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　４ 年　 ５ 月 ２６ 日現在

機関番号：１２６０１
研究種目：若手研究(B)
研究期間：2017～2021
課題番号：１７Ｋ１７６５９

研究課題名（和文）大規模計数時系列データのベイズ分析

研究課題名（英文）Scalable Bayesian analysis of high-dimensional streaming counts

研究代表者
入江　薫（Irie, Kaoru）

東京大学・大学院経済学研究科（経済学部）・講師

研究者番号：２０７８９１６９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：ウェブサイトへのアクセス数のデータに代表される、計数値のストリーミングデータの逐次分析に関する研究。次々にデータが観測される状況で、逐次的な事後・予測分布の計算が解析的に可能になるように、ポアソン・ガンマ型の共役性と呼ばれる統計的性質を活かした状態空間モデルを研究した。また、急激なアクセス数の増加に対応できるよう、当該のモデルを拡張するとともに、逐次モンテカルロ法と呼ばれる計算手法を適用できるようにした。

研究成果の学術的意義や社会的意義
計数時系列データは様々な分野で見られるデータの一形態であり、本研究課題で開発・提案した統計モデルおよび分析手法が広く応用されることが期待される。また、短時間で計算を完了しなければならないという設定は、実務上よくみられる状況であり、実データに対する予測や意思決定の問題に現実的な解答を与えている点でも本研究の成果には意義がある。本研究の成果をもとに多くの研究プロジェクトが派生していることから、純粋に統計学上の問題としての意義もおおいに認められる。

研究成果の概要（英文）：Research on sequential analysis of streaming counts. I developed a state-space model of the Poisson observations, utilizing the gamma-beta Markov chain as the state evolution. The conjugacy of gamma distributions enables the filtering of posterior and predictive distributions. I also considered an extension to the model with dynamic discount factors, providing a custom sequential Monte Carlo method. It is a slight extension, but crucial in adapting the predictions to an abrupt burst of counts, which is often observed in the access log data of websites.

研究分野：ベイズ統計学

キーワード：計数データ　時系列データ　状態空間モデル　逐次モンテカルロ法

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究課題は主たる研究（１）と、その更なる改善（２）からなる。（３）以降の事項は、それらから派生した研究である。

（１）本研究課題の申請以前より､私はウェブサイトのアクセスログのデータ分析を行ってきた。このデータはウェブサイトへのアクセスの数を計測したもので、計数（非負の整数値を取る）データであり、大規模・高次元な時系列データであり、かつ毎 30 秒ごとに新たな値が観測されるストリーミングデータであることが特徴である。統計分析上の主な困難は、次々に新たなデータが流れ込んでくる状況で、次のデータが取得されるまでの短い時間のあいだに、得られたデータにもとづいて統計分析の結果を更新しなければならないという、逐次分析の難しさにある。加えて、正規分布に基づく簡易な統計分析を直接適用できない計数値データであることも考慮しなければならない。これらの問題に対処するにあたり、複雑な統計モデルを時間のかかる計算統計学の手法で推定することをあきらめ、分析にかかる計算時間が現実的なものになるように、簡易かつなるべく柔軟なモデルを開発することを目指すに至った。

（２）実際のアクセスログのデータには急激なアクセス数の増加に代表される、特異な現象が見られる。そのため、（１）で開発したモデルをそのまま当てはめると予測の精度が悪くなる。その原因は、アクセス数が安定的である期間に最も適応したモデルは、過去のデータの値をより強力に分析結果に反映することにある。結果として、急激な変化に際しても過去のデータを尊重するあまり、予測に失敗するばかりか、その修正にも時間を要することになる。この観察から、モデルを改善するために以下の着想を得た。データの急激な変化においては、予測のバイアスを注視し、予測の著しい失敗に対応して過去のデータを「忘れる」ことで、現在時点の急激な変化に素早く適応することが必要である。一方で、そのような近視眼的な適応を常に行うことは予測精度の観点からは望ましくないため、一旦データの挙動が安定しはじめたら、モデルは過去のデータを再び蓄積し、予測の精度を高めるべきである。これらの着想を具体化する統計モデルの改良を目指す必要がある。

