる. 図 5 の右図は, 処理変数に $W_i^{(c)} = \mathbb{1}(\text{peak.vo2}_i < c)$ を設定して, 因果生存フォレストを適用して得た結果を, c を変化させて描いたものである. 丸点は, 各 c (x 軸) の値において, 条件付き処理効果 (CATE) $\tau_c(z_{md}) = E[\mathbb{1}(T_i^{*(1)} > 5) - \mathbb{1}(T_i^{*(0)} > 5) | Z_i = z_{md}]$ に対する 3.3 節で与えた推定値 $\hat{\tau}(z_{md})$ であり, $Z_i = z_{md}$ は (W_i 以外の) すべての共変量を中央値 z_{md} に設定することを意味する (実線の曲線はこれらの点の Lowess あてはめ, 破線の Lowess 曲線は対応する 95% 各点信頼区間を示す). 図 5 の右図の三角点は平均処理効果 (ATE) $\tau = E[\mathbb{1}(T_i^{*(1)} > 5) - \mathbb{1}(T_i^{*(0)} > 5)]$ の推定値を表す. ATE の推定量として, Cui et al. (2023) では, Robins et al. (1994) による拡張逆確率重み付け推定量 (augmented inverse propensity) を若干変形した 2 重ロバスト・スコア

$$(4.1) \quad \hat{\Gamma}_i = \hat{\tau}(Z_i) + \frac{\tilde{\psi}_{\hat{\tau}}(Z_i)(T_i, \Delta_i, Z_i, W_i, x; \hat{e}, \hat{m}, \hat{\lambda}_w^C, \hat{S}_w^C, \hat{Q}_w)}{\hat{e}(Z_i)(1 - \hat{e}(Z_i))}$$

に基づく平均の使用を示唆している (Cui et al. (2023) では CATE の議論に集中し, ATE に対する言及はないが, Athey and Wager (2019) では (4.1) に類似する形式を用いて ATE 推定量を提案している). ここで, $\tilde{\psi}_{\hat{\tau}}(\cdot)$ は (3.6) で定義されたものであり, $\hat{\tau}(Z_i)$ は因果生存フォレストのあてはめから得られる個人 i の共変量 Z_i に対する CATE 推定値である. Cui et al. (2023) では, CATE の異質性の解釈をなすために, 関心のある特徴量から $\{\hat{\Gamma}_i\}_{i=1}^N$ への回帰を行なうことも与えている. なお, 図 5 の右図のための計算は, R ライブラリの `grf` (Tibshirani et al., 2024) の `causal_survival_forest` 関数を用いた以下のコマンドで求めることができる (この例は処理変数 $W_i = \mathbb{1}(\text{peak.vo2} < c)$ の設定が $c = 20$ の場合. 5 年生存率の推測のため, `horizon=5`, `target="survival.probability"` に設定している) :

```
> Y <- peakV02m$ttodead; D <- peakV02m$died; Z <- peakV02m[,-c(30,32,33)]
> W <- ifelse(peakV02m$peak.vo2 < 20, 0, 1)
> csf <- causal_survival_forest(Z, Y, W, D, horizon=5, target="survival.probability")
> Zmed <- matrix(apply(Z, 2, median), nrow=1)
> predict(csf, newdata=Zmed, estimate.variance=TRUE) # Zmed に対する CATE
  predictions variance.estimates
  0.1839267  0.00386204
> average_treatment_effect(csf) # ATE
  estimate      std.err
  0.07390137  0.01650680
```

図 5 より, 5 年生存率の CATE や ATE の推定値 (右図) は, 最高酸素摂取量 (`peak.vo2`) の変化で平均的に 10% 程度変化すること (中央) を良好に捉えていることが確認できる. 最高酸素摂取量 (`peak.vo2`) は, Cox 回帰や生存フォレストにおいて予後への強い統計的有意性を示したが, 生存率を目的とした因果フォレストの 95% 各点信頼区間の下限がゼロ付近に留まっていることから, 因果フォレストの統計的検出力は低下していることが観察できる. 単純なシミュレーション実験から, 生存率を目的とした因果フォレストの検出力は, x (`horizon`) の設定値により大きく変化することに注意したい (Web Supplementary Materials B を参照). 生存率を目的とする因果生存フォレストは, 生存期間全体を用いる生存フォレストや Cox 回帰と比べ, 5 年生存率という 1 点だけの情報を縮約される点も検出力で不利に働く. なお, Athey and Wager (2019) では, 検討する変数が多い場合, 最初の因果フォレストで推定された変数重要度の低い

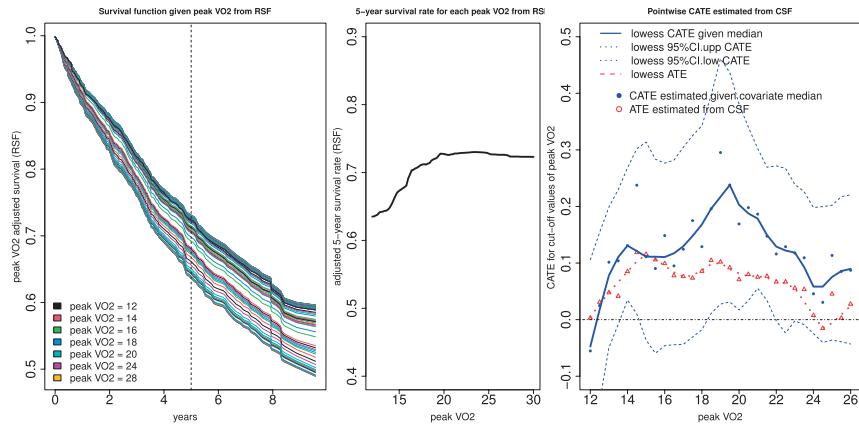


図 5. データ peakVO2(男性)における生存フォレスト(RSF)からの peak.vo2 に対する部分従属プロット(左), 部分従属プロットからの peak.vo2 による 5 年生存率の推移(中央), peak.vo2 のカットオフ値に対する因果生存フォレスト(CSF)からの CATE と ATE の推定結果(右).

変数を削除してから, 改めて因果フォレストをあてはめ直して, CATE や ATE を推定することを提示している(本節ではこのステップは省略した). 正直さ(honesty)の実現のために行うサンプル分割は検出力を下げるかもしれない. 因果生存フォレストに有利な状況であれば, ダブルサンプルに基づく検出力低下は, 実用的観点ではそれほど問題にならない可能性もある. いずれにしても予測指標や変数重要度指標の表示だけでなく, 因果フォレストによって実データから複雑な効果の異質性について有意義な結果を導くための統計的推測の実施のための基本が与えられた意義は極めて大きい. Web Supplementary Materials B では, 因果生存フォレストの統計的性能のさらなる数値的検討のため, シミュレーション実験の結果を与えていている.

5. 結びに代えて

本論文では, 機械学習・統計解析ツールとして, 広く応用されるようになっている決定木やフォレスト法などの樹木法における生存データへの適用と最近の因果フォレストへの展開を報告した. 近年, 様々な機械学習法を因果推論と融合させる方法論の研究が活性化している(Xu et al., 2023; Wan et al., 2024). 因果フォレスト法(Wager and Athey, 2018)の顕著な点は, 難解であったランダムフォレスト法の統計的漸近理論の構築を, 正直さ(honesty)などの着想を用いて与えたことである. 今後にかけて, 関係する理論のさらなる深耕, 方法の発展が期待される. 因果フォレスト法の適用可能なデータの範囲は徐々に拡大し, Athey and Wager (2019)ではクラスターデータに対する因果フォレストを議論し, Athey et al. (2019)では局所的推定方程式のアプローチによって因果フォレストの適用範囲を広げる一般化ランダムフォレストを提案し, 因果生存フォレストの展開が生み出されている. 現状の因果生存フォレストでは, CATE, ATE の推測において, RMST と各点生存率の利用に留まっているが, ハザード比や全体的な生存関数への適用可能にすることも興味のある展開の一つである. ダブルサンプルに基づく検出力の低下は予想外に小さい可能性があるがセミパラメトリック有効性にどの程度近づけられるかの検討は興味ある課題である. 因果生存フォレスト法の公式的利用は, 現時点では, 2 値処理変数の枠組みに限定されているが, 通常の因果フォレストの場合には連続処理変数への拡張も grf に既に実装されている. このような展開から, 依然として十分でないフォレ

スト法からの変数効果を調べるための有力なアプローチとしてさらなる研究の進展が期待される。樹木構造アプローチやランダムフォレストはこれまでも統計利用やデータ科学において有用なツールを与えてきたが、今後にかけても、有益な理論と方法論の研究が進展し、有意義な応用が展開されることが期待される。

参考文献

- Arano, I., Sugimoto, T., Hamasaki, T. and Ohno, Y. (2010). Practical application of cure mixture model to long-term censored survivor data from a withdrawal clinical trial of patients with major depressive disorder, *BMC Medical Research Methodology*, **10**, <https://doi.org/10.1186/1471-2288-10-33>.
- Athey S. and Wager S. (2019). Estimating treatment effects with causal forests: An application, *Observational Studies*, **5**, 36–51.
- Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests, *The Annals of Statistics*, **47**, 1148–1178.
- Bou-Hamad, I., Larocque, D. and Ben-Ameur, H. (2011). A review of survival trees, *Statistics Surveys*, **5**, 44–71.
- Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.
- Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.
- Breiman, L. (2003). How to use survival forests, Department of Statistics, University of California, Berkeley, California, http://www.stat.berkeley.edu/~breiman/SF_Manual.pdf (最終アクセス日 2025年3月18日).
- Breiman, L. (2004). Consistency for a simple model of random forests, Technical Report, **670**, University of California, Berkeley, California, <https://www.stat.berkeley.edu/~breiman/RandomForests/consistencyRFA.pdf> (最終アクセス日 2025年3月18日).
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman and Hall, Wadsworth, New York.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal*, **21**, 1–68.
- Ciampi, A., Thiffault, J., Nakache, J.-P. and Asselain, B. (1986). Stratification by stepwise regression, correspondance analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates, *Computational Statistics & Data Analysis*, **4**, 185–204.
- Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S. and Ruqiang, Z. (2023). Estimating heterogeneous treatment effects with right-censored data via causal survival forests, *Journal of the Royal Statistical Society Series B*, **85**, 179–211.
- Davis, R. B. and Anderson, J. R. (1989). Exponential survival trees, *Statistics in Medicine*, **8**, 947–961.
- Desai, R. J., Glynn, R. J., Solomon, S. D., Claggett, B., Wang, S. V. and Vaduganathan, M. (2024). Individualized treatment effect prediction with machine learning-salient considerations, *New England Journal of Medicine Evidence*, **3**, <https://doi.org/10.1056/EVIDoa2300041>.
- Efron, B. (2014). Estimation and accuracy after model selection (with discussion), *Journal of the American Statistical Association*, **109**, 991–1007.
- 江村剛志 (2023). 安定化スコア検定を用いた高次元生存データに基づく決定木の構築法, 日本統計学会誌, **52**, 373–390.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.
- Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis, *Cancer Treatment Reports*, **69**, 1065–1069.

- Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data, *Statistics in Medicine*, **18**, 2529–2545.
- Harrell, F., Califf, R., Pryor, D., Lee, K. and Rosati, R. (1982). Evaluating the yield of medical tests, *The Journal of the American Medical Association*, **247**, 2543–2546.
- He, J., Zhang, J. X., Chen, C. T., Ma, Y., De Guzman, R., Meng, J. and Pu, Y. (2020). The relative importance of clinical and socio-demographic variables in prognostic prediction in non-small cell lung cancer: A variable importance approach, *Medical Care*, **58**, 461–467.
- Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves, *Biometrics*, **61**, 92–105.
- Hothorn, T., Lausen, B., Benner, A. and Radespiel-Troger, M. (2004). Bagging survival trees, *Statistics in Medicine*, **23**, 77–91.
- Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A. and van der Laan, M. J. (2006a). Survival ensembles, *Biostatistics*, **7**, 355–373.
- Hothorn, T., Hornik, K. and Zeileis, A. (2006b). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics*, **15**, 651–674.
- Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H. and Lauer, M. S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests, *Circulation: Cardiovascular Quality and Outcomes*, **4**, 39–45.
- Ishwaran, H. and Lu, M. (2018). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival, *Statistics in Medicine*, **38**, 558–582.
- Ishwaran, H. and Lu, M. (2019). Random survival forests, *Wiley StatsRef: Statistics Reference Online*, 1–13, John Wiley & Sons, <https://doi.org/10.1002/9781118445112.stat08188>.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008). Random survival forests, *Annals of Applied Statistics*, **2**, 841–860.
- Ishwaran, H., Lu, M. and Kogalur, U. B. (2021). Variable importance (VIMP) with subsampling inference, <http://randomforestsrc.org/articles/vimp.html> (最終アクセス日 2025年3月18日).
- LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data, *Biometrics*, **48**, 411–425.
- LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split, *Journal of the American Statistical Association*, **88**, 457–467.
- LeBlanc, M. and Crowley, J. (1995). A review of tree-based prognostic models, *Journal of Cancer Treatment and Research*, **75**, 113–124.
- Liao, T., Su, T., Lu, Y., Huang, L., Wei, W. Y. and Feng, L. H. (2024). Random survival forest algorithm for risk stratification and survival prediction in gastric neuroendocrine neoplasms, *Scientific Reports*, **14**, <https://doi.org/10.1038/s41598-024-77988-1>.
- Liu, B., Niu, L., Boscoe, F. and Lee, F. F. (2021). Predictors of survival among male and female patients with malignant pleural mesothelioma: A random survival forest analysis of data from the 2000–2017 surveillance, epidemiology, and end results program, *Journal of Registry Management*, **48**, 118–125.
- Loh, W. Y. (2002). Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, **12**, 361–386.
- Mogensen, U., Ishwaran, H. and Gerds, T. (2012). Evaluating random forests for survival analysis using prediction error curves, *Journal of Statistical Software*, **50**, 1–23.
- Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, **58**, 415–434.
- Morvan, L., Carlier, T., Jamet, B., Bailly, C., Bodet-Milin, C., Moreau, P., Kraeber-Bodéré, F. and Mateus, D. (2020). Leveraging RSF and PET images for prognosis of multiple myeloma at diagnosis, *International Journal of Computer assisted Radiology and Surgery*, **15**, 129–139.
- 中村知繁 (2020). ランダムフォレストによる因果推論, 慶應義塾大学経済研究所, <https://ies.keio.ac.jp/>

- upload/20201201econo_nakamura_Slide.pdf (最終アクセス日 2025 年 3 月 18 日).
- Probst, P., Boulesteix, A.-L. and Bischl, B. (2019). Tunability: importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, **20**, 1–32.
- Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 846–866.
- Robinson, P. M. (1988). Root-n-consistent semiparametric regression, *Econometrica*, **56**, 931–954.
- Schick, A. (1986). On asymptotically efficient estimation in semiparametric models, *The Annals of Statistics*, **14**, 1139–1151.
- Scornet, E., Biau, G. and Vert, J.-P. (2015). Consistency of random forests, *The Annals of Statistics*, **43**, 1716–1741.
- Segal, M. R. (1988). Regression trees for censored data, *Biometrics*, **44**, 35–48.
- Segal, M. R. (1992). Tree-structured methods for longitudinal data, *Journal of the American Statistical Association*, **87**, 407–418.
- Segal, M. R. (2004). Machine learning benchmarks and random forest regression, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco, California, <https://escholarship.org/uc/item/35x3v9t4> (最終アクセス日 2025 年 3 月 18 日).
- 下川敏雄, 杉本知之, 後藤昌司 (2013). 『樹木構造接近法』, 共立出版, 東京.
- 杉本知之, 下川敏雄, 後藤昌司 (2005). 樹木構造接近法と最近の発展, 計算機統計学, **18**, 123–164.
- Therneau, T., Grambsch, P. and Fleming, T. (1990). Martingale-based residuals for survival models, *Biometrika*, **77**, 147–160.
- Tibshirani, J., Athey, S., Sverdrup, E. and Wager, S. (2024). grf: Generalized random forests, <https://github.com/grf-labs/grf> (最終アクセス日 2025 年 3 月 18 日).
- Wager, S. (2016). Causal inference with random forests, Ph.D. Thesis, Stanford University, Stanford, California.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association*, **113**, 1228–1242.
- Wager, S., Hastie, T. and Efron, B. (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife, *The Journal of Machine Learning Research*, **15**, 1625–1651.
- Wan, K., Tanioka, K. and Shimokawa, T. (2024). Survival causal rule ensemble method considering the main effect for estimating heterogeneous treatment effects, *Statistics in Medicine*, **43**, 5234–5271.
- Williamson, B. D., Gilbert, P. B., Carone, M. and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques, *Biometrics*, **77**, 9–22.
- Xu, Y., Ignatiadis, N., Sverdrup, E., Fleming, S., Wager, S. and Shah, N. H. (2023). Treatment heterogeneity for survival outcomes, *Handbook of Matching and Weighting Adjustments for Causal Inference* (eds. J. R. Zubizarreta, E. A. Stuart, D. S. Small and P. R. Rosenbaum), 445–482, Chapman and Hall/CRC, New York.
- Zhou, Y. and McArdle, J. J. (2015). Rationale and applications of survival tree and survival ensemble methods, *Psychometrika*, **80**, 811–833.

Tree-structured Approaches and Recent Advances in Survival Analysis

Tomoyuki Sugimoto¹, Kazushi Maruo² and Toshio Shimokawa³

¹Graduate School of Engineering Science, The University of Osaka

²Institute of Medicine, Tsukuba University

³Graduate School of Medicine, Wakayama Medical University

Tree-based methods such as decision trees and random forests are statistical and machine learning tools widely used in data science and many other fields. Over the past decade, there has been remarkable progress in methodologies and theories related to causal inference tools based on causal trees and causal forests, leveraging variable importance measures, the consistency of forests, and asymptotic normality. This paper investigates the recent developments in methodologies and theories of such tree-based methods, focusing on their application in survival analysis, and comprehensively reports on the characteristics of each method.

System signature の計算方法とその応用

—Consecutive- k -out-of- n :F システムを例として—

中村 太信¹・周 蕎²

(受付 2024 年 12 月 4 日；改訂 2025 年 2 月 5 日；採択 3 月 10 日)

要 旨

本論文では、コヒーレントシステムに対する system signature の算出方法を概説し、特に consecutive- k -out-of- n :F (Con/ k / n :F) システムに対する具体的な手法および計算例を示す。Con/ k / n :F システムは、 n 個のコンポーネントが線形に配置され、少なくとも k 個の連続したコンポーネントが故障した場合にシステム故障となるシステムモデルである。このモデルは石油パイプラインや高速道路の街灯、生産監視システムなど、様々な応用例を持つ。近年、信頼性工学の分野では、system signature の概念が注目されている。ここで、system signature とはシステムの構造にのみ依存する特性量であり、システム構造の比較や保全計画の立案に有用である。しかし、素朴な計算方法では、大規模なシステムに対して system signature を算出することが困難である。本論文では、Con/ k / n :F システムに対するピボット分解法に基づく system signature の算出方法を示す。この手法は規則的な構造を持つシステム群に適用可能である。その後、system signature の理論的応用の 1 つとして、故障コンポーネント数の期待値について述べる。最後に、汎用的な system signature の算出方法の可能性について議論する。

キーワード：コヒーレントシステム、system signature、consecutive- k -out-of- n :F システム、ピボット分解法。

1. はじめに

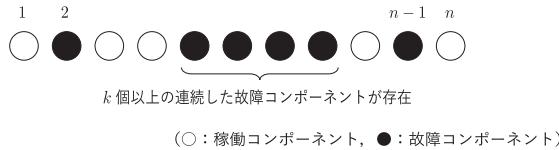
現代社会は、電力システム、交通システム、通信システムなど、様々なシステムに支えられている。これらの重要なインフラストラクチャの故障は、社会や経済に重大な影響を及ぼす可能性があるため、システムの信頼性向上と効率的な保全は、安全・安心な社会の実現に不可欠である。

信頼性工学(reliability engineering)は、各種製品やシステムの品質と信頼性を向上・保証し、信頼性・安全性上の問題の再発防止・未然防止を目的とする学問である(山本 他, 2021b)。この分野において、システムの寿命分布を導出し、信頼性などを評価することは重要な課題である。特に、現代のシステムは複雑化・大規模化しており、システム全体の信頼性を個々のコンポーネントの信頼性から導出することが必要となっている。そのため、システムの構造に基づいて信頼性を評価する手法が求められている。

近年、信頼性工学の分野では、system signature の概念が注目されている(Naqvi et al.,

¹ 東海大学 情報理工学部：〒259-1292 神奈川県平塚市北金目 4-1-1; nkmr@tokai.ac.jp

² 長岡技術科学大学 工学部：〒940-2188 新潟県長岡市上富岡町 1603-1; zhoullei@vos.nagaokaut.ac.jp

図 1. Con/ k/n :F システムの故障例。

2022). ここで, system signature とは, システムの構造のみに依存し, コンポーネントの寿命分布には依存しない特性量であり, システムの信頼性評価のための有用なツールとなっている (Samaniego, 2007). System signature を用いてさまざまなシステムの信頼性特性を評価し, システム構造を比較することが可能となる. また, system signature に基づいて最適な保全計画を立案することも可能である (Zhou and Yamamoto, 2022; Asadi et al., 2023). しかしながら, system signature の理論的な有用性にもかかわらず, その適用範囲は限定的である. その主な理由は, 素朴な計算方法では, コンポーネント数が少なくシステムの構造関数が極めて単純な場合を除いて, system signature の算出が困難であるためである.

現代のシステムは大規模で複雑であるが, 多くの場合, 特徴的な構造を有しているため, これらの特性を反映したシステムモデルが活用されている. 最も基本的なモデルとして, 直列システムと並列システムが挙げられる. 直列システムは, 少なくとも 1 つのコンポーネントが故障した場合にシステムが故障し, 一方, 並列システムは, すべてのコンポーネントが故障した場合にのみシステムが故障する. これらを一般化した k -out-of- n :F システムは, n 個のコンポーネントのうち少なくとも k 個が故障した場合にシステムが故障する. これらのシステムは比較的単純かつ一般的でありながら, 多様な問題に適用可能である.

