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　２ 年　 ６ 月　 １ 日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１２６５１

研究課題名（和文）幾何構造および代数構造に基づく統計的手法の研究

研究課題名（英文）Study on Statistical Methods based on Geometric and Algebraic Structures

研究代表者
小川　光紀（Ogawa, Mitsunori）

東京大学・大学院情報学環・学際情報学府・特任講師

研究者番号：５０７５８２９０

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：離散指数型分布族に含まれるパラメータの中の興味ある一部のパラメータのみを推定する手法について研究を行った．幾何学的概念である複合局所 Bregman ダイバージェンスの構成に，マルコフ基底をはじめとする代数統計由来の概念を用いることにより，実用に耐える推定手法を構築した．応用として，分割表の対数線形モデルにおける一部のパラメータのみを推定する問題に対し，本研究で得られた推定手法の具体的手順を整備し，数値実験によってその有効性を確認した．

研究成果の学術的意義や社会的意義
局外パラメータを含む統計モデルのパラメータ推定問題の歴史は長い．指数型分布族のように局外パラメータの十分統計量が存在する場合，十分統計量を所与とした条件付き分布に基づく推定量が統計的によい性質を持つことが知られている．しかし，条件付き分布の規格化定数は計算に適さない形であることが多く，実用上の障害になっていた．ダイバージェンスという幾何学的概念とマルコフ基底という代数統計由来の概念を組み合わせることにより，規格化定数の計算を経ないで興味あるパラメータを推定する方針が得られたことは，大きな意義も持つものである．

研究成果の概要（英文）：We developed a parameter estimation method for discrete exponential families under the presence of nuisance parameters. We used the framework of composite local Bregman divergences on discrete sample spaces and Markov bases from algebraic statistics. The resulting estimators are based on the conditional distributions given sufficient statistics for the nuisance parameters and do not require the calculations of the normalization constants of the conditional distributions. The consistency of the estimators is guaranteed by the connectivity of the underlying graph structures used in the construction of the divergences. We applied the proposed methods to the log-linear models of contingency tables and confirmed their usefulness.

研究分野：統計学

キーワード：統計数学　ダイバージェンス　マルコフ基底

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究課題には，局外パラメータを含む統計モデルに対する推測問題と，ダイバージェンスに基づく推測手法の理論，代数統計学におけるマルコフ基底という三つの異なる背景が存在する．

統計モデルのパラメータのうち，興味のないパラメータは局外パラメータとよばれる．局外パラメータを含む統計モデルにおける興味あるパラメータの推測問題は，統計学における古典的問題である．特に，観測ごとに局外パラメータの値が異なる場合の統計的推測は，Neyman-Scott 問題として知られ，Neyman and Scott (1948) で問題提起されて以来，多くの研究者によって研究が進められてきた．たとえば，Andersen (1970) では，ある種の指数型のモデルに対して，局外パラメータの十分統計量を所与とした条件付き尤度の最大化が与える推定量の漸近的性質を議論した．

この問題に対して，Amari とその共同研究者たちは，情報幾何学の観点から研究を行った．興味あるパラメータを推定するための推定方程式として，一致性をもつものが得られるための条件や，そのような推定方程式が存在する場合の最適な推定方程式の特徴付けなどを行った (Kumon and Amari (1984), Amari and Kumon (1988))．さらに，Amari and Kawanabe (1997) は，Neyman-Scott 問題をセミパラメトリックモデルの枠組みで定式化し，幾何学的考察を与えた．Amari らの一連の研究によって，局外パラメータの十分統計量が適切に得られる場合には，それを所与とした条件付き尤度に基づく推測が最適であることが明らかになった．

二つ目の背景として，用途に応じたダイバージェンスに基づく推測手法の発展がある．ダイバージェンスとは，確率分布間の乖離度の指標であり，統計学では古くから Kullback–Leibler ダイバージェンスをはじめとして様々なダイバージェンスが利用されてきた．近年では，ロバスト統計で用いられるべき密度ダイバージェンス(Basu et al. (1998))，γ-ダイバージェンス (Fujisawa and Eguchi (2008))のように，特定の目的を達成するために有効なダイバージェンスのクラスの提案や，関連する理論研究が盛んである．

三つ目の背景は，代数統計学とよばれる分野の発展である．近年の代数統計の発展は，2000 年手前の Pistone and Wynn (1996) によるグレブナー基底理論の実験計画への応用と，Diaconis and Sturmfels (1998) によるマルコフ基底を用いたマルコフ連鎖モンテカルロ法の開発に端を発するものである．本研究課題で重要となるのは，この二つのうちマルコフ基底に関連する先行研究の蓄積である．マルコフ基底は，数学的にはトーリックイデアルとよばれる代数構造の生成系に相当する． Diaconis and Sturmfels (1998) の研究以降，主にマルコフ連鎖モンテカルロ法の実装を目的として，マルコフ基底に関する数多くの研究がなされてきた．

