様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２２ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2022
課題番号：１７Ｋ００００６１

研究課題名（和文）ビッグデータ時代のグラフィカルモデル推測理論の新展開

研究課題名（英文）New developments in the theory of graphical model inference in the Big Data era

研究代表者
原　尚幸（Hara, Hisayuki）

京都大学・国際高等教育院・教授

研究者番号：４０３１２９８８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：空間疫学モデルを用いて、特定の疾患のホットスポットを検出する際に、罹患数の多さを表す検定統計量の多重性調整P値の計算が必要になるが、従来法では計算コストが高いことが問題とされていた。本研究では、地域間の隣接関係から定義される無向グラフの構造を利用して、そのグラフの分割を用いた分割統治的なアルゴリズムを提案し、その有用性を示した。
また、マルコフ基底や、グラフィカルモデルの推論、推測アルゴリズムに関する成果を、研究書にまとめて公刊をした。

研究成果の学術的意義や社会的意義
空間疫学モデルにおける多重性調整P値の正確計算は、ホットスポットを高精度に検出するために重要であるが、従来法では計算コストが高いという問題があった。また、近似アルゴリズムも存在はしているが、小標本のときに精度が悪いものであった。今回、地域間の空間的な隣接情報からグラフを定義し、そのグラフの分割を用いた分割統治アルゴリズムによって、正確なP値の計算が、実用時間内で可能になったことは、疫学の研究において意義があるものである。
また、近年の計算代数統計学による、グラフィカルモデルの推論に関する書籍は、和書ではまったくなかったなかで、書籍を刊行したことは、さらなるこの分野の発展に意義があることと考える。

研究成果の概要（英文）：For the computation of multiplicity-adjusted P-values in hotspot detection in the field of spatial epidemiology, we proposed an efficient calculation algorithm using the structure of graphs obtained from the adjacencies between regions.
We also published our results on Markov bases, graphical model inference, and inference algorithms in a research book.

研究分野：数理統計学

キーワード：グラフィカルモデル　計算代数統計学

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，ビッグデータ活用技術とその研究開発は，あらゆる産業で急速に普及・発展しており，データサイズは今後もさらに増大するとともに，これまで以上に複雑な自然・社会現象のメカニズムの解明が求められることが当初から現在に至るまで期待されている．そのための高度な分析技術の実用化のためには，背景の統計理論の整備を，古典的な数理統計的アプローチを越え，数理シーズ間で分野横断的に目指すアプローチが有用であると考えた．そこで，本研究課題では，ビッグデータ分析の基盤モデルであるグラフィカルモデルの推測問題に特に焦点を当て，数理統計学だけでなく，組合せ論，計算代数学などの方法論も駆使した新たなアプローチで，有用な推測技術の開発を目指すことにした．

２．研究の目的

本研究では，組合せ論，計算代数学の知見を，数理統計学の問題に適用し，高次元グラフィカルモデル，高次元の統計モデルの推測理論の発展を目指すものであった。当初は，潜在変数を含むグラフィカルモデルのパラメータが識別可能になるための十分条件の導出の問題を中心的に取り組むことを目的としてあげた．

本研究ではこれに加え，空間疫学モデルにおけるホットスポット検出の問題における多重性調整 P 値の正確計算の計算アルゴリズムの開発も行うことにした．ホットスポット検出には，地域の隣接構造と，特定の疾病への罹患数との関係が必要である．地域間の隣接構造から，自然にグラフの構造が誘導されることから，計算アルゴリズムの効率化にはグラフィカルモデルにおける知見が利用可能と考えた．この問題の解決も研究目的に加えた．

さらに，グラフィカルモデルは，その特殊ケースに統計的因果モデルを含む．前述の識別可能性の問題で扱う予定であったモデルも，非巡回有向グラフ(以下，DAG)が定義する因果モデルと解釈することもできる．そこで，特定の因果効果を識別するための十分条件の導出も，本課題の範囲内と考え，研究目的に加えることにした．

さらに，これまで行ってきた，グラフィカルモデルの代数的な考察や，計算代数的な統計手法の実装に関して，書籍にまとめ，公刊することも目的とした．

３．研究の方法

課題 1．パラメータの識別可能性の判定問題は，自明に識別が可能なパラメータと，所望の p らメータの間の写像が 1 対 1（あるいは有限対 1）であるかどうかを判定する問題であるので，代数的であると言える．この問題は，2000 年代後半から，計算代数統計学の枠組みでも議論され，近年大きく発展を遂げた分野である．