しかし, k -out-of- n :F システムでは, システムの状態が故障したコンポーネントの数のみに依存し, その位置には依存しない. 一方, 実際のシステムでは, 故障したコンポーネントの位置を考慮することが重要となる場合がある. 例えば, 実社会にはコンポーネントの集中故障がシステム故障を引き起こすようなシステムが存在する. このようなシステムとして, 1980 年に Kontoleon (1980) が初めて研究し, その後, Chiang and Niu (1981) によって正式に「consecutive- k -out-of- n :F システム」と命名されたモデルがある.

Consecutive- k -out-of- n :F (以下, Con/ k/n :F) システムとは, n 個のコンポーネントが線形に配置されており, 少なくとも k 個の連続したコンポーネントが故障した場合にシステムが故障するシステムモデルである(図 1). 以下に, Con/ k/n :F システムの応用例を示す.

応用例 1: 石油パイプラインシステム (Chiang and Niu, 1981). n 個のポンプステーションから構成されるパイプラインシステムを考える. 各ポンプステーションは隣接する k 個のステーションまで石油を輸送する能力を持つ. 1 つのポンプステーションが故障しても, 隣接するステーションが負荷を担うことができるため, 石油の流れは中断されない. しかし, k 個以上の連続したポンプステーションが故障した場合, 石油の流れが中断され, その結果, システム全体の故障となる. このような石油パイプラインシステムは Con/ k/n :F システムとしてモデル化できる.

応用例 2: 高速道路の街灯 (Dafnis et al., 2019). 一定の間隔で n 個の街灯が設置された高速道路を考える. 夜間ににおいて, k 個の連続する街灯が機能しないと, 特定区域の照明が不足し, 交通に影響を及ぼす可能性がある. このような街灯システムは Con/ k/n :F システムとしてモデル化できる.

応用例 3：生産監視システム (Zhao et al., 2019). n 個のカメラで構成された生産監視システムを考える。各カメラは隣接する k 個のユニットを監視可能である。1つのカメラが故障しても、隣接カメラが代替監視を行える。しかし、 k 個の連続するカメラが故障すると、監視不能領域が生じ、システム全体としては故障となる。このような生産監視システムは $\text{Con}/k/n:\text{F}$ システムとしてモデル化できる。

前述のとおり、一般的に system signature の計算は困難であり、このことは実践的な応用における重大な障壁となっている。本論文では、system signature の定義を述べた上で、さまざまな計算方法を紹介する。特に、 $\text{Con}/k/n:\text{F}$ システムに対する具体的な計算法を述べる。その後、system signature の理論的応用の1つとして、故障コンポーネント数の期待値について述べる。

2. コヒーレントシステムと system signature の定義

2.1 コヒーレントシステムとは

本節では、コヒーレントシステムの定義を述べる。 n 個のコンポーネントからなるシステムを考え、各コンポーネントは稼働状態または故障状態のいずれかであると仮定する。コンポーネント j の状態を示すために、指示変数 x_j を次のように定義する。

$$(2.1) \quad x_j = \begin{cases} 1, & \text{コンポーネント } j \text{ が稼働している場合} \\ 0, & \text{コンポーネント } j \text{ が故障した場合} \end{cases}$$

このとき、ベクトル $\mathbf{x} = (x_1, x_2, \dots, x_n)$ はすべてのコンポーネントの状態を表し、状態ベクトルと呼ばれる。

また、システムの状態はコンポーネントの状態によって決定されると仮定したとき、システムの状態を示す関数 φ は次式で定義される。

$$(2.2) \quad \varphi(\mathbf{x}) = \begin{cases} 1, & \text{状態ベクトル } \mathbf{x} \text{ のときシステムが稼働している場合} \\ 0, & \text{状態ベクトル } \mathbf{x} \text{ のときシステムが故障した場合} \end{cases}$$

ここで、 $\varphi(\mathbf{x})$ はシステムの構造関数と呼ばれる。

定義 1 にコヒーレントシステムの定義を示す。

定義 1. リレバントで単調なシステムをコヒーレント (coherent) システムと呼ぶ。

$a \in \{0, 1\}$ に対して、 $(a_j, \mathbf{x}) \equiv (x_1, x_2, \dots, x_{j-1}, a, x_{j+1}, \dots, x_n)$ とする。このとき、コンポーネント j が構造 φ に対してリレバントでないとは、任意の \mathbf{x} に対して $\varphi(1_j, \mathbf{x}) = \varphi(0_j, \mathbf{x})$ が成り立つことを意味する。これは、コンポーネント j の状態が変化することが、システムの状態が変化することに影響を及ぼさないことを意味する。それ以外の場合、コンポーネント j は構造 φ に対してリレバントであるとされる。

単調なシステムとは、システムの構造関数が各コンポーネントの状態に対して単調増加であることを指す。すなわち、 $\mathbf{y} = (y_1, y_2, \dots, y_n)$ を状態ベクトルとすると、 $\mathbf{x} \leq \mathbf{y}$ を満たす任意の \mathbf{x} および \mathbf{y} に対して $\varphi(\mathbf{x}) \leq \varphi(\mathbf{y})$ が成り立つ。これは、コンポーネントの状態が改善されたとき、システムの状態は悪くはならないことを意味する。

状態ベクトル \mathbf{x} が $\varphi(\mathbf{x}) = 1$ を満たすとき、 \mathbf{x} をパスベクトルと呼ぶ。さらに、全ての $\mathbf{y} < \mathbf{x}$ に対して $\varphi(\mathbf{y}) = 0$ が成り立つとき、 \mathbf{x} を極小パスベクトルと呼ぶ。 \mathbf{x} が極小パスベクトルであるとき、集合 $A = \{j \mid x_j = 1\}$ を極小パス集合と呼ぶ。すなわち、極小パス集合とは、シス

ムの稼働を保証する最小のコンポーネント集合である。

一方, 状態ベクトル x が $\varphi(x) = 0$ を満たすとき, x をカットベクトルと呼ぶ. さらに, 全ての $y > x$ に対して $\varphi(y) = 1$ が成り立つとき, x を極小カットベクトルと呼ぶ. x が極小カットベクトルであるとき, 集合 $C = \{j \mid x_j = 0\}$ を極小カット集合と呼ぶ. すなわち, 極小カット集合とは, システムの故障を引き起こす最小のコンポーネント集合である.

コヒーレントシステムやパスやカットに関する詳細は大鑄 (2019) を確認されたい.

2.2 システムの信頼度関数

本節では, システムの信頼度関数を説明する. システム寿命を表す確率変数を T とするとき, システムの信頼度関数(reliability function)は $P(T > t)$ と与えられ, これは, 時刻 t までにシステムが故障しない確率を意味する. システム全体の信頼度は, コンポーネントの信頼度とその構成方法から求められる. n 個のコンポーネントの寿命を表す確率変数を X_1, X_2, \dots, X_n とし, 独立同一分布に従うとする. また, $i = 1, \dots, n$ に対して, $\bar{F}(t) = 1 - F(t) = P(X_i > t)$ を時刻 t でのコンポーネントの信頼度とする. このとき, システムの信頼度関数は様々な表現で与えられる.

まず, $i = 0, 1, 2, \dots, n$ に対して, $r_i(n)$ を次のように定義する.

$$(2.3) \quad r_i(n) \equiv \left| \left\{ x \mid \varphi(x) = 1 \text{ and } \sum_{j=1}^n (1 - x_j) = i \right\} \right|$$

ただし, $|\cdot|$ は集合の要素数を表す. すなわち, $r_i(n)$ は n 個のコンポーネントを持つシステムにおいて, ちょうど i 個のコンポーネントが故障する場合にシステムが稼働状態となるシステムの状態数(以下, 稼働状態数)を意味する. 本論文では, 簡単のため, $r_i(n)$ は r_i と書くことがある. このとき, 寿命 T をもつシステムの信頼度関数は次のように表現できる.

$$(2.4) \quad P(T > t) = \sum_{i=0}^{n-1} r_i(\bar{F}(t))^i (\bar{F}(t))^{n-i}$$

次に, 独立同一分布に従う寿命をもつ n 個のコンポーネントからなるシステムが時刻 t_0 で稼働している確率を考える. このとき, $p = \bar{F}(t_0)$ および $q = F(t_0)$ とおくと, 式(2.4)から信頼度多項式 $h(q)$ を導出できる.

$$(2.5) \quad h(q) = \sum_{i=0}^{n-1} r_i q^i p^{n-i}$$

また, 包除原理(Inclusion-exclusion principle)を用いて, 信頼度関数を導出することもできる. コヒーレントシステムにおいて, 極小カットを次のように定義する.

$$(2.6) \quad \mathcal{C} = \{C_1, C_2, \dots, C_m\}$$

ここで, m は極小カットの数である. また, \tilde{C}_i ($i = 1, 2, \dots, m$) を極小カット C_i 内の全てのコンポーネントが故障する事象とする. 以後, 本論文では「事象 \tilde{C}_i が生起する」ことを「極小カット C_i が生起する」と記述することもある. このとき, システムの信頼度 R は

$$(2.7) \quad R = 1 - P \left(\bigcup_{i=1}^m \tilde{C}_i \right)$$

で得ることができ、包除原理を適用することで式(2.7)は

$$(2.8) \quad R = 1 - \sum_{i=1}^{2^m-1} (-1)^{|I_i|+1} P\left(\bigcap_{j \in I_i} \tilde{C}_j\right)$$

と書くことができる(弓削・柳, 2004). ここで I_i は \mathcal{C} の空でない部分集合である.

式(2.8)の各項は全て積事象の確率であり、それらは容易に求めることができるために、理論上システム信頼度は式(2.8)で算出できることになる. 例えば、Cowell (2015) は包除原理を用いて $\text{Con}/k/n:\text{F}$ システムを d 次元に拡張したシステムの信頼度関数を導出した. しかし、各項の数は極小カット数に対して指数的に増えるため、システムサイズが大きくなると信頼度算出時間も指数的に増えることになる.

ここで、式(2.8)は多くの互いに打ち消しあう項を含んでいることに注目し、それらを取り除くと、

$$(2.9) \quad R = 1 + \sum_i d_i P(E_i)$$

と書くことができる. ここで、 E_i は $\{C_1, C_2, \dots, C_m\}$ の部分集合に含まれるコンポーネントがすべて故障する事象であり、係数 d_i は domination と呼ばれる. 式(2.9)は式(2.8)を使用した場合に比べ、同じ事象の確率を複数回求める必要がないため、効率的にシステム信頼度を求めることができる.

2.3 System signature の定義

本節では、system signature の定義を述べる. n 個のコンポーネントからなるシステムの故障時間を T 、各コンポーネントの故障時間を X_1, X_2, \dots, X_n とする. また、 $X_{i:n}$ は X_1, \dots, X_n の中で i 番目に短い故障時間(i -out-of- n :F システムの故障時間)を表す.

このとき、Samaniego (1985) は独立同一分布に従う寿命をもつ n 個のコンポーネントを持つコヒーレントシステムの信頼度関数は次式で表されることを示した.

$$(2.10) \quad \begin{aligned} P(T > t) &= \sum_{i=1}^n P(T > t, T = X_{i:n}) \\ &= \sum_{i=1}^n P(T > t | T = X_{i:n}) P(T = X_{i:n}) \\ &= \sum_{i=1}^n s_i P(X_{i:n} > t) \\ &= \sum_{i=1}^n s_i \sum_{j=0}^{i-1} \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j} \end{aligned}$$

ここで、式(2.10)に現れる s_i ($i = 1, 2, \dots, n$) は system signature と呼ばれる確率ベクトルの要素であり、次のように定義される.

定義 2 (system signature). コヒーレントシステムにおける n 個のコンポーネントの寿命が独立かつ同一分布であると仮定する. system signature s は、 i 番目のコンポーネントの故障がシステムの故障を引き起こす確率 s_i を要素とする n 次元の確率ベクトルである. すなわち、

$$(2.11) \quad s_i = P(T = X_{i:n}),$$

である.

3. System signature の算出方法

本章では、system signature の算出方法について述べる。

System signature は、定義 2 とは異なる方法で定義することもできる。Kochar et al. (1999) は $i = 1, 2, \dots, n$ に対して、次のように定義している。

$$(3.1) \quad s_i = \frac{\text{システム故障が } i \text{ 番目のコンポーネント故障によって引き起こされる順列の数}}{n!}.$$

ここで、コンポーネント寿命が独立同一分布に従うと仮定されているため、これらの n 個の異なる故障時間の順列はすべて等確率で発生することに注意する。

ここで、式(3.1)を用いた system signature の計算例を示す。例として、Con/2/4:F システムを考える。このシステムの 4 つのコンポーネントの故障時間 X_1, X_2, X_3, X_4 は $4! = 24$ 通りの順序で並べることができ、独立同一分布の仮定により、これら 24 通りの順列は等確率で発生する。故障を引き起こすコンポーネント故障時間の順序とシステム寿命の関係を表 1 に示す。したがって、Con/2/4:F システムの system signature は

$$(3.2) \quad s = \left(0, \frac{12}{24}, \frac{12}{24}, 0 \right) = \left(0, \frac{1}{2}, \frac{1}{2}, 0 \right)$$

である。

一般的に、system signature はコンポーネント数が少ない、もしくは、システム構造が直列

表 1. Con/2/4:F システムにおけるコンポーネント故障時間の順序とシステム寿命 T .

コンポーネント故障時間の順序	システム寿命 T に等しい順序統計量
$X_1 < X_2 < X_3 < X_4$	$X_{2:4}$
$X_1 < X_2 < X_4 < X_3$	$X_{2:4}$
$X_1 < X_3 < X_2 < X_4$	$X_{3:4}$
$X_1 < X_3 < X_4 < X_2$	$X_{3:4}$
$X_1 < X_4 < X_2 < X_3$	$X_{3:4}$
$X_1 < X_4 < X_3 < X_2$	$X_{3:4}$
$X_2 < X_1 < X_3 < X_4$	$X_{2:4}$
$X_2 < X_1 < X_4 < X_3$	$X_{2:4}$
$X_2 < X_3 < X_1 < X_4$	$X_{2:4}$
$X_2 < X_3 < X_4 < X_1$	$X_{2:4}$
$X_2 < X_4 < X_1 < X_3$	$X_{3:4}$
$X_2 < X_4 < X_3 < X_1$	$X_{3:4}$
$X_3 < X_1 < X_2 < X_4$	$X_{3:4}$
$X_3 < X_1 < X_4 < X_2$	$X_{3:4}$
$X_3 < X_2 < X_1 < X_4$	$X_{2:4}$
$X_3 < X_2 < X_4 < X_1$	$X_{2:4}$
$X_3 < X_4 < X_1 < X_2$	$X_{2:4}$
$X_3 < X_4 < X_2 < X_1$	$X_{2:4}$
$X_4 < X_1 < X_2 < X_3$	$X_{3:4}$
$X_4 < X_1 < X_3 < X_2$	$X_{3:4}$
$X_4 < X_2 < X_1 < X_3$	$X_{3:4}$
$X_4 < X_2 < X_3 < X_1$	$X_{3:4}$
$X_4 < X_3 < X_1 < X_2$	$X_{2:4}$
$X_4 < X_3 < X_2 < X_1$	$X_{2:4}$

や並列のようにシンプルなものでない限り，算出するのは困難であることが知られている (Da et al., 2012)。そのため，効率的な system signature の算出方法が求められる。

3.1 稼働状態数を用いた system signature の計算

本節で，system signature を，稼働状態数を用いて計算する方法を示す (Boland, 2001)。

$a_i \equiv \binom{r_i}{i}$ とすると，式(2.5)は

$$(3.3) \quad h(q) = \sum_{j=0}^n a_j \binom{n}{j} q^j p^{n-j}$$

と書ける。

一方で，式(2.10)より，和の順序を入れ替えることで，

$$(3.4) \quad \begin{aligned} P(T > t) &= \sum_{j=0}^{n-1} \left(\sum_{i=j+1}^n s_i \right) \binom{n}{j} (F(t))^j (\bar{F}(t))^{n-j} \\ &= \sum_{j=0}^{n-1} \left(\sum_{i=j+1}^n s_i \right) \binom{n}{j} q^j p^{n-j} \end{aligned}$$

となる。

式(3.3)と式(3.4)の係数を比較することで， $i = 0, \dots, n-1$ に対して

$$(3.5) \quad a_i = \sum_{j=i+1}^n s_j$$

を得る。また，式(3.5)から，次のような s_i と a_i の関係式が導かれる (Boland, 2001)。

$$(3.6) \quad s_{i+1} = a_i - a_{i+1}$$

ただし， $a_{n+1} = 0$ とする。すなわち，「system signature を求める問題」は「稼働状態数 r_i を求める問題」に帰着する。

3.1.1 Con/k/n:F システムにおける稼働状態数の算出

本項では，Con/k/n:F システムにおける稼働状態数の算出について述べる。独立同一分布に従う寿命をもつコンポーネントからなる Con/k/n:F システムの信頼度 $R_L(k, n)$ は次式で表される。

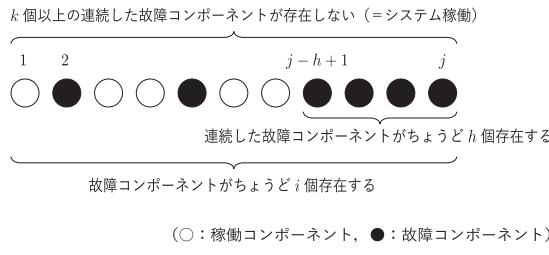
$$(3.7) \quad R_L(k, n) = \sum_{i=0}^n r_i(k, n) q^i p^{n-i}$$

ここで， $r_i(k, n)$ は k 個以上の連続した故障コンポーネントが存在しないような i 個の故障コンポーネントの並べ方の数を表す。Bollinger and Salvia (1982) は， $r_i(k, n)$ を次のように与えた。

$$(3.8) \quad r_i(k, n) = \begin{cases} 0, & i = n \geq k \text{ の場合} \\ \binom{n}{i}, & 0 \leq i \leq k-1 \text{ の場合} \\ \sum_{j=1}^k r_{i-j+1}(k, n-j), & k \leq i < n \text{ の場合} \end{cases}$$

3.1.2 ピボット分解法を用いた Con/k/n:F システムにおける稼働状態数の算出

本項では，Nakamura and Yamamoto (2022)に基づいて，式(3.8)とは異なる再帰方程式を用

図 2. $r_i(k, j; h)$ の説明.

いてシステムの稼働状態数を導出する. Nakamura and Yamamoto (2022) は $\text{Con}/k/n:\text{F}$ システムを 2 次元平面に拡張したシステムに対して、再帰的な稼働状態数の算出方法を提案したが、この方法は $\text{Con}/k/n:\text{F}$ システムに対しても適用可能である。この方法の基礎となっているのは、ピボット分解法 (Kuo and Zuo, 2003) である。ピボット分解法により、選択したコンポーネントの状態に基づいて信頼度を分解し、複雑なシステムの信頼度解析を体系的に行うことができる。

まず、次の記号を定義する。 $j = 1, 2, \dots, n$ に対して、 x_j を $\text{Con}/k/n:\text{F}$ システムの左から数えて j 番目のコンポーネントが稼働しているときに 0、故障したときに 1 となる 2 値変数とする。 $r_i(k, j; h)$ は、図 2 に示す、

- (1) 故障コンポーネントがちょうど i 個存在する
- (2) 右端(コンポーネント j 番目)から見て連続した故障コンポーネントがちょうど h 個存在する(すなわち、コンポーネント $j-h+1$ から j までが故障している)
- (3) k 個以上の連続した故障コンポーネントが存在しない(すなわち、システムは稼働している)

を満たす $\text{Con}/k/j:\text{F}$ システム内の j 個のコンポーネントの状態の総数である。

これらの記号を用いて、次の定理により $\text{Con}/k/n:\text{F}$ システムの稼働状態数 $r_i(k, n)$ を得ることができる。

定理 1. (a) $j = 1, 2, \dots, n$, $i = 0, 1, \dots, j$, および, $h = 0, 1, \dots, k-1$ に対して,

$$(3.9) \quad r_i(k, j; h) = \begin{cases} \sum_{g=0}^{k-1} r_i(k, j-1; g), & h = 0 \text{ の場合} \\ r_{i-1}(k, j-1; h-1), & 1 \leq h \leq k-1 \text{ の場合} \\ 0, & \text{それ以外} \end{cases}$$

であり、境界条件は、 $i = 0, 1, \dots, n$ および $h = 0, 1, \dots, n$ に対して、

$$(3.10) \quad r_i(k, 0; h) = \begin{cases} 1, & i = 0 \text{ かつ } h = 0 \text{ の場合} \\ 0, & \text{それ以外} \end{cases}$$

で与えられる。

(b) $i = 0, 1, \dots, n$ に対して、

$$(3.11) \quad r_i(k, n) = \sum_{h=0}^{k-1} r_i(k, n; h)$$

が成り立つ。

式(3.9)は、 $r_i(k, j; h)$ を次の 3 つの場合に分類して計算している。

• $h = 0$ の場合：この場合は、 j 番目のコンポーネントが稼働していることを意味する。右端の連続する故障数がゼロとなるため、 $(j-1)$ 番目までの状態で右端の連続故障数が g 個 ($0 \leq g \leq k-1$) である全ての状態の合計を取っている。これは、稼働コンポーネントが追加されることで、右端の連続故障数がリセットされることを意味している。

• $1 \leq h \leq k-1$ の場合：この場合は、 j 番目のコンポーネントが故障しており、右端に連続して h 個の故障コンポーネントが存在する場合である。故障コンポーネントが 1 つ増加し、右端の連続故障数も 1 つ増加するため、 $(j-1)$ 番目までの状態で故障コンポーネント数が $(i-1)$ 個、右端の連続故障数が $(h-1)$ 個である状態から導かれる。

• 上記以外の場合：これは、連続した故障コンポーネント数が k 個以上となり、システム故障となる場合や、論理的に存在し得ない状態を表している。したがって、そのような稼働状態数はゼロと定義する。

式(3.11)により、 $\text{Con}/k/n:\text{F}$ システムにおいて、故障コンポーネントがちょうど i 個である稼働状態数を求める。右端の連続故障数 h が 0 から $(k-1)$ までの全ての可能な状態について和を取ることで、システムが連続する k 個の故障を含まない条件下での全ての稼働状態数を計算している。

