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　 ２ 日現在

機関番号：３２６６０
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０５３７３

研究課題名（和文）高次元離散多変量解析の理論研究とその応用

研究課題名（英文）Theory and Application for high dimensional discrete data

研究代表者
田畑　耕治（Tahata, Kouji）

東京理科大学・理工学部情報科学科・准教授

研究者番号：３０４５３８１４

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：行と列が同じ分類からなる正方分割表において、様々な対称性や非対称性のモデルが提案されている。ある分割表データに対して、複数のモデルの当てはまりが良い場合にどのモデルを選択するかという問題が生じる。本研究では、非対称性のモデルの族を与え、その族から罰則付き尤度を用いてモデル選択する方法を提案した。この方法により、モデル選択とパラメータの推定を同時に行うことができるようになり、計算時間の短縮にも成功した。また、既存の非対称性のモデルに対して、情報理論的アプローチを用いて新しい解釈を与えることに成功した。この結果は、対称性の必要十分条件を考える場合に大変に有用な情報を与える。

研究成果の学術的意義や社会的意義
同じ分類からなる正方分割表データは、医学・薬学、政治学、心理学など量的に測ることのできない変量を扱う分野に現れる。分割表解析の大きな関心は、分類間の独立性であるが、同じ分類からなる正方分割表では、多くの場合に独立性は成り立たない。したがって、対称性の解析を行うことが多い。研究成果は、幅広い非対称性のモデルからデータに対して適切なモデルを自動的に判断することを可能にした。このことにより、専門的な知識のない一般ユーザにとって、対称性を用いたデータ解析が身近なものとなったと考える。

研究成果の概要（英文）：Various types of asymmetry models are proposed for the analysis of square contingency tables with ordinal categories. In this research, an asymmetry model family is given and models included in it are referred to as nonhierarchical models. Thus, we treat a problem of model selection because it is not easy to compare two models. For the problem, we employ the penalized likelihood approach and the simulation studies are given. Also, we show that each of asymmetry models can be interpreted as a property that it is the closest to the symmetry model in terms of the Kullback-Leibler divergence under some conditions. Moreover, we consider a model that indicates the structure of asymmetry for cell probabilities for square contingency tables. The model is the closest to the symmetry model in terms of the f-divergence under certain conditions and incorporates existing asymmetry models in special cases.

研究分野：カテゴリカルデータ解析

キーワード：離散多変量解析　分割表解析　スパース推定　モデル選択　情報理論的アプローチ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

## １．研究開始当初の背景

身長や体重などのように連続的な値をとる連続変数に対して、性別や支持政党のように、何らかの形で分類したものはカテゴリカル変数と呼ばれる。カテゴリカル変数に関するデータの集計結果は、多くの場合に表としてまとめられ、その表は分割表と呼ばれる。たとえば、表１はがん化学療法が施行されている貧血患者に対して、がん治療に伴う貧血を抑制するために開発された薬剤を投与し、ヘモグロビン濃度の推移を示した分割表である。

表１：がんの化学療法がおこなわれている貧血患者に対するヘモグロビン濃度の経時推移（Yamamoto et al., 2007）

| | | 8 週 | | |
|---|---|---|---|---|
| 投与前 | 4 週 | ≧10g/dl | 8-10g/dl | ＜8g/dl |
| ≧10g/dl | ≧10g/dl | 77 | 7 | 1 |
| | 8-10g/dl | 3 | 8 | 1 |
| | ＜8g/dl | 1 | 1 | 1 |
| 8-10g/dl | ≧10g/dl | 43 | 7 | 0 |
| | 8-10g/dl | 17 | 16 | 5 |
| | ＜8g/dl | 0 | 2 | 3 |
| ＜8g/dl | ≧10g/dl | 3 | 0 | 0 |
| | 8-10g/dl | 3 | 8 | 1 |
| | ＜8g/dl | 0 | 4 | 3 |

分割表解析における関心の一つは、分類間が独立かどうかである。しかし、同じ分類からなる分割表（これを特に正方分割表と呼ぶ）の解析においては、多くの場合に分類間の相互関連性が強く、統計的独立性は成り立たない。したがって、統計的独立性に代わって分類間の対称性に関する統計モデルを用いた解析が行われる。対称性のモデルとしては、例えば、Bowker (1948) の対称モデル、Stuart (1955) の周辺同等モデル、Caussinus (1965) の準対称モデルがある。一方、対称性のモデルが成り立たない場合に用いられる非対称性のモデルとしては、McCullagh (1978) の条件付対称モデル、Goodman (1979) の対角パラメータ対称モデルがある。また対称性に関する研究のレビュー論文（Tahata and Tomizawa, 2014）において、およそ１００編の参考文献が紹介されている。このように対称性のモデルが数多く提案されているため、データに対してどのモデルが適切なのか判断するためにモデル選択問題の解決が重要である。また、多様な対称性のモデルに対する新しい解釈を与えることもまた重要である。

