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　元 年　 ８ 月 ２９ 日現在

機関番号：１２６０１
研究種目：研究活動スタート支援
研究期間：2017～2018
課題番号：１７Ｈ０６５７０

研究課題名（和文）高次元状況下の転移学習における高速かつ高精度な分布予測手法の開発

研究課題名（英文）Efficient predictive density for transfer learning under high-dimensional settings

研究代表者
矢野　恵佑（Yano, Keisuke）

東京大学・大学院情報理工学系研究科・助教

研究者番号：２０８０６０７０

交付決定額（研究期間全体）：（直接経費）　　1,500,000円

研究成果の概要（和文）：本研究では，研究A)擬似ベイズ法に着目したベイズ法における計算コストの削減，研究B)高次元カウントデータのもつ疎性に着目した分布予測手法の構築,を行った.
研究Aでは，簡便な尤度を利用するベイズ法である擬似ベイズ事後分布の高次元状況下での性質を調べた．分布予測では現在はベイズ的な方法が主流であるが,擬似ベイズ法を利用することで計算コストを抑えつつ性能を劣化させない分布予測が行えることが明らかになった.
研究Bでは,疎性をもつ高次元カウントデータに対して高精度かつ計算コストが低い分布予測手法の構築を行った.疎性に着目することで高精度と低計算コストが両立可能であることが明らかになった.

研究成果の学術的意義や社会的意義
予測とは,現在の観測量をもとに予測したい量(予測量)の振る舞いを推測することである.地震予測,交通予測,遺伝子機能予測等,様々な予測が社会で活用されている.統計的な予測手法には,予測量の平均を推定する点予測と予測量の従う分布を推定する分布予測がある.予測量の従う分布が分かれば,検定や予測区間の構成ができるため,分布予測がより重要である.
転移学習とは,ある領域での観測量を利用して別の領域にある予測量を予測することである.転移学習は統計学と機械学習で近年注目されており,例えば、深層学習の精度向上に利用されている.転移学習の理論的性質が分かると,既存の学習手法の精度は飛躍的に向上するため重要である.

研究成果の概要（英文）：This research develops A)the strategies of reducing the computational cost of Bayesian methods by quasi-posteriors; B)the efficient predictive densities for sparse count data.
In Research A, theoretical properties of quasi-posteriors (posteriors based on handy or mis-specified likelihoods) such as information losses or performance in uncertainty quantification are studied under high dimensional settings. This research shows a possibility of constructing predictive densities with both low computational costs and high performance by leveraging quasi-posteriors.
In Research B, efficient predictive densities for sparse count data are constructed. This research shows compatibility of high performance and low computational cost in constructing predictive densities under high dimensional settings by focusing on sparsity or quasi-sparsity of data.

研究分野：統計的予測

キーワード：予測分布　高次元統計　擬似ベイズ　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

予測とは、現在の観測量をもとに予測したい量（予測量）の振る舞いを推測することである。地震予測・交通予測・遺伝子機能予測等、様々な予測が社会で活用されている。統計的な予測手法には、予測量の平均を推定する点予測と予測量の従う分布を推定する分布予測がある。予測量の従う分布が分かれば,検定や予測区間の構成ができるため、分布予測がより重要である。分布予測は点予測を包含する。しかし、点予測手法を用いて分布予測を行うと、予測量の分散が過少に評価されるため、精度が悪い。そこで、本研究では分布予測を扱う。

予測に関して、統計学と機械学習では転移学習が近年注目されている。転移学習とは、ある領域での観測量を利用して別の領域にある予測量を予測することである。より詳しくいうと、転移学習は観測量と予測量の分布が異なる状況での予測である。

応募者は、転移学習における分布予測のリスク解析を行った。これらの研究は、転移学習においては観測量と予測量の分布が異なるほど、ベイズ的な分布予測による予測精度が点予測による予測精度よりも極めて良くなることを示した点が画期的である。転移学習における分布予測の有効性は、これまで知られていなかった。しかし、転移学習における既存の分布予測手法は、