以上のように、定理 1 は、ピボット分解法 (Kuo and Zuo, 2003) の考え方を用いて、右端のコンポーネントの状態で場合分けを行うことで、稼働状態数を再帰的に計算する方法を与えている。この再帰的な計算により、大規模なシステムにおいても稼働状態数を効率的に計算することが可能となる。

3.1.3 計算例： $\text{Con}/2/5:\text{F}$ システムの場合

本項では、 $\text{Con}/k/n:\text{F}$ システムの system signature を算出する手順について述べる。例として、 $\text{Con}/2/5:\text{F}$ システムを考える。まず、以下のように、定理 1 を用いて稼働状態数 $r_i(2, 5)$ ($i = 0, 1, 2, 3, 4, 5$) を計算する。

ステップ 1：初期条件の設定。まず、式(3.10)から次のように境界条件を求める。

$$(3.12) \quad r_i(2, 0; h) = \begin{cases} 1, & i = 0 \text{ かつ } h = 0 \text{ の場合} \\ 0, & \text{それ以外} \end{cases}$$

ステップ 2：再帰的な $r_i(2, j; h)$ の計算。次に、式(3.9)から次のように $r_i(2, j; h)$ を再帰的に計算する。

$j = 1$ の場合：

$$r_0(2, 1; 0) = r_0(2, 0; 0) = 1$$

$$r_1(2, 1; 1) = r_0(2, 0; 0) = 1$$

$j = 2$ の場合：

$$r_0(2, 2; 0) = r_0(2, 1; 0) = 1$$

$$r_1(2, 2; 0) = r_1(2, 1; 0) + r_1(2, 1; 1) = 0 + 1 = 1$$

$$r_1(2, 2; 1) = r_0(2, 1; 0) = 1$$

$$r_2(2, 2; 1) = r_1(2, 1; 0) = 0$$

$j = 3$ の場合 :

$$\begin{aligned} r_0(2, 3; 0) &= r_0(2, 2; 0) = 1 \\ r_1(2, 3; 0) &= r_1(2, 2; 0) + r_1(2, 2; 1) = 1 + 1 = 2 \\ r_1(2, 3; 1) &= r_0(2, 2; 0) = 1 \\ r_2(2, 3; 1) &= r_1(2, 2; 0) = 1 \\ r_3(2, 3; 1) &= r_2(2, 2; 0) = 0 \end{aligned}$$

$j = 4$ の場合 :

$$\begin{aligned} r_0(2, 4; 0) &= r_0(2, 3; 0) = 1 \\ r_1(2, 4; 0) &= r_1(2, 3; 0) + r_1(2, 3; 1) = 2 + 1 = 3 \\ r_2(2, 4; 0) &= r_2(2, 3; 0) + r_2(2, 3; 1) = 0 + 1 = 1 \\ r_1(2, 4; 1) &= r_0(2, 3; 0) = 1 \\ r_2(2, 4; 1) &= r_1(2, 3; 0) = 2 \\ r_3(2, 4; 1) &= r_2(2, 3; 0) = 0 \end{aligned}$$

$j = 5$ の場合 :

$$\begin{aligned} r_0(2, 5; 0) &= r_0(2, 4; 0) = 1 \\ r_1(2, 5; 0) &= r_1(2, 4; 0) + r_1(2, 4; 1) = 3 + 1 = 4 \\ r_2(2, 5; 0) &= r_2(2, 4; 0) + r_2(2, 4; 1) = 1 + 2 = 3 \\ r_1(2, 5; 1) &= r_0(2, 4; 0) = 1 \\ r_2(2, 5; 1) &= r_1(2, 4; 0) = 3 \\ r_3(2, 5; 1) &= r_2(2, 4; 0) = 1 \end{aligned}$$

ステップ 3 : 稼働状態数の和の計算. 最後に, 式(3.11)を用いて, $i = 0, 1, 2, 3, 4, 5$ に対して, $r_i(2, 5) = r_i(2, 5; 0) + r_i(2, 5; 1)$ を計算する. 具体的には,

$$\begin{aligned} r_0(2, 5) &= r_0(2, 5; 0) + r_0(2, 5; 1) = 1 + 0 = 1 \\ r_1(2, 5) &= r_1(2, 5; 0) + r_1(2, 5; 1) = 4 + 1 = 5 \\ r_2(2, 5) &= r_2(2, 5; 0) + r_2(2, 5; 1) = 3 + 3 = 6 \\ r_3(2, 5) &= r_3(2, 5; 0) + r_3(2, 5; 1) = 0 + 1 = 1 \\ r_4(2, 5) &= r_4(2, 5; 0) + r_4(2, 5; 1) = 0 + 0 = 0 \\ r_5(2, 5) &= 0 \end{aligned}$$

この結果から, Con/2/5:F システムの稼働状態数を求めることができた.

次に, 求めた稼働状態数から system signature を求める. まず, $i = 0, 1, 2, 3, 4, 5$ に対して, $a_i = \frac{r_i(2, 5)}{\binom{5}{i}}$ より,

$$(3.13) \quad (a_0, a_1, a_2, a_3, a_4, a_5) = \left(1, 1, \frac{3}{5}, \frac{1}{10}, 0, 0\right)$$

であるので, 式(3.6)より,

$$(3.14) \quad (s_1, s_2, s_3, s_4, s_5) = \left(0, \frac{2}{5}, \frac{1}{2}, \frac{1}{10}, 0\right)$$

が得られた。

3.2 Domination を用いた system signature の計算

前節では、稼働状態数から system signature を計算する方法を説明したが、本節では、domination から system signature を計算する (Boland et al., 2003)。Boland et al. (2003) では、信頼度関数を p の多項式として表現しているが、 $\text{Con}/k/n\text{F}$ システムの場合は q の多項式としたほうが都合がよい。そこで、式(2.9)より、独立同一分布に従う寿命をもつ n 個のコンポーネントを持つコヒーレントシステムの信頼度関数 $h(q)$ を次式で与える。

$$(3.15) \quad h(q) = \sum_{r=0}^n d_r q^r$$

一方、式(3.3)において、 $p + q = 1$ より、

$$(3.16) \quad h(q) = \sum_{j=0}^{n-1} a_j \binom{n}{j} q^j (1-q)^{n-j}$$

となる。次に、二項定理を用いて $(1-q)^{n-j}$ を展開すると

$$(1-q)^{n-j} = \sum_{i=0}^{n-j} \binom{n-j}{i} (-1)^i q^i$$

が得られ、これを式(3.16)に代入すると、

$$(3.17) \quad h(q) = \sum_{j=0}^{n-1} a_j \binom{n}{j} q^j \left(\sum_{i=0}^{n-j} \binom{n-j}{i} (-1)^i q^i \right)$$

となる。内側の総和を展開して整理すると、

$$(3.18) \quad \begin{aligned} h(q) &= \sum_{j=0}^{n-1} \sum_{i=0}^{n-j} a_j \binom{n}{j} \binom{n-j}{i} (-1)^i q^{j+i} \\ &= \sum_{r=0}^n \left(\sum_{j=0}^{\min(r, n-1)} a_j \binom{n}{j} \binom{n-j}{r-j} (-1)^{r-j} \right) q^r \end{aligned}$$

を得る。

式(3.15)と式(3.18)の係数を比較することで、 $r = 0, 1, \dots, n$ に対して

$$(3.19) \quad d_r = \sum_{j=0}^{\min(r, n-1)} a_j \binom{n}{j} \binom{n-j}{r-j} (-1)^{r-j}$$

を得る。ここで、 $\mathbf{d} = (d_0, d_1, d_2, \dots, d_n)^\top$ とする。ただし、 $^\top$ は転置を表す。このとき、式(3.19)より、 \mathbf{d} と \mathbf{a} は線形関係にあることがわかるため、その関係は

$$(3.20) \quad \mathbf{d} = \mathbf{M}\mathbf{a}$$

と表現できる。ここで、 $\mathbf{a} = (a_0, a_1, a_2, \dots, a_n)^\top$ であり、 \mathbf{M} は次のように定義される $(n+1) \times (n+1)$ 行列である。

$$(3.21) \quad m_{rj} = \begin{cases} \binom{n}{j} \binom{n-j}{r-j} (-1)^{r-j}, & r \geq j \text{ の場合} \\ 0, & \text{それ以外} \end{cases}$$

ここで, $r, j = 0, 1, \dots, n$ である. また, 逆行列 \mathbf{M}^{-1} が存在すれば,

$$(3.22) \quad \mathbf{a} = \mathbf{M}^{-1} \mathbf{d}$$

として \mathbf{d} から \mathbf{a} を求めることも可能である. したがって, 式(3.6)と式(3.22)を用いることで, domination から system signature を算出することも可能である.

3.2.1 Con/k/n:F システムにおける domination の算出

本項では, Con/k/n:F システムにおける domination の算出について述べる. Kossow and Preuss (1989) は Con/k/n:F システムに関する domination に関する定理を示した.

定理 2. 極小カットで表現可能なシステム故障状態 X の domination を $d(X)$, 連続して故障したコンポーネント数を m とすると, 極小カットで表現可能なシステムの故障状態の domination は次式で求めることができる.

$$(3.23) \quad d(X) = \begin{cases} +1, & m \equiv 0 \pmod{k+1} \text{ の場合} \\ -1, & m \equiv k \pmod{k+1} \text{ の場合} \\ 0, & \text{それ以外} \end{cases}$$

ここで, mod は m を $(k+1)$ で割った場合の余りを表している.

式(3.23)により, 極小カットが繋がるか重なるかして表されている故障状態についての domination を求めることができる.

また, 極小カットセット同士が互いに重なりがない(非連結である)場合は, 次の定理によって domination を求めることができる.

定理 3 (極小カットセットが非連結である場合の domination). 故障状態 X は非連結な g 個のカットセット $X^{(i)}$ ($i = 1, 2, \dots, g$) からなるとすると, 故障状態 X の domination $d(X)$ は次式で得られる.

$$(3.24) \quad d(X) = \prod_{i=1}^g d(X^{(i)})$$

最終的に domination が 0 でない極小カットの組合せをすべて算出し, 各極小カットの組合せが生起する確率に domination を掛けた値の総和によってシステム信頼度が算出できる.

3.2.2 計算例: Con/2/5:F システムの場合

本項では, Con/k/n:F システムの system signature を算出する手順について述べる. 例として, Con/2/5:F システムを考える. 表 2 に, 各極小カットで表現可能なシステム故障状態に対する $d(X_i)$ ($i = 1, 2, \dots, 11$) を示す. 例えば, X_5 に着目すると, 5 個のコンポーネントが連続して故障しており, $m (= 5)$ を $k+1 (= 3)$ で割った場合の余りは 2 であるため, $d(X_5) = -1$ である. X_1 に着目すると, X_1 は非連結な 2 個のカットセット $X_1^{(1)}$ と $X_1^{(2)}$ からなる. $X_1^{(1)}$ と $X_1^{(2)}$ は $m = 2$ なので, $d(X_1^{(1)}) = d(X_1^{(2)}) = +1$ である. したがって,

$$(3.25) \quad d(X_1) = d(X_1^{(1)}) \times d(X_1^{(2)}) = +1$$

である.

Con/2/5:F システムのコンポーネントが共通したコンポーネント故障確率 q をもつとき, 信頼度関数 $R_L(2, 5)$ は式(3.15)より次式で表される.

$$(3.26) \quad R_L(2, 5) = 1 - 4q^2 + 3q^3 + q^4 - q^5$$

表 2. 故障状態と domination の対応.

故障状態	x_1	x_2	x_3	x_4	x_5	$d(X_i)$
X_1	1	1	0	1	1	+1
X_2	1	1	0	0	0	-1
X_3	1	1	1	0	0	+1
X_4	1	1	1	1	0	0
X_5	1	1	1	1	1	-1
X_6	0	1	1	0	0	-1
X_7	0	1	1	1	0	+1
X_8	0	1	1	1	1	0
X_9	0	0	1	1	0	-1
X_{10}	0	0	1	1	1	+1
X_{11}	0	0	0	1	1	-1

よって、 $d = (1, 0, -4, 3, 1, -1)^\top$ である。

式(3.21)より、行列 M は次のようにになる。

$$(3.27) \quad M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -5 & 5 & 0 & 0 & 0 & 0 \\ 10 & -20 & 10 & 0 & 0 & 0 \\ -10 & 30 & -30 & 10 & 0 & 0 \\ 5 & -20 & 30 & -20 & 5 & 0 \\ -1 & 5 & -10 & 10 & -5 & 1 \end{pmatrix}.$$

式(3.22)より、 a は次のように求まる。

$$a = \left(1, 1, \frac{3}{5}, \frac{1}{10}, 0, 0 \right)^\top.$$

これは、式(3.13)と一致する。よって、3.1.3 項と同様に Con/2/5:F システムの system signature が計算でき、式(3.14)を得ることができる。

3.3 他の system signature 算出方法

上記では、稼働状態数や domination を用いた system signature の算出方法について述べた。本節では、これらとは異なるアプローチによる system signature 算出方法を紹介する。

まず、一般的なシステムに対する計算手法として、Reed (2017) は縮約二分決定図(Reduced Ordered Binary Decision Diagram, ROBDD)を用いた効率的なアルゴリズムを提案した。これは従来の列挙的アプローチと比較して、計算効率を大幅に改善している。

特定のシステム構造に対する計算手法も開発されている。Con/ k/n :F システムについては、Triantafyllou and Koutras (2008) が母関数(generating function)アプローチを用いて system signature の再帰方程式を導出した。さらに、Yi and Cui (2018) は、マルコフ過程に基づく新しい算出方法を提案している。これらの手法は、特定のシステム構造に特化することで、より効率的な計算を実現している。

大規模システムを複数のサブシステムに分割して、system signature を計算する方法も提案されてきた。Da et al. (2012) は、2 つのサブシステムからなるコヒーレントシステムについて、それぞれのサブシステムの system signature を用いて、直列および並列構造の system signature

を計算する手法を提案した。その後, Da et al. (2014) は, この手法を一般化し, n 個のサブシステムからなる k -out-of- n システムの system signature の計算方法を確立した。このようなサブシステムに基づく計算手法は, 他の手法と組み合わせることで計算過程を大幅に簡略化できるという利点を持つが, システムをサブシステムに分割できる場合に限り適用可能である。

4. 故障コンポーネント数の期待値算出

本章では, system signature の理論的応用の 1 つとして, 故障コンポーネント数の期待値について述べる。故障コンポーネント数の期待値を求ることで, 保守計画の最適化や予備のコンポーネント管理のための定量的な基準を提供することが可能となる。

4.1 パス列挙による故障コンポーネント数の期待値の算出

本節では, Endharta et al. (2016) によって提案された, システムが故障するパスを列挙することで故障コンポーネント数の期待値を算出する方法を説明する。この方法は system signature を用いないが, 後述する system signature を用いた算出方法との比較のため, 本論文で取り上げる。

まず, システム故障時の故障コンポーネント数の期待値について述べる。システムの可能なパスの総数を P , パス j に従ってシステムが故障する確率を π_j , パス j における故障までのステップ数を W_j とする。このとき, システム故障時の故障コンポーネント数を表す確率変数 M の期待値は次式で与えられる。

$$(4.1) \quad E[M] = \sum_{j=1}^P \pi_j \cdot W_j$$

次に, 時刻 t 以前にシステム故障が発生する場合の故障コンポーネント数の期待値について述べる。時刻 t 以前のシステム故障時の故障コンポーネント数を表す確率変数を $M(t)$ とし, α_{ji} をパス j におけるステップ i の稼働コンポーネントの故障率の和とする。このとき, $M(t)$ の期待値は次のように表される。

$$(4.2) \quad E[M(t)] = \sum_{j=1}^P \pi_j \cdot W_j \cdot F_j(t)$$

ここで,

$$(4.3) \quad F_j(t) = 1 - \sum_{i=0}^{W_j-1} A_{ji} e^{-\alpha_{ji} t}$$

$$(4.4) \quad A_{ji} = \prod_{m=0, m \neq i}^{W_j-1} \frac{\alpha_{jm}}{\alpha_{jm} - \alpha_{ji}}$$

である ($i = 0, 1, \dots, W_j - 1$).

最後に, 時刻 t でシステムが稼働している場合における故障コンポーネント数の期待値について述べる。時刻 t においてシステムが稼働している場合における故障コンポーネント数を表す確率変数を $S(t)$ とすると, その期待値は次式で与えられる。

$$(4.5) \quad E[S(t)] = \sum_{j=1}^P \pi_j \sum_{i=0}^{W_j-1} \sum_{m=0}^{i-1} i A_{jm} \frac{\alpha_{jm}}{\alpha_{ji} - \alpha_{jm}} (e^{-\alpha_{jm} t} - e^{-\alpha_{ji} t})$$

ここで,

$$(4.6) \quad A_{jm} = \prod_{l=0, l \neq m}^{i-1} \frac{\alpha_{jl}}{\alpha_{jl} - \alpha_{jm}}$$

である。

例として、故障率 λ の指数分布に従う寿命をもつコンポーネントからなる Con/2/4:F システムを考える。図 3 に Con/2/4:F システムにおける故障までのパスを示す。ここでは、1 はコンポーネント稼働、0 はコンポーネント故障を表す。例えば、Path 2 ((1, 1, 1, 1) → (1, 1, 1, 0) → (1, 0, 1, 0) → (1, 0, 0, 0)) は最初に左から 4 番目のコンポーネント、次に左から 2 番目のコンポーネント、最後に左から 3 番目のコンポーネントが順番に故障するパスを表す。図 3 に示すように、Con/2/4:F システムは 18 個のパスを持つ。このとき、表 3 の値を用いると、以下の結果が得られる。

$$\begin{aligned} E[M] &= \frac{5}{2} \\ E[M(t)] &= \frac{5}{2} - \frac{3}{2}e^{-4\lambda t} + 8e^{-3\lambda t} - 9e^{-2\lambda t} \\ E[S(t)] &= 2e^{-4\lambda t} - 8e^{-3\lambda t} + 6e^{-2\lambda t} \end{aligned}$$

4.2 System signature を用いた故障コンポーネント数の期待値の算出

前節で紹介したパス列挙法は故障コンポーネント数の算出を可能とするが、各パスに対して W_j と α_{ji} を計算する必要があり、コンポーネント数の増加に伴い計算量が著しく増大する。この課題に対し、Eryilmaz (2012) と Zhou et al. (2020) は system signature を用いた算出方法

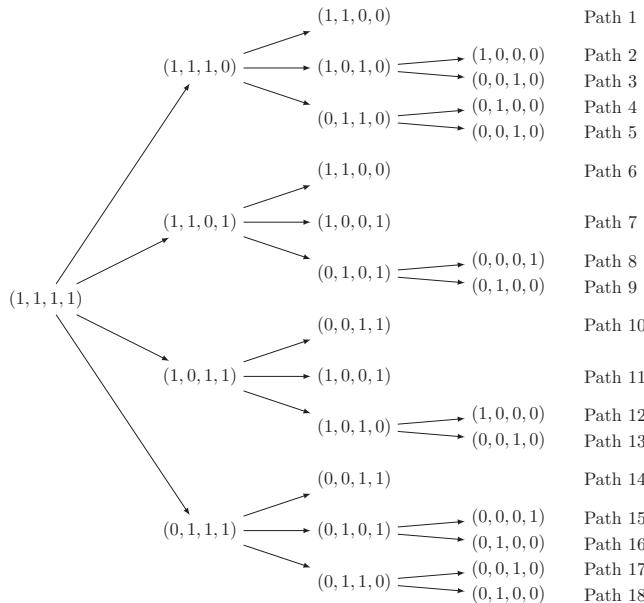


図 3. Con/2/4:F システムにおける故障までのパス。

表 3. Con/2/4:F システムの各パスに対する W_j と π_j .

Path	j	W_j	π_j	Path	j	W_j	π_j
1	2	1/12		10	2	1/12	
2	3	1/24		11	2	1/12	
3	3	1/24		12	3	1/24	
4	3	1/24		13	3	1/24	
5	3	1/24		14	2	1/12	
6	2	1/12		15	3	1/24	
7	2	1/12		16	3	1/24	
8	3	1/24		17	3	1/24	
9	3	1/24		18	3	1/24	

を導出した. 本手法は system signature が得られれば, パスの列挙を必要とせずに期待値を計算できるという利点を持つ.

