

令和 6 年 6 月 16 日現在

機関番号：14201

研究種目：基盤研究(C) (一般)

研究期間：2021～2023

課題番号：21K11783

研究課題名(和文) 不完全情報を伴う多次元の事象時間データの推測と機械学習法

研究課題名(英文) Inferences and machine learning methods for multivariate time-to-event data with incomplete information

研究代表者

杉本 知之 (Sugimoto, Tomoyuki)

滋賀大学・データサイエンス学系・教授

研究者番号：70324829

交付決定額(研究期間全体)：(直接経費) 3,100,000円

研究成果の概要(和文)：(1)層別解析における試験間変動の正規ランダム効果モデルの推測理論の精緻化を研究した。2値データのMantel-Haenszel型の層別解析法において計算代数統計による正確計算の研究をログランク検定に応用した。(2)Cox回帰での時間変化共変量の利用のための多重代入法の研究を行い、2変量生存時間モデルに対する群逐次デザインの研究を発展させた。コピュラ型の相関構造をもつ2変量イベント時間データに対して、セミ競合リスク問題の推測の理論と方法を研究し、NPMLEのための計算理論、ログランク統計量の修正に関する研究を行った。これらの研究における諸問題にいくつかの解決法を与え、研究論文等にまとめた。

研究成果の学術的意義や社会的意義

層別解析は交絡調整の伝統的な統計手法だがランダム効果モデルといった多用される方法論においても、その標本分布は精緻化されていない側面があったが、本研究によりそれらの問題を有意義に解消する基盤が創出できた。多次元データのとり扱いに必要となる多変量分布論において多変量正規分布を有意義に超えるものは多くないが、本研究ではコピュラ型相関をもつ2変量分布をイベント時間データに定式化し、医学統計の応用において興味ある展開のいくつかを惹き出した。今後において決定木やランダムフォレストといった機械学習の方法に層別解析の方法論を融合させていくための足掛かりを得ることができ、学術的および社会的意義をもつ。

研究成果の概要(英文)：(1) We studied the refinement of the inference theory of the normal random effects model of between-trial variation in stratified analysis, and applied our research on exact computation by computational algebra statistics to the log-rank test in the Mantel-Haenszel type stratified analysis method for binary data. (2) We studied multiple assignment methods for the use of time-varying covariates in Cox regression, and extended our research on group sequential design for bivariate survival time models. For bivariate event-time data with copula-type correlation structure, we studied the theory and methods of inference for semi-competitive risk problems, and conducted research on computational theory for NPMLE and modification of log-rank statistics. Some solutions to various problems in these studies were given and summarized in research papers and other publications.

研究分野：医学統計，数理統計，機械学習

キーワード：生存時間解析 コピュラモデル 計数過程 ランダム効果解析 層別解析 木構造モデリング 多変量標本分布 計算機統計

科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

1. 研究開始当初の背景

統計理論や方法論の研究、樹木アプローチに基づく機械学習の研究を通じて、がん・循環器病などの医学研究では、複数種類のイベントによる多次元事象時間データの影響を評価すること、経済データでは、複数の企業イベントにより休業や倒産の確率を予測することが求められ、カレンダー時間により変化をする年代的ハザードと個人生存時間ハザードの分離モデリングの応用例に直面したことがあげられる。ビックデータにより機械学習の性能の高まりに関心が集まるようになったが、以前より樹木アプローチによる機械学習法を取り入れ、これまでに方法論や計算環境の整備を行ってきた。ただし、医療、経済のデータ分析、異常検知分析などにおいて、予測だけではなく、より解釈可能な統計的モデリングの構築の実践は求められる。

多次元事象時間データの推測では、点過程に対するマルチンゲール接近法を、イベントの種類といった方向にも展開する必要がある。ただし、通常の変量マルチンゲール中心極限定理は、変量といっても同一時間軸に対して構築されたものであり、異なる方向の時間軸に対して適用可能なものはあまり整備されていなかったといえるが、これまでの研究を通して、そのための理論的ツールは揃ってきている。

実地の問題に適合するように、分析モデルのデータ適用範囲を拡大したり、拡張を行えば、不完全データを伴う形で推測しなければならないことが多いことを経験している。このとき、3タイプの部分尤度法による方法論開発、推定量の漸近性質や有限標本挙動を捉える研究を行い、不完全情報を伴う場合の本研究を実施するための準備を行っていた。

2. 研究の目的

不完全情報を伴うときの多次元の事象時間データの推測の理論、方法論、その機械学習法の展開を如何にすればより有意義に構築できるかを明らかにしていくことを研究の目的として遂行していく。

いわゆる正規分布に基づく「多変量解析」では、多変量正規の性質の良さからかなりの発展的領域に到達している。一方、事象時間データの多次元推測では、ノンパラメトリック推測が基本であり、点過程マルチンゲール接近法を、時間方向だけでなく、イベントの種類といった多次元方向にも展開する必要がある。多変量マルチンゲール中心極限定理は、通常、同一時間系列に対して構築されたものであり、多次元の時間事象データにとって必要な異なる時間軸に対して適用可能なものは、例えば、Sugimoto et al(2020)において理論的ツールを整備している。また、多次元化のモデリングには、現実的なデータに照らし合わせれば、いくつかの不完全情報を含むことも必須となる。このようなことを考慮して、より有意義な多次元推測と機械学習を組み込む展開を研究する。