問題点 a) 理論的な予測精度は良いが計算コストが大きい

問題点 b) 冒頭のゲノムワイド回帰のような高次元モデルで予測精度が劣化する

といった欠点をもち、実用化に至っていない。問題点 a について、転移学習において高精度な分布予測手法はブートストラップを用いた安定化を必要とするため、点予測手法に比べ計算コストが非常に大きい。予測は地震予測のようにある程度即時性が重要であるため、計算コストが大きいことは実用化の妨げとなっている。問題点 b について、高次元状況下では多くの分布予測手法が点予測手法 LASSO より精度が悪くなってしまうことが知られている。この問題は、観測量と予測量の分布が等しい場合には Mukherjee and Johnstone (2015)により解決されたが、転移学習においては未解決である。実データでは高次元モデルが頻繁に現れるため、高次元モデルで性能が劣化することは好ましくない。

２．研究の目的

以上の経緯から本研究では、高次元モデルでの転移学習において「低計算コストかつ高精度」な分布予測手法の開発を行う。具体的には、転移学習において、

課題 A) どのような状況で低計算コストと高予測精度が両立できるか?

また、両立するような手法は何か?

課題 B) 高次元モデルで性能が劣化しないような手法は何か?

を明らかにし、手法の実用化を目指す。計算コストと予測精度はトレードオフの関係にあるため、一般には両立し得ない。従って、「低計算コストかつ高精度」な分布予測手法を開発するにあたって、そもそも両立が可能であるかを調べることは重要である。

３．研究の方法

応募者は研究目的を達成するため、研究を以下の 3 つのフェイズに分けて研究を進めていった：

(P1)予測精度研究(課題 A の解決)：転移学習における分布予測について、計算量制約のもとでの予測精度の限界を調べる。さらに、予測精度の上限を達成する手法をいくつか開発する。

(P2)高次元拡張研究(課題 B の解決)：P1 で開発した手法のうち、あるいはそれらを改良し、高次元モデルで性能が劣化しない手法を開発する。

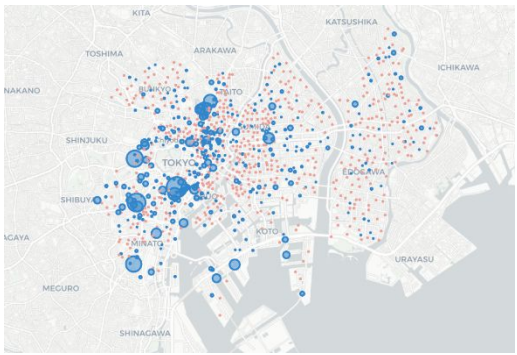
(P3)実用化: P1 と P2 で考察・構築した手法の汎用パッケージの作成を行う。また、開発した手法を実データに適用する。

４．研究成果

本研究では、 研究 A)擬似ベイズ法に着目したベイズ法における計算コストの削減、研究 B)高次元カウントデータのもつ疎性に着目した分布予測手法の構築を行った。

研究 A では、 簡便な尤度を利用するベイズ法である擬似ベイズ事後分布の高次元状況下での性質を調べた。 特に、Bernstein–von Mises の定理と呼ばれるベイズ版の中心極限定理の有限標本での精緻化および高次元中心極限定理の適用により、「線形回帰において背後にある真の誤差分布が何であっても正規分布の尤度を用いた疑似ベイズ法が不確実性評価において一定の精度をもつ」ことが示せた。この研究で開発された Bernstein–von Mises の定理は事前分布に置く仮定が少なく、非常に汎用的である。更に、線形回帰における分散未知の状況に対応している点も利点として挙げられる。また、理論面でも、高次元中心極限定理のベイズの理論解析での有用性を示しており、他の研究に与える影響は大きい。分布予測では現在はベイズ的な方法が主流であるが、擬似ベイズ法を利用することで計算コストを抑えつつ性能を劣化させない分布予測が行えることが明らかになった。この研究は論文としてまとめられ、論文誌 Bernoulli にて査読中である。

