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　　５ 月 ２８ 日現在

機関番号：１２１０２
研究種目：挑戦的研究（萌芽）
研究期間：2017～2018
課題番号：１７Ｋ１９９５６

研究課題名（和文）非スパースモデリングによるビッグデータの新展開

研究課題名（英文）New developments for big data by non-sparse modeling

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）　　4,900,000円

研究成果の概要（和文）：本研究は、スパース性に基づいた学術の体系を大きく見直して、高次元データの非スパース性に着目することで、広汎なビッグデータから高速かつ高精度に最大限の情報を抽出するための新たな技術の開発と、科学技術・産業への革新的展開を目指したものである。次の3つの研究成果を得た。(1)非スパース性の評価基準と潜在構造分析の基礎的方法論の開発。(2)非スパースなノイズ構造をスパース化させるデータ変換法の構築。(3)非スパースモデリング技法の確立とビッグデータ解析の新展開。

研究成果の学術的意義や社会的意義
ビッグデータ解析は、様々な都合から、スパース性を仮定したスパースモデリング(SM)が主流である。しかし実際には、スパース性が成立しないビッグデータも多く、SMは間違った結果を与え得る。本研究は、非スパース性に立脚した非スパースモデリングという、ビッグデータの新たな解析技法を確立する。ビッグデータの本質に合ったモデリング技法を提供することで、学術上の突破口を切り拓くこととなり、波及効果は極めて大きい。非スパースモデリングは、高精度かつ高速で汎用性が非常に高い方法論であるため、科学技術・産業への革新的なインパクトや貢献が期待できる。

研究成果の概要（英文）：In this study, we reviewed the previous academic systems based on sparsity and focused on non-sparsity of high-dimensional data. By using the non-sparsity, we aimed to develop new technology that can extract the maximum information at high speed and with high accuracy from a wide range of big data, and aimed for innovative development in science, technology and industry. We produced the following significant results: (1) Developments of a criteria for non-sparsity and basic methodologies for latent structure analysis. (2) Construction of data transformation from non-sparse noise into sparse noise. (3) Establishment of non-sparse modeling techniques and new development of big data analysis.

研究分野：統計科学

キーワード：非スパースモデリング　スパイクノイズ　ビッグデータ　人工知能　ディープラーニング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

ビッグデータ時代の到来と叫ばれたのは、2012 年。これまで、国内外でビッグデータを解析する研究開発が盛んに進められてきた。特にスパースモデリング（以降、SM と記す）による高次元データの解析事例は数多く報告され、人工知能とも強く結び付いて、ビッグデータ解析に SM を用いることは一つのコンセンサスにもなっている。しかしながら、SM が前提としているスパース（疎）性が、現実の高次元データには必ずしも成立せず、それを知らずに SM を使った結果、ビッグデータが有する豊富な情報を全く抽出できていない解析事例が、しばしば見られることも現実である。

研究代表者の青嶋は、統計学の立場で、理論と方法論の両面から、高次元データを研究してきた。青嶋と矢田（分担者）が共同で創生した高次元統計解析を用いて実際に解析してみると、高次元における多くのビッグデータについて、以下のような事実を突き止めた。

(1) 潜在空間がデータの次元数に依存して膨張する。つまり、非スパースな潜在構造をもつ。
(2) ノイズ空間内の相関が非常に強く、複雑である。つまり、非スパースなスパイクノイズの構造をもつ。

これらは、SM の要となるスパース性の仮定が、高次元のビッグデータには必ずしも成立しないことを意味する。スパース性が成立しないとなると、巨大なスパイクノイズが残留するために、単純な SM の使用は、本質的な情報を有する非スパース潜在構造を破壊することになる。

以上から、高次元のビッグデータが本質的に抱える上記(1)と(2)の問題に対応できるような、新たなモデリングの技法と統計解析法の開発が急務と考え、科学技術・産業への革新的な展開をもたらすべく、非スパースモデリング技法によるビッグデータの新展開を構想するに至った。

２．研究の目的

本研究は、これまでのスパース性に基づいた学術の体系を大きく見直して、高次元データの非スパース（稠密）性に着目することで、広汎なビッグデータから高速かつ高精度に最大限の情報を抽出する新たな技術の開発と、科学技術・産業への革新的展開を目指すものである。