例えば，Leung，Drton and Hara(2016，LDH2016)では，DAG が定義するガウスグラフィカルモデルの一つのソースノードに対応する変数が潜在変数であるようなモデルが識別可能になるための十分条件を導出した．一方，同じ設定で，変数が離散であるようなモデルの識別可能性については,変数の数が 4 までの不十分な考察しか存在しない（Allman et al.(2015)）．本研究では，LDH2016 の議論を離散のグラフィカルモデルに直接的に拡張できるかどうかを検証した．具体的には，変数の数がpの場合に識別可能であることが既知のモデルに，観測可能なソースノードやシンクノードを加えたモデルの識別可能性を評価することを考えた．これを示すことで，識別が可能であるための十分条件を得ることができれば，任意の次元の離散モデルについて，その十分条件が満たされるかどうかを検証することによって，識別可能 or 不能の判別が可能になる．

課題 2．空間疫学モデルにおいては，自治体などの各地域における特定疾患への罹患数に基づいて、ホットスポットの検出を行う．各地域の度数にはポアソン分布のモデルを想定する．ホットスポットは，期待値パラメータがすべての地域で等しいかどうかの検定や，特定地域が他の地域より平均が高いことを示すような検定を用いることによって検出される．その際に，検定の多重性の調整が必要になるが，多重性調整 P 値を直接的に計算することは計算コストが高いことが知られている．ここでは，地域の隣接関係が定義するグラフを利用して，そのグラフのクリークセパレータによる分割を利用して，P 値の計算問題を小さいモデルの確率計算に帰着させることで，計算効率の高いアルゴリズムを導出することを考えた．

課題 3．マーケティングサイエンスにおける因果効果の識別問題は，課題 1 の問題とも深く関連する．ここでは，課題 1 の副産物として，以下の 2 つの問題にも取り組んだ．

(1) 広告の出稿の効果は，2 時点で調査が行われ，2 時点間における広告接触の有無と，その間消費行動に関するアンケートデータを用いて識別を試みることがしばしば行われている．しかし，多くの調査では，1 時点目の前に，すでに広告出稿がなされており，その場合は通常，1 時点目より前に広告接触をしたかどうかが未観測となってしまう．こうした状況で，1 時点目の広告接触の有無によって，広告効果を区別して識別する問題を考える．これは，すでに商品認知をしている人と，していない人での広告効果を区別して識別することにあ

たり，実用上意義があることと考える．この問題は，変数を離散として，モデルをグラフィカルモデルとして捉えれば，課題１の識別問題と関連することがわかり，課題１での成果を用いれば，識別可能になるための十分条件を導出することが可能と考えた.さらに，変数が連続の場合を含むより一般の場合への一般化も目指すことにした．これらの問題では，処置の有無の 2 値変数が未観測であることから，分析上許容可能な条件を一つ付与することで因果効果は識別可能なると考えられる．したがって，許容可能な条件の導出と，その条件下での一致推定量の導出が主たる課題となる．

(2) 広告出稿では，観測された処置変数が，広告接触の有無を表さないケースも存在する.TVCM への接触は，その CM が出稿されているテレビ番組の視聴の有無によって判断されることが多いが、出稿されているテレビ番組を視聴しても、CM を注視しているかどうかまでは，厳密にはわからない．そこで，テレビ視聴の有無のデータのみを用いて，CM 接触の有無による広告効果を識別することを考えた．これは，前述の問題における考察が利用できると考えた．

４．研究成果

課題１．当初の想定通り，LDH2016 におけるガウスモデルにおける識別可能であるための十分条件は，モデルが離散 2 値の場合でも成立することがわかった（Hara(2017)）．また，LDH2016 では議論されていない十分条件の導出にも成功した（Naito and Hara(2018)）．Naito and Hara(2018)では，変数が多値の離散の場合への一般化へも成功した．一方，離散の場合とガウスの場合は，完全にパラレルではなく，ガウスで識別不能なモデルが，離散であれば識別可能であるといった差異についても確認することができた．現時点では，求められた十分条件の強さの評価（求められた十分条件で評価が可能なグラフの割合など）を十分に評価しきれていない点が今後の課題であると言える．

