科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２０　日現在

機関番号：　１２１０２
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５０１４２

研究課題名（和文）高速で頑健かつ高精度な多変量統計手法の新展開

研究課題名（英文）New Developments of Multivariate Statistical Methodologies - High Speed, Robustness, and High Accuracy

研究代表者
青嶋　誠（AOSHIMA, Makoto）

筑波大学・数理物質系・教授

研究者番号：９０２４６６７９

交付決定額（研究期間全体）：（直接経費）　　2,600,000円、（間接経費）　　780,000円

研究成果の概要（和文）：本研究は、高速・頑健・高精度な新しい多変量統計手法の開発を目指すものである。我々は、非ガウスで外れ値が混入するデータに対しても、低計算コストで安定した高い精度を保証する、3つの多変量統計手法を開発した。研究成果は次の通りである。(1) 高次モーメントを利用した高速かつ高精度な判別方法の開発。(2) 高速計算を可能にする高精度変数選択と外れ値検出法の開発。(3) 外れ値が混在するデータ空間における潜在空間の精密な解析。

研究成果の概要（英文）：In this research project, we aim to develop new multivariate statistical methods satisfying the criteria of high speed, robustness and high accuracy for inferences on modern data. We provided three multivariate statistical methods to ensure robustness and high accuracy with low computational cost even for non-Gaussian, contaminated models. The findings of this research are as follows: (1) Developments of high-speed and highly accurate classification methods using higher moments. (2) Developments of high-speed and highly accurate variable selection and outlier detection methods. (3) Intrinsic space analysis in a contaminated data space.

研究分野：　多変量解析

科研費の分科・細目：　情報学・統計科学

キーワード：　異常値　クラスター分析　判別分析　多変量解析　クロスデータ行列法　高次モーメント　非ガウス　回帰分析

１．研究開始当初の背景

(1) データの科学と情報化の進展に伴い、機械学習の発展が目覚ましい近年の統計学において、その発展を支えてきたのは、2000 年代の計算機の発展だといえる。しかしながら、最近の計算機の発展は、徐々に鈍化しているように思え、それに反し、データの形態はますます複雑化を増し、データの高次元化はいっそう進んでいる。今後は、計算機の性能に頼るだけでは統計学の発展を支えることが出来ないと予想される。

(2) 機械学習の手法は汎用性と引き換えに、カーネル関数の選び方に結果が大きく影響される。さらに、カーネル関数のパラメータは交差検証法等で決まるため、膨大な計算コストが掛かる。また、解に理論的な保証があるとは言えず、得られた出力だけで良し悪しを判断しているように思える。一方、数理統計学の方法は、精度保証と引き換えにガウス性等の数学的な仮定に根ざしている。それらは、しばしば非現実的であり、汎用性に劣る。

(3) 近年の統計学は、複雑化の一途を辿り、計算コストは爆発的に増大している。アルゴリズムの開発で計算コストの問題に取り組む研究も盛んに行われているが、根本的な解決になっているとは言い難い。ゲノムなどの複雑なデータを解析する場面では、もはや、アルゴリズムの開発だけでは対応できない。さらに、複雑な統計手法はモデルに多くの制約を必要とし、実用面において疑問と不安を抱かせるものである。

２．研究の目的

2010 年代のデータ科学は、高速・頑健・高精度の３つがキーワードになると思われる。本研究は、計算コストを抑え、頑健かつ高精度な解析を理論的に保証する、新しい統計的方法論の開発を目指す。本研究は、次の 3 つを目的とする。

(1) 高次モーメントを利用した高速かつ高精度な判別方法の開発。

機械学習において、データを超高次元特徴空間へ写像して判別するサポートベクトルマシン(SVM)があるが、この方法が理論的な保証を与えているとは言い難い。本研究は、数理統計学における２次判別法に着目する。古典的な判別方法であるが、これに高次モーメントを導入することで、非ガウスデータにも高速で高精度な判別を可能にする。

(2) 高速計算を可能にする高精度変数選択と外れ値検出法の開発。

機械学習において、モデルの推定と変数選択を兼ねた LASSO タイプの方法があるが、平滑化パラメータの選択や、それに伴う計算コストを考えると、十分なパフォーマンスとは言い難い。本研究は、精度に関する理論的な保証から、高速な変数選択法を開発する。外れ値を分離する潜在空間への射影という考え方から、高速かつ高精度な外れ値検出法を開発する。

(3) 外れ値が混在するデータ空間における潜在空間の精密な解析。

計算コストを削減するための単純な方法として、データを低次元の潜在空間へ射影することが考えられる。高次モーメントの利用と、外れ値を分離する射影法を融合させることによって、非ガウス性をもつ多次元データの潜在空間に、高速で頑健かつ高精度な推測方法を理論的に構築する。