（３）以上の研究においては、のちに述べるように、基本的な確率分布であるガンマ分布が大きな役割を果たす。これらの研究で得られるガンマ分布に関する発見や知見は、他の統計分析においても有用となることが期待される。たとえば、金融商品のリターンの時系列データにおいてはボラティリティと呼ばれる分散の時間的変動が興味の対象となるが、正規分布の分散（の逆数）の統計モデルとして相性がいい共役な分布として知られているのは、ガンマ分布である。実際、金融時系列の統計モデルとして、（１）と類似のモデルが知られている。これまでの研究の知見を活かし、このような関連する統計分析上の問題について考察する。

２．研究の目的

（１）計数ストリーミングデータの分析に有用なモデルを開発し、その逐次分析の計算手法を明らかにする。実際のウェブサイトのアクセス数のデータに応用し、ウェブページ間の関係性や予測精度について議論する。

（２）急激なアクセス数の増加の可能性を考慮して、上記のモデルを拡張するとともに、計算のための手法を考案する。

（３）統計学上の関連する問題に対して、得られた知見を活かして研究を行う。

３．研究の方法

（１）計数データの標本モデルとして用いられる確率分布のうち、最も簡単なものはポアソン分布である。ガンマ分布は、ポアソン分布の平均パラメータの共役な事前分布として知られる。つまり、データを得た後の平均パラメータの事後分布もまたガンマ分布となり、平均・分散・分位点などの計算が容易になる。このような計算上の利点から、時系列データのモデリングにおいても、周辺分布が常にガンマ分布となるようなモデルが求められる。そこで、計数時系列データの分析に用いられてきたモデルであるガンマ・ベータ型のマルコフ連鎖を利用して、ポアソン標本モデルに関する状態空間モデルを構成する。統計分析上必要な事後分布、データを得た時の更新式、予測分布などを導出し、実際のアクセスログデータへ応用する。

（２）上記のガンマ・ベータ型のマルコフ連鎖のうち、割引因子と呼ばれるパラメータを固定せず、自己回帰過程に従う動的な状態変数に変更する。この変更に伴い共役性は失われるため、代わりに逐次モンテカルロ法と呼ばれる逐次分析の手法を考案する。単に既存の一般的な方法を適用するのではなく、モデルの簡易さを最大限に利用し、本モデルに特化した方法を提示する。具体的には、提案分布として事前分布を用いるのではなく、目標分布をよく近似するものを解析的に導出して使用する。実データに適用し、急激なアクセス数の増加に適応するための一時的な割引因子の低下が、期待された通りに見られるかを検討する。

４．研究成果

（１）提案する統計モデルおよび計算手法は実際のニュース配信ウェブサイトのアクセスログのデータに適用された。このウェブサイトはニュース記事のジャンルに応じて複数のカテゴリーに分かれており、元となるアクセスログからは各カテゴリーへの流入者数やカテゴリー間の移動者数からなる高次元の計数時系列データが構成される。提案する手法によって、各時系列の逐次的分析や予測のみならず、カテゴリー間の相対的な人気度や、カテゴリー間の移動の特異なパターンの検出など、応用上重要となるデータの性質を明らかにすることができた。また、実データ特有の問題、とくに異常値を処理するための実用的なフローチャートを統計的意思決定理論にもとづき提案した。これらの研究を共同で行い、その成果は Chen et al. (2018)として、Journal of American Statistical Association 誌に採択された。

（２）急激なアクセス数の増加に対して、動的な割引因子を用いる提案手法は期待通りの適応的反応をみせた。また、逐次モンテカルロ法特有の問題として知られる粒子退化の問題は見られなかった。これは計算手法を本モデルの特徴に合わせて特化させたことによると考えられる。この拡張されたモデルは急激な変化が存在しないデータや、計数の規模が小さい（0,1,2 などの値を多く取る）データに対しても有効であった。これらの発見を共同研究者とともに論文にまとめた。論文は Irie et al. (2022)として Annals of Applied Statistics 誌に採択された。

（３）ガンマ・ベータ型のマルコフ連鎖を多変量に一般化したものとして、ウィシャート・行列ベータ型のモデルが知られている。これを金融商品のボラティリティの逐次分析に用いる研究がふたつ存在する。（１）の研究で得られた知見をもとに、両モデルの差異を明らかにする研究を行った。特に、両者は別のモデルでありながら、まったく同一の予測分布をもち、そのため通常の方法(事後モデル確率)によっては峻別できないことが分かった。そして両モデルの違いは、過去のボラティリティの値を振り返った時に、想定する不確実性の大きさに表れることを示した。予測分布や事後分布の計算手法を与えて、実際の為替データを例に両モデルの比較を行った。この共同研究は Pena and Irie (2021)として Journal of Time Series Analysis に採択された。