次の 3 つを研究目的とする。

(1) 非スパース性の評価基準と潜在構造分析の基礎的方法論の開発
(2) 非スパースなノイズ構造をスパース化するデータ変換法の構築
(3) 非スパースモデリング技法の確立とビッグデータ解析の新展開

３．研究の方法

研究目的(1)について、潜在空間とノイズ空間からなるデータ空間において、ノイズの完全除去と潜在情報の完全抽出が同時に可能となるようなモデルを考える。SM の精度は、標本数が小さい場合、スパース性の仮定に強く依存する。このことから、スパース性と非スパース性の境界は標本数に依存すると考えられる。高次元のビッグデータを理論的に扱うための道具として、ランダム行列理論はスパース性を前提としているために適当でない。ここでは、非スパース性に対処できる新しい高次元漸近理論を構築し、スパース性と非スパース性の境界を標本数と次元数と S/N の観点から導き出し、非スパース性の評価基準を与える。さらに、非スパースな潜在構造について、低ランク性をもとにデータの特異値と非スパース性の評価基準を検討して、潜在構造のランクを推定する。高次元における巨大なノイズに頑健で、計算コストの低い高次元主成分分析として、青嶋と矢田が開発したクロスデータ行列法がある。これにランクの推定を導入して、膨大な潜在情報を余すことなく高速に抽出する。

研究目的(2)について、非スパースなノイズの処理に、青嶋と矢田が開発したノイズ掃き出し法を再考する。ノイズ掃き出し法は、高次元小標本空間の双対空間における幾何学的表現を用いて高次元における巨大なノイズの半径を見積もり、ノイズを掃き出して潜在構造を推定する方法論である。本研究では、これを、ノイズ構造の推定に応用することを考える。非スパースノイズは、スパイクノイズと非スパイクノイズからなるので、非スパイクノイズの幾何学的表現を理論的に導出して、その大きさを見積もることで、非スパイクノイズの影響を掃き出してスパイクノイズ構造を浮き彫りにする方法を試みる。さらに、スパイクノイズを無効化するような空間を探索し、その空間にデータを射影する変換を考える。この変換を施すことで、巨大なスパイクノイズを取り除いて、スパース化されたノイズをもつビッグデータを構成する。

研究目的(3)について、目的(2)のデータ変換法による誤差の精密な分布を導出し、スパース化されたノイズをもつ縮小データの精密なモデルを与える。目的(1)で開発する非スパースな潜在構造の分析法を縮小データの精密なモデルに適用し、非スパースな潜在構造とノイズ構造をもつビッグデータにモデリング技法を確立する。非スパースな潜在構造とノイズ構造の同時探索は、困難を極めると予想される。最近、プリンストン大学の Fan 教授らは、青嶋と矢田が開発したノイズ掃き出し法を応用し、単純な非スパースモデルにおける非スパース共分散構造推定を考えた。これを、一般の非スパースモデルに拡張することは、潜在構造とノイズ構造の同時探索に繋がるものと考えられる。非スパースモデリング技法を使えば、高次元におけるビッグデータの非スパースノイズが除去され潜在情報が浮き彫りになり、高精度な推定・検定、判別分析、クラスター分析等が可能になる。例えば、目的(1)のランク推定でクラスター数と潜在構造を同時推定し、その上で(2)の変換を施せば、クラスター構造が浮き彫りになる。非スパースモデリング技法は、高精度かつ高速で汎用性が非常に高い方法論であるため、人工知能におけ

る特徴量の抽出にも、非スパースモデリングを用いた高速化が期待できる。

４．研究成果

本研究は、平成 29 年度と 30 年度の 2 年間で計画され、各年度の研究成果は次の通りである。

(1) 初年度に当たる平成 29 年度は、潜在空間とノイズ空間からなるデータ空間において、非スパースなノイズ構造をスパース化するデータ変換法を構築した。非スパースなノイズは、スパイクノイズと非スパイクノイズからなる。青嶋と矢田が開発したノイズ掃き出し法のアイディアを応用して、非スパイクノイズの幾何学的表現を理論的に導出し、その大きさを見積もることで、非スパイクノイズの影響を掃き出してスパイクノイズ構造を浮き彫りにすることに成功した。青嶋と矢田と石井は、スパイクノイズの構造解析を行い、スパイクノイズを無効化するような空間を探索し、その空間にデータを射影する変換を考案した。この変換を施すことで、巨大なスパイクノイズを取り除いて、スパース化されたノイズをもつビッグデータを構成することが可能となる。世の中に広く普及している SM は、標本数が十分ではない場合、データにスパース性の仮定が崩れると精度が酷く悪くなる。実際、現実の高次元データは必ずしもスパース性が成り立たず、ビッグデータが有する豊富な情報を抽出できていない事例がしばしば見られる。本研究で開発したデータ変換法は、この問題に根本的な解決を与えるものである。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。青嶋は、台湾で開催された国際学会での招待講演や、日本統計学会賞受賞者記念講演など、国内外で多くの招待講演を行い、大きな反響を呼んでいる。