３．研究の方法

(1) 研究目的の(1)について、多変量解析の従来の推測には余り用いられてこなかった高次モーメントに着目する。これらを通して非ガウス性に関する情報を取り入れることで、非ガウスデータを判別する。青嶋と矢田は、非ガウスデータの判別のために、その目的に合った高次モーメントの適切な推定を考える。数理統計学における２次判別を再考する。従来の２次判別は、マハラノビス距離によって精度が保証され、計算コストが掛からないものの、２次までの特徴量でデータを判別しようとする方法と言える。本研究は、非ガウスデータに対する判別方法を開発するために、高次モーメントを判別関数に組み込む。機械学習のサポートベクトルマシンとの関連性も研究する。赤平とも意見交換を密にとり、非正則構造を明確にして、非ガウスデータに高速かつ高精度な判別を実現する。

(2) 研究目的の(2)について、外れ値を分離するために有効な空間を探索し、その空間にデータを射影することで外れ値の検出を考える。青嶋と赤平は、情報量規準と平滑化パラメータの関連から、非正則モデルにおける変数選択法を考える。青嶋と矢田は、潜在分布と外れ値を分離させる潜在空間への射影を考え、高精度な外れ値検出法を開発する。そのために、Yata and Aoshima (2010, Journal of Multivariate Analysis)で考案したクロスデータ行列法に着目し、外れ値が混在する場合に方法論を拡張する。外れ値を含んだ多変量非正則分布における漸近理論を研究し、クロスデータ行列法を使って潜在分布と外れ値が分離する方向を推定することで、低計算コストかつ高精度な外れ値検出を実現する。

(3) 研究目的の(3)について、前年度までに開発した高次モーメントを利用する方法論

と、外れ値を分離する射影法を融合させる。ノイズや外れ値の影響を排除した非ガウスデータの潜在空間を推測することで、大幅な計算コストの削減と頑健かつ高精度な統計解析を実現する。そのために、青嶋と矢田は、確率論におけるランダム行列理論を支えている数学的な（それゆえに非現実的な）仮定を緩め、統計学的な一般モデルのもとで理論を再構築する。また、赤平とも意見交換を密にとり、LASSO をβダイバージェンスと融合させて外れ値に影響を受けずに変数選択を行い、その後で潜在空間を解析する方法も考える。

４．研究成果

(1) 研究目的の(1)について、研究計画の初年度に取り組んだ。青嶋と矢田は、高次モーメントの情報を取り入れた非ガウスデータの判別方法を考案した。一般に、マハラノビス距離に基づく判別方法は、標本共分散行列の逆行列の計算が不安定になる。そのため、これを縮小型の推定量やナイーブ・ベイズな推定量で代替する先行研究が多く見られる。青嶋と矢田は、これらの推定量が必ずしも好ましい性質をもたないことを示し、固有空間の幾何学的表現に基づく逆行列の新しい推定を与え、誘導される判別方法の性能が先行研究に優ることを示した。また、判別関数そのものを非ガウスデータの幾何学的表現から導き出し、それが漸近正規性と一致性を有することを証明し、提案する判別方法に精度保証を与えた。機械学習のサポートベクトルマシンおよび関連ベクトルマシンとの比較を行い、提案手法が計算コストの削減にも成功していることを確認した。

青嶋と矢田は、高次モーメントの利用をクラスター分析にも応用し、非ガウスデータの幾何学的表現に基づくクラスタリング方法を提案した。

また、青嶋と矢田は赤平と意見交換を密にとり、外れ値が混入した非正則な状況下での回帰分析を考えた。外れ値に対して頑健なモデル構築を行うためのダイバージェンスを導入し、これに基づくモデル選択基準を与えた。提案する回帰分析法が真のモデルを選択する確率を数値的に評価し、外れ値に対して頑健なモデル構築とモデル選択を実現することを確認した。さらに、一連の分析を高速に処理するためのアルゴリズムも開発した。

なお、本研究の一部については、国内外の学会やシンポジウム等で成果発表をしている。成果を取り纏めたものは、国際学術誌に投稿し、現在、改訂の段階である。

(2) 研究目的の(2)について、研究計画の 2 年目に取り組んだ。青嶋と矢田は、高速計算を可能にするための外れ値検出法について、多次元データ空間において潜在分布と外れ値が分離するような射影を考えた。Yata and Aoshima (2010, Journal of Multivariate Analysis)で考案したクロスデータ行列法を用いて射影方向を決める方法を考案し、外れ値が混在するデータで解析を試みた。その結果、クロスデータ行列の 2 分割の構成法が、外れ値に対する頑健性の決め手になることを突き止めた。実験によると、ある条件の下、外れ値と潜在分布は明確に分離される。すなわち、データを射影して得られる潜在空間において、外れ値が明確に可視化される。この実験結果は、低計算コストかつ高精度な異常値検出の可能性を示唆している。