２．研究の目的

本研究の目的は，局外パラメータを含む統計モデルにおいて，興味あるパラメータの推測手法として現実的な方法を確立することである．特に，有限標本空間上の離散指数型分布族において，興味あるパラメータに対する計算効率のよい推定手法を構築することを目指した．前述の通り，局外パラメータを含む統計モデルに対しては，局外パラメータの適切な十分統計量が得られるのであれば条件付き尤度に基づく推測が統計的によい性質をもつ．条件付き尤度の最大化に基づいて定まる推定量を扱うためには，条件付き尤度の規格化定数を計算する必要が生じる．一般に，確率分布における規格化定数が計算困難であることは統計学ではよくあることであるが，今回の研究対象である離散指数型分布族の条件付き尤度においては非常に単純なモデルであってもこの問題に直面する．Amari らの研究が，Neyman-Scott 問題に対する理論的理解を主としていたのに対し，本研究では計算の観点から実用的な推定手法を与えることを主眼としている．

３．研究の方法

離散指数型分布族の中でも，局外パラメータに関するモデル構造がトーリックモデルとよばれるクラスに属するモデルクラスを考察対象とした．このクラスに属するモデルでは，局外パラメータに対して代数統計の観点からよい十分統計量が存在する．既存研究から，この十分統計量を所与とした条件付き尤度に基づく推測手法には利点があるため，本研究でも条件付き尤度のもとになっている条件付き分布に基づく推定手法の構築を目指した．

実用的な方法の障害となっているのは条件付き分布の規格化定数の部分である．一方，規格化定数を除いた条件付き分布の形は，今回の考察対象の範囲では簡便な形をしていることが自然に想定できる．そこで，規格化定数部分を除いた非正規化モデルに基づく推測手法の観点から推定方法を構築することとした．特に，非正規化モデルに適用可能なダイバージェンスのクラスとして，Kanamori and Takenouchi (2017)によって整備された複合局所 Bregman ダイバージェンスの枠組みを利用することを考えた．この枠組みは，凸関数をもとに構成される Bregman ダイバージェンスを基礎としている．具体的には，標本空間にグラフ構造を定め，そのグラフ構造に応じた近傍系を考える．そして，近傍ごとに局所的な Bregman ダイバージェンスを定義し，それらを複合することで最終的なダイバージェンスを得る．Kanamori and Takenouchi (2017)では，グラフ構造に基づく複合局所 Bregman ダイバージェンスのクラスを提案し，得られたダイバージェンスが一致公理を満たすための条件を議論している．ダイバージェンスの一致公理は，そのダイバージェンスの最小化によって定まる推定量が一致性をもつために重要な

性質である．特に，標本空間に定めるグラフ構造について，ある種の連結性を満たすことが一致公理のための必要十分条件であることが示されている．

４．研究成果

有限標本空間上の離散指数型分布族であって，特に局外パラメータに関するモデル構造がトーリックモデルであるものについて，興味あるパラメータを推定するための現実的な手法を構築した．提案手法は，局外パラメータに対する十分統計量を与えた条件付き分布を考え，その条件付き分布にとっての標本空間上の確率分布について，複合局所 Bregman ダイバージェンスを構成することによって得られるものである．個々の条件付き標本空間に対する複合局所 Bregman ダイバージェンスが一致公理を満たすのであれば，Neyman-Scott 問題のように観測ごとに局外パラメータの値が異なる状況であっても，興味ある共通のパラメータに関する推定量として一致性をもつことが示される．残る問題は，一致公理を満たすような複合局所 Bregman ダイバージェンスを構成することが現実に可能かどうかである．

前述の通り，複合局所 Bregman ダイバージェンスの一致公理は，ダイバージェンスを構成する際に近傍系を定めるグラフ構造の連結性と密接に関連している．先行研究の Kanamori and Takenouchi (2017)では，非正規化モデルの研究としてよく取り上げられる制限付きボルツマンのパラメータ推定を念頭に，対応する標本空間上のグラフ構造として適切なものの例が与えられている．この場合の標本空間は，規格化定数の直接計算を行うには規模として大きすぎるものの，構造そのものは簡明であることから，直感的に適切な連結性をもつグラフ構造を導入できる．一方，今回の研究対象である離散指数型分布族の場合には，対応する標本空間は代数統計分野でファイバーとよばれる十分統計量を観測値で固定したデータの集合であり，その構造は通常は非常に複雑である．

本研究では，代数統計分野で研究されてきたマルコフ基底を用いて，適切なグラフ構造を得る方法を提案した．マルコフ基底を用いれば，十分統計量の任意の観測値に対するファイバー上で，一致公理を満たす複合局所 Bregman ダイバージェンスを構成するために必要な連結性を有するグラフ構造を構成することができる．原理的には，与えられた局外パラメータに対するトーリックモデルの構造に応じて，計算機によってマルコフ基底を求めることが可能であるが，その計算コストは膨大であり多くの場合に現実的でない．しかし，代数統計分野ではこのことを動機付けとして具体的な統計モデルに付随するマルコフ基底の構造解析の研究が数多く行われており，それらの結果を活用することができる．特に，適切なダイバージェンスを構成するという目的のためにはマルコフ基底の要素を全列挙する必要はなく，実際の観測値に対する近傍の列挙を効率よく行えればよい．マルコフ基底の構造が具体的に特徴付けられた結果があれば，近傍列挙を行うアルゴリズムを構成することは比較的容易である．

