TSUBAME 共同利用　令和 5 年度　学術利用　成果報告書

臨床情報統合データベースの機械学習解析
Machine Learning for Integrated Database of Clinical Information

浅井　聰
Satoshi Asai

日本大学　医学部　生体機能医学系　薬理学分野
Department of Pharmacology, School of Medicine, Nihon University
https://www.med.nihon-u.ac.jp/department/pharmacology/

臨床情報統合データベースに含まれる薬物投与・過去の診断や検査結果の履歴に応じて疾患の発症・合併症進行リスクがどのように変動するのかといった効果推定を行うための複数のガウス過程を用いた階層ベイズモデルを構築し、前年度の結果に基づきそのためのモンテカルロアルゴリズムを開発した。また前年度開発した大規模時系列データのためのマルチカーネル法のプログラムコードを用いて、抗コリン薬の認知症リスクを見積もった。効率的な階層モデルにおけるベイズマルチカーネル法のアルゴリズムと、抗コリン薬とその他の薬理作用の交絡を示唆する所見を得た。

We developed a Monte-Carlo algorithm for hierarchical Bayesian model based on multiple Gaussian processes by extending the algorithm developed in the previous year, in order to estimate effects of drugs, past diseases and blood-test results on the risk of disease onset. We also applied previously developed efficient programs for risk estimation in a large clinical time-series dataset, in order to estimate the risk of dementia onset associated with use of anticholinergics. We succeeded to develop an efficient algorithm for hierarchical Bayesian multiple-kernel models and identified the risk associated with anticholinergics together with results indicating confounding with other factors.

Keywords: electronic medical records, pharmacoepidemiology, Bayesian analysis, multiple-kernel learning, Monte-Carlo method

背景と目的

本研究の背景及び目的は前年度・前々年度の我々の同名の課題と概ね同じであり、まずこれを以下に引用する。（**引用開始**）近年、電子カルテデータなどを含む大規模な医療データベースから薬剤の効果・疾患リスクなどの医学的知見を機械学習・人工知能技術を用いて抽出する試みが注目を集めている。実際に、世界各国の大学病院や地域中核病院の電子カルテデータを用いた解析が行われはじめており[1,2]、さらに適切な倫理的な枠組みのもと、電子カルテデータに患者から採取した遺伝·生化学的測定データを組み合わせ、これを患者個別医療に役立てようとする試みがはじまっている。

しかしながら、近年の機械学習·人工知能技術の発展にもかかわらず、上記のような大規模医療データから有益な医学的知見を抽出するためには解決すべき技術的課題が複数存在する。例えば、医療データベースの解析では薬剤投与・検査値推移・疾患発症など数千以上の予測変数から疾患進行を予測しようとするが、実際に疾患進行に影響する変数は少数であり、変数選択を効率的に行う必要がある。また臨床医が判断する重症度のような潜在変数の情報をアルゴリズムによって推定できることも必要だ。これらを変数間の非線形関係も考慮しながら行うことは，潜在変数付きノンパラメトリックベイズモデルでスパースな変数選択を行う問題に帰着される(次ページ図参照)。上記のような潜在変数付きのノンパラメトリックベイズの枠組みにおいて、医療データベースのような大規模データ上でスパースかつ効率的な推定に成功した研究はまだ存在しない。先行研究のほとんどで潜在変数のモデリングは巧妙に避けられている。実際、医療データに限らずノンパラメトリックベイズモデルにおける潜在変数の推定やスパース変数選択は、機械学習一般の問題として依然として困難な問題の一つであり、効率的な解法が模索されている[3]。申請者のグループは、下にも述べるように、この問題を解決しうる新奇なアルゴリズムを得たので、本課題

において TSUBAME を利用した大規模実装を行いこの手法の有用性を示そうとする。**（引用終了）**

本年度では上記の文脈において、(1) 前年度までに開発したベイズマルチカーネル法のアルゴリズムの有用性を基本的な応用例において示すこと、特に実世界の問題において必要とされる階層的モデリングに適用可能とすること、(2)前年度までに開発した大規模臨床時系列データに適用可能な高効率マルチカーネルアルゴリズムを用いた抗コリン薬の認知症リスク推定を実行すること、を目的とした。

概要

隠れ状態を含む階層的な因果グラフにおいて、各因子の効果推定を複数のガウス過程を用いて行うアルゴリズムを開発した。前年度開発したアニールド重点サンプリング修正ヘッセ多様体ハミルトンモンテカルロ法を改変して実装したところ、既存のライブラリよりも計算量オーダーの少ない実装が得られ、効率的に推定を行うことに成功した。

また前々年度開発したマルチカーネル法の大規模実装を用いて、倫理審査を経て 20 年分の電子カルテデータから抗コリン薬に関連する認知症リスクの非線形推定を行なった。推定の結果、抗コリン薬そのものに対する関連は有意には見られず、抗コリン薬の適応となる疾患や抗コリン薬の持つ別の受容体作用に対して有意に関連する認知症リスクが推定された。

