

令和 5 年 6 月 23 日現在

機関番号：32660

研究種目：若手研究

研究期間：2019～2022

課題番号：19K14597

研究課題名（和文）ロバストな擬似事後分布に対する漸近的な解析に関する研究

研究課題名（英文）A study on asymptotic analysis for robust quasi-posterior distributions

研究代表者

中川 智之（Nakagawa, Tomoyuki）

東京理科大学・理工学部情報科学科・嘱託特別講師

研究者番号：70822526

交付決定額（研究期間全体）：（直接経費） 3,300,000円

研究成果の概要（和文）：大規模なデータを扱う際には、外れ値やモデルの誤特定の問題は非常に深刻である。外れ値はデータ生成過程からは大きく外れたデータであり、推論に大きな影響を与える。一方で、外れ値はモデルの誤特定を示唆している場合があり、外れ値を扱いは重要である。ベイズ統計においてもモデル誤特定や外れ値の混在の問題は古くから議論されている。本研究では、外れ値に影響されにくいベイズ法の開発とその理論的性質について研究を行った。具体的にはダイバージェンスと呼ばれる分布間の擬距離を用いてロバストなベイズ推定を構築し、MCMCなどを用いて推論を行う際の漸近的性質を導出した。

研究成果の学術的意義や社会的意義

近年は膨大な数のデータが取れるため、その中から外れ値を見つけることは大変困難である。また外れ値に影響されるような手法は、しばしば誤った解析結果を誘導することがある。一方で外れ値を含む場合は仮定したモデルが誤っている可能性もあるため、外れ値の扱いは重要である。特に外れ値に影響を受けにくい解析は、現在のモデルとデータで説明できる部分の結果を返してくれる。そのため本研究は、データに外れ値が含まれていても影響を受けにくいベイズ法とその理論的性質の導出を行ったことで、推定だけでなく予測や不確実性の評価も外れ値の影響を受けにくくすることができる。

研究成果の概要（英文）：When dealing with large amounts of data, issues such as outliers and model misspecification can be very serious. Outliers are data points that deviate greatly from the data generating process and can have a significant impact on inference. On the other hand, outliers can also suggest model misspecification, making it important to handle them appropriately. The problem of model misspecification and the presence of outliers has been discussed in Bayesian statistics for a long time. In this study, we developed a Bayesian method that is robust to outliers and investigated its theoretical properties. Specifically, we constructed a robust Bayesian estimator using a pseudo-distance between distributions called divergence and derived its asymptotic properties when performing inference using methods such as MCMC.

研究分野：数理統計

キーワード：ベイズ統計 ロバスト統計 漸近理論

1. 研究開始当初の背景

外れ値を含むデータのように、解析したいデータに何らかの要因で異なるデータが混入することがある。このような場合に事後平均などの既存のベイズ推定ではバイアスが大きくなるなど、外れ値の影響を大きく受けてしまう。また近年、膨大かつ複雑なデータが多くなり、解析者が外れ値を除くことが困難であるため、外れ値の影響を受けにくい推定方法が必要になってきた。これまで外れ値に関する研究は世界的に多くなされており、特に頻度論の観点からロバストなダイバージェンスの研究が盛んに行われている (Basu et al., 1998, *Biometrika*; Fujisawa and Eguchi, 2008, *J. Multivariate Anal.*). 一方で、ベイズ法は MCMC 法などの強力な計算手法があり、推定だけでなく区間推定も容易に行うことができる。さらに事前分布という方法で解析者の経験を反映することができるなど様々な利点がある。そのため、Ghosh and Basu (2016, *Ann. Inst. Statist. Math.*) では density power divergence を用いて外れ値にロバストな擬似事後分布を提案した。さらに申請者たちは外れ値の割合が大きいデータにも適用するためにガンマダイバージェンスを用いて擬似事後分布を同様に提案した。