（４）ガンマ分布は縮小事前分布に対してもよく用いられる。特に金融商品のポートフォリオ選択問題に関して、ガンマ分布による縮小効果によってポートフォリオの変動を減らしたり、使用する商品を減らすなどの工夫を可能にした。この研究成果は共同で Irie and West (2019)として Bayesian Analysis 誌に採択された。

（５）時系列に限らない一般の計数データの分析において、推定値を少なく見積もってしまう現象は縮小効果として知られる。使用されるガンマ分布の形状パラメータにさらにガンマ分布を使用する階層構造を導入することで、大きな観測値についてはこの縮小効果を限りなく小さくできるという望ましい性質（tail-robustness）を達成できることを示した。この共同研究の結果は Hamura et al. (2022)として Bayesian Analysis 誌に採択された。

（６）ガンマ分布の形状パラメータの重要性を認識するにつれ、同パラメータは点推定を行うのが困難なパラメータであることに思い至った。同じ問題意識を共有する数理統計の研究者の誘いをうけ、同問題の理論的研究に取り組んだ。研究成果は二本の論文にまとめられ、Tamae et al. (2019)および Tamae et al. (2022)としてそれぞれ Japanese Journal of Statistics and Data Sciences および Communication in Statistics 誌に掲載された。

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　4件／うちオープンアクセス　6件）

１．著者名 Chen Xi, Irie Kaoru, Banks David, Haslinger Robert, Thomas Jewell, West Mike	４．巻 113
２．論文標題 Scalable Bayesian Modeling, Monitoring, and Analysis of Dynamic Network Flow Data	５．発行年 2018年
３．雑誌名 Journal of the American Statistical Association	６．最初と最後の頁 519-533
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/01621459.2017.1345742	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 Irie Kaoru, West Mike	４．巻 14
２．論文標題 Bayesian Emulation for Multi-Step Optimization in Decision Problems	５．発行年 2019年
３．雑誌名 Bayesian Analysis	６．最初と最後の頁 137-160
掲載論文のDOI（デジタルオブジェクト識別子） 10.1214/18-BA1105	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 該当する

１．著者名 Tamae Hiromasa, Irie Kaoru, Kubokawa Tatsuya	４．巻 3
２．論文標題 A score-adjusted approach to closed-form estimators for the gamma and beta distributions	５．発行年 2020年
３．雑誌名 Japanese Journal of Statistics and Data Science	６．最初と最後の頁 543-561
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-019-00071-x	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 Pena Victor, Irie Kaoru	４．巻 43
２．論文標題 On the Relationship between Uhlig Extended and beta‐Bartlett Processes	５．発行年 2021年
３．雑誌名 Journal of Time Series Analysis	６．最初と最後の頁 147-153
掲載論文のDOI（デジタルオブジェクト識別子） 10.1111/jtsa.12595	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 該当する

1．著者名 Irie Kaoru, Glynn Chris, Aktekin Tevfik	4．巻 16
2．論文標題 Sequential modeling, monitoring, and forecasting of streaming web traffic data	5．発行年 2022年
3．雑誌名 The Annals of Applied Statistics	6．最初と最後の頁 300-325
掲載論文のDOI（デジタルオブジェクト識別子） 10.1214/21-AOAS1505	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 該当する

1．著者名 Hamura Yasuyuki, Irie Kaoru, Sugasawa Shonosuke	4．巻 17
2．論文標題 On Global-Local Shrinkage Priors for Count Data	5．発行年 2022年
3．雑誌名 Bayesian Analysis	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） 10.1214/21-BA1263	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 Tamae Hiromasa, Irie Kaoru, Kubokawa Tatsuya	4．巻 -
2．論文標題 Score-adjusted methods for estimation of shape parameters in Gamma-Poisson and Beta-Binomial distributions	5．発行年 2022年
3．雑誌名 Communications in Statistics - Simulation and Computation	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） 10.1080/03610918.2022.2044051	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21件（うち招待講演　12件／うち国際学会　15件）

