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

平成　２９　年　　６　月　　９　日現在

機関番号：　３４３１０
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３００３７

研究課題名（和文）組合せ代数的手法によるビッグデータ時代の統計的推測理論と推測アルゴリズムの研究

研究課題名（英文）A study on statistical inference and inference algorithm by using combinatorial algebra for big data analyses

研究代表者

原　尚幸（Hara, Hisayuki）

同志社大学・文化情報学部・准教授

研究者番号：４０３１２９８８

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：　非巡回有向グラフが定義し、潜在変数を1つ含むような、因子分析型のガウスグラフィカルモデルの識別可能性について、先行研究で得られている十分条件を大きく改良する、有用な組合せ論的な十分条件の導出を行った。また、この一部の結果の離散のグラフィカルモデルの識別可能性への拡張を行った。
　また、空間疫学モデルのホットスポット検出の多重性調整p値の正確計算アルゴリズムを、グラフ理論を用いて改良し、実データを用いてその有用性を確認した。

研究成果の概要（英文）：We studied parameter identifiability of directed Gaussian graphical models with one latent variable. In the scenario we consider, the latent variable is a confounder that forms a source node of the graph and is a parent to all other nodes, which correspond to the observed variables. We give some useful graphical conditions that is sufficient for the model to be identifiable.
 We also studied the problem of the evaluation of multiplicity-adjusted p-value of scan statistics in spatial epidemiology. We use some notions on graph theory and proposed an efficient algorithm to compute multiplicity-adjusted p-value of the exact distribution of scan statistics. We also implemented the proposed algorithm and confirm the usefulness of it through some real data examples.

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　計算機統計学

１．研究開始当初の背景
ビッグデータ時代となり、古典的な統計学的手法では扱いきれなかった複雑な統計モデルを用いたデータ分析手法の開発は、理論、応用の両面で喫緊の課題であった。理論面では、スパース推定のように、古典的な数理統計理論を越えた理論開発が望まれており、代数学、組合せ論など、シーズ間の異分野交流の中でのブレイクスルーが期待されていた。

２．研究の目的
本研究では、計算代数学、組合せ論を用いて、大規模グラフィカルモデルの推測理論の構築を目指した。特に、以下の問題に焦点をあてて研究を行った。

(1) グラフィカルモデルは、分散構造が複雑になると、モデルの識別可能性が自明でなくなる。グラフィカルモデルの識別可能性の条件は、多項式方程式の解が一意的であることと等価であることが多く、したがって代数的な問題と言える。ここでは複雑なグラフィカルモデルの識別可能性問題を、組合せ代数の手法を駆使して考察を行った。

(2) 空間疫学では、ポアソンモデルを用いて、多重性検定を行うことで、ホットスポットの検出を行う。その際に問題となるのは、スキャン統計量を用いた検定において、正確分布に基づく多重性調整 p 値の計算が困難であるということであった。本研究では、空間的な相関構造をグラフととらえ、グラフ理論的なアプローチでアルゴリズムの改良を試みた。

(3) 社会ネットワークの基本的な統計モデルに p1 モデルがある。p1 モデルは、0 か 1 の値のみを頻度に持つ分割表のモデルに対応する。したがって、本質的に小標本で、標準的な漸近分布に基づく適合度検定を用いることは好ましくない。そこで、マルコフ基底を用いた正確検定アルゴリズムの開発を試みた。

(4)　離散の説明変数を持つロジットモデルは、代数学の用語でローレンス持ち上げといい、そのマルコフ基底の構造が複雑になることが知られている。そのため、説明変数が１つの非常に簡単なモデルを除けば、これまで理論的な成果はほとんど得られていなかった。本研究では、代数学におけるアプローチにより、ロジットモデルのマルコフ基底の導出を行い、有用な正確検定アルゴリズムの導出を目指した。

３．研究の方法
(1) グラフィカルモデルの識別可能性の判定は、代数的アルゴリズムにより理論的には可能であるが、アルゴリズムが NP hard であるため、低次元の場合のみに限られる。本研究では、まず低次元の識別判定結果を代数ソフトを用いて計算し尽くし、その判定結果を見て、識別可能性のための十分条件に関する予想を立て、理論的な検証を行った。