まず, Eryilmaz (2012) が導出したシステム故障時の故障コンポーネント数の期待値について述べる. このとき, M の期待値は次式で与えられる.

$$(4.7) \quad E[M] = \sum_{i=1}^n i \cdot s_i$$

次に, Zhou et al. (2020) が導出した時刻 t 以前のシステム故障における故障コンポーネント数の期待値について述べる. $M(t)$ の期待値は以下のように展開される.

$$(4.8) \quad \begin{aligned} E[M(t)] &= \sum_{i=1}^n i \cdot P(T = X_{i:n}, T \leq t) \\ &= \sum_{i=1}^n i \cdot P(T \leq t | T = X_{i:n}) \cdot P(T = X_{i:n}) \\ &= \sum_{i=1}^n i \cdot P(X_{i:n} \leq t) \cdot s_i \\ &= \sum_{i=1}^n i \cdot s_i \sum_{j=i}^n \binom{n}{j} F(t)^j \bar{F}(t)^{n-j} \end{aligned}$$

最後に, Zhou et al. (2020) が導出した時刻 t におけるシステム稼働時の故障コンポーネント数の期待値について述べる. まず, 以下の事象を定義する.

$$(4.9) \quad B_i = \{X_{i:n} \leq t < X_{(i+1):n}\}$$

このとき, 条件付き期待値 $E[S(t)|B_i]$ は, 時刻 t までにちょうど i 個のコンポーネントが故障するという条件下で, 時刻 t でシステムが稼働している場合の故障コンポーネント数の期待値を表す. これを用いると,

$$(4.10) \quad \begin{aligned} E[S(t)] &= \sum_{i=0}^{n-1} E[S(t)|B_i] \cdot P(B_i) \\ &= \sum_{i=0}^{n-1} i \cdot P(T > t | B_i) P(B_i) \end{aligned}$$

と書ける。ここで、時刻 t までにちょうど i 個のコンポーネントが故障した場合のシステム信頼度は次式で表される。

$$(4.11) \quad P(T > t | B_i) = \sum_{j=i+1}^n s_j$$

したがって、式(4.10)と式(4.11)より、以下の結果を得る。

$$(4.12) \quad E[S(t)] = \sum_{i=0}^{n-1} i \cdot \left(\sum_{j=i+1}^n s_j \right) \binom{n}{i} F(t)^i \bar{F}(t)^{n-i}$$

例として、故障率 λ の指数分布に従う寿命をもつコンポーネントからなる Con/2/4:F システムを考える。式(3.2)で算出した system signature を用いて計算すると、

$$\begin{aligned} E[M] &= \frac{5}{2} \\ E[M(t)] &= \frac{5}{2}(1 - e^{-\lambda t})^4 + 10(1 - e^{-\lambda t})^3 e^{-\lambda t} + 6(1 - e^{-\lambda t})^2 e^{-2\lambda t} \\ E[S(t)] &= 4(1 - e^{-\lambda t}) e^{-3\lambda t} + 6(1 - e^{-\lambda t})^2 e^{-2\lambda t} \end{aligned}$$

となる。これらの結果は、Endharta et al. (2016)によるパス列挙法で得られた値と一致する。

Endharta et al. (2016)によるパス列挙法は、システムの全故障パスを探索する必要があり、システムの規模が大きくなるにつれて計算量が指数的に増大するため、大規模システムへの適用には計算時間の点において課題がある。一方、system signature を用いた手法は、system signature が得られれば、パスの列挙を必要とせず、より効率的に期待値を算出できる。

計算例として、図 4 から図 7 に、故障率 $\lambda = 1$ の指数分布に従う寿命をもつコンポーネントからなる Con/ k/n :F システム ($k = 3, 4, 5, 10$) の(a)時刻 t 以前のシステム故障における故障コンポーネント数の期待値 ($E[M(t)]$) と、(b)時刻 t におけるシステム稼働時の故障コンポーネント数の期待値 ($E[S(t)]$) を示す。

5. おわりに

本論文では、Con/ k/n :F システムを中心にコヒーレントシステムに対する system signature の算出方法について概説した。

一般的に、system signature の算出においてシステム構造の情報を利用することで、算出ア

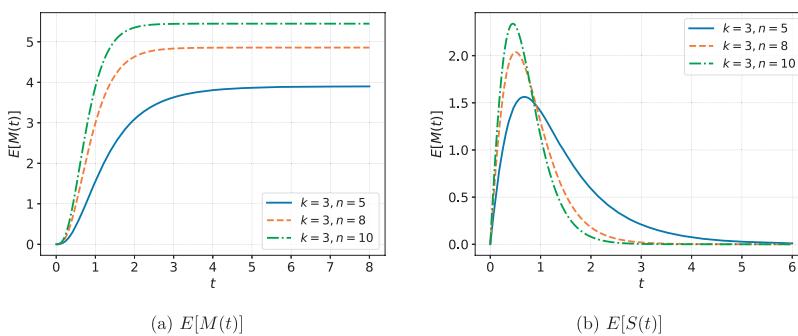
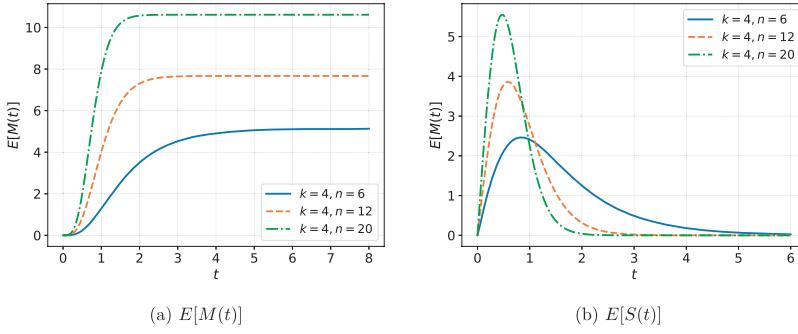
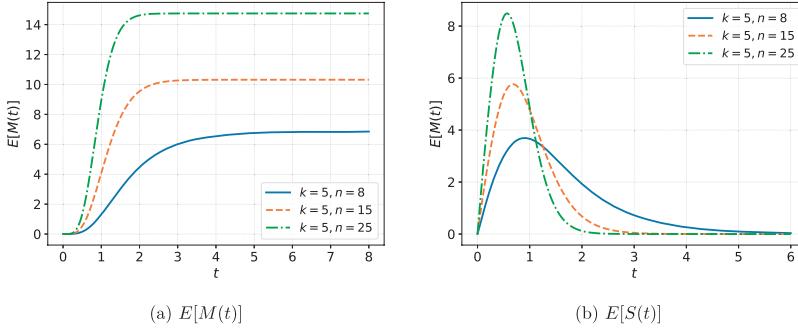
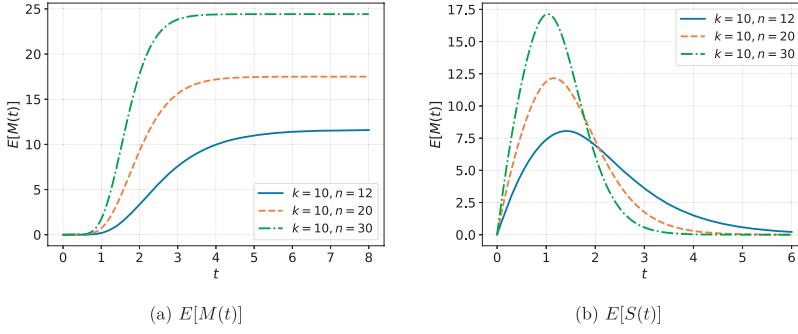


図 4. Con/3/n:F システムにおける故障コンポーネント数の期待値。

図 5. $Con/4/n:F$ システムにおける故障コンポーネント数の期待値.図 6. $Con/5/n:F$ システムにおける故障コンポーネント数の期待値.図 7. $Con/10/n:F$ システムにおける故障コンポーネント数の期待値.

ルゴリズムの性能向上が期待できる。本論文で取り上げた $Con/k/n:F$ システムは，“規則的な”極小カットを持つという特徴的な構造を有しており (山本 他, 2021a), この構造を活用することで効率的な system signature の算出が可能となった。しかし, $Con/k/n:F$ システムのみに特化した手法では, 汎用性に欠けるという課題がある。

実際, 現実のシステムをより忠実に表現するために, $Con/k/n:F$ システムを以下のような方向で一般化したシステムが提案してきた (Cui and Dong, 2019).

- ・故障基準(单一故障基準と複数故障基準)
- ・システムの幾何学的構造
- ・コンポーネントおよびシステムの状態
- ・各コンポーネントの重み
- ・コンポーネント間の従属性

また, Triantafyllou (2015) では, 一般化された $Con/k/n:F$ システムの関係が体系的に図示されている。このような多様なシステムに対して, 個別に算出方法を開発することは現実的とは言えない。したがって, 性能と汎用性の両面をバランスよく実現する方法の開発が求められる。

性能と汎用性のトレードオフに対して, 定理 1 のピボット分解法に基づく方法は一つの解決策を提示していると考える。この手法は Bollinger and Salvia (1982) が導出した式(3.8)と比べて複雑であるものの, $Con/k/n:F$ システムに限らず, “規則的” という特徴を持つシステム群に幅広く適用可能であると考えている。実際に, Nakamura and Yamamoto (2022) は $Con/k/n:F$ システムを 2 次元平面に拡張したシステムに適用した。

将来的には, system signature の様々な算出方法を体系的に整理し, 対象とするシステムの特徴に応じて適切な手法を選択できる枠組みの構築を目指している。これにより, 「規則的な構造を持つシステム群」や「サブシステムに分割可能なシステム群」に限らず, さまざまな特徴に基づきグループ化したシステム群に対しても system signature を効率的に算出することが可能となると考える。このような性能と汎用性をバランスよく実現する手法の確立は, システムの信頼性評価や構造比較, さらには保全計画の立案など, system signature の実践的な応用の発展に大きく貢献すると期待される。

謝 辞

本研究は JSPS 科研費 JP21K14370, JP24K17469 の助成を受けたものです。また, 原稿を注意深くお読みいただき, 適切な助言をいただいたことに対し, 匿名査読者および編集委員に感謝を申し上げます。

参 考 文 献

- Asadi, M., Hashemi, M. and Balakrishnan, N. (2023). An overview of some classical models and discussion of the signature-based models of preventive maintenance, *Applied Stochastic Models in Business and Industry*, **39**(1), 4–53.
- Boland, P. J. (2001). Signatures of indirect majority systems, *Journal of Applied Probability*, **38**(2), 597–603.
- Boland, P., Samaniego, F. and Vestrup, E. (2003). Linking dominations and signatures in network reliability theory, *Mathematical and Statistical Methods in Reliability* (eds. B. H. Lindqvist and K. A. Doksum), 89–103, World Scientific, New Jersey.
- Bollinger, R. and Salvia, A. (1982). Consecutive- k -out-of- $n:F$ networks, *IEEE Transactions on Reliability*, **31**(1), 53–56.
- Chiang, D. T. and Niu, S.-C. (1981). Reliability of consecutive- k -out-of- $n:F$ system, *IEEE Transactions on Reliability*, **30**(1), 87–89.
- Cowell, S. (2015). A formula for the reliability of a d -dimensional consecutive- k -out-of- $n:F$ system, *International Journal of Combinatorics*, **2015**, <https://doi.org/10.1155/2015/140909>.
- Cui, L. and Dong, Q. (2019). Consecutive k and related models — A survey, *Stochastic Models in Reliability, Network Security and System Safety* (eds. Q.-L. Li, J. Wang and H.-B. Yu), 3–18, Springer, Singapore.

- Da, G., Zheng, B. and Hu, T. (2012). On computing signatures of coherent systems, *Journal of Multivariate Analysis*, **103**(1), 142–150.
- Da, G., Xia, L. and Hu, T. (2014). On computing signatures of k -out-of- n systems consisting of modules, *Methodology and Computing in Applied Probability*, **16**(1), 223–233.
- Dafnis, S. D., Makri, F. S. and Philippou, A. N. (2019). The reliability of a generalized consecutive system, *Applied Mathematics and Computation*, **359**, 186–193.
- Endharta, A. J., Yun, W. Y. and Yamamoto, H. (2016). Preventive maintenance policy for linear consecutive- k -out-of- n :F system, *Journal of the Operations Research Society of Japan*, **59**(4), 334–346.
- Eryilmaz, S. (2012). The number of failed components in a coherent system with exchangeable components, *IEEE Transactions on Reliability*, **61**(1), 203–207, <https://doi.org/10.1109/TR.2011.2167781>.
- Kochhar, S., Mukerjee, H. and Samaniego, F. J. (1999). The “Signature” of a coherent system and its application to comparisons among systems, *Naval Research Logistics*, **46**(5), 507–523.
- Kontoleon, J. M. (1980). Reliability determination of a r -successive-out-of- n :F system, *IEEE Transactions on Reliability*, **29**(5), <https://doi.org/10.1109/TR.1980.5220921>.
- Kossow, A. and Preuss, W. (1989). Reliability of consecutive- k -out-of- n :F systems with nonidentical component reliabilities, *IEEE Transactions on Reliability*, **38**(2), 229–233.
- Kuo, W. and Zuo, M. J. (2003). *Optimal Reliability Modeling: Principles and Applications*, John Wiley and Sons, Hoboken.
- Nakamura, T. and Yamamoto, H. (2022). Algorithm for computing the system signatures of linear and circular connected-(1, 2)-or-(2, 1)-out-of-(m, n):F lattice systems, *International Journal of Reliability, Quality and Safety Engineering*, **29**(04), <https://doi.org/10.1142/S0218539321410072>.
- Naqvi, S., Chan, P. S. and Mishra, D. B. (2022). System signatures: A review and bibliometric analysis, *Communications in Statistics — Theory and Methods*, **51**(7), 1993–2008.
- 大鑄史男 (2019). 『システム信頼性の数理』, シリーズ情報科学における確率モデル, コロナ社, 東京.
- Reed, S. (2017). An efficient algorithm for exact computation of system and survival signatures using binary decision diagrams, *Reliability Engineering & System Safety*, **165**, 257–267.
- Samaniego, F. J. (1985). On closure of the IFR class under formation of coherent systems, *IEEE Transactions on Reliability*, **34**(1), 69–72, <https://doi.org/10.1109/TR.1985.5221935>.
- Samaniego, F. J. (2007). *System Signatures and Their Applications in Engineering Reliability*, Springer, New York.
- Triantafyllou, I. S. (2015). Consecutive-type reliability systems: An overview and some applications, *Journal of Quality and Reliability Engineering*, **2015**(1), <https://doi.org/10.1155/2015/212303>.
- Triantafyllou, I. S. and Koutras, M. V. (2008). On the signature of coherent systems and applications, *Probability in the Engineering and Informational Sciences*, **22**(1), 19–35, <https://doi.org/10.1017/S0269964808000028>.
- 山本久志, 秋葉知昭, 中村太信, 周 蕎 (2021a). システムの信頼度評価と最適設計, 電子情報通信学会基礎・境界ソサイエティ Fundamentals Review, **14**(4), 287–296, https://doi.org/10.1587/essfr.14.4_287.
- 山本久志, 秋葉知昭, 竹ヶ原春貴, 深津 敦, 古井光明 (2021b). 『確率・統計から始めるエンジニアのための信頼性工学: 身近な故障から宇宙開発まで』, コロナ社, 東京.
- Yi, H. and Cui, L. (2018). A new computation method for signature: Markov process method, *Naval Research Logistics*, **65**(5), 410–426.
- 弓削哲史, 柳 繁 (2004). 2次元連続 k -out-of- n :F システムの信頼度下限値, 電子情報通信学会技術研究報告, **104**(141), 1–6.
- Zhao, J., Si, S. and Cai, Z. (2019). A multi-objective reliability optimization for reconfigurable systems considering components degradation, *Reliability Engineering and System Safety*, **183**, 104–115.
- Zhou, L. and Yamamoto, H. (2022). Number of failed components in consecutive- k -out-of- n :G systems and their applications in optimization problems, *IEICE Transactions on Fundamentals of*

- Electronics, Communications and Computer Sciences*, **105**(6), 943–951.
- Zhou, L., Yamamoto, H. and Xiao, X. (2020). Number of failed components in coherent systems, *Journal of Japan Industrial Management Association*, **71**(2E), 92–98.

Computation of the System Signature and Their Application —Case of Consecutive- k -out-of- n :F Systems—

Taishin Nakamura¹ and Lei Zhou²

¹School of Information Science and Technology, Tokai University

²Faculty of Engineering, Nagaoka University of Technology

This paper introduces methods for calculating the system signature of coherent systems, specifically presenting concrete techniques and calculation examples for Consecutive- k -out-of- n :F (Con/ k / n :F) systems. A Con/ k / n :F system is a system model where n components are arranged linearly, and the system fails when at least k consecutive components fail. This model has various real-world applications, such as oil pipelines, highway street lighting, and production monitoring systems. The system signature is a characteristic that depends solely on the system's structure, making it useful for comparing system structures and planning maintenance policies. However, naive calculation methods make it difficult to compute the system signature for large-scale systems. We present a method based on the pivot decomposition method for calculating the system signature of Con/ k / n :F systems. This technique is applicable to systems with regular structures. Subsequently, we discuss the expected number of failed components as one of the theoretical applications of system signatures. Finally, we discuss the potential for a general method to calculate the system signature.

多様体学習を用いた銀河進化の新しい定量化

竹内 努^{1,2}・クレ スチエータ^{3,4}・山形 大青¹・曹 愛奈^{1,6}・内田 舜也¹・
池田 思朗²・福水 健次²・加納 龍生¹・大森 清顕 クリストファ^{1,5}・
馬 海霞¹・施 文¹・松井 瀬奈¹

(受付 2024 年 2 月 13 日；改訂 2025 年 2 月 25 日；採択 2 月 25 日)

要 旨

宇宙に存在する物質は、空間的にはほぼ一様の状態から始まった。その中のかすかに密度の高い領域が重力で収縮し、最終的に銀河へと成長した。宇宙物理学は、この 130 億年を超える宇宙の歴史における銀河の形成と進化を引き起こす複雑な物理現象を、物理の第一原理から解き明かそうと試みてきた。しかし、説明変数が 100 を超える現在の天文学ビッグデータに対し、このような従来の方法は限界を迎えている。そこで我々は、従来の第一原理からの理論構築とは異なる方法でこの問題へのアプローチを行っている。銀河の多波長光度および宇宙年齢が張る高次元特徴空間中での銀河分布に対し、我々はデータ科学で最近発展してきた多様体学習を適用し、銀河の進化の特徴づけを行った。これにより、我々は銀河多様体と呼ばれる、多波長光度空間内のデータ点に埋め込まれた低次元非線型構造を発見した。そして紫外線、光学、近赤外線の光度空間における銀河の進化が、銀河多様体上の星形成と星の質量進化という 2 つのパラメーターによってよく記述されることを発見した。これら銀河多様体座標を物理量に結び付ける方法についても議論する。

キーワード：銀河進化、銀河形成、星形成率、星質量、多波長光度、多様体学習。

1. はじめに

1.1 大規模銀河探査時代の銀河進化研究

銀河とは、星と星間物質(ガスとダストの混合流体)、暗黒物質からなる巨大な天体であり、観測可能な宇宙の範囲に数千億個におよぶ銀河が存在している。宇宙は 138 億年前に誕生したが、初期の宇宙の物質はほぼ一様に分布しており、銀河のような天体は存在していなかった。

¹ 名古屋大学 素粒子宇宙物理学専攻：〒 464-8602 愛知県名古屋市千種区不老町; tsutomu.takeuchi.ttt@gmail.com, yamagata.taisei.p4@s.mail.nagoya-u.ac.jp, so.aina.t6@s.mail.nagoya-u.ac.jp, uchida.shunya.i4@s.mail.nagoya-u.ac.jp, kano.ryusei.z5@s.mail.nagoya-u.ac.jp, mhx11235@gmail.com, shiwenbaobao0223@gmail.com, matsui.sena.x7@s.mail.nagoya-u.ac.jp

² 統計数理研究所：〒 190-8562 東京都立川市緑町 10-3; shiro@ism.ac.jp, fukumizu@ism.ac.jp

³ 国立天文台 科学研究部：〒 181-8588 東京都三鷹市大沢 2-21-1

⁴ スタンフォード大学 カブリ素粒子宇宙物理学・宇宙論研究所：452 Lomita Mall, Stanford, CA 94305-4085; cooray@nagoya-u.jp

⁵ セント・マリー大学 天文学・物理学部：Halifax Nova Scotia, B3H 3C3 Canada; k.omori116@gmail.com

⁶ 学習院大学 理学部物理学科：〒 171-8588 東京都豊島区目白 1-5-1

つまり、銀河は形成し、現在の姿に進化してきた時間的に動的に進化する存在である。時間進化が銀河の本質であり、これを定量化する銀河形成進化の研究は半世紀以上にわたり銀河研究の中心であり続けている。

銀河の進化を物理法則から定量的に説明する試みは 1970 年代に始まった。銀河が単一の巨大なガス雲から形成されたという仮定の下で、星の形成とそれに関連する重元素合成の歴史を扱う理論の開発が試みられた。この方向の研究は 1980 年代前半に Tinsley (1980) によって一旦理論体系としては完成されたものの、これで銀河進化の研究が終了とはならなかった。時を同じくして進められてきた宇宙論研究により、銀河は合体して成長することが明らかになってきた。これは、銀河の進化が周囲の銀河の密度とガス密度に大きく依存する非常に複雑なプロセスであることを示している。このように、銀河の進化は周囲の銀河の密度やガス密度など、銀河が置かれた環境に大きく依存する非常に複雑な過程であることが判明した。新たな銀河進化を記述する方程式は記号的には次のように表される。

$$\begin{aligned}
 \text{SFR}(t) &= f_1(\text{SFR}, \mathcal{M}_*, \mathcal{M}_{\text{mol}}, \mathcal{M}_{\text{HI}}, \mathcal{M}_{\text{dust}}, \mathcal{M}_{\text{halo}}, \delta_{\text{gal}}, \dots), \\
 \mathcal{M}_*(t) &= f_2(\text{SFR}, \mathcal{M}_*, \mathcal{M}_{\text{mol}}, \mathcal{M}_{\text{HI}}, \mathcal{M}_{\text{dust}}, \mathcal{M}_{\text{halo}}, \delta_{\text{gal}}, \dots), \\
 \mathcal{M}_{\text{mol}}(t) &= f_3(\text{SFR}, \mathcal{M}_*, \mathcal{M}_{\text{mol}}, \mathcal{M}_{\text{HI}}, \mathcal{M}_{\text{dust}}, \mathcal{M}_{\text{halo}}, \delta_{\text{gal}}, \dots), \\
 (1.1) \quad \mathcal{M}_{\text{HI}}(t) &= f_4(\text{SFR}, \mathcal{M}_*, \mathcal{M}_{\text{mol}}, \mathcal{M}_{\text{HI}}, \mathcal{M}_{\text{dust}}, \mathcal{M}_{\text{halo}}, \delta_{\text{gal}}, \dots), \\
 \mathcal{M}_{\text{dust}}(t) &= f_5(\text{SFR}, \mathcal{M}_*, \mathcal{M}_{\text{mol}}, \mathcal{M}_{\text{HI}}, \mathcal{M}_{\text{dust}}, \mathcal{M}_{\text{halo}}, \delta_{\text{gal}}, \dots), \\
 \mathcal{M}_{\text{halo}}(t) &= f_6(\text{SFR}, \mathcal{M}_*, \mathcal{M}_{\text{mol}}, \mathcal{M}_{\text{HI}}, \mathcal{M}_{\text{dust}}, \mathcal{M}_{\text{halo}}, \delta_{\text{gal}}, \dots), \\
 \delta_{\text{gal}}(t) &= f_7(\text{SFR}, \mathcal{M}_*, \mathcal{M}_{\text{mol}}, \mathcal{M}_{\text{HI}}, \mathcal{M}_{\text{dust}}, \mathcal{M}_{\text{halo}}, \delta_{\text{gal}}, \dots), \\
 &\vdots
 \end{aligned}$$