これらの先行研究により、“独立同一分布に従う”データに対しては外れ値にロバストなベイズ推定が可能になった。しかしながら、実際のデータ解析の場面では回帰分析や時系列などのより複雑なモデルが想定され、多くの統計的モデルが“独立に同一の分布に従う”という仮定を満たしていない。Ghosh and Basu (2017) では、density power divergence を用いた擬似事後分布に対し、回帰モデルや時系列モデルに拡張して漸近性質を導出している。本研究では回帰モデルや時系列モデルにガンマダイバージェンスを用いた擬似事後分布の拡張を行い、漸近性質の導出とロバスト性の評価を行う。

2. 研究の目的

本研究では『多種多様な統計的モデルについてガンマダイバージェンスを用いた擬似事後分布の拡張を行い、漸近性質の導出を行うこと』を目的とする。本研究では、回帰モデルや時系列モデル等の様々なモデルに対し、ガンマダイバージェンスを用いた擬似事後分布の適用を行い、サンプルサイズが大きくなる場合や外れ値が大きく外れていく場合のような漸近的な挙動を考え、近似的な事後分布の性質を導出することで、ガンマダイバージェンスを用いた擬似事後分布による推論の妥当性を与える。本研究課題は理論研究が主なため実問題への応用は今後の課題であるが、本研究が成功すれば様々な分野において適用が可能になると考えている。

3. 研究の方法

本研究では、「サンプルサイズが大きくなる場合」や「外れ値が大きく外れていく場合」の2つの漸近論を考え、回帰モデルや時系列モデルなど様々なモデルにおけるガンマダイバージェンスを用いた擬似事後分布による推論の漸近的な挙動を調べる。以下の2つの方法を用いて研究を実行する。

- (1) 「サンプルサイズが大きくなる場合」は大標本漸近理論や高次元漸近理論に基づいて、一貫性や漸近正規性が成り立つための統計モデルや事前分布の条件の導出を行い、その妥当性を評価する。また高次のオーダーの挙動を調べるために漸近展開を与える。まず、回帰モデルにおける漸近性質を Ghosh and Basu (2017) や Ghosal (1999) などを参考に事前分布などの条件を調べる。次に、Ghosh and Basu (2017) を参考に、AR モデルや拡散過程モデルについての漸近性質についての条件を調べる。また、それらの条件がどの程度妥当な条件なのかを正規分布などの具体的な分布などを用いて検証する。さらに、それらの有限標本の場合での近似精度について数値実験などを用いて、検証する。時間があれば、導出した漸近展開を用いて、Reference priors や Moment matching priors などの客観事前分布の導出も行う。
- (2) ロバスト性の指標としては、感度分析などで用いられる影響関数の有界性を用いた議論が多くなされている。ガンマダイバージェンスを用いた擬似事後分布においても Nakagawa and Hashimoto (2020) で影響関数の導出を行った。本研究では、この影響関数に関して、外れ値が大きく外れていく場合に影響関数が0に収束する再下降性が成り立つ条件を調べる。これは大きな外れ値が入った場合でも擬似事後分布からの推論にその外れ値からの影響が全くないことを示している。しかしながら、外れ値が大きく外れていく場合に外れ値が含まれる場合の擬似事後分布と外れ値が含まれない場合の擬似事後分布が一致するための条件を考える。