研究 B では、疎性をもつ高次元カウントデータに対して高精度かつ計算コストが低い分布予測手法の構築を行った。右図は疎性をもつ高次元カウントデータの例である。この研究では、観測データと予測データの標本サイズの不均一性というある種の転移学習の状況を扱っている。高次元状況下におけるベイズ統計学では Spike-and-slab 事前分布や horse shoe 事前分布といった事前分布が利用されるが、この研究ではそこに「予測の観点からの尺度の調整」という新しいアイデアを導入し、非常に高精度な予測分布を構成した。更に、spike-and-slab prior の slab prior に広義の事前分布を利用するというアイデアにより予測分布構成時の計算コストを下げることに成功した。結果として、高精度と低計算コストが両立可能であることが明らかになった。更に、従来の研究では疎性を表すハイパーパラメタは既知として理論解析が進められていたが、我々はこの仮定を外すことに成功した。アイデアとしては、疎性を表すハイパーパラメタに対する簡単な推定量をプラグインするという簡単なものであるが、提案する予測分布が簡単な形をしていることから、予測分布が疎性に対して適応的であること理論的に示すことができた。この研究の有用性は東京都の犯罪データおよび遺伝子の発現に関する実データを用いて確かめられている。公開パッケージに関しては製作中である。



図：疎性をもつ高次元データの例：東京都の町村ごとのスリ件数.青い円の半径は件数に比例する．赤い町村ではスリがない.

以上のように、研究期間内に(P1)-(P3)の全てのフェイズが順調に進み、高次元モデルでの転移学習において「低計算コストかつ高精度」な分布予測手法の開発が行えた。

５．主な発表論文等

〔雑誌論文〕（計 1 件）

①Yuya Takasu, Keisuke Yano, and Fumiyasu Komaki, Scoring Rules for Statistical Models on Spheres, *Statistics and Probability Letters,* vol. 138, pp. 111-115, 2018 (doi:10.1016/j.spl.2018.02.054).

〔学会発表〕（計９件）

①Ryoya Kaneko, Keisuke Yano, and Fumiyasu Komaki, Sparse Poisson sequence model with different sample sizes, CMStatistics 2018, Italy, 2018.

②Keisuke Yano and Fumiyasu Komaki, Non-asymptotic minimax adaptiation and weak admissibility using random sieve priors, The 5th Institute of Mathematical Statistics Asia Pacific Rim Meeting, Singapore, June, 2018.

③矢野恵佑，駒木文保，Non-asymptotic Bayesian minimax adaptation in several nonparametric models，企画セッション「New trends in Bayesian perspective」，2018 年度統計関連学会連合大会，東京，2018.

④矢野恵佑, Inequalities for minimax Renyi divergence, 2018 年度統計関連学会連合大会，東京，2018.

⑤矢野恵佑, Divergence for statistical analysis of spherical data，シンポジウム「統計・機械学習の交わりと広がり」，東京, 2018.

⑥矢野恵佑，駒木文保，Non-asymptotic minimax Bayesian nonparametric estimation based on invariance，Current topics on algebraic statistics and related fields，兵庫，2018.

⑦矢野恵佑，駒木文保，Weak admissibility in high-dimensional and nonparametric statistical models，2017 年度統計関連学会連合大会，愛知，2017.

⑧今泉允章，矢野恵佑，Nonparametric regression for manifold data via embedding distance，2017 年度統計関連学会連合大会，愛知，2017.

⑨矢野恵佑，加藤賢悟，Finite sample bound for the Bernstein-von Mises theorem，2017 年度統計関連学会連合大会，愛知，2017.

〔その他〕

ホームページ等

６．研究組織

(1)研究分担者
(2)研究協力者
研究協力者氏名：アンドリュー バロン
ローマ字氏名： (Barron, Andrew)
研究協力者氏名： 加藤賢悟
ローマ字氏名： (Kato, Kengo)
研究協力者氏名： 駒木 文保
ローマ字氏名：(Komaki, Fumiyasu)
研究協力者氏名：ゴウラ ムカジー
ローマ字氏名：(Mukherjee, Gourab)