課題 2．当初の想定通り，分割統治アルゴリズムを用いた多重性調整 P 値のアルゴリズムの導出に成功した．計算オーダーは指数オーダーにはなるが，グラフの分割の手続きを含めても大幅に改善することを示すことができた．また、提案手法を，山形県の胆嚢がんのホットスポット検出の問題に適用し，有用性の確認を行った．成果の詳細については，Kuriki，Takahashi and Hara(2018)を参照されたい．

課題 3．

(1) 課題１との関連を考察することで，１時点目での接触の有無別に広告効果を識別するための条件を導出した（内藤，原(2019)）．具体的には，ある種の条件を満たす，追加的な変数の利用によって，識別が可能になるというものである．しかし，ここでの結果は，すべての変数が離散，交絡をもたらす共変量が存在しないという，強い条件のもとでの結果であった．そこで，それを一般化し，交絡をもたらす共変量が存在し，処置変数以外に連続変数を許容するような場合における，因果効果識別のための十分条件の導出を行った（原，富山(2022)）．

(2) (1)における知見を発展させ，処置変数の遵守が十分でない場合の因果効果の識別のための十分条件の導出に成功した．

さらに，今回の成果を一部含む，グラフィカルモデルの代数的考察，計算代数的なデータ分析手法の実装に関する書籍をまとめた（青木・竹村・原(2019)）．ここでは，マルコフ基底，imset，実験計画など，今後もグラフィカルモデルとの関連で発展が期待される分野を網羅的に解説した内容となっている．日本ではこのような題材を扱った書籍は現時点では他に存在しない．

本研究に関わる成果物は以下の通りである．

論文などの主な成果物

1. Satoshi Kuriki, Kunihiko Takahashi and Hisayuki Hara (2018). Multiplicity adjustment for temporal and spatial scan statistics using Markov property. Japanese Journal of Statistics and Data Science, 1, 191-213.
2. Ruriko Yoshida, Hisayuki Hara and Patric Saluke (2019). Sequential importance sampling for logistic regression model. In Computational Models for Biomedical Reasoning and Problem Solving (Chen, C. and Cheung, S. S. eds.), 231-255.
3. 青木敏，竹村彰通，原尚幸．代数的統計モデル．共立出版，2019.
4. Hisayuki Hara. Identifiability of binary Bayesian networks with one latent variable. CMStatistics 2017, Dec. 2017, University of London, U.K.
5. Hiroaki Naito and Hisayuki Hara. Identifiability of discrete Bayesian network with

a latent source. CMStatistics 2018, Dec. 2018, Pisa, Italy.

6. Hiroaki Naito and Hisayuki Hara. Uplift modeling for panel data using switch doubly robust method, Joint Statistical Meeting 2020, Aug. 2020, online.
7. Hiroaki Naito and Hisayuki Hara. Uplift modeling with multitreatment for observational pretest-posttest data. CMStatistics 2020, Dec. 2020, online.
8. 内藤宏明, 原尚幸. 離散ベイジアンネットワークのパラメータの識別可能性. 計算機統計シンポジウム(Nov 2018 滋賀大学)
9. 内藤宏明, 原尚幸. 事前に処置を受けた対象を考慮したDID 推定法. 統計関連学会連合大会(Sep. 2019 滋賀大学)
10. 内藤宏明, 原尚幸. Transformed Outcome Method を用いたパネルデータのためのUplift Modeling. 計算機統計シンポジウム(Nov 2019 青山学院大学)
11. 内藤宏明, 原尚幸. 観察研究から得られた処置前後データのためのUplift Modeling.統計関連学会連合大会(Sep. 2020 オンライン)
12. 原尚幸, 富山慶. 未観測の処置変数を含む場合のATTの識別. 統計関連学会連合大会(Sep. 2022 オンライン)

