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　２ 年　 ５ 月 １５ 日現在

機関番号：１２１０２
研究種目：基盤研究(A)（一般）
研究期間：2015～2019
課題番号：１５Ｈ０１６７８

研究課題名（和文）大規模複雑データの理論と方法論の総合的研究

研究課題名（英文）Theories and Methodologies for Large Complex Data

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）　　33,300,000円

研究成果の概要（和文）：高次元データのノイズを解析すると、高次元球面付近でのノイズの漸近的な振る舞いについて、高次元中心極限定理が成立することを証明した。高次元における統計量の漸近分布を導き、データ数が少ない状況でも推測の精度を保証するための、高次元小標本漸近理論を開拓した。巨大なノイズを自動除去するデータ変換を考案し、非スパースな大規模複雑データにも高次元中心極限定理を成立させる技術を開発して、これを非スパースモデリングと名付けた。これら理論と方法論による新しい推測を体系化し、高次元統計解析を構築した。

研究成果の学術的意義や社会的意義
高次元中心極限定理の成立は、ノイズのスパース性に依存する。ゲノムや金融等の従属データは、高次元ノイズが巨大で非スパースとなり、高次元中心極限定理は成立しない。そこで、非スパースモデリングによって、巨大なノイズを除去して潜在情報の幾何的構造を浮き彫りにし、推測の精度向上と計算コストの大幅削減に繋げている。非スパースモデリングの精度は、高次元小標本漸近理論で保証できる。これら理論と方法論を装備した新しい解析技術が、高次元統計解析である。

研究成果の概要（英文）：By analyzing the noise of high-dimensional data, we proved that the high-dimensional central limit theorem holds for the asymptotic behavior of noise near the high-dimensional sphere. We provided asymptotic distributions of statistics in high dimensions, and developed the high-dimension, low-sample-size asymptotic theory to guarantee the accuracy of inference even when the sample size is low. We devised a data transformation that automatically removes enormous noise, and developed a technique for establishing the high-dimensional central limit theorem even for non-sparse, large complex data. We called the technique "Non-sparse Modeling". We systematized new inferences based on these theories and methodologies, and established the high-dimensional statistical analysis.

研究分野：統計科学

キーワード：高次元データ　データサイエンス　統計数学　ゲノム　マイクロアレイ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 近年、急速に発展した遺伝子解析は、遺伝子ネットワーク等のモデルの構築と評価が益々重要になっている。その実践は、計算機の性能に頼るところが大きく、理論と方法論の研究が十分に追いついていない。ゲノムデータのような数万次元の高次元データに対して、従来の統計学は理論が破綻する。方法論についても、例えば次元圧縮法として知られる主成分分析など、それそのものが次元の呪いを受けて間違った解析結果を出力することが、研究代表者等によって証明された。計算機の発展は一頃と比べ鈍化しており、理論と方法論の構築が急務であった。これは、遺伝子解析だけでなく大規模複雑データを扱う領域全般にわたって共通の要請であり、理論と方法論を総合的に研究することが、学術的に見て推進すべき重要な研究課題であった。

(2) 研究代表者の青嶋は、分担者の矢田との共同研究で、高次元データの固有値・固有ベクトルを双対空間で可視化する技術を開発し、球面集中現象と座標軸集中現象という 2 つの幾何学的表現を発見した。高次元データの固有値に、ランダム行列理論のスパイクモデルは適切ではないことを示し、次元数と共に固有値が発散するパワースパイクモデルを提唱し、サイズが発散する高次元分散共分散行列の固有値の分布を導出した。ノイズに埋もれた潜在空間を推定するための高次元主成分分析として、ノイズ掃き出し法とクロスデータ行列法を開発した。大標本漸近論に替わる高次元小標本漸近論を展開し、高次元データの各種の統計的推測に対して、幾何学的表現による統計量の導出、一致性と漸近正規性の証明、推測の精度保証など、基礎となる理論と方法論の先駆的成果をあげていた。

(3) 青嶋と矢田の研究成果は、母集団分布に正規性を必要とせず、高次元における非正規性と巨大なノイズの扱いを理論と方法論から総合的に研究したものであった。本研究課題は、これを大規模複雑データの理論と方法論に拡大・発展させることを目指した。大規模複雑データの多様性を如何に扱うか。必ずしも数学的枠組みが当てはまらない従属性・非線形性・非正則性に対して、統計数理を如何に整備するか。大規模複雑データの解析に、推測の精度を如何に保証するか。これらが、研究開始当初の問いであった。

２．研究の目的

本研究は、大規模複雑データの統計数理を理論と方法論から総合的に研究し、モデルの構築と評価を担うモデリング技法の開発を目的とする。大規模複雑データの統計数理は十分に開拓されておらず、学術的に見て推進すべき重要な研究課題である。個別のテーマを世界的レベルでリードしてきた研究者達が本研究課題のもと一堂に会し、連携・融合・発展することで、科学技術・社会・経済・産業の要請に多大な貢献をもたらすことを目指す。次の 3 つを研究テーマとする。