3. 研究の方法

具体的な研究の方向性は、Cox 回帰やその拡張型を、不完全データを伴う場合の推測の理論、方法論、応用面を多次元の事象時間データの上に展開すること、さらにその中に樹木構造アプローチなどの機械学習法を接続・融合させる研究を行う。実地における有意義な統計現象の問題が扱えるような、Cox 型計数過程モデルの多次元化をとりあげ、データ分析のために有益なツールの開発を行うことである。とくに、コピュラ型の相関構造をもつ 2 変量事象時間データに対して、セミ競合リスク問題の推測の研究をさらに発展させる。群逐次型 2 変量ログランク統計量によるデザインと解析、さらに情報寄与ありの打ち切りに対する多次元ログランク統計量や多次元 Cox 回帰の修正版を研究する。周辺プロフィール尤度による推測に問題が生じるときは、周辺部分尤度からの解法も求め、両者のハイブリッドを検討する。また、社会変化をとり入れ経済的な分析にも拡張できるように、イベント時間とカレンダー時間のハザード関数を有意義に分離して表現した 2 変量ハザード・モデルのセミパラメトリック推測の研究も発展させていく。

4. 研究成果

Cox 回帰モデルの解法である部分尤度法は、層別分析法の一つの帰結であると見ることでもできる。本研究を通じて、層別解析における試験間変動の正規ランダム効果モデルの推測理論の精緻化を研究し、その応用として、ランダム効果メタアナリシスの DerSimonian-Laird 法に対する正確な推測法をまとめた。2 値データの Mantel-Haenszel 型の層別解析法において計算代数統計による正確計算の研究を行い、その結果を、ログランク検定に応用してまとめた。Cox 回帰での時間変化共変量の利用のために、ガウス過程回帰と多重代入法の研究を行い、その成果をまとめた。また、2 変量生存時間モデルに対する群逐次デザインの研究を進展させ、いくつかの理論結果と、数値結果をまとめている。コピュラ型の相関構造をもつ 2 変量イベント時間データに対して、セ

ミ競合リスク問題の推測の理論と方法を研究し, NPML のための計算理論, ログランク統計量の修正に関する研究を行った. これらの研究における諸問題にいくつかの解決法を与え, 研究論文等にまとめた.

5. 主な発表論文等

〔雑誌論文〕 計3件（うち査読付論文 2件 / うち国際共著 0件 / うちオープンアクセス 0件）

1. 著者名 Hanada Keisuke, Sugimoto Tomoyuki	4. 巻 75
2. 論文標題 Inference using an exact distribution of test statistic for random-effects meta-analysis	5. 発行年 2022年
3. 雑誌名 Annals of the Institute of Statistical Mathematics	6. 最初と最後の頁 281 ~ 302
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s10463-022-00844-4	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 杉本 知之、田中 健太	4. 巻 52
2. 論文標題 2変量生存時間モデルにおけるコピュラとその利用	5. 発行年 2023年
3. 雑誌名 日本統計学会誌	6. 最初と最後の頁 153 ~ 176
掲載論文のDOI (デジタルオブジェクト識別子) 10.11329/jjssj.52.153	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 水間 浩太郎、杉本 知之	4. 巻 52
2. 論文標題 計算代数統計を用いたログランク検定	5. 発行年 2023年
3. 雑誌名 日本統計学会誌	6. 最初と最後の頁 355 ~ 371
掲載論文のDOI (デジタルオブジェクト識別子) 10.11329/jjssj.52.355	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

〔学会発表〕 計5件（うち招待講演 2件 / うち国際学会 0件）

1. 発表者名 川崎大輔, 杉本知之
2. 発表標題 2変量2値データに対する群逐次同時信頼区間について
3. 学会等名 日本計算機統計学会第36回シンポジウム
4. 発表年 2022年

1. 発表者名 杉本知之
2. 発表標題 多変量生存時間モデル
3. 学会等名 2022年度計量生物セミナー「生存時間解析」(招待講演)
4. 発表年 2022年

1. 発表者名 田中健太, 杉本知之
2. 発表標題 時間共変量を含むCox回帰モデルにおけるガウス過程回帰と多重代入法の利用について
3. 学会等名 2021年度統計関連学会連合大会
4. 発表年 2021年

1. 発表者名 杉本知之
2. 発表標題 2変量事象時間アウトカムと群逐次ログランク法について
3. 学会等名 2021年度統計関連学会連合大会(招待講演)
4. 発表年 2021年

1. 発表者名 小畑 諒人, 杉本 知之
2. 発表標題 経時測定データに対する樹木モデルの構築法とその性能評価
3. 学会等名 日本計算機統計学会第37回シンポジウム
4. 発表年 2023年

〔図書〕 計1件

1. 著者名 杉本知之	4. 発行年 2021年
2. 出版社 朝倉書店	5. 総ページ数 240
3. 書名 生存時間解析	

〔産業財産権〕

〔その他〕

-

6. 研究組織

氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
---------------------------	-----------------------	----

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------