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

平成　２９　年　　６　月　　９　日現在

機関番号：　１２１０２
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６５４００１０

研究課題名（和文）ビッグデータの統計学：理論の開拓と3Vへの挑戦

研究課題名（英文）Statistics for Big Data: Development of Theories and Tackling the 3Vs

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）　　2,700,000円

研究成果の概要（和文）：本研究は、ビッグデータの統計理論を世界に先駆けて開拓することを目指したものである。我々は、異常値や欠損値が混入する非正則で非ガウスなビッグデータに対して、潜在構造分析の新しい理論と方法論を開発した。それらは、低計算コストで安定した高い精度を保証するためのものである。研究成果は次の通りである。(1) 多様性をもつ大規模データの非正則推定論の開拓。(2) 高速かつ高精度な潜在構造分析の開拓。(3) 異常値・欠損値に頑健な潜在構造分析の開拓。

研究成果の概要（英文）：In this research project, we aim to pioneer new statistical theories for big data, ahead of the world. We have developed new theories and methodologies in latent structural analysis for big data: irregular and non-Gaussian data contaminated with outliers and missing values. New theories and methodologies guarantee stable and high accuracy at low computational cost. The findings of this research project are as follows: (1) Developments of the irregular inference theory for big data with diversity. (2) Developments of high-speed and highly accurate latent structural analysis for big data. (3) Pioneering latent structural analysis robust against outliers and missing values.

研究分野：　統計科学

キーワード：　ビッグデータ　潜在構造分析　異常値　欠損値　非正則推定論

１．研究開始当初の背景

ここ数年の間に、ビッグデータ解析の必要性と重要性は、世界中で広く認識されたように思われる。ビッグデータの特性は、大規模(Volume)・多様性(Variety)・高頻度(Velocity)のいわゆる 3V である。これら 3 つの特性ゆえに、ビッグデータの解析は従来の統計学で対処できない様々な問題が発生する。しかしながら、ビッグデータに対する統計理論は、未だ開拓されていないのが現状である。そのため、従来の統計学の方法論を組合せて、理論的に破綻していることに気付かないまま、間違った解析をしている事例が多く見られる。例えば、3V の特性の一つをもつ高次元データの解析に、標本共分散行列の固有値や固有ベクトルを使用しているものが、その例である。これらは理論的に不一致性が証明されている (Yata and Aoshima, 2009, Comm. Statist. Theory Methods)。研究代表者の青嶋と研究分担者の矢田は、高次元データに対して最新の理論と方法論を構築し、p>>n 問題を解決する一定の成果をあげた (日本統計学会研究業績賞(2012), Abraham Wald Prize in Sequential Analysis (2012), 共に共同受賞)。青嶋と矢田の研究成果は、高次元データが独立同分布(IID)標本で得られる各種推測に、高速かつ高精度な処理を可能にした。これは、高頻度に発生するビッグデータの扱いに自ずと要求される高速処理法を開発する際のヒントになる。ビッグデータの文脈では、データは多様性をもち IID が成立しないので、大規模データの非正則推定論を開拓する必要がある。研究分担者の赤平は、非正則推定論の世界的大家である（文部科学大臣表彰科学技術賞(2013)）。以上の学術的背景のもとで、本研究は、研究組織 3 名の最先端の理論研究を結集させ、大規模・多様性・高頻度の 3V に真っ向から挑戦し、ビッグデータの統計理論を世界に先駆けて開拓するものである。

２．研究の目的

ビッグデータの特性は、大規模(Volume)・多様性(Variety)・高頻度(Velocity)のいわゆる 3V である。これら 3 つの特性ゆえに、ビッグデータの解析は、従来の統計学で対処できない様々な問題が発生する。しかしながら、ビッグデータに対する統計理論は、未だ開拓されていないのが現状である。本研究は、次の 3 つ目的を具体的に掲げ、3V に真っ向から挑戦し、ビッグデータの統計理論を世界に先駆けて開拓する。