５．主な発表論文等

〔雑誌論文〕　計2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　1件）

１．著者名 Ruriko Yoshida, Hisayuki Hara, Patrik Saluke	４．巻 1
２．論文標題 Sequential importance sampling for logistic regression model	５．発行年 2019年
３．雑誌名 Computational Models for Biochemical Reasoning and Problem Solving	６．最初と最後の頁 231-253
掲載論文のDOI（デジタルオブジェクト識別子） 10.4018/978-1-5225-7467-5.ch001	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 Satoshi Kuriki, Kunihiko Takahashi and Hisayuki Hara	４．巻 1
２．論文標題 Multiplicity adjustment for temporal and spatial scan statistics using Markov property	５．発行年 2018年
３．雑誌名 Japanese Journal of Statistics and Data Science	６．最初と最後の頁 191-213
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-018-0007-5	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計16件（うち招待講演　2件／うち国際学会　7件）

１．発表者名 原尚幸
２．発表標題 処置割り当てが立ちの処置前後データにおけるUplift modeling
３．学会等名 統計関連学会連合大会
４．発表年 2021年

１．発表者名 Hiroaki Naito and Hisayuki Hara
２．発表標題 Uplift modeling for panel data using switch doubly robust method
３．学会等名 Joint Statistical Meeting 2020（国際学会）
４．発表年 2020年

１．発表者名
Hiroaki Naito and Hisayuki Hara

２．発表標題
Uplift modeling with multitreatment for observational pretest-posttest data

３．学会等名
CMStatistics 2020（国際学会）

４．発表年
2020年

１．発表者名
内藤宏明，原尚幸

２．発表標題
観察研究から得られた処置前後データのためのUplift Modeling

３．学会等名
2020年度統計関連学会連合大会

４．発表年
2020年

１．発表者名
松田周也，原尚幸

２．発表標題
キーワード検索数とツイートの情報を用いたビットコイン価格の騰落予測

３．学会等名
人口知能学会金融情報研究会

４．発表年
2021年

１．発表者名
戎達也，原尚幸

２．発表標題
大喜利における回答の面白さに関する定量的考察ーお題と回答の意味的類似度からの考察ー

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
宮城夏帆，阪田真己子，原尚幸

２．発表標題
漫才対話におけるマルチモーダル情報の動的構造分析

３．学会等名
情報処理学会第83回全国大会

４．発表年
2021年

１．発表者名
内藤 宏明 , 原 尚幸

２．発表標題
事前に処置を受けた対象を考慮したDID推定法

３．学会等名
統計関連学会連合大会

４．発表年
2019年

１．発表者名
Hisayuki Hara

２．発表標題
Multiplicity adjustment with Markov property in temporal and spatial epidemiology

３．学会等名
DSSV2019（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
内藤 宏明 , 原 尚幸

２．発表標題
Transformed Outcome Methodを用いたパネルデータのためのUplift Modeling,

３．学会等名
計算機統計学会シンポジウム

４．発表年
2019年

１．発表者名
内藤宏明、原尚幸

２．発表標題
離散型ベイジアンネットワークのパラメータの識別可能性

３．学会等名
計算機統計学会シンポジウム

４．発表年
2018年

１．発表者名
Hiroaki Naito and Hisayuki Hara

２．発表標題
Identifiability of discrete Bayesian network with a latent source

３．学会等名
CMStatistics 2018（国際学会）

４．発表年
2018年

１．発表者名
原尚幸

２．発表標題
Identifiability of Directed Graphical Models with a Latent Source

３．学会等名
AMBN2017（国際学会）

４．発表年
2017年

１．発表者名
原尚幸

２．発表標題
Data Science Education of Faculty of Culture and Information Science in Doshisha University

３．学会等名
Conference on Education of Data Science（招待講演）（国際学会）

４．発表年
2017年

１．発表者名 原尚幸
２．発表標題 Identifiability of binary Bayesian networks with one latent variable
３．学会等名 CM Statistics 2017（国際学会）
４．発表年 2017年

１．発表者名 原尚幸
２．発表標題 同志社大学におけるデータサイエンス教育
３．学会等名 日本統計学会春季集会
４．発表年 2018年

〔図書〕　計2件

１．著者名 日本統計学会（編）（原尚幸ら5名が編集委員）	４．発行年 2020年
２．出版社 学術図書出版社	５．総ページ数 330
３．書名 統計学実践ワークブック	

１．著者名 青木敏，竹村彰通，原尚幸	４．発行年 2019年
２．出版社 共立出版	５．総ページ数 300
３．書名 代数的統計モデル	

〔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