4. 研究成果

本研究では、以下の3つの研究成果があった。

- (1) Nakagawa and Hashimoto (2020)では、漸近展開を用いてガンマダイバージェンスを用いた疑似事後分布の高次のオーダーの評価を導出し、Reference priors と Moment matching priors の2つを導出した。疑似事後分布の事前分布の決定は常に問題になっている。特にロバストなダイバージェンスを用いた場合は、共役事前分布が存在しないことが多く今回の研究で1つの指標を与えたことになった。また客観事前分布を用いたことで、数値実験により、MSE など意味で一様分布などを用いるよりも安定した推定が可能になったことがわかった。
- (2) データ生成過程から大きく外れたデータに関しては、ガンマダイバージェンスを用いた疑似事後分布を用いることで除くことができることがわかった。これは density power divergence を用いた疑似事後分布では常に起きるとは限らない性質であり、通常の事後分布でも仮定する尤度の裾がかなり重くないと達成することができない。一方で、ガンマダイバージェンスを用いた疑似事後分布はかなりの一般的な場合でも成り立つことが確認できた。
- (3) 順序カテゴリデータに対する回帰モデルに関して、通常のベイズ推定では外れ値が混入する場合は、外れ値の影響を受けてしまう。Scalra et al. (2021) では、最尤推定量において裾の重い分布の分布関数をリンク関数に用いることで、影響関数が有界であることを示している。一方で、再下降性に関しては $(\log x)^{-1}$ より裾の重い分布を用いる必要があるが、このような分布は密度関数を持たない。つまり、通常のベイズ推定はデータ生成過程から大きく外れたデータを完全に除去できず、一定程度影響が残ってしまうことが示すことができた。そのため、本研究では、density power divergence, ガンマダイバージェンスを用いた疑似事後分布を用いることによって、データ生成過程から大きく外れたデータを完全に除去できることがわかった。特に、リンク関数の誤特定は外れ値があるかないかにかかわらず、推論に大きな影響を与えてしまう。つまり、裾の重い分布のリンク関数を用いるとリンク関数の誤特定を引き起こす可能性が上がる。一方で、ダイバージェンスを用いることにより、リンク関数は自由に選択でき、柔軟なリンク関数の選択と外れ値にも頑健なベイズ推定法を提案できた。

5. 主な発表論文等

〔雑誌論文〕 計6件（うち査読付論文 6件／うち国際共著 1件／うちオープンアクセス 2件）

1. 著者名 Nakagawa Tomoyuki, Ohtsuka Shuntaro	4. 巻 16
2. 論文標題 An Asymptotic Expansion for the Distribution of Euclidean Distance-Based Discriminant Function in Normal Populations	5. 発行年 2022年
3. 雑誌名 Journal of Statistical Theory and Practice	6. 最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42519-022-00292-6	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 Sugasawa Shonosuke, Nakagawa Tomoyuki, Solvang Hiroko Kato, Subbey Sam, Alrabeei Salah	4. 巻 -
2. 論文標題 Dynamic spatio-temporal zero-inflated Poisson models for predicting capelin distribution in the Barents Sea	5. 発行年 2022年
3. 雑誌名 Japanese Journal of Statistics and Data Science	6. 最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-022-00183-x	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 該当する
1. 著者名 Hyodo Masashi, Watanabe Hiroki, Nakagawa Shigekazu, Nakagawa Tomoyuki	4. 巻 -
2. 論文標題 Normalizing transformation of Dempster type statistic in high-dimensional settings	5. 発行年 2022年
3. 雑誌名 Communications in Statistics - Theory and Methods	6. 最初と最後の頁 1~18
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/03610926.2022.2056749	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 Nakagawa Tomoyuki, Watanabe Hiroki, Hyodo Masashi	4. 巻 184
2. 論文標題 Kick-one-out-based variable selection method for Euclidean distance-based classifier in high-dimensional settings	5. 発行年 2021年
3. 雑誌名 Journal of Multivariate Analysis	6. 最初と最後の頁 104756~104756
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2021.104756	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 Momozaki Tomotaka, Nakagawa Tomoyuki, Ishii Aki, Saigusa Yusuke, Tomizawa Sadao	4. 巻 13
2. 論文標題 Two-Dimensional Index of Departure from the Symmetry Model for Square Contingency Tables with Nominal Categories	5. 発行年 2021年
3. 雑誌名 Symmetry	6. 最初と最後の頁 2031 ~ 2031
掲載論文のDOI (デジタルオブジェクト識別子) 10.3390/sym13112031	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 -

1. 著者名 Nakagawa Tomoyuki, Hashimoto Shintaro	4. 巻 23
2. 論文標題 On Default Priors for Robust Bayesian Estimation with Divergences	5. 発行年 2020年
3. 雑誌名 Entropy	6. 最初と最後の頁 29 ~ 29
掲載論文のDOI (デジタルオブジェクト識別子) 10.3390/e23010029	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15件 (うち招待講演 5件 / うち国際学会 5件)