(1) 多様性をもつ大規模データの非正則推定論の開拓。
(2) 高速かつ高精度な潜在構造分析の開拓。
(3) 異常値・欠損値に頑健な潜在構造分析の開拓。

３．研究の方法

研究目的の(1)について、多様性をもつ大規模データに対応するために、IID の枠組みを外し、データ空間が膨張することを考慮した漸近理論を構築する。これは、IID の枠組みのもとで青嶋と矢田が構築した高次元小標本漸近理論を、多様性の観点から拡張することに対応する。青嶋と矢田の漸近理論は標本数が有限個のもとで構築されており、たった１つの高次元データでも成立する。そこで、青嶋と矢田は、高次元データをデータ空間に対応させて、データ空間を膨張させる新たな漸近理論を構築する。データの潜在空間を覆う巨大なノイズ空間を精確に捉えるために、青嶋と矢田によって発見されたデータの幾何学的表現を IID の枠組みを外して再考し、ノイズ空間の漸近的な挙動を幾何学的表現で捉える。青嶋と矢田と赤平は、多様性をもつ大規模データの潜在空間を浮き彫りにするための非正則推定論を研究する。

研究目的の(2)について、データの潜在空間を探索する手法として、伝統的に主成分分析(PCA)が知られる。しかしながら、大規模データに対しては、潜在空間が巨大なノイズに覆われるために、PCA は不一致性をもつことが理論的に知られている。従来型の PCA に替わる新しい PCA として、青嶋と矢田はクロスデータ行列法を開発した。クロスデータ行列法は、高次元データが IID 標本として得られる場合に、潜在空間に一致性を保証する推定を与えるための高速かつ高精度なノンパラメトリック手法である。青嶋と赤平は、多様性をもつ大規模データに対応させるために、研究目的(1)で開拓する非正則推定論に基づいて IID の枠組みを外す。青嶋と矢田は、クロスデータ行列法の考え方を拡張して、膨張するデータ空間の巨大なノイズを除去し、多様性をもつ大規模データの潜在空間に対して、高速かつ高精度な潜在構造分析の開拓に挑む。

研究目的の(3)について、ビッグデータにおいては、大量に発生する異常値の検出や欠損値の補填は、理論面からも計算コストの面からも問題が生じる。本研究は、異常値や欠損値は最初から取り込んで考え、非ガウス分布に対応する潜在構造分析の開拓に挑む。青嶋と矢田は、異常値の頻度や欠損値の割合から異常値や欠損値の構造を抽出し、それらに柔軟かつ高速に対応できるように、(2)で応用したクロスデータ行列法の理論を深める。青嶋と赤平は、非正則な分布における潜在構造分析の推測の精度を、非正則推定論を用いて精密に計算する。これらの研究により、異常値・欠損値が混入したビッグデータの潜在構造分析に、高速かつ高精度な処理を提案する。

得られた結果を取り纏め、国内外の学会やシンポジウムで成果の発表を行い、国際学術雑誌に投稿する。

４．研究成果

(1) ビッグデータは、大規模・多様性・高頻

度の特性をもつ。これらの特性ゆえに、ビッグデータの解析には、従来の統計学では対処できない様々な問題が発生する。多様性をもつ大規模データの非正則推定論の開拓に鍵となるのは、IID の枠組みを外し、また、確率過程の枠組みも外し、膨張するデータ空間の漸近理論を如何に構築するかである。青嶋と矢田は、赤平と意見交換を行うことで、青嶋と矢田が一連の共同研究で構築してきた高次元小標本漸近理論を双対空間で展開するというアイディアに至った。高次元小標本漸近理論は、たった一つの高次元データでも成立するので、これを膨張するデータ空間に対応させて漸近理論を展開することを考えた。高次元データの非スパース性が長期記憶に対応するため、従属データを扱う時系列解析に、双対空間から新たなアプローチが開拓できる。その際に、データの潜在空間を覆う巨大なノイズ空間の漸近的な挙動を解析的にどう捉えるかが問題になる。ここでは、データ空間の幾何学的表現によって漸近的な挙動を捉えるというアイディアを思いつき、幾何学的表現を得るための数学的な条件を導き出した。その結果、ビッグデータの実用的な観点から比較的緩い条件のもとで、多様性をもつ大規模データの潜在空間を浮き彫りにできることが分かった。