(1) 高次元における最適性理論と精度保証付きモデリング
(2) 高次元における非線形構造と従属構造の統計数理
(3) 大規模複雑データの非正則推定論と相補的モデリング

３．研究の方法

4 つの研究班を編成する。高次元漸近論グループ、機械学習グループ、時系列解析グループ、計算機統計解析グループ。これら 4 つの研究班が、研究目的に掲げた 3 つの研究テーマについて連携して研究を推進する。その際に、若手研究者も巻き込み、人材育成を推進する。各研究班で毎年シンポジウムを開催し、研究成果や問題提起について活発な意見交換の場とする。研究代表者は研究全体を総括し、各研究班の研究状態と問題の所在を明確に把握し、解決に向けて対策を講ずる。海外の先端研究者とも連携し、随時、招聘して協業を図り、世界トップレベルの先端研究を推進する。2 年目と 4 年目に国際シンポジウムを開催し、さらに最終年度には、本研究課題の成果の総括と将来展望を討論するための国際シンポジウムを開催する。本研究課題で得られる成果は、国内外の学会で広く発表し、論文・図書を出版して、社会に向けて発信する。

４．研究成果

(1)　2015 年度において、青嶋と矢田は、高次元空間のクラス識別について、高次元データの巨大なノイズと潜在空間の大きさを理論的に評価し、検定統計量に漸近正規性が成立するための条件を導いた。この条件を満たさない場合、ノイズ掃き出し法を使った高次元データの変換を考え、漸近正規性を有するクラスで最適性理論を展開した。井元は、ゲノムビッグデータの解析として、薬剤感受性の予測、HLA 領域の解析、体細胞変異の同定に関して新たな方法を構築した。鈴木は、構造的正則化手法を効率的に計算するための確率的最適化手法について理論的な結果

を強め、低ランクテンソル推定においてベイズ推定量がミニマックスレートを達成すること、及び、非線形ノンパラメトリックモデルに拡張できることを示した。蛭川は、高次元局所定常時系列因子モデルの漸近理論を考え、平均 2 乗誤差と漸近正規性を導き、金融データや地震波データへの応用を考えた。関連するシンポジウムを東京大学・富山県民会館・筑波大学・東京工業大学で開催した。大規模複雑データを扱う様々な分野から多くの参加者が集まり、問題提起など活発な意見交換がなされ、今後の共同研究が生まれるなど、実り多いものであった。

(2) 2016 年度において、青嶋と矢田は、高次元特有のデータ構造の膨張という問題を解決するために、推測の最適性をもった識別関数のクラスに高次元データを変換する方法を考え、小標本でも高精度を保証する精度保証付きモデリングを構築した。植木は、ゲノムデータにおける連鎖不平衡に起因する相関構造の予測についてアルゴリズムを構築し、さらに、隠れた遺伝要因を双方向グラフによって検出する統計手法を開発した。金森は、離散ビッグデータに対する統計的手法を開発して統計的性質を理論的に解析し、これまで深く考察されてこなかった離散データに対するロバスト推定に応用した。星野は、コルチンモデル確率分割族について、周辺分布を求める方法を開発し、疎な分割表モデルを用いて匿名データを作成する理論を構築した。関連するシンポジウムを久留米シティプラザ・筑波大学・金沢大学・名古屋大学で開催した。機械学習分野から大変多くの参加があり、研究成果や問題提起など活発な情報交換と意見交換の場となった。

(3) 2017 年度において、青嶋と矢田は、高次元混合データの幾何学的一致性というクラスターの最適配置を考え、次元数の増加とともに識別性能が向上するクラスタリング手法を開発した。また、サポートベクターマシンの高次元空間における巨大なバイアスを理論的に解明し、そのバイアスを補正する手法を与え、さらに、高次元データの非線形構造とカーネルの性能に関係性を見い出した。竹之内は、大規模な離散確率モデルのパラメータを効率的に推定すべく、局所性を用いた正規化項の計算を必要としない手法を提案し、一致性と効率性を示した。松井は、経時観測される大規模複雑データを分析するためのモデリング手法を開発し、医学や植物などのデータ解析へ応用した。関連するシンポジウムを公立はこだて未来大学・新潟大学・筑波大学・滋賀大学で開催した。生物分野から大変多くの参加があり、研究成果や問題提起など活発な情報交換と意見交換の場となった。