ここで $\text{SFR}(t)$, $\mathcal{M}_*(t)$, $\mathcal{M}_{\text{mol}}(t)$, $\mathcal{M}_{\text{HI}}(t)$, $\mathcal{M}_{\text{dust}}(t)$, $\mathcal{M}_{\text{halo}}(t)$, $\delta(t)$ はそれぞれ時刻 t での星形成率、星質量、分子ガス質量、水素原子ガス質量、ダスト質量、暗黒物質ハロー質量、周囲の銀河密度超過をそれぞれ表す。右辺に含まれる変数は象徴的に書かれたもので、それぞれの変数の過去の歴史全てに依存することを表している。

銀河の進化を定式化するには、このような巨大な方程式系を決定する必要がある。天体物理学者はこれまで第一原理の物理法則から支配方程式を構築してきたが、量空間が 10 次元を超えると、そのような方法はもはや現実的ではなくなる。1970 年代から 1980 年代半ばにかけて、主成分分析(PCA)などの古典的な多変量解析手法が、高次元空間で銀河の物理量を結合するために使用された。これにより、さまざまな(対数)線形関係、いわゆる銀河スケーリング関係が発見されている。スケーリング関係を統一して基本的な関係を見つけるための研究により、銀河多様体 (Brosche, 1973; Djorgovski, 1992) の概念が生まれた。しかし、古典的 PCA が扱えるのは線型関係のみであり、銀河の(対数)線型関係を探索的に検証するには今でも有用であるものの、銀河多様体は極めて限定された概念に留まり、一旦ほとんど忘れられた (Hunt et al., 2012; Zhang and Zaritsky, 2016; Ginolfi et al., 2020)。時は流れ、21 世紀の銀河調査では数億個の銀河について数百の物理量が得られ、まさに質・量ともに典型的なビッグデータとなっている。解析対象となる銀河の特徴空間は 100 次元を超える。したがって、銀河進化の特徴づけは、物理的直観に頼った従来の方法では不可能で、根本的に異なった新たな発想による方法が必要である。

そこで我々は、これに代わる現代的手法による銀河進化の議論に着手した (Siudek et al., 2018)。具体的には、紫外から近赤外までの 12 の波長(波長 $\lambda = 150 \text{ nm} - 2.2 \mu\text{m}$)と各宇宙年齢ごとの光度を含む 13 次元特徴空間を構築し、教師なし機械学習の方法であるフィッシャー EM

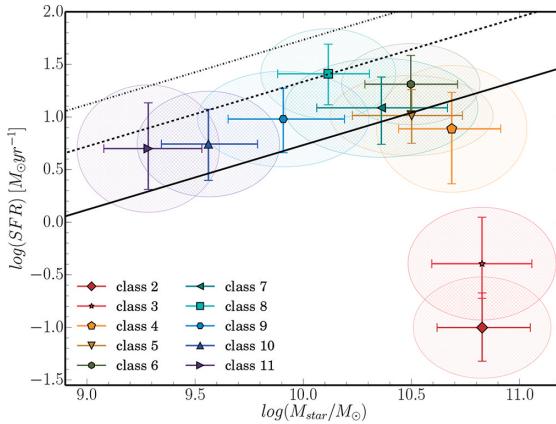


図 1. フィッシャー EM アルゴリズム (FEM) によって求められた星質量 (M_*)–SFR 関係 (Siudek et al., 2018, Fig. 7 を再掲). FEM が分類した銀河のクラスの $\log \text{SFR}$, $\log M_*$ の中央値, および分散が示されている. 梱円の面積は絶対偏差の中央値に対応する. 黒い実線は $z = 0.7$ での星形成銀河主系列を表す (Whitaker et al., 2012). 破線と一点鎖線はそれぞれ活動的な星形成銀河とスターバースト銀河の星形成主系列に対応している (Rodighiero et al., 2011).

アルゴリズム (FEM: Bouveyron and Brunet, 2012) を適用した. その結果, FEM は銀河の総星質量 M_* と星形成率 SFR の間に見られる星形成銀河主系列と呼ばれる関係を恣意的なサンプル選択なしで発見することに成功した(図 1). また, FEM は星形成銀河主系列はある総星質量を超えると星形成を停止した銀河の系列に連続的につながっていくことも発見した. これは, 銀河の星形成が急激に停止し, 不連続に星形成を停止した銀河に遷移するという仮説に反する構造で, 多波長光度空間の情報を全て活用することで初めて発見することができたものである. この銀河の連続体こそが, 銀河進化の基本を表す銀河多様体の一つの射影になっている. 多波長光度空間の銀河多様体はその非線型空間構造のため, 古典的 PCA に基づく以前の研究では発見され得なかった.

しかし, 宇宙物理学の研究は銀河多様体の定量的記述のみでは満足せず, その構造を完全に理解し, 銀河進化の物理学を支配する(おそらくいくつかの)パラメータへの依存性を解明することが課せられる. この更なる目標のためには, より洗練された方法が必要である.

1.2 多波長光度空間における銀河多様体

単位時間あたりにどのくらいの質量の星が形成されるかを星形成率とよび, [太陽質量 年 $^{-1}$] ($[M_\odot \text{yr}^{-1}]$ と表記)で測る. 星形成率の時間発展を星形成史といい, これが銀河の進化を決める重要な要素の 1 つである. 紫外線から近赤外線の波長では, 銀河の放射スペクトルは星とガスの寄与が支配的である. 星の温度および寿命は星の質量に強く依存しており, 大質量の星ほど明るく高温で, 寿命は短い. 高温の星は紫外線を大量に放射するが, 低温の星は紫外線では暗く, 近赤外線で光る. 定量的には, 星が安定して定常的に核融合をする段階である主系列星でいられる時間を τ_{MS} , 星の表面温度を T , 星の光度を L とおくと

$$(1.2) \quad \tau_{\text{MS}} \propto M^{-2.5},$$

$$(1.3) \quad L \propto M^{3.5}$$

$$(1.4) \quad L \propto T^4$$

と近似できる。この帰結として、高温の星から先に寿命が尽きるため、星形成史は銀河のスペクトルに直接反映される。つまり、星形成史は銀河の多波長(バンド)光度が張る空間において特徴が適切に表れていると期待される。

従来の天文学では、さまざまな波長での光度の比によって、多波長光度空間における進化を特徴づける方法が使われてきた。天文学ではこの比を色(color)と呼ぶ。異なる 2 つの波長 λ_1, λ_2 ($\lambda_1 < \lambda_2$) における天体の単色光度のペア L_{λ_1} と L_{λ_2} を考える。もし $L_{\lambda_1} < L_{\lambda_2}$ であればその天体は「赤い」、 $L_{\lambda_1} > L_{\lambda_2}$ の場合は「青い」と表現する。銀河の光度(絶対等級¹⁾)と色の関係をプロットすると(色-等級図)，明らかな 2 つの系統が現れる。これは銀河の色の二峰性(bimodality)と呼ばれている。具体的には、赤い銀河のタイトな系列(レッドシークエンス: red sequence)と、より広がった青い銀河の系列(ブルークラウド: blue cloud)が普遍的に存在している。レッドシークエンスとブルークラウドの間の領域に存在する銀河は相対的に少ないため、緑の谷(green valley)と呼ばれることもある(たとえば Blanton, 2006)。ブルークラウドの銀河は星形成が活発で、短寿命かつ高温度の大質量星が存在しているが、レッドシークエンスは星形成が停止し、小質量で低温の小質量星が卓越する。銀河の進化はブルークラウドからレッドシークエンスに移行すると考えられているが、この遷移がどのように起きるかは長らく未解決問題として残っていた。最近の先行研究で、ブルークラウドとレッドシークエンスの間は不連続ではなく、色-色-絶対等級の 3 次元空間において連続的に接続する構造が存在することが示唆された(たとえば Chilingarian and Zolotukhin, 2012)。

しかし、従来の色に基づく銀河進化の評価方法には潜在的な問題がいくつか存在する。あらゆる天文学探査データに共通するのが、データは観測装置の検出限界(detection limit)よりも明るい天体しか含まれないという偏りである。振動数 ν での天体の等級(magnitude)を m_ν とすると、観測データに含まれるのは $m_\nu < m_\nu^{\text{lim}}$ の天体のみである。これが等級選択効果(magnitude selection effect)である。前述したように、色とは 2 波長の光度比であるため、観測における選択効果は入り組んだ形で現れ、単純なコンプリートネス²⁾の検証はほぼ不可能である。色-等級図上での銀河進化の研究は、この複雑な選択効果が物理的な特徴と分離できず、混乱した議論が続いていた。しかし、色は光度の比であることから、多波長の光度(絶対等級)が張る多次元空間に戻った議論も可能である。この方法ならば、選択効果は直接的な形で評価できる利点がある。たとえば色-等級図における二峰性は、元の多次元光度空間にも対応するピーク構造が存在するはずである。よって、我々は多波長光度の高次元空間内で銀河が形作る構造に注目する。

我々が Siudek et al. (2018) で発見した銀河多様体は非線型な構造を持っている。さらに驚くべきことに、この銀河多様体を構成するサンプル銀河のスペクトルは、いくつかの広帯域バンド光度³⁾の情報のみで区別され、より複雑な物理量の組み合わせは必要なかった。この事実は、紫外線から可視光、近赤外線での銀河の多波長光度は、せいぜい数個の物理量で説明できることを示唆している。これは、従来の方法では決して見つけることができなかつた銀河進化の新しい特徴付けである。この発見をきっかけとして、我々は銀河多様体をさらに追求し、銀河多様体から銀河進化の物理を支配するパラメータ(おそらく多くてもいくつか)の依存性を解明し、銀河進化の支配方程式を導出する研究に着手した。このため我々は、従来の天文学的方法論とはまったく異なる、データサイエンスの最新手法の 1 つである多様体学習(たとえば Ma and Fu, 2012)として知られる手法に注目し、更なる解析を進めている。本論文ではこの一連の研究の現状報告と将来展望について紹介する。結果の一部は Cooray et al. (2023) にて公表済みであるが、本稿でより一般的な解説を試みる。

本論文では、観測データに関するすべての計算において宇宙論パラメータ $h = H_0/(100 \text{ [km s}^{-1} \text{ Mpc}^{-1}]) = 0.7$, $\Omega_{\Lambda 0} = 0.7$, $\Omega_{M0} = 0.3$, 曲率パラメータ $\Omega_{K0} = 0$ とした。これらのパラメータの意味は付録 A で説明している。

2. データ

本研究で用いたデータは Reference Catalog of Galaxy Spectral Energy Distribution (RCSED: Chilingarian et al., 2017) である。RCSED は、紫外線天文衛星 GALEX の全天探査カタログ、可視光大規模分光測光探査プロジェクト SDSS のカタログ、および近赤外線広域探査 UKIDSS カタログを統合し、最新の天文学スペクトル分析方法を用いて構築されている。RCSED は全天の約 25% をカバーし、数百万個の銀河の FUV, NUV, $u, g, r, i, z, Y, J, H, K$ の 11 バンドで k 補正⁴⁾された測光データ、および関連する物理量の情報が記載されている。さらに、いくつかの公開データを再処理した関連情報が測光カタログに追加されている。基本となる天体リストは SDSS データリリース 7(DR7) (Abazajian et al., 2009) のうち、赤方偏移 $0.007 < z < 0.6$ の範囲にある活動銀河ではない銀河の分光サンプル⁵⁾である。このデータには 800,299 個の銀河が含まれている。

全サンプルのうち、11 バンドすべてで測光値を持つ銀河を抽出すると、90,565 個の銀河が得られる。そして赤方偏移⁶⁾の信頼度が ≤ 0.5 の銀河を取り除くと、銀河数は 90,460 個となる。親サンプルと比べて大きく銀河数が減ったのは、主に UKIDSS サンプルとのクロスマッチが可能な天域が狭いことによる。

本研究の主目的は銀河光度空間における普遍的な関係の発見と定量化である。等級選択効果を回避するため、上記のサンプルを SDSS g バンドの等級をもとにコンプリートなサンプルを構築した⁷⁾。 g バンドでの限界等級 $m_{AB,g} = 18.3$ ⁸⁾ から算出した限界絶対等級曲線を用い、最終サンプルの銀河数が最大となるように絶対等級を求めた。これにより、最終的に赤方偏移 $z_{\text{lim}} < 0.097$, $M_{\text{lim},g} \leq -20.016$ の範囲にある 27,056 個の銀河からなるサンプルが構築された。以降の解析は全てこの volume-limited サンプルに基づく。

多様体学習においては、データの値の範囲を最適化するための前処理が重要となる。本研究では、各バンドでの銀河の光度は絶対等級の単純平均で中心化し、分散を 1 に規格化した解析と、リスケールせず絶対等級の値そのものを用いた解析を行った。これら 2 つの解析の結果は(各軸を再び絶対等級にスケールしなおせば)定量的にもほとんど差がなかった。よって本稿では、サンプルの絶対等級はリスケールせずにそのままデータの特徴空間とした結果のみ示す。これは、今回のサンプルでは、どのバンドでも放射には銀河内の星からの寄与が支配的であることから、絶対等級は各バンドで互いに似た範囲に収まったことが理由である。しかし、赤方偏移を加えるなど今後の他の物理量も含めた解析のためには、適切な規格化が必要となることに注意しておく。

3. 方法：多様体学習による銀河多様体の定量化

3.1 多波長光度空間での銀河多様体

2 章で構築したサンプルの基本的検証のため、まず Siudek et al. (2018) と同様に FEM をデータに適用し、このデータからも 11 次元の多波長光度空間内で銀河が形作る低次元の構造が得られることを確認した。この銀河多様体は FEM によって抽出されたクラスタの空間配置を解析することにより、多波長光度の特徴空間において 2 次元曲面をなすことが明らかになった。RCSED の銀河多様体を図 2 に示す。

我々の銀河多様体は 11 次元光度空間内での低次元部分空間で表現されているが湾曲した構

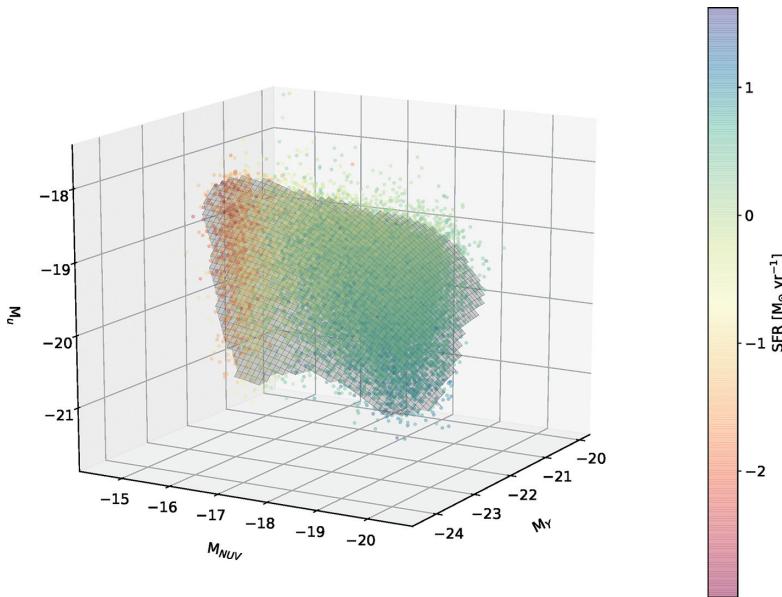


図 2. 多次元特徴空間で発見された銀河多様体。元の特徴空間は 11 次元であるが、多様体は 2 次元構造のみを持ち、基本的に紫外、可視光、および近赤外の光度の 3 次元空間に埋め込まれている。多様体は湾曲した形状をしているため、古典的 PCA のような線型関係を分析する手法では決して見つけることができない。カラーコーディングはサンプル銀河の星形成率を表している。

造を持っており、視覚的に構造を把握するのも難しく、またそれ以上の定量化も容易ではない。古典的なスケーリング法則における多峰性と分散は、銀河多様体の非線型構造を反映しない、最適ではない射影の結果であることが多い。1 章でも述べたように、銀河進化に関する観測データは巨大化する一方である。このため、これから研究においては情報量を最大限に活かす簡潔な記述方法が必要である。

我々は、銀河多様体をより有効に利用して定量化するため、次元削減と呼ばれる一群の方法に着目した。具体的には、銀河多様体から銀河進化の物理学を支配する（おそらくは高々数個の）物理量への依存性を解明するため多様体学習を適用した。我々はこの方法を用い、古典的な天文学における第一原理的理論構築とは全く異なる、銀河進化の新しい定量化を試みている。銀河多様体を定量的に表現するもう一つの大きな利点は、例えば観測した天体に欠けていいる光度などの観測量や、SFR や星の質量 M_* などの物理量を多様体上の位置から直接推定することが可能になることである。これは銀河多様体を光度空間に引き戻すことで実現できる。この引き戻しのためには、多様体上の距離をもとの多波長光度空間の計量でも記述しておけば便利である (Lin et al., 2017)。天文観測は一般に観測限界ぎりぎりを用いる困難なデータ解析が多く、観測量の予測や推定、補間に用いることのできる銀河多様体の定量表現は今後の天文学研究における強力な武器となる。

3.2 多様体学習

多様体学習では、データは測地線距離 d^M で決まる計量を持つ、滑らかな d 次元多様体 M からランダムに抽出された有限個の点集合 $\{y_i\}$ ($i = 1, \dots, N$) と考える。これらのデータ点は、滑

らかな写像 ψ によってユークリッド計量 $\|\cdot\|_{\mathcal{X}}$ の特徴空間(あるいは入力空間) $\mathcal{X} = \mathbb{R}^n$ ($d \ll n$) に埋め込まれている。埋め込まれた特徴空間内でのデータ点を $\{x_i\}$ ($i = 1, \dots, N$) とすると、埋め込み写像は $\psi: \mathcal{M} \rightarrow \mathcal{X}$ であり、多様体上の点 $y_i \in \mathcal{M}$ は

$$(3.1) \quad y_i = \psi^{-1}(x_i), \quad x_i \in \mathcal{X}$$

と表される。多様体学習の目的は、特徴空間の(入力)データ点集合 $\{x_i\} \in \mathcal{X}$ が与えられたときに、多様体 \mathcal{M} および写像 ψ の具体的な形を求める、元のデータ点集合 $\{y_i\} \in \mathcal{M}$ を再構成することである。多様体学習アルゴリズムを入力データ点集合に適用すると、 \mathbb{R}^n のデータ集合を隣接点との関係を維持しながら低次元空間 \mathbb{R}^d ($d \ll n$) に写像する。すなわち

$$(3.2) \quad x_i \mapsto \hat{y}_i = (\psi_1^{-1}(x_i), \dots, \psi_d^{-1}(x_i))^{\top} \in \mathbb{R}^d$$

を通じて本来の $\{y_i\} \subset \mathbb{R}^d$ の推定値としての $\{\hat{y}_i\} \subset \mathbb{R}^d$ が得られる。ここで、 \top はベクトルの転置を表す。これは、データ集合は高次元の特徴空間内で低次元の部分多様体上に分布するという「多様体仮説」という仮定(たとえば Goodfellow et al., 2016)に基づいてデータの次元を縮小する方法で、非線型次元削減と呼ばれる方法論の一つである。多様体の次元 d も \hat{d} を通じてデータから推定できれば理想的である。しかし実際の解析では我々が \hat{d} を設定し、いくつかの基準から最適な \hat{d} を選ぶという手順を取る(第 4.1 章)。

多様体学習に関する研究の萌芽は散発的に発表されてきた 90 年代の研究まで遡るが、2 編の独創的な論文(Roweis and Saul, 2000; Tenenbaum et al., 2000)が出版されて以降人気が高まり、積極的に研究されるようになった。多様体学習アルゴリズムは、特徴空間内で複雑な形状を持つ多様体を「展開」し、その上に局所座標系を提供できる(Tenenbaum et al., 2000; Roweis and Saul, 2000)。重要なのは、次元削減後のデータ点どうしの繋がりが元の高次元空間でのデータ点どうしの繋がりをなるべく忠実に表現することで、そのためにはアルゴリズムがデータの形を学習する必要がある。これが多様体学習という言葉の由来である。

古典的 PCA などの線型な方法がデータ構造の大域的構造を表すのに有効なのに対し、非線型な方法は局所的な構造の表現に効力を発揮する。一方、非線型手法の多くはデータ点の近傍の位置関係の低次元表現に注目して次元削減を行うため、大域的構造が失われることがある。よって、多様体学習を用いる場合、目的によって適切なアルゴリズムを選択する必要がある。また、多様体学習が与える低次元表現の座標系は、直感的あるいは物理的意味を持つことが必ずしも保証されないことに注意が必要である(たとえば Liu et al., 2017)。この点については 4 章で議論する。

3.3 Isomap および UMAP アルゴリズム

本研究では、多様体学習のアルゴリズムとして Isomap (isometric feature mapping: Tenenbaum et al., 2000) および UMAP (uniform manifold approximation and projection: McInnes et al., 2018, 2020) を採用する。本研究の目的は銀河進化の物理量依存性の定量化であり、元の高次元特徴空間で連結な構造は多様体としても連結した構造に写像されることが必要である。2 つのアルゴリズムはともに、元のデータ点分布の持つ連結性を保つ性質を持っており、理想的な方法である。これらの計算には scikit-learn (Pedregosa et al., 2011) を使用した。

3.3.1 Isomap

Isomap アルゴリズムは滑らかな多様体 \mathcal{M} が \mathbb{R}^d ($d \ll n$) の測地的凸(geodesically convex)領域であり、埋込み写像 $\psi: \mathcal{M} \rightarrow \mathcal{X}$ が等長写像(isometry)であると仮定する。まず測地的凸性を以下で定義する(Tenenbaum et al., 2000)。

定義 1. 測地的凸(geodesically convex) (\mathcal{M}, g) をリーマン多様体とする。 \mathcal{M} の部分集合 \mathcal{U} は、 \mathcal{U} 内の任意の 2 点を結ぶ \mathcal{U} 内の最短測地線が一意に存在するとき、測地的凸集合であるという。