また、判別の問題では、高次モーメントを利用する方法について研究を続行した。Aoshima and Yata (2011, Sequential Analysis, Editor’s Invited Paper)で与えた幾何学的判別法について、Aoshima and Yata (2013, Annals of the Institute of Statistical Mathematics)は判別確率に関する一致性を理論的に証明し、従来の判別方法よりも精度に優ることを数値的にも示した。赤平は、ピアソンのカイ 2 乗適合度検定とステューデント t 分布を再考し、2 標本問題において、基盤となる非心 t 分布に関する高次近似とそれに伴う高速計算法を与えた。

なお、本研究の一部については、国内外の学会やシンポジウム等で成果発表をしている。成果を取り纏めたものは、下記の主な発表論文等の[雑誌論文]に記載した通り、国際学術誌に掲載されている。

(3) 研究目的の(3)について、研究計画の最終年に取り組んだ。外れ値が混在するデータ空間における潜在空間の解析について、青嶋と矢田は、外れ値が混入する多変量モデルに対する潜在空間の推定を考え、従来型の推定とクロスデータ行列法による推定を比較し、それぞれの方法が推測の精度を保証するための条件を理論的に明らかにした。本研究は、クロスデータ行列法が本来威力を発揮する高次元データの枠組みではなく、あくまで、通常の多変量解析の枠組みで扱い、その性質を研究した。その結果、従来型の推定法は、母集団分布のガウス性・ノイズの大きさ・外れ値の混入率に精度が著しく影響されるのに対して、クロスデータ行列法は、それらに極めて頑健に推測の精度を保証することが分かった。さらに、クロスデータ行列法は、従来の多変量解析法と比べて計算コストを大幅に軽減させることも分かった。以上から、高速で頑健かつ高精度なクロスデータ行列法は、多変量解析の枠組みにおいても有力なノンパラメトリック法となり得るであろう。実際、青嶋と矢田は、赤平とも意見交換を行い、クラスター分析と判別分析におけるクロスデータ行列法による新しいアプローチを考え、その方法論を理論的に纏めた。

なお、本年度に得られた研究成果の一部は、国内外の学会やシンポジウム等で成果発表を行い、国際学術誌に投稿準備中である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 10 件）

① Aoshima, M., Yata, K. A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. Annals of the Institute of Statistical Mathematics,査読有, 2013, 印刷中.
DOI: 10.1007/s10463-013-0435-8

② Akahira, M., Ohyauchi, N., Kawai, S. A higher order approximation to a percentage point of the distribution of a noncentral t-statistic without the normality assumption. Commun. Statist.-Simula. 査読有, 42, 2013, pp. 2086-2105.
DOI: 10.1080/03610918.2012.695841

〔学会発表〕（計 14 件）

① Yata, K. Effective Classifiers for High-Dimensional Data. Workshop on Statistics for High-Dimensional and Dependent Data, 2014 年 3 月 21 日, National Taiwan University （中華民国）.

② 赤平 昌文. Asymptotic comparison of the MLE and MCLE of a natural parameter up to the second order for a truncated exponential family of distributions. 日本数学会 2014 年度年会, 2014 年 3 月 17 日, 学習院大学（東京都）.

③ Aoshima, M. Effective Methodologies for High-Dimensional Data and their Applications. STOR Colloquium, 2013 年 7 月 15 日, University of North Carolina （アメリカ合衆国）.

④ Aoshima, M. Effective Classification for High-Dimension, Non-Gaussian Data. The 2nd IMS Asia Pacific Rim Meeting, 2012 年 7 月 3 日, つくば国際会議場（茨城県）.

⑤ Aoshima, M. Discussion on Professor Shelemyahu Zacks' Talk. The Sixth International Workshop on Applied Probability, 2012 年 6 月 13 日, Inbal Hotel Jerusalem （イスラエル国）.

〔その他〕

ホームページ等
http://www.math.tsukuba.ac.jp/~aoshima-lab/

６．研究組織

(1)研究代表者

青嶋 誠 (AOSHIMA, Makoto)
筑波大学・数理物質系・教授
研究者番号：90246679

(2)研究分担者

矢田 和善 (YATA, Kazuyoshi)
筑波大学・数理物質系・助教
研究者番号： 90585803

赤平 昌文 (AKAHIRA, Masafumi)
筑波大学・ ・名誉教授
研究者番号：70017424