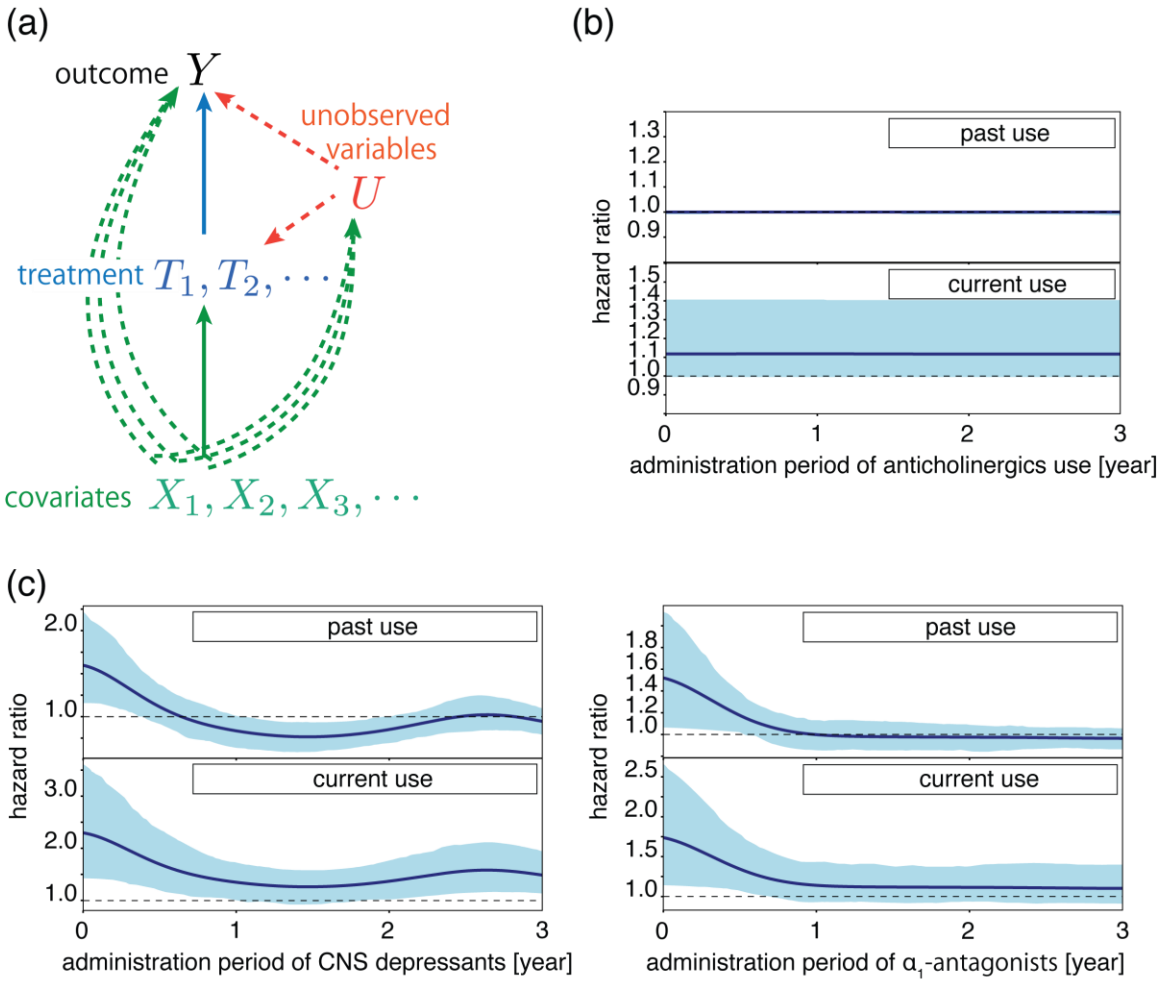

図 1 (a) 潜在変数を含む階層モデルを表す因果グラフ．各効果がガウス過程によってモデリングされる．(b) 推定された抗コリン薬の過去/現在の投与及びその投与期間と認知症発症リスクの関連．(c) 推定された中枢神経抑制薬と抗アドレナリン $\alpha 1$ 拮抗薬の過去/現在の投与及びその投与期間と認知症発症リスクの関連．