測地的凸なリーマン多様体は、測地線距離に関して凸な計量空間である。

これより、Isomap の仮定は以下のように表現される。

凸性 \mathcal{M} は \mathbb{R}^d の測地的凸な部分集合である。

等長性 測地線距離は写像 ψ の下で保存される。多様体 \mathcal{M} 上のあらゆる 2 点 $y, y' \in \mathcal{M}$ について、これらの間の測地線距離が対応する埋め込まれた \mathbb{R}^n の 2 点 $x = \psi(y), x' = \psi(y') \in \mathcal{X}$ のユークリッド距離に等しくなる。即ち

$$(3.3) \quad d^{\mathcal{M}}(y, y') = \|x - x'\|_{\mathcal{X}}.$$

Isomap は \mathcal{M} が測地的凸の領域、 ψ が等長変換であるという仮定を用いて多次元尺度構成法 (multidimensional scaling: MDS) を一般化したアルゴリズムである。MDS はデータの 2 点間のユークリッド距離を保存しつつ、データ点が分布するより低次元の部分空間を探す方法である。MDS は線型次元削減法の 1 つであり、曲がった領域ではうまく機能しない。Isomap は MDS の思想を踏襲し、データのすべてのペア間の \mathcal{M} 上での測地線距離を近似することで、非線型多様体の大域的幾何構造を最大限に保存するアルゴリズムである。この意味で、Isomap は多様体学習としては局所的方法でもあり、同時に大域的方法でもある。

Isomap アルゴリズムは 5 つのステップから構成される。

(1) 最近傍探索

ある整数 K あるいは $\epsilon > 0$ を取る。特徴空間 \mathcal{X} 内全てのデータ点ペア $x_i, x_j \in \mathcal{X}$, $(i, j = 1, \dots, n)$ 間の距離

$$(3.4) \quad d_{ij}^{\mathcal{X}} \equiv d^{\mathcal{X}}(x_i, x_j) = \|x - x'\|_{\mathcal{X}}$$

を計算する。距離としては一般にユークリッド距離を取る。 K 番目に近い点 (K -nearest neighbor) まで、あるいは半径 ϵ の球内にあるすべての点を結合することにより、 \mathcal{M} 上の隣接点を決める。Isomap の性能は K ないし ϵ の選択によって定まる。

Isomap は、効率的な近傍検索のため `sklearn.neighbors.BallTree` を使用する。

(2) グラフ距離の算出

入力データ点群 $\{x_i\}$ ($i = 1, \dots, N$) に対し、重み付き近傍グラフ (weighted neighborhood graph) $\mathcal{G} = \mathcal{G}(\mathcal{V}, \mathcal{E})$ を構成する。グラフの頂点集合 (vertices) \mathcal{V} はデータ点群 $\{x_1, \dots, x_N\}$ で、辺集合 (edges) \mathcal{E} はデータ間の近傍関係を示す辺 e_{ij} である。辺 e_{ij} には 2 点間の距離 $d_{ij}^{\mathcal{X}}$ に対応する重み w_{ij} が与えられている。2 点 x_i, x_j が直接辺で接続していないければ重みは ∞ となる。

\mathcal{M} 上の 2 点間の測地線距離 $\{d_{ij}^{\mathcal{X}}\}$ を、グラフ \mathcal{G} によるグラフ距離 (graph distance) $d_{ij}^{\mathcal{G}}$ によって推定する。グラフ距離 $d_{ij}^{\mathcal{G}}$ とは、グラフ \mathcal{G} 上の 2 点間の最短の道 (path) の長さで定義する。近傍にない 2 点は最近傍をつないだ最短距離の道で接続し、道の長さは重みの和で与える。この長さが離れた 2 点の測地線距離の近似を与える。

データ点が多様体 \mathcal{M} 上で定義される確率分布から抽出されたとすると、多様体が平坦ならば $N \rightarrow \infty$ でグラフ距離 $d^{\mathcal{G}}$ は測地線距離 $d^{\mathcal{M}}$ に収束する (Bernstein et al., 2001)。このための最も効率的なアルゴリズムとして Floyd-Warshall アルゴリズム (Floyd, 1962; Warshall, 1962) や Dijkstra のアルゴリズム (Dijkstra, 1959) が用いられる。前者はグラフが密な場合、後者は疎な場合に有効であることが知られている (たとえば Ma and Fu, 2012)。

(3) MDS によるスペクトル埋め込み(spectral embedding)

距離行列 $D^G \equiv (d_{ij}^G)$ を考える ($N \times N$ 対称行列). 行列 D^G に古典的 MDS を適用し, 多様体 \mathcal{M} 上のデータ点間の測地線距離ができるだけ保存されるように d 次元空間 \mathcal{Y} を再構成する. $S^G \equiv ((d_{ij}^G)^2)$ をグラフ距離の 2 乗を成分とする $N \times N$ 対称行列とする. これを以下によって二重中心化する

$$(3.5) \quad K_N^G = -\frac{1}{2}HS^G H,$$

$$(3.6) \quad H \equiv \mathbb{I}_N - \frac{1}{N}1_N.$$

ここで \mathbb{I}_N は $(N \times N)$ 単位行列, 1_N は $(N \times N)$ の成分がすべて 1 の対称行列である.

(4) 埋め込みベクトル $\{\hat{y}_i\}$ を $\|K_N^G - K_N^Y\|$ を最小化するように選ぶ. ここで

$$(3.7) \quad K_N^Y = -\frac{1}{2}HS^Y H.$$

ここで $S^Y = ((d_{ij}^Y)^2)$, $d_{ij}^Y = \|y_i - y_j\|$ は y_i と y_j のユークリッド距離である. K_N^G を固有値行列 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$, 固有ベクトル行列 $V = (v_1, \dots, v_N)$ によって固有値分解すると

$$(3.8) \quad K_N^G = V\Lambda V^\top.$$

最適解は K_N^G の大きい方から d 個の固有値 $\lambda_1 \geq \dots \geq \lambda_d$ に対応する固有ベクトル v_1, \dots, v_d によって得られる.

(5) グラフ \mathcal{G} は $d \times N$ 行列

$$(3.9) \quad Y \equiv (\hat{y}_1, \dots, \hat{y}_N) = (\lambda_1^{\frac{1}{2}} v_1, \dots, \lambda_d^{\frac{1}{2}} v_d)$$

によって d 次元部分空間 \mathcal{Y} に埋め込まれる.

その構造上, Isomap は 2 点間の計量を保持するため, 特徴空間内のデータ点の多様体上での「表面密度」を保存する. すなわち, データ点が密に存在する領域は多様体上でも密になり, 疎な領域は多様体上でも疎となる. Isomap は多様体 \mathcal{M} がユークリッド空間の測地的凸な部分多様体であること, および ψ の等長性を仮定するため, 曲率が大きすぎる, 多様体に穴がある, 非凸であるなどの場合にはうまく機能しない. 実践上の問題として, ノイズがある, すなわちデータ点が必ずしも多様体上に分布していない場合, Isomap のパフォーマンスは近傍の取り方に依存する. ノイズが大きすぎなければ, Isomap は基本的にノイズの影響に対してある程度頑健である. 本研究では, Isomap の近傍は $K = 5$ に取った.

3.3.2 UMAP

UMAP(uniform manifold approximation and projection) は 2018 年に提案された比較的新しい手法で, 微分幾何学と代数トポロジーに基づいている. UMAP では, 元の特徴空間で近いデータ点は多様体上でも近くなる. 実行時間が速いことから計算コストが低減でき, 4 次元以上の多様体への次元削減も可能である. UMAP は位相的データ解析とリーマン幾何学が元になっているアルゴリズムである. このアルゴリズムは以下の 3 つの仮定に基づく: 1) データはリーマン多様体上に一様に分布している, 2) リーマン計量は局所的に一定である(そのように近似できる), 3) 多様体は局所連結性を持つ. これらの仮定から, ファジーな位相構造をもつ多様体をモデル化することが可能である. データ点が可能な限り均一に分布するように多様体を定義するため, Isomap とは違いデータ点の表面密度は保存されないことに注意する. UMAP アルゴリズムは次の 3 つの段階で構成される:

- (1) リーマン多様体の推定,
- (2) 距離空間のファジートポロジーによる表現,
- (3) 次元削減.

UMAP の核となる概念はファジートポロジー表現であるが、圈論を用いて説明されていることもあって簡潔な記述は難しいため、ここでは概略を示すにとどめる。詳細は関連文献を参照されたい。UMAP はその構成法から、Isomap よりもノイズに対して頑健である。本研究では、UMAP の近傍は $K = 50$ に取った。

4. 結果と議論

4.1 結果: Isomap および UMAP 銀河多様体

Isomap および UMAP によって得られた銀河多様体を図 3 に示す。異なったアルゴリズムである Isomap と UMAP が定性的にかなり似通った 2 次元多様体を与えることは注目すべき点である(図 3)。2 つの方法で推定された銀河多様体の違いは図 3 からはっきり見て取れる。

Isomap はデータ点群の密度を保存するため、多様体には密度構造、つまり多様体上の密な領域と疎な領域がある。対照的に、UMAP は多様体の構築において密度を可能な限り均一にするため、UMAP 多様体はかなり一様な密度分布となっている。つまり、Isomap 多様体で密度が高い領域はそのまま現れるが、UMAP 多様体では面積が拡大する。

Isomap および UMAP によって抽出された多様体の次元を推定するため、再構成誤差および情報量基準をもとに評価した⁹⁾。Isomap、UMAP とも再構成誤差は次元数を 2 から順に増加させても数値誤差の範囲で変化がみられない。また Akaike information criterion (Akaike, 1974), Bayesian information criterion (Schwarz, 1978) からはともに次元数は 2 が選ばれた。多波長光度空間での結果と合わせ、このデータから得られた銀河多様体の次元は 2 であると結論される。

銀河多様体に銀河進化に関する情報がどう表れるかを検証するため、図 4 で SFR と星の質量を多様体上の関数として比較する。2 次元銀河多様体の座標軸 1 および 2 と星形成率、星質量の各相関を Isomap および UMAP それぞれについて示したのが図 5 および図 6 である。ここで、Isomap 多様体の軸は第 1、第 2 固有値に対応する固有ベクトルであった。UMAP でも 2 次元

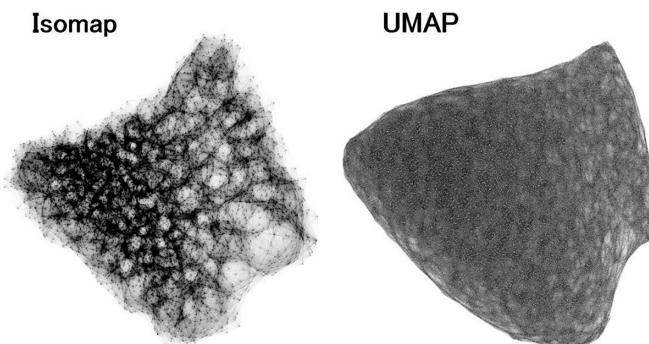


図 3. 多様体学習アルゴリズム Isomap および UMAP による「展開された」銀河多様体。左側と右側のパネルは、それぞれ Isomap と UMAP からの多様体を示している。この空間上の多様体の構造は、図 2 よりもはるかに簡単に認識できる。全般的な形状は互いにわずかに異なるが、統く解析で示すように星形成率や星質量の多様体上の分布などの特徴は共通している。

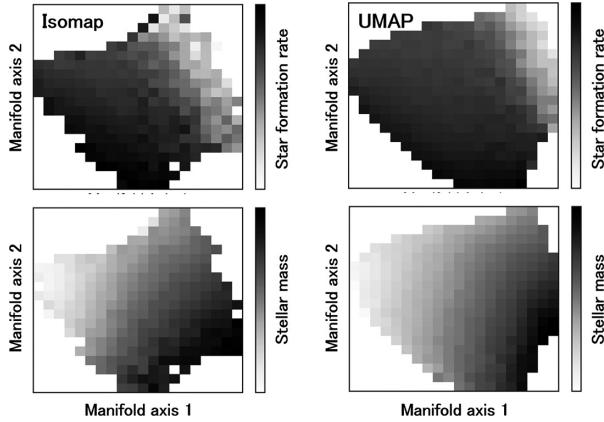


図 4. 2 つの異なる多様体学習アルゴリズム, Isomap と UMAP によって取得された銀河多様体. 星形成率 SFR と星質量 M_* は多様体上の関数として表される. 左パネルは, それぞれ Isomap によって得られた多様体上の星形成率と星質量を示している. 右パネルは, UMAP によって得られた多様体で, カラーコーディングは左パネルと同じである.

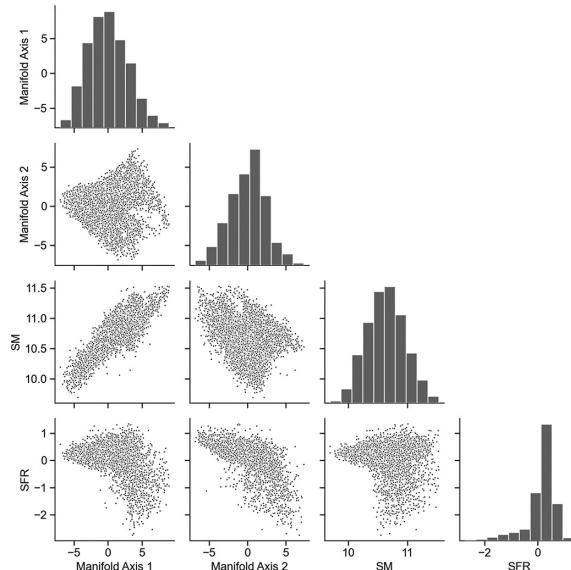


図 5. Isomap による 2 次元銀河多様体の座標軸と星形成率(SFR), 星質量(SM)の散布図. 多様体座標 1 は星質量と, 座標 2 は星形成率と強く相関している.

的な多様体が得られるが, その意味は明確には与えられない. Isomap 多様体との対応は, ここで物理量の多様体上での分布を評価することで明らかになる. 以下, 座標軸はこれによって描えた形で議論する. 星形成率 SFR と星質量 M_* の挙動は 2 つの図で定性的に非常に似ており, 推定された多様体構造がロバストであることを示唆している. これは, 多様体学習が実際に多波長光度空間における銀河進化の重要な特徴を「学習」したことを意味する. 図 5, 図 6 から, 多様体座標 1 は星質量と, 座標 2 は星形成率と強く相関していることが分かる. 我々はすでに可視光度空間における銀河多様体が基本的に 2 次元であることを見た. これはすなわち, 紫

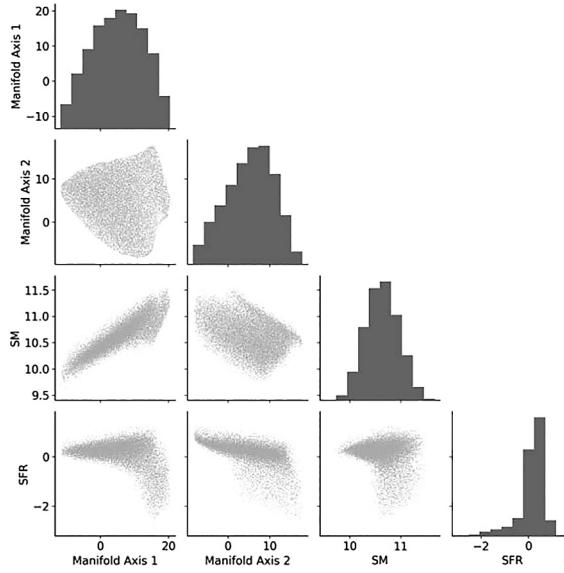


図 6. UMAP による 2 次元銀河多様体の座標軸と星形成率(SFR), 星質量(SM)の散布図. 図 5 と同様, 多様体座標 1 は星質量と, 座標 2 は星形成率と強く相関している.

外線・可視光線・近赤外線の波長における銀河進化は星形成率, 星質量のわずか 2 つの物理量で過不足なくあらわされることを意味しており, 銀河進化理論に強い制限を与える重要な発見である (Cooray et al., 2023).

このように, 多様体学習は銀河多様体と星形成率や星質量などの物理量を結び付けることができる. これをさらに押し進めることで, 多様体上の銀河進化をパラメータ化することも原理的には可能である. 星が形成されると, 星質量, つまり蓄積された星の総質量が増加する. これは銀河進化の基本的な側面の 1 つであり, この進化を多様体上のベクトル場として可視化できる. 星形成のベクトル場を図 7 および図 8 に示す. 銀河進化の「速度場」はこれら 2 つの図から明らかに見て取れる. 小質量の銀河は急速に進化して星形成率が減少し, 星質量が大きくなる(図上左から下への向き). 大質量の銀河は進化が遅く, 多様体上の同じ位置(右上)に長く滞在する.

4.2 銀河多様体と観測量

銀河多様体から入力空間である多波長光度空間における情報に引き戻して理解するため, 多様体の軸と観測された光度の間のペアプロットを図 9 に示す. 多様体座標 1 は, 銀河体の骨格を構成する古い恒星集団の寄与が大きい可視光長波長域から近赤外線, r, i, Y, I, J, H, K バンドの光度と密接に相関している. 対照的に, 多様体座標 2 は紫外線 FUV, NUV から可視光短波長側 u, g バンドと密接に相関しており, 現在あるいは非常に最近の星形成活動を表している. 可視光での光度に注目すると, それらが互いに非常にタイトに相関している. これは, 可視光の波長範囲で複数の光度を解析に含めても, 銀河の特性に関する情報は本質的に追加されないことを意味する. 対照的に, 紫外線光度は散布図で自明ではない非線型相関を示しており, 紫外線と可視光バンドの組み合わせが多様体の構造に関する基本的な情報を提供することを示している. 近赤外線バンドの光度は可視光線光度と似た振舞いをするものの, この相関関係は多様体の構造に関する追加情報がまだ存在することを示唆している. したがって, 11 個

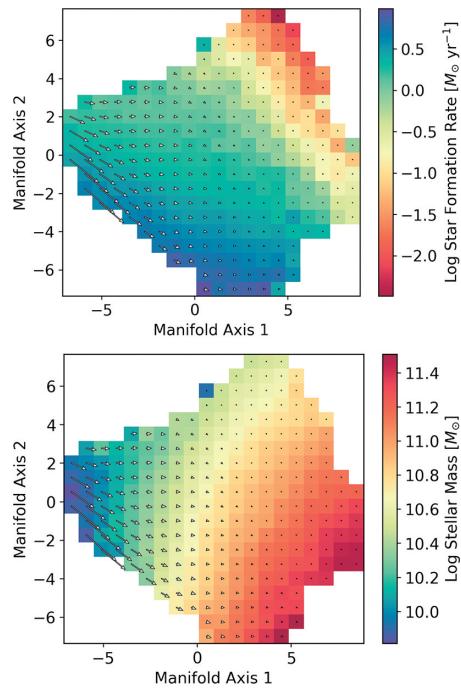


図 7. Isomap 銀河多様体上の星形成率・星質量進化のベクトル場. 上パネルのカラーバーは現在の銀河の星形成率, 下パネルは星質量を表している.

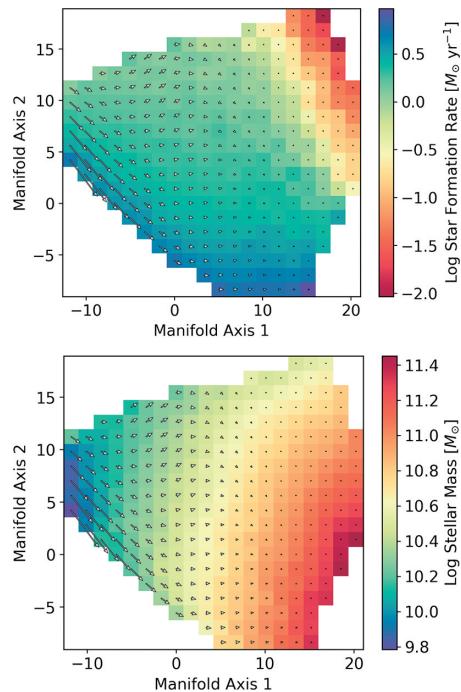


図 8. UMAP 銀河多様体上の星形成率・星質量進化のベクトル場.

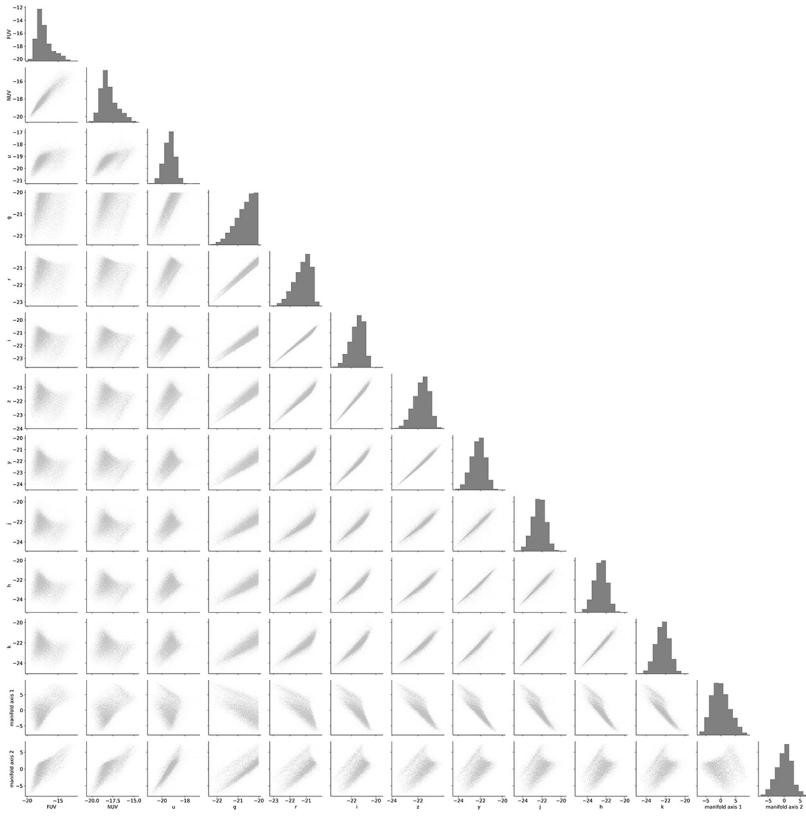


図 9. 各バンドでの銀河の光度と銀河多様体座標 1 および 2 との相関.