マルコフ基底の構造解析研究で扱われる対象への批判として，考察するモデルが非常に単純であることが挙げられる．一般に，マルコフ基底の構造は非常に複雑であるため，構造が理論的に特徴付けられるモデルが単純なものに限られることは自然である．一方，今回の研究で考察対象とするモデルにおいては，局外パラメータに関するモデル構造に付随するマルコフ基底が必要であり，想定するモデル全体はより複雑なものであってもよい．

応用として，分割表の対数線形モデルの場合について，具体的な推定手順を整備し，数値実験によって提案手法の有効性を確認した．また，ランダムグラフの統計モデルについても，局外パラメータに対するモデル構造がベータモデルの場合について，具体的な手順を整備した．ベータモデルは，各頂点の次数が十分統計量となるようなランダムグラフの基本的なモデルである．ここで考えた設定は，頂点に割り振ったパラメータを局外パラメータとして，さらに複雑な構造を表すパラメータに関心がある状況を想定したものである．ランダムグラフの場合には，通常のマルコフ基底ではファイバーに対するグラフ構造として適切な一致性を保証することができない．マルコフ基底よりも大きな集合であるグレイバー基底を用いれば一般のベータモデルに対して必要な連結性を担保できる．しかし，グレイバー基底の構造はマルコフ基底よりもさらに複雑であるため，近傍系の列挙の目的に限っても現実的で無い．一方，ランダムグラフの台となるグラフを完全グラフの場合に限定すれば，4-サイクルに対応する要素のみからなる集合で連結性を保証できる．

本研究で得られた推定手法は，局外パラメータに対する十分統計量を所与とした条件付き分布に基づく頻度論的手法であるが，規格化定数計算を回避したことで計算の観点からより実用的なものになっている．また，本研究で用いたマルコフ基底やグレイバー基底は，従来の代数統計ではこれまでは正確検定を実施する計算手段であるマルコフ連鎖モンテカルロ法の構成要素として研究されてきたものである．本研究によって，これらの概念や関連理論が検定だけでなく推定の文脈でも重要な役割を果たす可能性が示唆された．

引用文献


① Amari, S. and Kawanabe, M. (1997). Information geometry of estimating functions in semi-parametric statistical models. Bernoulli, 3(1):29–54.

② Amari, S. and Kumon, M. (1988). Estimation in the presence of infinitely many

nuisance parameters---geometry of estimating functions. Ann. Statist., 16(3): 1044–1068.

③ Andersen, E.B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. J. Roy. Statist. Soc. Ser. B, 32:283–301.

④ Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. Biometrika, 85(3):549–559.

⑤ Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. Ann. Statist., 26(1):363–397.

⑥ Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. J. Multivariate Anal., 99(9):2053–2081.

⑦ Kanamori, T. and Takenouchi, T. (2017). Graph-based composite local Bregman divergence on discrete sample spaces. Neural Networks, 95:44–56.

⑧ Kumon, M. and Amari, S. (1984). Estimation of a structural parameter in the presence of a large number of nuisance parameters. Biometrika, 71(3): 445–459.

⑨ Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. Econometrica, 16:1–32.

⑩ Pistone, G. and Wynn, H.P. (1996). Generalised confounding with Grobner bases. Biometrika, 83(3):653–666.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1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<br>Mitsunori Ogawa, Kazuki Nakamoto and Tomonari Sei | ４．巻<br>－ |
| ２．論文標題<br>On the fractional moments of a truncated centered multivariate normal distribution | ５．発行年<br>2020年 |
| ３．雑誌名<br>Communications in Statistics - Simulation and Computation | ６．最初と最後の頁<br>－ |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1080/03610918.2020.1725821 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　4件）

| |
|---|
| １．発表者名<br>Mitsunori Ogawa, Kazuki Nakamoto and Tomonari Sei |
| ２．発表標題<br>On the fractional moments of a truncated centered multivariate normal distribution |
| ３．学会等名<br>2019 Joint Statistical Meetings（国際学会） |
| ４．発表年<br>2019年 |

| |
|---|
| １．発表者名<br>Mitsunori Ogawa |
| ２．発表標題<br>Parameter estimation for discrete exponential families under the presence of nuisance parameters |
| ３．学会等名<br>International Conference on Statistical Distributions and Applications（国際学会） |
| ４．発表年<br>2019年 |

| |
|---|
| １．発表者名<br>Mitsunori Ogawa, Kazuki Nakamoto and Tomonari Sei |
| ２．発表標題<br>The holonomic gradient method for the moments of truncated centered multivariate normal distribution |
| ３．学会等名<br>The 5th Institute of Mathematical Statistics Asia Pacific Rim Meeting（国際学会） |
| ４．発表年<br>2018年 |

| |
|---|
| １．発表者名<br>Mitsunori Ogawa |
| ２．発表標題<br>Composite local Bregman divergences for conditional discrete exponential families |
| ３．学会等名<br>Joint Statistical Meetings（国際学会） |
| ４．発表年<br>2018年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

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