(2) ホットスポット検出のような分析では、自治体ごとにデータを採取し分析を行う。ここでは自治体の隣接の構造をグラフととらえることにより、そのグラフの構造を利用して、スキャン統計量の多重性 p 値の計算アルゴリズムの改良を試みた。

(3) 代数学におけるグレーバー基底という概念を適用し、その一部分を推移子にしたマルコフ連鎖モンテカルロ法により正確検定の実装が可能になると考え、グレーバー基底の理論的な導出を目指した。

(4) 代数幾何学における、ローレンス持ち上げのグレーバー基底という概念が、ロジットモデルのマルコフ基底に対応する。ローレンス持ち上げのグレーバー基底は、代数幾何学の文脈で、いつかの理論結果が存在する。そこでの結果と、統計モデルとの関連を考察することで、ロジットモデルのマルコフ基底と、正確検定の実装アルゴリズムの導出を目指した。

４．研究成果
(1) 本研究では、特に、非巡回有向グラフが定義し、潜在変数を一つ含むグラフィカルモデルに着目し、その十分条件の導出を行った。前述の通り、識別可能性条件は、多項式連立方程式の解の一意性に帰着するが、これは多項式連立方程式のヤコビ行列が列フルランクであることに対応する。本研究では、このヤコビ行列が列フルランクになるための十分条件の導出を行い、多項式時間で計算可能な有用な組合せ論的条件の導出に成功した。具体的には、モデルを定義するグラフの補グラフが奇サイクルを持つという、非常にシンプルな十分条件が得られた。
　また、多項式連立方程式の組合せ論的な構造を利用して、再帰的に識別可能性をチェックするアルゴリズムの導出も行った。具体的には、モデルを定義するグラフにある sink node が、他のすべての node に隣接していない場合、その node を除去した部分グラフが定義するモデルが識別可能であれば、元のグラフが定義するモデルも識別可能であることを示した。この事実を用いれば、上の条件を満たす sink node を順に除いて行って、低次元の識別可能モデルにたどり着けば、元のモデルも識別可能ということになる。この再帰的アルゴリズムを用いると、ヤコビ行列に基づく条件では識別可能と判定できなかったモデルでも、識別可能と判定できるケースが存在する。
　本研究での結果は、Stanghellini and Wermuth(2005)の既存研究の結果を優越するもので、有用な結果であると言える。また、

ここでの結果の一部は、離散のグラフィカルモデルの場合にも適用が可能であることを示すことにも成功した。しかし、やや結果が不完全であるので、この点は今後の課題である。

以上は、2013 年 9 月から 2014 年 9 月までに米国ワシントン大学で、Mathias Drton 氏、Dennis Leung 氏と行った共同研究の結果である。

(2) 多重調整 p 値は、条件付き確率の計算問題に帰着する。しかし、高次元の離散分布の場合、標本点が爆発的に増大するため、期待値における和の計算が困難になる。本研究では、地域間の隣接構造をグラフととらえ、そのグラフの構造、特にグラフの分解を利用して、期待値の和の計算を分解し、計算量的に大きく改善するアルゴリズムの導出に成功した。また、胆嚢がんの疫学データなどの実データに適用することで、その有用性を確認した。

以上は、連携研究者である、統計数理研究所の栗木哲氏との共同研究による。

(3) p1 モデルのグレーバー基底があれば、そのうちの平方自由の成分のみを抽出することにより、MCMC 法による正確検定の実装が可能になる。平方自由な要素のみを用いる理由は、p1 モデルが、0 か 1 の頻度を持つ分割表に対応することによる。実装アルゴリムズの実現のためには、グレーバー基底の理論的導出と、平方自由の成分のみをサンプリングするアルゴリズムの開発が必要であった。本研究では、まず p1 モデルのグレーバー基底の理論的導出に成功した。さらには、そのグレーバー基底の平方自由の要素のみを効率的にサンプリングするアルゴリズムを開発し、p1 モデルの適合度検定の実装アルゴリズムの提案を行った。また実データに適用し、その有用性を確認した。