の多波長光度にまたがる銀河の特徴空間は、必然的により低次元の部分多様体によって表される。これが、発見された 2 次元銀河多様体の背後にある天体物理的基礎である。

4.3 定量化から定式化へ

残された課題は、多様体上の銀河進化の軌跡をどのように記述し、解釈するかである。無論これは簡単ではなく、さらなる議論が必要である。Cooray et al. (2023)において、我々は銀河の化学進化の古典的な理論モデルを適用した。化学進化は、恒星の進化理論に基づいて銀河内の元素の形成と進化を扱う銀河物理学の分野である¹⁰⁾。重要な物理プロセスは、星の中心部での核融合によって生じる元素合成である。Lilly et al. (2013) によって提案された物質の流出を伴う単純なモデルを採用する。

$$(4.1) \quad \mathcal{M}_*(t_{n+1}) = \mathcal{M}_*(t_n) + (1 - r) \text{SFR}(t_n) \Delta t ,$$

$$(4.2) \quad \mathcal{M}_{\text{ISM}}(t_{n+1}) = \mathcal{M}_{\text{ISM}}(t_n) - (1 - r + \zeta) \text{SFR}(t_n) \Delta t$$

ここで r はリターンドマスフラクション (returned mass fraction: ガスが星間物質に戻る割合)、 ζ はマスローディングファクター (mass loading factor: 星形成率に対する質量流出の比)、 Δt は時刻のタイムステップを表す (Cooray et al., 2023)。図 10 に示すように、式(4.1) と (4.2) から銀河の理論的な進化軌跡を計算できる。銀河多様体上のベクトル場の解釈は、式(4.1) および

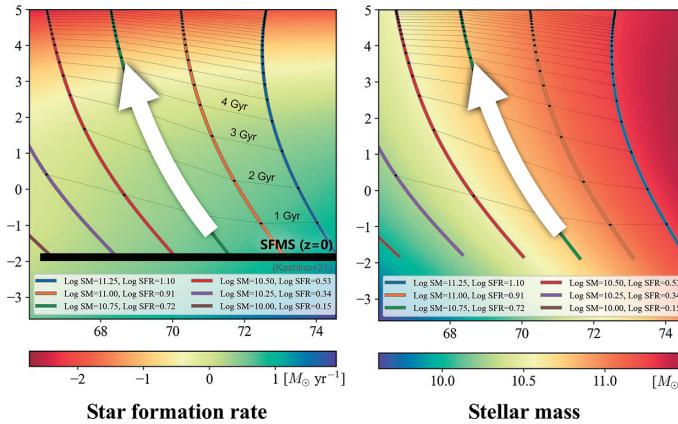


図 10. 銀河の古典的な化学進化モデルによって予測される理論的な進化の軌跡. 太実線は異なるった星質量を初期値としたときの銀河進化の軌跡を示す. 細線は銀河年齢. SFMS は星形成銀河主系列(連続的に星形成をしている銀河の経験則)を表している.

(4.2)を図 7 あるいは図 8 と比較することで得られる. 図 9 に関する第 4.2 章での議論でみたように多波長光度は互いに強く相関しており、大きく星形成率 SFR と星質量 M_* に関連するものに分けられる. とはいっても(1.1)式でみたように、理論のみから第一原理によって独立な物理量の数を決定することは、天体物理のような複雑な系では非常に難しい. 物理学的な議論から予想される自由度の幅はかなり大きく、多様体学習などデータ科学的方法によって得られる次元の情報をもとに絞り込むのが現状では最善の手段である.

この現状の方法はしかし、銀河多様体自体から進化方程式が確定できないという点で不満が残る. より直接的な銀河多様体の解釈・定式化には本質的に洗練された方法論が必要であり、我々は現在シンボリック回帰(たとえば Cranmer, 2023)などの方法によってこの問題に取り組んでいる.

5. 結論と展望

本研究では、銀河の多波長光度が張る高次元特徴空間中での銀河分布に対し、データ科学で最近発展してきた多様体学習を適用し、銀河の進化の特徴づけを行った. これにより、我々は銀河多様体と呼ばれる、多波長光度空間内のデータ点に埋め込まれた低次元非線型構造を発見した. これが銀河多様体である. そして紫外線、光学、近赤外線の光度空間における銀河の進化が、銀河多様体上の星形成と星の質量進化という 2 つのパラメーターによってよく記述されることを発見した. これは、銀河進化研究への多様体学習の有効性を示す結果である.

本研究で用いたデータは紫外線から近赤外線の波長域であった. この波長の放射は基本的に星からの寄与で、ガスの放射がそれに加わっている. しかし、さらに可視光から離れた波長の放射、たとえば γ 線、X 線など短波長側、あるいはダストの放射が卓越する中間・遠赤外線、そして原子ガス、分子ガスの放射が見えてくる電波に拡張すれば、より多くの物理過程を考慮することができる. 我々は現在、中間赤外線データを加えた多波長光度空間で同じ解析を行い、得られた多様体が銀河の持つダスト粒子の放射の情報を持つという示唆を得ている. 短波長側の放射は高エネルギー現象と密接にかかわっており、ブラックホールへの物質降着による超高エネルギー放射天体である活動銀河核の進化も含めた理論構築も視野に入れることができる.

また、ここでは議論を多波長測光調査に限定したが、この方法は分光探査データにも用いる

ことができる。さらに放射だけでなく、力学的、構造的、環境的特性など、銀河の他の特性も含めることにより、銀河の形成や相互作用、合体など、より力学的な物理現象も含めた銀河進化の大統一理論の構築も視野に入ってくる。

本稿では銀河進化を銀河多様体上のベクトル場として記述したが、銀河多様体を宇宙年齢ごとに構成できるデータが得られれば、多様体上の進化ベクトルとしてではなく、銀河多様体自身の進化として銀河進化を記述することが可能になる。遠方宇宙まで十分なデータが取得できるような将来の赤方偏移探査が遂行された暁には、この方法はさらに強力な方法論となると期待される。

このように、多様体学習は銀河の形成と進化に関する研究に根本的に新しい洞察を提供する強力な方法である。しかしこれは、数ある物理学への応用の可能性のうち、特にシンプルな例の一つである。多様体学習は物理現象の記述にとどまらず、広く物理法則の発見と統一のための新たな方法論となるであろう。

注.

1) 可視光天文学(紫外線、近赤外線を含む)では、天体の光度を L としたとき

$$M \equiv -2.5 \log_{10} L + \text{銀河までの距離に寄らない定数}$$

を絶対等級とよび、これが光度の代替量として広く用いられる。等級についての正確な定義は付録 B で説明する。

- 2) 天体探査データにおいて、装置の検出限界まで天体が漏れなく検出されているとき、データはコンプリートであるという。
- 3) ある広い波長範囲で測定された光度を広帯域バンド光度と呼ぶ。
- 4) 銀河のスペクトルは複雑な波長依存性を持つため、宇宙論的赤方偏移によって波長が伸びるだけでなくスペクトルの観測する波長も変わる。この変化分の補正を k 補正と呼ぶ。詳しい定式化は付録 B を参照。
- 5) 詳しくは、SDSS の分光分類指標(specclass)で GAL_EM および GALAXY とマークされた銀河。つまり、クエーサーや 1 型セイファート銀河など、ブラックホール起源の放射が支配的な銀河はサンプルから除去されている。
- 6) この場合は距離の指標として用いられている。詳細は A にて解説している。
- 7) 天文学ではこのようなデータを volume-limited と呼ぶ。定訳はないが、考へている体積内で、ある光度 L_ν^{lim} よりも明るい天体を数え落としなく含んだサンプルを意味する。
- 8) 添字 AB は物理的な定義に基づく AB 等級という等級であることを表す(付録 B)。
- 9) この目的で広く用いられている Farahmand-Szepesvári-Audibert (FSA) dimension estimator (Farahmand et al., 2007) による評価も試みたが、検証してみたところ近傍点の設定によって著しく次元の推定値が変わることが判明したため、本論文では FSA の結果は議論に用いない。FSA 自体の性能の問題については更なる検証が必要である。
- 10) 「化学」進化という奇妙な用語は、この理論が銀河内の星と星間物質(ISM)の化学組成の解析に用いられてきたことに由来している

謝 辞

本論文は統計数理特集号『諸科学における統計数理モデリングの拡がり II』のために執筆したが、責任著者のスケジュール上の都合により投稿が間に合わなかったため、一般論文として投稿したものである。特集号担当編集者の島谷健一郎氏の示唆に深く感謝する。本研究は日本学

術振興会(JSPS 科研費補助金(24H00247, 21H01128, 19K03937, および JP17H06130)の補助を受けている。また本研究の一部は、住友財団平成 30 年度基礎科学的研究費補助金(180923), 統計数理研究所共同研究費「データサイエンスによる銀河進化研究の新展開」の支援も受けた。

本論文執筆のきっかけとなるアイディアを頂き、本稿にも貴重なコメントをして頂いた栗木哲氏に心から御礼申し上げる。有益な示唆を頂いた今泉允聰氏、矢野恵佑氏に深く感謝する。

付 錄

A. 宇宙論の基礎

A.1 Friedmann-Lemaître-Robertson-Walker 計量とスケール因子

一般相対性理論は、時空がその幾何学的構造によって特徴付けられることを示し、微分幾何が使用するべき基本的な枠組みであることを示した。微分幾何において、多様体上の局所的性質を記述する基本的な概念が線素(line element)あるいは計量(metric)である。Friedmann-Lemaître-Robertson-Walker (FLRW) 計量は、均一で等方的な時空の(局所的)幾何構造を表すために提唱された計量で、宇宙論のモデルとして広く用いられている。

$$(A.1) \quad ds^2 = g_{\mu\nu} dx_\mu dx_\nu = -c^2 dt^2 + a^2(t) \left[\frac{dr^2}{1 - Kr^2} + r^2(d\theta^2 + \sin^2 \theta d\phi^2) \right]$$

ここで $g_{\mu\nu}$ が計量テンソル、 t は時刻、 r, θ, ϕ は極座標の動径距離および角度方向の座標を表す。また K はガウス曲率、 $a(t)$ はスケール因子である。スケール因子は宇宙膨張による空間のスケールの変化を表現しており、宇宙論の慣習上 $a_0 \equiv a(t_0) = 1$ (t_0 : 現在の宇宙年齢) と規格化して用いられる。ここでは K は(長さ) $^{-2}$ の次元を持つように定義を採用する。この場合 r は長さの次元を持ち、 $a(t)$ は無次元になる。

スケール因子を導入すると、宇宙の進化による天体間の距離の変化から宇宙膨張による「自明な」距離の変化を除いた正味の変化分を表現することができる。即ち、距離をあらわすベクトルを \vec{r} とすると

$$(A.2) \quad \vec{r} = a(t) \vec{x}$$

と書ける。この \vec{x} は宇宙膨張の影響を除いた距離の指標であり、共動座標(comoving coordinate)と呼ばれる。

A.2 宇宙論的赤方偏移

過去のある時刻 $t = t_{\text{em}}$ で共動座標 $(r, \phi, \theta) = (r_{\text{e}}, 0, 0)$ で放射された光が、時刻 $t = t_0$ において原点 $(r, \phi, \theta) = (0, 0, 0)$ に位置する観測者に到達する状況を考える。相対性理論では、光はヌル測地線(null geodesic)に沿って移動する。ヌル測地線とは $ds^2 = 0$ を満たす経路のことで、最小作用の原理で実現する軌道と考えてよい。式(A.1)を用いれば、距離と時間を

$$(A.3) \quad \frac{cdt}{a(t)} = \frac{dr}{\sqrt{1 - Kr^2}}$$

のように関係付けられる。宇宙膨張の光への影響を議論するため、変数 r を

$$(A.4) \quad \chi \equiv \int_0^{r_{\text{em}}} \frac{dr}{\sqrt{1 - Kr^2}},$$

によって座標距離 χ に、時間 t を

$$(A.5) \quad \eta \equiv \int^t \frac{dt'}{a(t')},$$

によって共形時間 η に変換すると、光の満たすべき方程式は

$$(A.6) \quad c(\eta_{\text{obs}} - \eta_{\text{em}}) = \chi,$$

となる。ここで c は光速、 η_{obs} は光が観測者に到達した時刻を表す。

時刻 $\eta = \eta_0$ および $\eta = \eta_0 + \delta\eta_0$ で放出された光が、それぞれ η_1 および $\eta_1 + \delta\eta_1$ で観測者に到達したとする。式(A.6)の右辺は η によらないことから

$$(A.7) \quad \delta\eta_0 = \delta\eta_1$$

すなわち

$$(A.8) \quad \frac{\delta t_0}{a(t_0)} = \frac{\delta t_1}{a(t_1)}$$

となる。時間間隔 δt_0 および δt_1 において、光の位相が保存することから、

$$(A.9) \quad a(t_0)\nu_0 = a(t_1)\nu_1 \iff \frac{\lambda_0}{a(t_0)} = \frac{\lambda_1}{a(t_1)}$$

が得られる。ここで ν は光の振動数で、波長とは $\lambda\nu = c$ の関係がある。

膨張する宇宙では $a(t_0) > a(t_1)$ なので、観測される時刻での波長は放射のときよりも長くなる、即ち $\lambda_0 > \lambda_1$ となる。この波長の伸びが宇宙膨張による赤方偏移と呼ばれている。赤方偏移 z は

$$(A.10) \quad z \equiv \frac{\lambda_0 - \lambda_1}{\lambda_1} = \frac{a(t_0)}{a(t_1)} - 1$$

と定義される。

ここで、上記のようにスケール因子 $a(t)$ を $a(t_0) = 1$ となるよう規格化する。このとき $a(t)$ と z との関係は

$$(A.11) \quad z = \frac{1}{a(t)} - 1,$$

あるいは

$$(A.12) \quad a(t) = \frac{1}{1+z}$$

となる。

A.3 フリードマン方程式と宇宙論パラメータ

ここまで準備、およびインシュタイン方程式(Einstein's equation)を用いれば、膨張する宇宙を記述するための「運動方程式」を導くことができる。インシュタイン方程式とは、時空の歪みとエネルギー密度を関連付けた一般相対性理論の基本方程式であり、

$$(A.13) \quad R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu} + \Lambda g^{\mu\nu} = \frac{8\pi G}{c^4}T^{\mu\nu}$$

と書かれる。ここで $R^{\mu\nu}$ はリッチテンソル(Ricci tensor)、 $R \equiv R^\alpha_\alpha$ はそのトレースでリッチスカラー(Ricci scalar)、 G はニュートンの重力定数、 Λ は宇宙定数である。詳細は相対性理論の文献に譲るが、リッチテンソルは計量 $g_{\mu\nu}$ とその微分量の関数であり、時空の歪みを表現する2階の共変テンソルである(たとえば Foster and Nightingale, 2010)。FLRW 計量の場合、時空の運動方程式は

$$(A.14) \quad \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{c^2 K}{a^2} + \frac{c^2 \Lambda}{3}$$

$$(A.15) \quad \frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \left(\rho + \frac{3p}{c^2}\right) + \frac{c^2 \Lambda}{3},$$

となる。これはフリードマン方程式(Friedmann equations)と呼ばれている。ここで ρ は宇宙の密度, p は圧力を現す。式(A.14)はAINシュタイン方程式(式(A.13))の00成分, 式(A.15)はトレースから求められる。

さらに, 宇宙論パラメータと呼ばれる以下の量を定義すると, 宇宙膨張の物理量依存性がより具体的に記述できる。

- ハップブルパラメータ :

$$(A.16) \quad H(t) \equiv \frac{\dot{a}(t)}{a(t)}.$$

これは時間依存する量であるが, 観測的には現在値 H_0 が用いられることが多い。また無次元化した $h = H_0/100$ もよく用いる。

- 密度パラメータ :

$$(A.17) \quad \Omega_M(t) \equiv \frac{\rho(t)}{\rho_c(t)} \equiv \frac{8\pi G\rho(t)}{3H^2(t)},$$

ここで ρ_c は臨界密度パラメータと呼ばれる量で, 現在の宇宙では $\rho_c = 1.88 \times 10^{-29} h^2$ [g cm⁻³] と測定されている。

- 無次元宇宙項パラメータ :

$$(A.18) \quad \Omega_\Lambda(t) \equiv \frac{c^2 \Lambda}{3H^2(t)}$$

- 曲率パラメータ :

$$(A.19) \quad \Omega_K(t) \equiv -\frac{c^2 K}{a(t)^2 H^2(t)}$$

式(A.16)–(A.19)を用いると, 式(A.14)は

$$(A.20) \quad \Omega_M(t) + \Omega_\Lambda(t) + \Omega_K(t) = 1$$

と表せる。これは一般の宇宙年齢 t で成り立つの, 現在の宇宙年齢 t_0 では

$$(A.21) \quad \Omega_{M0} + \Omega_{\Lambda0} + \Omega_{K0} = 1$$

となる。本論文で採用した宇宙論パラメータ $h = H_0/100$ [km s⁻¹ Mpc⁻¹] = 0.7, $\Omega_{\Lambda0} = 0.7$, $\Omega_{M0} = 0.3$, 曲率パラメータ $\Omega_{K0} = 0$ は最新の精密宇宙論的観測から支持されるもので, 現在の宇宙が空間的に平坦で, 宇宙定数(または暗黒エネルギー)により加速膨張している宇宙モデルを示唆している。

B. 等級

検出器の単位面積を単位時間に通過するエネルギーを放射流束(flux)と呼び, さらに単位振動数当たり(あるいは単位波長当たり)の量を放射流束密度(flux density)と呼ぶ。そして天文学では, 放射流束あるいは放射流束密度の代わりに, その対数の -2.5 倍に対応する数値である等級(magnitude)を用いるのが習慣となっている。これは次のように表される。

$$(B.1) \quad m_{\nu_{\text{obs}}} = -2.5 \log_{10} S_{\nu_{\text{obs}}} + \text{定数}.$$

歴史的に、等級の定義に現れる定数はベガ(こと座 α 星)を 0 となるように定義されたが、これには物理的な根拠はなく、長らく混乱の元となってきた(たとえば Bessell, 2005)。

AB 等級(AB magnitude)は、放射流束密度から等級への系統的かつ物理的な変換のために Oke and Gunn (1983) によって導入された。AB 等級は次のように表される。

$$(B.2) \quad m_{\text{AB}, \nu_{\text{obs}}} = -2.5 \log_{10} S_{\nu_{\text{obs}}} - 48.60.$$

ここで、 S_{ν} の単位は $[\text{erg s}^{-1} \text{cm}^{-2} \text{Hz}^{-1}]$ である。下付きの “obs” は、その振動数が地球の静止座標系で観測される値であることを示す。また、下付き “em” はそれが光が放射されたときの銀河の静止系での振動数であることを表している。よって、宇宙論的赤方偏移 z を用いると、 $\nu_{\text{em}} = (1+z)\nu_{\text{obs}}$ となる。また定数 -48.60 は観測される振動数 ν にも波長 λ にも依存しないことに注意してほしい。したがって、絶対等級(absolute magnitude) $M_{\text{AB}, \nu}$ と単色光度(monochromatic luminosity) L_{ν} の関係は次のようにになる。

$$\begin{aligned} M_{\text{AB}, \nu_{\text{em}}} &= m_{\text{AB}, \nu_{\text{obs}}} - 25 - 5 \log_{10} d_L(z) \\ &= -2.5 \log_{10} \left[\frac{(1+z)L_{\nu_{\text{em}}}}{4\pi d_L(z)^2} \right] - 48.60 - 25 - 5 \log_{10} d_L(z) \\ (B.3) \quad &= -2.5 \log_{10} [(1+z)L_{\nu_{\text{em}}}] - 48.60 - 25 + 2.5 \log_{10}(4\pi). \end{aligned}$$

ここで

$$(B.4) \quad d_L(z) = \frac{c}{H_0} \int_0^z \frac{dz'}{\sqrt{\Omega_M(1+z')^3 + \Omega_\Lambda}} [\text{Mpc}]$$

は銀河の光度距離、 z はその赤方偏移、

$$(B.5) \quad \frac{c}{H_0} = 3000h^{-1} [\text{Mpc}],$$

はハップル長、そして

$$(B.6) \quad 1 [\text{pc}] = 3.086 \times 10^{18} [\text{cm}]$$

である(例えは Peebles, 1993)。

しかし、赤方偏移の効果は単純な観測波長の伸びだけではなく、観測する波長と放射した波長が異なることによる単色光度の変化分も考慮する必要がある。単位振動数当たりの放射流束密度は単色光度を用いて

$$(B.7) \quad S_{\nu_{\text{obs}}} d\nu_{\text{obs}} = \frac{L_{\nu_{\text{em}}} d\nu_{\text{em}}}{4\pi d_L(z)^2} = \frac{L_{\nu_{\text{obs}}(1+z)} d(1+z)\nu_{\text{obs}}}{4\pi d_L(z)^2} = \frac{(1+z) L_{\nu_{\text{obs}}} d\nu_{\text{obs}}}{4\pi d_L(z)^2}$$

となるので、

$$(B.8) \quad S_{\nu_{\text{obs}}} = \frac{(1+z) L_{\nu_{\text{obs}}}}{4\pi d_L(z)^2}$$

が得られる。よって単位波長当たりの放射流束密度 S_{λ} は

$$(B.9) \quad S_{\lambda_{\text{obs}}} = \frac{L_{\lambda_{\text{obs}}/(1+z)}}{4\pi d_L(z)^2(1+z)}$$

となる。正確には、観測される放射流束はある波長範囲だけの放射を透過させるフィルターを通した量である。観測するフィルターの有効波長 λ_0 における観測バンド流束 $S^{[\lambda_0]}$ $[\text{erg s}^{-1} \text{cm}^{-2}]$ を

$$(B.10) \quad S^{[\lambda_0]} \equiv \int S_\lambda R_\lambda^{[\lambda_0]} d\lambda$$