1．発表者名 Kaoru Irie
2．発表標題 On the conjugate multivariate stochastic volatility processes
3．学会等名 Webinar of Bayesian Econometrics 2020（招待講演）（国際学会）
4．発表年 2020年

１．発表者名
Kaoru Irie

２．発表標題
Robust regression with log-regularly varying error distributions

３．学会等名
日本統計学会春季大会（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
Kaoru Irie

２．発表標題
Bayesian dynamic fused LASSO

３．学会等名
EAC-ISBA 2019 (The 4th Eastern Asia meeting on Bayesian Statistics)（国際学会）

４．発表年
2019年

１．発表者名
Kaoru Irie

２．発表標題
Bayesian dynamic fused LASSO

３．学会等名
Japanese Joint Statistical Meeting （日本統計関連学会連合大会）（国際学会）

４．発表年
2019年

１．発表者名
Kaoru Irie

２．発表標題
Bayesian dynamic fused LASSO

３．学会等名
CFE 2019 (The 13th International Conference on Computational and Financial Econometrics)（国際学会）

４．発表年
2019年

１．発表者名
入江　薫

２．発表標題
縮小事前分布と状態空間モデル

３．学会等名
科研費シンポジウム「多様な分野における統計科学に関する諸問題」

４．発表年
2019年

１．発表者名
入江　薫

２．発表標題
繰り返し対数関数を用いた縮小事前分布のクラスについて

３．学会等名
科研費研究集会「ベイズ計量経済学研究集会」

４．発表年
2019年

１．発表者名
Kaoru Irie

２．発表標題
Adaptive Monitoring of Steaming Counts by Poisson-gamma State Space Models

３．学会等名
International Society of Bayesian Analysis, world meeting (ISBA2018)（国際学会）

４．発表年
2018年

１．発表者名
Kaoru Irie

２．発表標題
On-line analysis of count-valued time series by dynamic discount factors

３．学会等名
2018年度統計関連学会連合大会（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
入江薫

２．発表標題
状態空間モデルにおけるスパース性

３．学会等名
科研費研究集会「ベイズ計量経済分析研究集会」

４．発表年
2018年

１．発表者名
入江薫

２．発表標題
割引因子ポアソン-ガンマ状態空間モデルによる計数時系列モデルの逐次分析

３．学会等名
科研費研究集会「融合する統計科学」（招待講演）

４．発表年
2018年

１．発表者名
Kaoru Irie

２．発表標題
Filtering for stochastic volatility with leverage by mixture approximation

３．学会等名
The 12th International Conference on Computational and Financial Econometrics (CFE 2018)（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Kaoru Irie

２．発表標題
Fox News Network Data Analysis: Bayesian Dynamic Modeling

３．学会等名
2017 CSA-KSS-JSS International Conference（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
Kaoru Irie

２．発表標題
Scalable Bayesian modeling, monitoring and analysis of dynamic network flow data

３．学会等名
International Workshop on Bayesian Econometric Analysis（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
入江　薫

２．発表標題
ベイジアン・モデリングによるポートフォリオ最適化

３．学会等名
2017年度統計関連学会連合大会

４．発表年
2017年

１．発表者名
Kaoru Irie

２．発表標題
Bayesian Emulation for High-Dimensional Portfolio Problems

３．学会等名
The 1st International Conference on Econometrics and Statistics (EcoSta 2017)（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
Kaoru Irie

２．発表標題
Bayesian Emulation for Multi-Step Optimization in Portfolio Decisions

３．学会等名
The Third International Conference on Engineering and Computational Mathematics (ECM2017)（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
Kaoru Irie

２．発表標題
Bayesian dynamic fused lasso

３．学会等名
Bayesian Inference in Stochastic Processes 12 (BISP12)（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
Kaoru Irie

２．発表標題
Robust regression with log-regularly varying error distributions

３．学会等名
Eastern Asia Chapter of the International Society for Bayesian Analysis (EAC-ISBA 2021)（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
Kaoru Irie

２．発表標題
On the conjugate multivariate stochastic volatility processes

３．学会等名
The 14th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2021)（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
入江薫

２．発表標題
単調回帰のための事前分布

３．学会等名
2021年度統計関連学会連合大会

４．発表年
2021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			
米国	University of New Hampshire	The City University of New York	Duke University	