## ２．研究の目的

高次元データに基づく線形回帰モデルの推定と変数選択において、それらを同時に実行可能な手法として L1 型正則化法（Lasso）に基づくモデリングの研究が盛んに行われている。たとえば、川野ら（2010）、Zou and Hastie（2005）、Tibshirani（1996, 2011）。二元分割表および多元分割表における対称性や非対称性のモデルは数多く提案されており、あるデータに対して複数のモデルの適合度が良い場合に、どのモデルを選択するかという問題が生じる。各モデル間に包含関係が存在する場合には、尤度比検定統計量の差を用いてモデル間の比較をすることが可能であることが知られている。一方で、各モデル間に包含関係が存在しない場合には、AIC や BIC に代表されるモデル選択基準を用いることが一般的である。

本研究では、分割表解析におけるモデル選択の方法として、冒頭で述べた L1 型正則化法を応用することを考える。実際に、多元分割表解析において独立性および条件付き独立性などに関するモデル選択問題に L1 型正則化法を用いた研究（Nardi and Rinaldo, 2012）が存在し、その効果が実証されている。また機械学習の分野では、分割表解析においてベイズ論的アプローチと Lasso を組み合わせた研究が行われている（Raman, Fuchs, Wild, Dahl and Roth, 2009）。さらに高次元データ特有の問題として、多くのセル度数が 0 となる問題において、Lasso を用いた応用研究として Dahinden, Parmigiani, Emerick and Buhlmann (2007) がある。これらの先行研究は、すべて分割表解析における独立性や条件付き独立性に関連するモデルの研究である。分割表のモデルを対数線形モデルで表現するとき、独立性や条件付き独立性などのモデルの表現と対称性や非対称性のモデルの表現は大きく異なることが多い。そのため、先行研究をそのまま適用することが難しいため、本研究に取り組み多元分割表統計解析における対称性や非対称性のモデルを用いた解析方法を発展させる。また、既存のモデルに対して情報理論的アプローチを用いた新しい解釈を与えられるかどうかについても検討する。

## ３．研究の方法

(1) スパース推定を実装し、その推定量の性質を調べる。L1 型正則化法は、高次元データに基づく線形回帰モデルの推定と変数選択において、それらを同時に実行可能な手法として広く研究がなされている。たとえば、川野ら（2010）、Zou and Hastie（2005）、Tibshirani（1996, 2011）。これらの論文や関連する文献について調査し、2016 年 10 月時点で大学院生の協力のもと分割表のスパース推定実装まであと一歩のところまで到達した。

(2) Lasso などを分割表解析に応用した研究への理解を深める。分割表解析によく用いられる対数線形モデルに対して、正則化法の研究論文がいくつか発表されている（Nardi and Rinaldo, 2012; Raman, Fuchs, Wild, Dahl and Roth, 2009; Dahinden, Parmigiani, Emerick and Buhlmann, 2007)。これらの手法と対称性や非対称性のモデルの共通点と相違点を整理した結果、多元分割表統計解析における対称性・非対称性のモデル構築とモデル選択に Lasso はうまく機能することが予想される。また、ベイズ論的アプローチと Lasso を組み合わせた手法（Raman, Fuchs, Wild, Dahl and Roth, 2009）についても、対称性の解析に応用可能かどうか検討する。

(3) (1)と(2)から対称性のモデル構築及びモデル選択の手法を開発する。分割表解析において、対数線形モデルを用いた独立性や条件付き独立性の表現と、対数線形モデルを用いた対称性や非対称性の表現は、その性質が大きく異なる。この部分が正則化法の適用に関する一番の問題であると考えられる。しかし、対称性や非対称性のモデルの別表現などを模索することで、問題点が解決すると予想される。また、正則化法の適用がうまく実現できた場合には、従来のモデル選択基準である AIC や BIC との比較、さらにはモデル間に包含関係が存在する場合に尤度比検定統計量を用いた場合との比較などの考察により性能評価を行う。さらに、提案手法の数学的妥当性とパラメータの推定精度やその性質に関することも考察する。