以上は、連携研究者である、滋賀大学データサイエンス学部教授の竹村彰通氏、首都大学東京の小川光紀氏（現在東京大学に在籍）との共同研究である。

(4) Petrovic(2008)は、ある種のローレンス持ち上げのグレーバー基底を代数幾何的に導出している。本研究では、このローレンス持ち上げが、実用的な計画行列を持つロジットモデルに対応することを示し、そのモデルでは説明変数が複数ある場合も含めて、マルコフ基底を明示的に導出できることを示した。また、その結果を用いて、正確分布に基づいた適合度検定のアルゴリズムの導出を行った。

さらに、各説明変数の組に対して、1 以上の標本が得られているという条件の元では、さらに簡単な構造を有するマルコフ基底の部分集合を用いることにより、より広いクラスのロジットモデルで正確検定を実装することが可能であることを示すとともに、実装アルゴリズムの提案を行った。

以上は、Illinois Institute of Technology の Sonja Petrovic 氏、Dane Wilburne 氏との共同研究により。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計４件）

① Mitsunori Ogawa, Hisayuki Hara and Akimichi Takemura (2013).Graver basis for an undirected graph and its application to testing the beta model of random graph. Annals of Institute of Statistical Mathematics. 65, 1, 191-212. (査読有).

② Akikuni Matsumoto, Hisayuki Hara and Kazumitsu Nawata.(2014). Contract workers in Japan's nuclear utility industry: can we maintain safety and health standards at nuclear power plants?, Journal of Reviews on Global Economics, 3, 401-414.(査読有).

③ Satoshi Kuriki, Hisayuki Hara and Kunihiko Takahashi.(2015). Recursive computation for evaluating the exact p-values of temporal and spatial scan statistics. arXiv:1511.00108.

④ Dennis Leung, Mathias Drton and Hisayuki Hara.(2016). Identifiability of directed graphical models with one latent source. Electoric Journal of Statistics.10. 394-422. (査読有).

〔学会発表〕（計５件）

① Hisayuki Hara.(2013). Running Markov chain without Markov bases. 8th IASC-ARS Conference. Yonsei University, Seoul, Korea.(2013 年 8 月 22 日).

② Hisayuki Hara.(2013). Efficient computation of maximum likelihood estimator of hierarchical subspace model,59th ISI WSC, Hong Kong, China.(2013 年 8 月 26 日).

③ Hisayuki Hara.(2014). Running Markov chain with and without Markov bases. Algebraic Statistics 2014. Illinois Institute of Technology.U.S.A.(2014 年 5 月 20 日).

④ Hisayuki Hara. (2015). Identifiability of Gaussian DAG models with one latent source", IASC-ARS 2015, National University of Singapore, Singapore. (2014 年 12 月 17 日).

⑤ Hisayuki Hara.(2016). Markov bases

for logit models with some designs, IMS-APRM 2016, Chinese University of Hong Kong, Hong Kong, China.(2016年6月28日).

〔図書〕(計1件)

① 日比孝之，竹村彰通，原尚幸，東谷章弘，清智也，他，『グレブナー教室』(共立出版)，2015.(総ページ数205ページ(164ページ～176ページを担当)).

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
原　尚幸（HARA, Hisayuki）
同志社大学・文化情報学部・准教授
研究者番号：４０３１２９８８

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
竹村　彰通（TAKEMURA, Akimichi）
滋賀大学・データサイエンス学部・教授
研究者番号：１０１７１６７０

栗木　哲（KURIKI Satoshi）
統計数理研究所・数理推論研究系・教授
研究者番号：９１０９５５４５

二宮　嘉行（NINOMIYA, Yoshiyuki）
九州大学・数理学研究科・准教授
研究者番号：５０３４３３３０

小林　景（KOBAYASHI, Kei）
慶應義塾大学・理工学部・准教授
研究者番号：９０４６５９２２

(4)研究協力者
(　　)