(4) 2018 年度において、青嶋と矢田は、高次元スパース PCA を考え、正則化パラメータを自動で決定し、固有ベクトルの推定に一致性をもつ新たなスパース PCA を開発した。また、高次元データの非線形構造をカーネル法で捉え、カーネルの最適選択を研究した。柳原は、高次元重回帰モデルの一般化リッジ回帰について、GCV 規準が最小となるリッジパラメータを陽な形で与えた。また、目的変数が高次元の多変量回帰モデルの変数選択について、真の分布が非正規分布のとき高次元大標本漸近論の枠組みで一致性を保証する一般化 GCp 規準を提案した。小森は、大規模な医療データ解析への応用を考え、統計解析アルゴリズムの開発を行った。星野は、大規模複雑データの生成構造を、平均と分散が別母数の一般化多項分布で捉え、性質のよいサンプリングアルゴリズムを開発し、プライバシー保護の研究分野に応用した。関連するシンポジウムを筑波大学・成蹊大学・金沢大学・広島大学で開催した。特に、筑波大学で開催した国際シンポジウムは、国内外の先端研究者が集い、先端研究に関する討論の場となった。

(5) 2019 年度において、青嶋と矢田は、高次元の従属標本を膨張する巨大な１つのデータと捉え、双対空間上で高次元小標本漸近論を展開することで巨大なノイズを除去するデータ変換法を考え、精度保証付きモデリングを構築した。蛭川は、長期にわたって観測される大規模時系列データについて、時間とともに変化するスペクトル構造を局所定常残差の緩やかな爆発モデルで捉え、その漸近理論を構築し、バブル期の始まりと終焉の検出に応用した。金森は、大規模データの非一様なデータ構造を適切に扱うために、マルチドメイン環境から得られるデータに対して転移学習アルゴリズムを提案し、その数理的基盤を与えることで、非一様データの利活用に信頼性を向上させた。青嶋は、大規模複雑データの非正則性を分類し、本研究課題で得られた戦略を相補的に活用するモデリングを整備し、矢田と廣瀬と連携し、実用化に向けた高速処理アルゴリズムを開発した。関連するシンポジウムを、九州大学・新潟大学・東京工業大学・秋田大学で開催し、多方面からの発表と意見交換があった。さらに、本研究課題の最終年度として、国際シンポジウムをつくば国際会議場で開催し、国内外から招聘した先端研究者と参加者たちが本研究課題の将来展望について活発な討論を行い、大変に有益で好評な締め括りの場となった。

なお、本研究課題における成果の一部分を纏め、社会に向けて啓蒙書①を出版した。

＜引用文献＞

① 青嶋 誠、矢田和善：「高次元の統計学」共立出版、2019 年

５．主な発表論文等

〔雑誌論文〕　計42件（うち査読付論文　38件／うち国際共著　4件／うちオープンアクセス　31件）

１．著者名 Aoshima Makoto、Yata Kazuyoshi	４．巻 21
２．論文標題 High-dimensional quadratic classifiers in non-sparse settings	５．発行年 2019年
３．雑誌名 Methodology and Computing in Applied Probability	６．最初と最後の頁 663～682
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11009-018-9646-z	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 Aoshima Makoto、Yata Kazuyoshi	４．巻 71
２．論文標題 Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models	５．発行年 2019年
３．雑誌名 Annals of the Institute of Statistical Mathematics	６．最初と最後の頁 473～503
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10463-018-0655-z	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 Yata Kazuyoshi、Aoshima Makoto	４．巻 －
２．論文標題 Geometric consistency of principal component scores for high‐dimensional mixture models and its application	５．発行年 2019年
３．雑誌名 Scandinavian Journal of Statistics	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） 10.1111/sjos.12432	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 Aoshima Makoto、Yata Kazuyoshi	４．巻 23
２．論文標題 Two-sample tests for high-dimension, strongly spiked eigenvalue models	５．発行年 2018年
３．雑誌名 Statistica Sinica	６．最初と最後の頁 43-62
掲載論文のDOI（デジタルオブジェクト識別子） 10.5705/ss.202016.0063	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 青嶋　誠	4．巻 48
2．論文標題 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	5．発行年 2018年
3．雑誌名 日本統計学会誌	6．最初と最後の頁 89-111
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y., Marron, J. S.	4．巻 60
2．論文標題 A survey of high dimension low sample size asymptotics	5．発行年 2018年
3．雑誌名 Australian & New Zealand Journal of Statistics	6．最初と最後の頁 4-19
掲載論文のDOI（デジタルオブジェクト識別子） 10.1111/anzs.12212	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 Nakayama, Y., Yata, K., Aoshima, M.	4．巻 191
2．論文標題 Support vector machine and its bias correction in high-dimension, low-sample-size settings	5．発行年 2017年
3．雑誌名 Journal of Statistical Planning and Inference	6．最初と最後の頁 88-100
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jspi.2017.05.005	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 Yata, K., Aoshima, M.	4．巻 151
2．論文標題 High-dimensional inference on covariance structures via the extended cross-data-matrix methodology	5．発行年 2016年
3．雑誌名 Journal of Multivariate Analysis	6．最初と最後の頁 151-166
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2016.07.011	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 Aoshima, M., Yata, K.	4．巻 17
2．論文標題 Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions	5．発行年 2015年
3．雑誌名 Methodology and Computing in Applied Probability	6．最初と最後の頁 419-439
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11009-013-9370-7	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74件（うち招待講演　44件／うち国際学会　37件）