結果および考察

(1)階層モデルにおけるベイズマルチカーネル法とその医療統計への応用

前年度に開発したアニールド重点標本修正ヘッセ多様体ハミルトニアンモンテカルロ法の性能を実証するべく、図 1(a)の因果グラフによって示されるような階層モデルにおける問題設定に応用できるよう実装した。これは治療変数 Tの結果変数 Yへの効果を共変量 Xと非観測変数 U の調整下に見積もるという、医療統計における典型的な問題設定となっている(例えば[4])。この伝統的な問題設定において依然として伝統的な医療統計の手法(線形推定に基づく傾向スコアとそれによる重み付けによる方法等)が用いられる理由は、機械学習アルゴリズムでは効果の有意性を議論できないことにある。すなわち共変量 X=($X_1, X_2, \cdots, X_d$) のうちどの変数に対して調整をかけるべきか、あるいは治療変数 T=($T_1, T_2, \cdots, T_k$)のうちどの変数に対する効果がありどの変数に対してないのか、という問題に答えるには、Bayesian Model Evidence に基づいて、変数選択の良し悪しを評価する必要があるが, このような評価は従来の機械学習研究では与えられていないことが多い(例えば[5])。この問題は因果探索と呼ばれる問題設定に該当し、線形推定の枠組みでは LiNGAM[6]と呼ばれる効率の良い推定方法が知られるが、非線形の場合は限定的な結果が知られるにとどまる[7]。前年度に実装したアルゴリズムを改変し、L 階層のモデルにおいて因子間の因果効果を有限 M 次元で打ち切ったガウス過程を用いてモデリングし、N 標本を用いて推定する問題において、1 モンテカルロステップあたり $O(LNM^2)$の計算量のアルゴリズムを得た。これは、従来用いられてきた自動微分に基づくリーマン多様体上のハミルトンモンテカルロ法(例えば[8])が O(LNM^3)の計算量を必要とすることと対照的である。実際、小規模な問題においてアルゴリズムを動作させることに成功し、Bayesian Model Evidence を見積もることに成功した。この結果をさらに発展させ、非線形因果探索を推し進めていくことができると考えられる。

(2)非ベイズマルチカーネル法を用いた抗コリン薬使用に関連する認知症リスク推定

抗コリン薬の使用は従来認知症発症リスクを上昇させると考えられてきた[9,10]。しかしながら、抗コリン薬を使用する対象の疾患である 鼻アレルギー・うつ・不安障害・膀胱障害・パーキンソン病などの疾患自体が認知症発症のリスクでもあり、また抗コリン薬の多くがヒスタミン受容体・ドーパミン受容体・セロトニン受容体など他の受容体にも作用するため、抗コリン作用が認知症発症リスクを上昇させているのかどうかが定かでない、という問題点もあった。そこで本研究では、抗コリン作用の適応のある疾患の有無や罹患期間、抗コリン薬の持つ他の受容体作用を持つ薬剤の内服の有無や投与機関といった複数の変数を用いて、前々年度に開発したマルチカーネル法のプログラムコード(詳細は[11]を参照)を用いて認知症リスクの非線形推定を行なった。その結果、図 1(b)の推定結果に示されるように、抗コリン薬の使用と罹患期間に対しては有意な認知症リスクの関連が見られず、むしろ図 1(c)に示されるようにドーパミン受容体・中枢神経抑制薬全般・などに対する関連が見られた。これらの結果をまとめて、論文発表を準備中である。

まとめ、今後の課題

本課題の結果により、隠れ状態を含む階層的なモデルのような特異モデルにおいても効果推定及び Bayesian Model Evidence に基づく推定結果のエビデンス評価が可能になった。現状では小規模なモデルに対してのみの結果であるが、今後より大規模なモデル・データセットへの適用範囲の拡大の余地があり、これを進めていく。特に、長い時系列に対する潜在変数の推定方法である祖先サンプリング付き粒子ギブス法との組み合わせのためのプログラムコードの開発を進めており、これを引き続き行なっていく。規模を拡大する際には計算量の側面からの制約も想定されるが、全体としてはアニールド重点サンプリングによる特異性の取り扱いを行いながらも部分的にラプラス近似ができる局面ではこれを用いて計算量を減らす工夫等を取り入れていけば解決できると考えている。

抗コリン薬の認知症リスクに関しては、抗コリン薬の適応疾患や抗コリン作用以外による交絡が疑われる結果が得られた。この結果に対してさらにベイズマ

ルチカーネル法による検証や伝統的な因果推論の手法(時間変化する因子と傾向スコアを用いた方法)を機械学習化した方法による検証などにより、さらに認知症発症の精密なリスク評価を行なっていく。

参考文献

[1]S. Lee and H.-S. Kim J Lipid, Atheroscler. (2021) 10(3):282-290
[2]M. Chowdhury et al., Front. Psychiatry (2021) 738466
[3]M. Gönen, Proceedings of ICML (2012)
[4]K. Imai and D.A. van Dyk, JASA (2004)
[5]W. Shi and W. Xu, Proceedings of UAI (2023)
[6]清水昌平「統計的因果探索」(2017) 講談社
[7]K. Zhang and A. Hyvärinen, "Statistics and Causality: Methods for Applied Empirical Research", (2016), John Wiley & Sons, Inc.
[8]A. Cobb, HamilTorch, (2023), Proceedings in CPS-IoT Week Workshops '23
[9]S.L. Gray et al. (2015) JAMA Int. Med.
[10]K. Richardson et al. (2018) BMJ
[11]T. Hayakawa et al. (2023), Digital Health