(2) 最終年度に当たる平成 30 年度は、非スパースモデリング技法の確立に取り組んだ。青嶋と矢田と赤平は、前年度までの研究によって、ビッグデータの潜在構造とノイズ構造について、非スパース性の評価基準を与え、非スパースなノイズ構造をスパース化するデータ変換法を開発した。青嶋と矢田は、一般に、ビッグデータのノイズが非スパース構造をもつ場合、データ変換による前処理を行わないと、潜在構造分析における様々な統計的推測に漸近正規性が成立しないことを証明した。データ変換を施すことで、非スパースノイズが除去されて潜在情報が浮き彫りになり、潜在構造の非スパース性を利用した非スパースモデリング技法を確立するに至った。さらに、青嶋と矢田と石井は、非スパースなノイズが、ある特殊な構造をもつ場合には、潜在構造とノイズ構造を同時に解析することで、データ変換よりも優れた性能をもつ非スパースモデリング技法が構築できることを示した。非スパースモデリング技法は、高精度かつ高速で汎用性が非常に高く、特徴量の抽出にも高速化が期待できる。本研究の成果は世界的に注目され、多数の招待講演を行った。特に、青嶋は台湾 Academia Sinica で開催された国際学会で基調講演を行い、矢田はハンガリーで開催された IWAP の国際学会で招待講演、石井は日本数学会年会で特別講演を行った。本研究の成果の一部分は、青嶋・矢田による著書「高次元の統計学」（共立出版）の一部分になっている。

５．主な発表論文等

〔雑誌論文〕（計 9 件）

① Ishii, A., Yata, K., Aoshima, M. Equality tests of high-dimensional covariance matrices under the strongly spiked eigenvalue model. Journal of Statistical Planning and Inference, 査読有, 202, 2019, pp. 99-111.
DOI:10.1016/j.jspi.2019.02.002

② 青嶋　誠. 日本統計学会賞受賞者特別寄稿論文：高次元統計解析: 理論と方法論の新しい展開. 日本統計学会誌, 査読有, 48, 2018, pp. 89-111. http://www.terrapub.co.jp/journals/jjssj/pdf/4801/48010089.pdf

〔学会発表〕（計 18 件）

① 石井　晶. 強スパイク固有値モデルにおける高次元統計的推測. 日本数学会 2019 年度年会, 2019.

② Aoshima, M. High-Dimensional Statistical Analysis: Non-Sparse Modeling, Geometric Representations and New PCAs. 2018 Workshop on High-Dimensional Statistical Analysis, 2018.

③ Yata, K. Inference on high-dimensional mean vectors under the strongly spiked eigenvalue model. The Ninth International Workshop on Applied Probability, 2018.

④ Aoshima, M. High-dimensional statistical analysis under spiked models. The Fourth Conference of the International Society for Nonparametric Statistics, 2018.

⑤ Aoshima, M. High-dimensional Statistical Analysis for the SSE Model. A Symposium on Complex Data Analysis 2017, 2017.

〔図書〕（計 1 件）

青嶋　誠, 矢田和善. 共立出版. 高次元の統計学. 2019. 120 ページ.

〔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)研究分担者

研究分担者氏名： 矢田　和善
ローマ字氏名： (YATA, Kazuyoshi)
所属研究機関名： 筑波大学
部局名： 数理物質系
職名： 准教授
研究者番号（8 桁）：90585803

研究分担者氏名： 石井　晶
ローマ字氏名： (ISHII, Aki)
所属研究機関名： 東京理科大学
部局名： 理工学部
職名： 助教
研究者番号（8 桁）：20801161

研究分担者氏名： 赤平　昌文
ローマ字氏名： (AKAHIRA, Masafumi)
所属研究機関名： 筑波大学
部局名： 数理物質系
職名： 名誉教授
研究者番号（8 桁）：70017424