得られた結果は、論文に纏められ、現在投稿中である。また、成果の一部分について、京都大学数理解析研究所 RIMS 研究集会で発表した。

(2) ビッグデータの潜在構造分析に取り組み、ビッグデータに含まれる潜在構造をモデル化し、巨大なノイズに埋もれた潜在空間に高速かつ高精度な推測法を確立した。従来の潜在構造分析は、スパース性とノイズの正規性や成分間の IID といった、ビッグデータの特徴をまったく捉えてない非現実的な仮定のもとで展開されていた。青嶋と矢田は、こういった非現実的な仮定を一切入れず、多様なビッグデータに十分対応できる柔軟な潜在構造を考えた。まず、ビッグデータを巨大なたった一つの高次元データ行列と解釈し、潜在構造とノイズを柔軟な一つの高次元モデルとして捉えた。潜在構造の非スパース性に着目して、ビッグデータの特異値にパワースパイクモデルを提唱し、理論的な考察と実際のビッグデータ解析でモデルの妥当性を検証した。さらに、青嶋と矢田は、ビッグデータの特異値が従来の方法では推定できないことを証明した。巨大なノイズが推定量の不一致性を招くためである。青嶋と矢田は、赤平と意見交換を密に行い、ノイズ掃き出し法を理論的に拡張した方法を考え、巨大なノイズを除去することで特異値を推定し、非スパースな特異値モデルにもとづく高速かつ高精度な潜在構造の推定法を開発することに成功した。

得られた結果は、学術論文として纏められ、既に出版されている。また、オーストリアで開催された国際学会での招待講演をはじめ、日本統計学会・日本数学会・京都大学数理解析研究所 RIMS 研究集会などで多くの学会発表を行った。

(3) ビッグデータを扱う上で、大量に発生する異常値や欠損値に対して、それらを逐一検出したり補填したりすることは、理論面からも計算コストの面からも実用的ではない。本研究は、異常値や欠損値はビッグデータに当然あるものと考え、非正則な非ガウス分布に対する潜在構造分析を考えた。青嶋と矢田によって提唱されたクロスデータ行列法を、ある方法でビッグデータに適用すると、ビッグデータの潜在構造が浮き彫りになり、その際に異常値を自動検出できることを発見した。青嶋と矢田は、この方法論が理論的に推測の精度を保証するものであることを証明し、論文に纏めて国際学術誌に投稿中である。青嶋と矢田は、赤平と連絡を密に取りながら、非正則な分布における潜在構造分析の推測の精度を非正則推定論を用いて精密に計算し、異常値・欠損値が混入したビッグデータの信号行列を、高速かつ高精度に再構成する方法を開発した。得られた結果は、ビッグデータの巨大なノイズを除去して潜在構造を高い精度で分析するための、統一的な方法論を提供する可能性を秘めている。今後、ビッグデータの解析に、飛躍的な発展が期待できるだろう。さらに、青嶋は、推測の誤差限界について理論的な定式化を行うことにも成功した。これらの成果は、現在、投稿準備中である。

得られた結果について、日本統計学会やトルコで開催された国際学会などで招待講演を行った。さらに、関連する国際シンポジウムを筑波大学で開催し、著名な研究者を国内と海外から招聘して、ビッグデータを扱う様々な分野の研究者から高い関心を集めた。研究成果や問題提起について活発な意見交換がなされ、大変に盛況な国際シンポジウムとなった。

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 Yata, K., Aoshima, M. High-dimensional inference on covariance structures via the extended cross-data -matrix methodology. Journal of Multivariate Analysis, 査読有, 151, 2016, pp. 151-166.
DOI: 10.1016/j.jmva.2016.07.011

② Yata, K., Aoshima, M. Reconstruction of a high-dimensional low-rank matrix. Electronic Journal of Statistics, 査読有, 10, 2016, pp. 895-917.
DOI: 10.1214/16-EJS1128

〔学会発表〕（計 20 件）
① 青嶋　誠．高次元固有空間の推測と高次元統計解析．第 11 回日本統計学会春季集会．2017 年 3 月 5 日．政策研究大学院大学（東京都港区）．
②青嶋　誠．High-dimensional two-sample tests under strongly spiked eigenvalue models．研究集会「大規模統計モデリングと計算統計 III」．2016 年 9 月 27 日．東京大学（東京都目黒区）．
③　矢田和善．Reconstruction of a high-dimensional low-rank matrix．2016 年度統計関連学会連合大会．2016 年 9 月 7 日．金沢大学（石川県金沢市）．
④ Yata, K. Effective Classifiers for High-Dimensional Non-Sparse Data. International Conference on Information Complexity and Statistical Modeling in High Dimensions with Applications. 2016 年 5 月 20 日. Cappadocia (Turkey).
⑤ Aoshima, M. Statistical Methods for Heterogeneous Data. ISNPS Meeting “Biosciences, Medicine, and novel Non-Parametric Methods”. 2015 年 7 月 15 日. Graz (Austria).

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

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・准教授
研究者番号：９０５８５８０３

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