で定義する。ここで $R_\lambda^{[\lambda_0]}$ はフィルターの波長感度特性を表す関数で、応答関数と呼ばれる。観測される放射流束は式(B.9)より

$$(B.11) \quad S^{[\lambda_0]} = \frac{1}{4\pi d_L(z)^2(1+z)} \int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} d\lambda = \frac{\int L_\lambda R_\lambda^{[\lambda_0]} d\lambda}{4\pi d_L(z)^2(1+z)} \frac{\int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} d\lambda}{\int L_\lambda R_\lambda^{[\lambda_0]} d\lambda}$$

と表せる。ここで、次の項

$$(B.12) \quad \frac{\int L_\lambda R_\lambda^{[\lambda_0]} d\lambda}{4\pi d_L(z)^2(1+z)} \equiv \tilde{S}^{[\lambda_0]}$$

は宇宙論的赤方偏移がない場合に得られる仮想的なバンド放射流束を表している。式(B.11)の残りの部分は、赤方偏移によって天体のスペクトルの異なる波長範囲が観測されることで生じる放射流束の変化を表している。これが k 補正である。等級で表せば、

$$(B.13) \quad \begin{aligned} m_{\lambda_0} - M_{\lambda_0} &= -2.5 \log S^{[\lambda_0]} + 2.5 \log \left[\frac{4\pi d_L(z)^2(1+z)}{4\pi(10 \text{ [pc]})^2} \tilde{S}^{[\lambda_0]} \right] \\ &= 2.5 \log [d_L(z)^2(1+z)] - 2.5 \log \left[\frac{\int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} d\lambda}{\int L_\lambda R_\lambda^{[\lambda_0]} d\lambda} \right] + 25 \\ &= 5 \log d_L(z) + 2.5 \log(1+z) - 2.5 \log \left[\frac{\int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} d\lambda}{\int L_\lambda R_\lambda^{[\lambda_0]} d\lambda} \right] + 25 \end{aligned}$$

となる。式(B.13)の第2項と第3項は等級で表した k 補正(K 補正と書かれる)を表す(Oke and Sandage, 1968)。第2項は帯域幅の伸びの影響、第3項はバンド放射流束の変化分を表している。

C. 銀河の星形成史と可視光スペクトル

ここでは星形成史と銀河の可視光(紫外線から近赤外線まで)スペクトルの関係について定量的に説明する。簡単のため、銀河の可視光スペクトルが星のみの寄与からなると仮定すると、ある時刻 t における銀河のスペクトル $L_\lambda(t)$ は

$$(C.1) \quad L_\lambda(t) = \int_0^t \int_{\mathcal{M}_{\text{low}}}^{\mathcal{M}_{\text{up}}} \text{SFR}(t-\tau) \mathcal{F}_{\lambda, Z_{(t-\tau)}}(m, \tau) \Phi(m) dm d\tau,$$

で与えられる。ここで \mathcal{M}_{up} , \mathcal{M}_{low} は形成される星の質量の上限と下限、 $\Phi(m)$ は初期質量関数(initial mass function: IMF)、 $\mathcal{F}_{\lambda, Z_{(t-\tau)}}$ は質量 m , 金属量(metallicity) $Z(t-\tau)$ の星のスペクトルを表す。初期質量関数とは、集団的に形成される星の質量の分布関数(確率密度関数)に比例する量である。なお、ここで規格化は

$$(C.2) \quad \int_{\mathcal{M}_{\text{low}}}^{\mathcal{M}_{\text{up}}} m \Phi(m) dm = 1 [\mathcal{M}_\odot]$$

としている。即ち $\Phi(m)$ はどのくらいの質量の星がどのくらいの数形成されるかを表す。金属

量とは、星の中心における核融合によって生成された重元素が全星間物質の質量に占める割合である。時刻 t で観測される銀河の中の星はそれ以前の時刻に形成されたことを考慮するため、スペクトルおよび金属量が時刻 t ではなく $t - \tau$ で評価され、積分されている。

より詳細な理論モデルではこれに加えてガスからの放射を考慮し、星などから形成されたダスト(炭素やケイ素などの重元素からなる塵の微粒子)による減光を加味する必要があるが、本研究の主要な成果を理解するためには式(C.1)を考えれば十分である。しかし、これは古典的銀河進化理論、すなわち孤立した銀河が進化していくことを念頭に置いた理論によって与えられているが、1章で述べたように、銀河は内部での星形成だけでなく合体による成長によっても進化してゆく。つまり式(C.1)が表現できるのは合体していく銀河の断片の内的進化のみであり、銀河の統計的進化の記述のためには本研究のようなアプローチが必要になることを強調しておく。

参考文献

- Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Allende Prieto, C., An, D., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., Bailer-Jones, C. A. L., Barentine, J. C. and Bassett, B. A. (2009). The seventh data release of the Sloan digital sky survey, *Astrophysical Journal Supplement Series*, **182**(2), 543–558, <https://dx.doi.org/10.1088/0067-0049/182/2/543>.
- Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), 716–723, <https://dx.doi.org/10.1109/TAC.1974.1100705>.
- Bernstein, M., Silva, V. D., Langford, J. C. and Tenenbaum, J. B. (2001). Graph approximations to geodesics on embedded manifolds, Unpublished Technical Report, Stanford University, California.
- Bessell, M. S. (2005). Standard photometric systems, *Annual Review of Astronomy and Astrophysics*, **43**(1), 293–336, <https://dx.doi.org/10.1146/annurev.astro.41.082801.100251>.
- Blanton, M. R. (2006). Galaxies in SDSS and DEEP2: A quiet life on the blue sequence?, *Astrophysical Journal*, **648**(1), 268–280, <https://dx.doi.org/10.1086/505628>.
- Bouveyron, C. and Brunet, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace, *Statistics and Computing*, **22**(1), 301–324, <https://dx.doi.org/10.1007/s11222-011-9249-9>.
- Brosche, P. (1973). The manifold of galaxies. galaxies with known dynamical parameters, *Astronomy and Astrophysics*, **23**, 259–268.
- Chilingarian, I. V. and Zolotukhin, I. Y. (2012). A universal ultraviolet-optical colour-colour-magnitude relation of galaxies, *Monthly Notices of the Royal Astronomical Society*, **419**(2), 1727–1739, <https://dx.doi.org/10.1111/j.1365-2966.2011.19837.x>.
- Chilingarian, I. V., Zolotukhin, I. Y., Katkov, I. Y., Melchior, A.-L., Rubtsov, E. V. and Grishin, K. A. (2017). RCSED—A value-added reference catalog of spectral energy distributions of 800,299 galaxies in 11 ultraviolet, optical, and near-infrared bands: Morphologies, colors, ionized gas, and stellar population properties, *Astrophysical Journal Supplement Series*, **228**(2), <https://dx.doi.org/10.3847/1538-4365/228/2/14>.
- Cooray, S., Takeuchi, T. T., Kashino, D., Yoshida, S. A., Ma, H.-X. and Kono, K. T. (2023). Characterizing and understanding galaxies with two parameters, *Monthly Notices of the Royal Astronomical Society*, **524**(4), 4976–4995, <https://dx.doi.org/10.1093/mnras/stad2129>.
- Cranmer, M. (2023). Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, arXiv, <https://doi.org/10.48550/arXiv.2305.01582>.
- Dijkstra, E. W. (1959). A note on two problems in connexion with graphs, *Numerische Mathematik*, **1**(1), 269–271.
- Djorgovski, S. (1992). *Morphological and Physical Classification of Galaxies* (eds. G. Longo, M.

- Capaccioli and G. Busarello), 337–356, Springer Netherlands, Dordrecht.
- Farahmand, A. m., Szepesvári, C. and Audibert, J.-Y. (2007). Manifold-adaptive dimension estimation, *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, 265–272, Association for Computing Machinery, New York, <https://dx.doi.org/10.1145/1273496.1273530>.
- Floyd, R. W. (1962). Algorithm 97: Shortest path, *Communications of the ACM*, **5**(6), 345.
- Foster, J. and Nightingale, J. (2010). *A Short Course in General Relativity*, Springer, New York.
- Ginolfi, M., Hunt, L. K., Tortora, C., Schneider, R. and Cresci, G. (2020). Scaling relations and baryonic cycling in local star-forming galaxies, I. The sample, *Astronomy and Astrophysics*, **638**, <https://dx.doi.org/10.1051/0004-6361/201936304>.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*, MIT Press, Cambridge, Massachusetts.
- Hunt, L., Magrini, L., Galli, D., Schneider, R., Bianchi, S., Maiolino, R., Romano, D., Tosi, M. and Valiante, R. (2012). Scaling relations of metallicity, stellar mass and star formation rate in metal-poor starbursts—I. A Fundamental Plane, *Monthly Notices of the Royal Astronomical Society*, **427**(2), 906–918, <https://dx.doi.org/10.1111/j.1365-2966.2012.21761.x>.
- Lilly, S. J., Carollo, C. M., Pipino, A., Renzini, A. and Peng, Y. (2013). Gas regulation of galaxies: The evolution of the cosmic specific star formation rate, the metallicity-mass-star-formation rate relation, and the stellar content of halos, *The Astrophysical Journal*, **772**(2), <https://dx.doi.org/10.1088/0004-637X/772/2/119>.
- Lin, L., St. Thomas, B., Zhu, H. and Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data, *Journal of the American Statistical Association*, **112**(519), 1261–1273, <https://dx.doi.org/10.1080/01621459.2016.1208615>.
- Liu, S., Maljovec, D., Wang, B., Bremer, P.-T. and Pascucci, V. (2017). Visualizing high-dimensional data: Advances in the past decade, *IEEE Transactions on Visualization and Computer Graphics*, **23**(3), 1249–1268, <https://dx.doi.org/10.1109/TVCG.2016.2640960>.
- Ma, Y. and Fu, Y. (2012). *Manifold Learning Theory and Applications*, CRC Press, Boca Raton.
- McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection, *Journal of Open Source Software*, **3**(29), <https://dx.doi.org/10.21105/joss.00861>.
- McInnes, L., Healy, J. and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for dimension reduction, arXiv, <https://dx.doi.org/https://doi.org/10.48550/arXiv.1802.03426>.
- Oke, J. B. and Gunn, J. E. (1983). Secondary standard stars for absolute spectrophotometry, *Astrophysical Journal*, **266**, 713–717, <https://dx.doi.org/10.1086/160817>.
- Oke, J. B. and Sandage, A. (1968). Energy distributions, K corrections, and the Stebbins-Whitford effect for giant elliptical galaxies, *The Astrophysical Journal*, **154**, <https://dx.doi.org/10.1086/149737>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in python, *Journal of Machine Learning Research*, **12**, 2825–2830.
- Peebles, P. J. E. (1993). *Principles of Physical Cosmology*, Princeton University Press, Princeton, New Jersey, <https://dx.doi.org/10.1515/9780691206721>.
- Rodighiero, G., Daddi, E., Baronchelli, I., Cimatti, A., Renzini, A., Aussel, H., Popesso, P., Lutz, D., Andreani, P., Berta, S., Cava, A., Elbaz, D., Feltre, A., Fontana, A., Förster Schreiber, N. M., Franceschini, A., Genzel, R., Grazian, A., Gruppioni, C., Ilbert, O., Le Floc'h, E., Magdis, G., Magliocchetti, M., Magnelli, B., Maiolino, R., McCracken, H., Nordon, R., Poglitsch, A., Santini, P., Pozzi, F., Riguccini, L., Tacconi, L. J., Wuyts, S. and Zamorani, G. (2011). The lesser role of starbursts in star formation at $z = 2$, *The Astrophysical Journal*, **739**(2), <https://dx.doi.org/10.1088/2041-8205/739/2/L40>.
- Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding,

- Science*, **290**(5500), 2323–2326, <https://dx.doi.org/10.1126/science.290.5500.2323>.
- Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464, <https://dx.doi.org/10.1214/aos/1176344136>.
- Siudek, M., Małek, K., Pollo, A., Krakowski, T., Iovino, A., Scodeggio, M., Moutard, T., Zamorani, G., Guzzo, L., Garilli, B., Granett, B. R., Bolzonella, M., de la Torre, S., Abbas, U., Adami, C., Bottini, D., Cappi, A., Cucciati, O., Davidzon, I., Franzetti, P., Fritz, A., Krywult, J., Le Brun, V., Le Fèvre, O., Maccagni, D., Marulli, F., Polletta, M., Tasca, L. A. M., Tojeiro, R., Vergani, D., Zanichelli, A., Arnouts, S., Bel, J., Branchini, E., Coupon, J., De Lucia, G., Ilbert, O., Haines, C. P., Moscardini, L. and Takeuchi, T. T. (2018). The VIMOS Public Extragalactic Redshift Survey (VIPERS), The complexity of galaxy populations at $0.4 < z < 1.3$ revealed with unsupervised machine-learning algorithms, *Astronomy & Astrophysics*, **617**, <https://dx.doi.org/10.1051/0004-6361/201832784>.
- Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**(5500), 2319–2323, <https://dx.doi.org/10.1126/science.290.5500.2319>.
- Tinsley, B. M. (1980). Evolution of the stars and gas in galaxies, *Fundamentals of Cosmic Physics*, **5**, 287–388.
- Warshall, S. (1962). A theorem on boolean matrices, *Journal of the ACM*, **9**(1), 11–12.
- Whitaker, K. E., van Dokkum, P. G., Brammer, G. and Franx, M. (2012). The star formation mass sequence out to $z = 2.5$, *The Astrophysical Journal*, **754**(2), <https://dx.doi.org/10.1088/2041-8205/754/2/L29>.
- Zhang, H. and Zaritsky, D. (2016). Examining early-type galaxy scaling relations using simple dynamical models, *Monthly Notices of the Royal Astronomical Society*, **455**(2), 1364–1374, <https://dx.doi.org/10.1093/mnras/stv2413>.

New Quantification of Galaxy Evolution with Manifold Learning

Tsutomu T. Takeuchi^{1,2}, Sucheta Cooray^{3,4}, Taisei D. Yamagata¹, Aina May So^{1,6},
 Shun-Ya S. Uchida¹, Shiro Ikeda², Kenji Fukumizu², Ryusei R. Kano¹,
 Kyoaki Christopher Omori^{1,5}, Hai-Xia Ma¹, Wen Shi¹ and Sena A. Matsui¹

¹Division of Particle and Astrophysical Science, Nagoya University

²The Institute of Statistical Mathematics

³Division of Science, National Astronomical Observatory of Japan

⁴Kavli Institute Particle Astrophysics and Cosmology, Stanford University

⁵Department of Astronomy and Physics, Saint Mary's University

⁶Department of Physics, Gakushuin University

Matter in the early Universe was almost uniform, and a slightly dense region grew by gravity, finally into a galaxy. Astrophysics has been trying to unveil the complex physical phenomena that caused the formation and evolution of galaxies through the history of the Universe of 13 billion years from the first principles of physics. However, since present-day astrophysical big data contain more than 100 explanatory variables, such a conventional methodology faces a limit to deal with such data. We, instead, elucidate the physics of galaxy evolution by applying manifold learning, one of the latest methods of data science, to a feature space spanned by galaxy luminosities and cosmic time. We discovered a low-dimensional nonlinear structure of data points in this space, referred to as the galaxy manifold. We found that the galaxy evolution in the ultraviolet–optical–near infrared luminosity space is well described by two parameters, star formation and stellar mass evolution, on the manifold. We also discuss a possible way to connect the manifold coordinates to physical quantities.

2025年4月改訂

「統計数理」投稿規程

1. 「統計数理」は、統計科学の深化と発展、そして統計科学を通じた社会への貢献を目指すものである。投稿原稿は、統計科学に関連した内容を持つもので、和文の原稿に限る。
2. 投稿原稿は次の6種とする。
 - a. 原著論文 (Paper)
統計科学の発展に貢献すると考えられる研究結果。
 - b. 総合報告 (Review Article)
特定の主題に関する一連の研究およびその周辺領域の発展を著者の見解に従って総括的、かつ体系的に報告したもの。
 - c. 研究ノート (Letter)
研究速報、新しい発想、提言、問題提起、事例報告など研究上、記録にとどめておく価値があると認められるものや、既発表の論文等に対するコメントで、研究上、記録にとどめておく価値があると認められるもの。
 - d. 研究詳解 (Research Review)
特定の研究領域における理論的あるいは応用的成果を、最近の結果や知見を加えてわかりやすく説明したもの。
 - e. 統計ソフトウェア (Statistical Software)
有用な計算法や解析法に関する短いプログラムおよびサブルーチンのリスト、利用手引き、実行例など。
 - f. 研究資料 (Research Archives)
歴史的なデータ、入手困難なデータや統計的手法の比較検討のために有用なデータ、あるいは、歴史的文献の翻訳や解説など。
3. いずれも原則として、未発表のものに限る。
4. 投稿された原稿は、編集委員会が選定・依頼した査読者の審査を経て、掲載の可否を決定する。
5. 投稿原稿は電子投稿査読システム <https://www.editorialmanager.com/toukei/> より投稿するものとする。原稿はpdfファイルとし、必要なフォントはすべて埋め込み、原稿全体を一つのファイルにまとめるとしている。論文が採択になった場合、著者は最終稿のソースファイルとハードコピーを提出するものとする。
6. 著作権
 - (1) 掲載される論文等の著作権はその採択をもって統計数理研究所に帰属するものとする。統計数理研究所は、紙媒体の「統計数理」のほか電子媒体などを通じて論文等を公表することができる。特別な事情がある場合は、著者と本編集委員会との間で協議の上措置する。
 - (2) 投稿原稿の中で引用する文章や図表の著作権に関する問題は、著者の責任において処理する。
 - (3) 著者が自分の論文等を複製、転載、翻訳、翻案等の形で利用するのは自由である。この場合、著者は掲載先に出典を明記する。
7. 原稿は次の執筆要項に従って作成する。

「統計数理」執筆要項

1. 原稿はA4用紙に1行36字から40字で1行おき、1頁あたり22行程度とする。原稿の長さは原則として表・図を含めて30頁相当以内とし、各ページにページ番号を付す。図表は別紙にまとめ、本文中には挿入箇所のみを指定する。LaTeXで原稿を作成する場合は、「統計数理」スタイルファイルの使用を推奨する。
<https://www.ism.ac.jp/editsec/toukei/>
2. 文体は「である体」とし、句読点は「、」「。」を用いる。
3. 原稿は以下の順に書くものとする。

[第1頁] 標題、著者名、所属名および所在地、メールアドレス、和文要旨(500字程度、文献の引用および数式は原則として避ける)、和文キーワード(6語以内)。

[第2頁] 英語による標題、著者名、所属名、Abstract(450ワード程度)、Key words(6 words and phrases以内)。Abstractは、問題の所在と得られた結果等がそれだけで理解できるようなものとする。

[第3頁以降]

- ① 本文：章、節の番号は、第1章にあたるものは、“1.”、第1章第1節にあたるものは、“1.1”というようにつける。また、式の番号は、章ごとに(2.1)、(2.2)のようにし、式の左側に配置する。
 - ② 数式：数式は簡明さを心がけ、添字にさらに添字をつけるのはなるべく避ける。
 - ③ 参考文献：書き方は本要項第4項を参照。
 - ④ 表：一枚の用紙に一つの表を書く。表の番号は論文中に現れる順に従って、表1、表2、…または、Table 1, Table 2、…のようにする。
 - ⑤ 図：一枚の用紙に一つの図を描く。図はそのまま写真製版できる鮮明なものを用意する。大きさは印刷出来上がりの1~2倍とし、トレースが必要な場合は原則として著者が行うものとする。図の番号は論文中に現れる順に従って、図1、図2、…または、Fig. 1, Fig. 2、…のようにする。
 - ⑥ 注：本文中の注釈は極力避ける。やむを得ず注釈をつける場合は脚注とせず、論文末尾に後注とする。後注は、順番に“1, 2, …”の番号を付け、本文中では上付きで示す。
4. 本文中の参考文献の引用は、著者名(出版年)とする。たとえば、Efron (1982), 清水・湯浅 (1984), Cox and Snell (1981), 坂元 他 (2004), Nakano et al. (2000)。
5. 参考文献の書き方

[雑誌の場合] いざれの場合も雑誌名は省略しないものとする。

① ページ番号がある文献

著者名(出版年)、標題、雑誌名、巻、ページ〔始-終〕。

【例】Chernoff, H. (1973). The use of faces to represent points in k -dimensional space graphically, *Journal of the American Statistical Association*, **68**, 361–368.

【例】Bligh, E. G. and Dyer, W. J. (1959). A rapid method of total lipid extraction and purification, *Canadian Journal of Biochemistry and Physiology*, **37**, 911–917, <https://doi.org/10.1139/o59-099>.

② ページ番号がない文献

著者名(出版年)、標題、雑誌名、巻に続けて、DOIがあるものはDOIを入れ、DOIがない場合はURLと最終アクセス日を記載する。

【例】Lien, G. Y., Kalnay, E. and Miyoshi, T. (2013). Effective assimilation of global precipitation: Simulation experiments, *Tellus A*, **65**, <https://doi.org/10.1175/WAF-D-13-00032.1>.

【例】Kiraly, F. J. and Oberhauser, H. (2019). Kernels for sequentially ordered data, *Journal of Machine Learning Research*, **20**(31), <http://jmlr.org/papers/v20/16-314.html> (最終アクセス日 2023年10月1日)。

[叢書の中の一巻の場合]

著者名(出版年)、書名(編集者名)、叢書名、発行所名、発行地名。

【例】Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1983). *Akaike Information Criterion Statistics*, Mathematics and Its Applications, Reidel, Dordrecht.

[単行本等の場合]

著者名(出版年)、書名、発行所名、発行地名。

【例】Cressie, Noel (1993). *Statistics for Spatial Data*, Wiley, New York.

[編集書の中の一部の場合]

著者名(出版年)、標題、編集書名(編集者名)、巻、ページ、発行所名、発行地名。

【例】Akaike, H. (1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 143–166, University Press, Valencia, Spain.

なお、同じ著者によるものが同一年に複数個現れる場合には、(1980a), (1980b)などとして区別する。文献は、日本人も含め、著者名のアルファベット順に並べる。

6. 著者校正は原則として一回とする。その際、印刷上の誤り以外の字句や図版の訂正、挿入、削除等は原則として認めない。