４．研究成果

(1) ステップ１：分割表解析におけるモデル選択の方法として、L1 型正則化法を応用することを考えた。実際に、多元分割表解析において独立性および条件付き独立性などに関するモデルの選択問題に L1 型正則化法を用いた研究（Nardi and Rinaldo, 2012）が存在し、その効果が実証されている。これらの先行研究は、すべて分割表解析における独立性や条件付き独立性に関連するモデルの研究である。これらの方法の肝は、独立性に関する各種モデルを対数線形モデルで表現した際にパラメータが 0 かどうかでモデルを選択できる点であった。したがって、対称性のモデルの族を考え、その族に含まれる各種モデルがパラメータ 0 と対応する表現を与え、対称性の問題に適用可能にした。実際に、Lasso と Elastic Net を実装した。さらに、それらの拡張である Adaptive Lasso と Adaptive Elastic Net も実装した。それぞれの罰則項を用いて、実データ解析を行い良好な結果を得ることに成功し、これまでよりも詳細な分割表データの解析が可能になった。

(2) ステップ２： Adaptive Lasso と Adaptive Elastic Net を実装することに成功したが、その理論的な性質の解明やシミュレーションによる方法論の妥当性の検証など多くの課題が残った。それらを解決するために、先行研究を参考に本課題の場合に関する詳細な計算を行い、罰則付き推定量の導出に成功した。この結果を実装したことにより、数値計算の実行時間短縮につながった。また、別の罰則項を用いた場合についても同様に計算可能であることが予想された。さらに、プログラムの精査を行なったことにより、数値計算の実行時間を大幅に改善することに成功した。プログラムに用いていた組み込み関数を変更することにより、計算精度および実行時間に大きな変化があることを確認した。方法論の妥当性については、シミュレーションを行うことで、方法論の良い点・悪い点が浮き彫りになった。

(3) ステップ３：シミュレーション研究の充実を図った。その結果、モデルのパラメータの大きさ、パラメータの数、観測度数の合計が結果の良し悪しに大きな影響を与えることが確認できた。これらの研究成果を論文としてまとめ、現在投稿中である。この研究は、次に述べる情報理論的アプローチとも関連が深いことがわかった。

(4) 情報理論的アプローチを用いた分割表解析に関する論文が掲載された（Tahata, 2019, Japanese Journal of Statistics and Data Science)。この論文の中で、正方分割表解析でこれまでに提案されてきた非対称性のモデルは、ある条件のもとで f-divergence に関して対称性にもっとも近い性質を持つことが示された。この結果は、Ireland, Ku and Kullback (1969) の結果を大きく拡張したものであり、先行研究で提案されたモデルに対して情報理論的な新しい解釈を与えるものである。

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>Kouji Tahata | ４．巻<br>- |
| ２．論文標題<br>Separation of symmetry for square tables with ordinal categorical data | ５．発行年<br>2019年 |
| ３．雑誌名<br>Japanese Journal of Statistics and Data Science | 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>10.1007/s42081-019-00066-8 | 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5件（うち招待講演　0件／うち国際学会　3件）

| |
|---|
| １．発表者名<br>田畑耕治 |
| ２．発表標題<br>正方分割表における対称性のモデリング |
| ３．学会等名<br>統計的モデルの新展開 |
| ４．発表年<br>2020年 |

| |
|---|
| １．発表者名<br>Kouji Tahata |
| ２．発表標題<br>Asymmetry Models for Square Contingency Tables with Ordinal Categories |
| ３．学会等名<br>10th International Workshop on Simulation and Statistics（国際学会） |
| ４．発表年<br>2019年 |

| |
|---|
| １．発表者名<br>Kouji Tahata and Ukyo Matsushima |
| ２．発表標題<br>On model selection via penalized likelihood for square contingency tables |
| ３．学会等名<br>CMStatistics 2018（国際学会） |
| ４．発表年<br>2018年 |

| |
|---|
| １．発表者名<br>松島右京、田畑耕治 |
| ２．発表標題<br>順序カテゴリ正方分割表における正則化法を用いたモデル選択 |
| ３．学会等名<br>第12回日本統計学会春季集会 |
| ４．発表年<br>2018年 |

| |
|---|
| １．発表者名<br>Kouji Tahata |
| ２．発表標題<br>On testing marginal homogeneity for square contingency tables with ordinal categories |
| ３．学会等名<br>Biometrics by the Border（国際学会） |
| ４．発表年<br>2017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 |
|---|---|---|---|