1. 発表者名 中川智之, 橋本真太郎, 菅澤翔之助
2. 発表標題 外れ値に対する一般化事後分布の頑健性について
3. 学会等名 日本計算機統計学会 第 36 回大会 (招待講演)
4. 発表年 2022年

1. 発表者名 中川智之
2. 発表標題 ロバストダイバージェンスを用いたベイズ推論について.
3. 学会等名 科研費シンポジウム 「大規模複雑データの理論と方法論-新たな発展と関連分野への応用-」
4. 発表年 2022年

1. 発表者名 桃崎智隆, 中川智之
2. 発表標題 ロバストダイバージェンスを用いたordinal response model に対するベイズ推定
3. 学会等名 2022年度統計関連学会連合大会
4. 発表年 2022年

1. 発表者名 桃崎智隆, 中川智之
2. 発表標題 ロバストダイバージェンスによる ordinal response model に対するロバスト推定
3. 学会等名 日本計算機統計学会 第36 回大会
4. 発表年 2022年

1. 発表者名 桃崎智隆, 中川智之
2. 発表標題 Ordinal response model におけるロバストダイバージェンスを用いた推定
3. 学会等名 RIMS 共同研究『種々の統計的モデルにおける推測方式の有効性』（招待講演）
4. 発表年 2023年

1. 発表者名 桃崎智隆, 中川智之
2. 発表標題 ダイバージェンスを用いた ordinal response model におけるロバ ストな推定
3. 学会等名 広島大学 統計グループ 金曜セミナー（招待講演）
4. 発表年 2022年

1. 発表者名 T. Nakagawa
2. 発表標題 Comparison of various Robust Bayesian Inference against the outliers.
3. 学会等名 Australian and New Zealand Virtual Statistical Conference (国際学会)
4. 発表年 2021年

1. 発表者名 Tomotaka Momozaki, Koji Cho, T. Nakagawa, Sadao Tomizawa
2. 発表標題 Choice of the Dirichlet parameter for estimation of measures in two-way contingency tables.
3. 学会等名 The 11th Conference of the IASC-ARS The Asian Regional Section of the International Association for Statistical Computing (国際学会)
4. 発表年 2021年

1. 発表者名 中川智之, 桃崎智隆, 長光司, 富澤貞男
2. 発表標題 Choice of the Dirichlet parameter to estimate measures for square contingency tables
3. 学会等名 RIMS 共同研究『ベイズ法と統計的推測』(招待講演)
4. 発表年 2021年

1. 発表者名 桃崎智隆, 長光司, 中川智之, 富澤貞男
2. 発表標題 ベイズ法を用いた正方分割表における尺度の推定
3. 学会等名 日本計算機統計学会 第 35 回大会
4. 発表年 2021年

1. 発表者名 Tomoyuki Nakagawa
2. 発表標題 Robust Bayesian estimation by using the quasi-posterior with divergence
3. 学会等名 12th International Conference of the ERCIM WG on Computational and Methodological Statistics (国際学会)
4. 発表年 2019年

1. 発表者名 Tomoyuki Nakagawa, Shintaro Hashimoto
2. 発表標題 Robust Bayesian Inference using ϕ -divergence
3. 学会等名 10th International Workshop on Simulation and Statistics (国際学会)
4. 発表年 2019年

1. 発表者名 中川智之
2. 発表標題 ダイバージェンスを用いた外れ値に頑健なベイズ推定
3. 学会等名 計量経済学ワークショップ (招待講演)
4. 発表年 2019年

1. 発表者名 Tomoyuki Nakagawa, Shintaro Hashimoto
2. 発表標題 Objective Priors in Robust quasi-Bayesian Inference using the Divergences
3. 学会等名 THE 4TH EASTERN ASIA MEETING ON BAYESIAN STATISTICS, EAC-ISBA 2019 (国際学会)
4. 発表年 2019年

1. 発表者名 中川智之, 橋本真太郎
2. 発表標題 -divergence を用いたロバストなベイズ推測
3. 学会等名 応用統計学会2019年年会
4. 発表年 2019年

〔図書〕 計0件

〔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
---------	---------