1．発表者名 Aoshima Makoto
2．発表標題 High-Dimensional Statistical Analysis: Non-Sparsity, Strongly Spiked Noise and HDLSS (Keynote Speech)
3．学会等名 The 7th International Workshop in Sequential Methodologies（招待講演）（国際学会）
4．発表年 2019年

1．発表者名 Aoshima Makoto
2．発表標題 New Techniques in High-Dimensional Statistical Analysis: SSE vs. NSSE and Data Transformation (Keynote Speech)
3．学会等名 2018 Workshop on High-Dimensional Statistical Analysis（招待講演）（国際学会）
4．発表年 2018年

1．発表者名 Aoshima Makoto
2．発表標題 High-dimensional statistical analysis: Spiked models and data transformation (Special Invited Lecture)
3．学会等名 The 2nd International Conference on Econometrics and Statistics（招待講演）（国際学会）
4．発表年 2018年

１．発表者名
青嶋　誠

２．発表標題
高次元統計解析：理論・方法論とその周辺（日本統計学会賞受賞者記念講演）

３．学会等名
2017年度統計関連学会連合大会（招待講演）

４．発表年
2017年

１．発表者名
青嶋　誠

２．発表標題
高次元の統計学

３．学会等名
日本数学会2016年度年会市民講演会（招待講演）

４．発表年
2016年

１．発表者名
Aoshima, M.

２．発表標題
High-Dimensional Quadratic Classifiers in Non-Sparse Settings under Heteroscedasticity (Plenary Lecture)

３．学会等名
ISNPS Meeting “Biosciences, Medicine, and novel Non-Parametric Methods”（招待講演）（国際学会）

４．発表年
2015年

〔図書〕　計6件

１．著者名	４．発行年
青嶋 誠、矢田 和善	2019年
２．出版社	５．総ページ数
共立出版	120
３．書名	
高次元の統計学	

〔産業財産権〕

〔その他〕

青嶋研究室ホームページ
http://www.math.tsukuba.ac.jp/~aoshima-lab/jp/
青嶋研究室ホームページ　科研費基盤研究(A) シンポジウム
http://www.math.tsukuba.ac.jp/~aoshima-lab/jp/kiban_A.html

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
研究分担者	矢田　和善 (YATA Kazuyoshi) (90585803)	筑波大学・数理物質系・准教授 (12102)	
研究分担者	蛭川　潤一 (HIRUKAWA Junichi) (10386617)	新潟大学・自然科学系・准教授 (13101)	
研究分担者	金森　敬文 (KANAMORI Takafumi) (60334546)	東京工業大学・情報理工学院・教授 (12608)	
研究分担者	星野　伸明 (HOSHINO Nobuaki) (00313627)	金沢大学・経済学経営学系・教授 (13301)	
研究分担者	井元　清哉 (IMOTO Seiya) (10345027)	東京大学・医科学研究所・教授 (12601)	

6．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	鈴木　大慈 (SUZUKI Taiji) (60551372)	東京工業大学・情報理工学研究科・准教授 (12608)	
研究分担者	植木　優夫 (UEKI Masao) (10515860)	久留米大学・付置研究所・准教授 (37104)	
研究分担者	松井　秀俊 (MATSUI Hidetoshi) (90633305)	滋賀大学・データサイエンス学部・准教授 (14201)	
研究分担者	竹之内　高志 (TAKENOUCHI Takashi) (50403340)	公立はこだて未来大学・システム情報科学部・准教授 (20103)	
研究分担者	小森　理 (KOMORI Osamu) (60586379)	成蹊大学・理工学部・准教授 (32629)	
研究分担者	柳原　宏和 (YANAGIHARA Hirokazu) (70342615)	広島大学・理学研究科・教授 (15401)	
研究分担者	廣瀬　慧 (HIROSE Kei) (40609806)	九州大学・マス・フォア・インダストリ研究所・准教授 (17102)	
研究分担者	宇野　力 (UNO Chikara) (20282155)	秋田大学・教育文化学部・教授 (11401)	

6．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	佐藤　美佳 (SATO Mika)		
研究協力者	日野　英逸 (HINO Hideitsu)		
研究協力者	赤平　昌文 (AKAHIRA Masafumi)		
研究協力者	谷口　正信 (TANIGUCHI Masanobu)		
研究協力者	水田　正弘 (MIZUTA Masahiro)		