様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ４ 月 ３０ 日現在

機関番号：１２１０２
研究種目：基盤研究(B)
研究期間：2010～2014
課題番号：２２３０００９４

研究課題名（和文）高次元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 High-Dimensional Data Analysis

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）　　13,800,000円

研究成果の概要（和文）：ノイズ掃き出し法とクロスデータ行列法という2つの高次元PCAを考案した。固有値にパワースパイクモデルを提唱し、固有値・固有ベクトル・主成分スコアの一致推定を与えた。バンド幅信頼領域・2標本問題・判別分析・変数選択・回帰分析・パスウエイ解析等の推測に、先駆的成果をあげた。高次元の特徴量に不偏推定を低コストで与える拡張クロスデータ行列法を考案し、相関係数ベクトルの検定に応用した。多群判別分析を考え、線形判別・幾何学的判別・特徴抽出判別がスパース正則化分類器に優り、誤判別率は高次元で零になることを証明した。研究成果は医学やビッグデータ等多くの領域に応用でき、既存の方法よりも低コスト且つ高精度である。

研究成果の概要（英文）：We created two high-dimensional PCAs which we called the noise-reduction methodology and cross-data-matrix methodology. We proposed a new model, the power spiked model, for eigenvalues and gave consistent estimators of the eigenvalues, eigenvectors and PC scores. We did pioneering work on band-width confidence regions, two-sample problems, classification, variable selection, regression, pathway analysis and so on. We created the extended cross-data-matrix methodology which gives an unbiased estimator at low cost and applied it to the test of correlations. We considered multiclass discriminant analysis and showed that the distance-based classifier, geometric classifier and feature selection by DQDA are superior to sparse regularized classifiers. We proved their misclassification rates go to zero in high-dimension, non-sparse settings. Our work can be applied to many fields, such as medicine and big data, and has much lower computational costs with higher accuracy than existing methods.

研究分野：統計科学

キーワード：高次元データ解析　多変量解析　主成分分析　判別分析　クラスター分析　ノイズ掃き出し法　クロスデータ行列法　マイクロアレイデータ

１．研究開始当初の背景

国内外の研究動向は、工学の分野における機械学習グループの研究開発に目覚しいものがあった。欧米では、統計学グループが機械学習グループと共同研究をすることで、理論と方法論が融合した成果が生まれていた。一方、日本では、こういった研究スタイルが活発であるとは言い難く、旧来型の多変量データ解析法を（不適切であるにも関わらず）高次元データに当てはめようとする研究が根強かった。

現代科学のデータは、もはや旧来型の統計理論や方法論では歯が立たないほど高次元化・複雑化しており、従来の大標本漸近理論に換わる新たな漸近理論の構築が急務である。研究代表者は、先の基盤研究(B)において、数理統計学・確率論・幾何学の立場から、高次元データの数学的特徴を捉えるための基礎研究を行い、高次元データの幾何学的構造の解明と、それに基づく高次元漸近理論の確率論的構成と統計的推定論の基礎を確立していた。ここで基礎理論を構築した 4 名に、新たに機械学習から 2 名を加えて、研究組織を構成した。

２．研究の目的

近年、画像データや遺伝子発現データなど、データの高次元化が目覚しい。データはますます複雑になり、高次元データを扱うための特有な推測技法を開発することが重要である。本研究は、大きく 3 つのテーマからなり、それぞれ次のような研究目的をもつ。

(1) 高次元データの固有空間の統計的推測
- ① 高次元データの幾何学的表現を用いた固有空間の新たな推定法を構築する。
- ② 高次元データの特異値分解に基づいた固有空間の新たな推定法を構築する。
- ③ 高次元データの推測の精度を保証するための標本数決定法を構築する。

(2) 高次元データのパターン認識の研究
- ① 高次元データの幾何学的表現に基づくパターン認識を構築する。
- ② 高次元非線形データの高次モーメントに基づくパターン認識を構築する。

(3) 高次元データのグラフィカルモデル。
- ① 関連変数をグループにした変数選択とモデル選択の理論と方法論を構築する。
- ② 高次元データの部分ネットワークに関するグラフィカルモデルを構築する。

３．研究の方法

研究目的(1)

① 高次元データの幾何学的表現に着目した。青嶋と矢田は、先の基盤研究(B)において、高次元小標本データの固有空間には特有の幾何学的構造が見られることを実験レベルで確認していた。高次元空間の球形度がある閾値を越えると、固有空間は球面から座標軸に大きく挙動を変化させる。この性質を理論的に解明して、ノイズに埋もれた固有空間からノイズを掃き出して推定の精度を格段に改善する。ノイズの大きさを理論的に如何に見積もるかが鍵となる。高次元空間の固有値の精確な推定は、固有ベクトルや主成分スコア等の精確な推定を導くことが可能になる。

② 高次元空間の球形度がある閾値を越えると、固有空間は球面から座標軸に大きく挙動を変化させる。一般の高次元データの球形度はこの閾値を超えることが多く、その場合に座標軸に挙動する固有空間は、高次元小標本において、もはや従来の標本共分散行列では挙動を捉えることができない。そのことは、幾何学的な側面とは別に解析的な側面から、Yata and Aoshima (2009)によって証明されていた。本研究では、標本共分散行列に替わる新たな方法として「クロスデータ行列法」を考案する。これは、データ行列を２分割し、それらを掛け合わせてクロスデータ行列を定義し、その特異値分解に基づいて固有値と固有ベクトルを推定するというアイディアである。ノンパラメトリックな方法なので、高次元空間の球形度の閾値を気にする必要もない。クロスデータ行列法を高次元データの固有空間の推定に応用して、高次元空間における主成分分析を創生する。

③ 推定の有効性と母数・次元数・標本数の関係を明らかにして、標本数の決定を推測に組み込むことを考える。ノイズに埋もれた高次元空間において、如何に推測を考えるかが重要になる。①と②の研究で得られる成果をもとにして、高次元小標本データ空間の幾何学的表現とクロスデータ行列法を用いることで、高次元空間に特有な推定量や検定統計量を構築する。高次元漸近理論を開拓することで、推測の一致性や漸近正規性を証明する。高次元空間を潜在空間とノイズ空間に分離するために、これらを母数と次元数の関数として定義し、潜在空間における推測の精度を保証するための標本数の条件式を導く。

研究目的(2)

① 高次元データの判別分析を考える。いわゆるベイズ最適ルールは定義式に共分散行列の逆行列を含み、その推定が問題になる。例えば、標本共分散行列は逆行列が存在しないため、それに替わる推定を考える。その際、クラスの識別情報をできるだけ取り入れて判別器の性能を上げるためにも、クラス間に共分散行列の同等性は仮定しない。識別精度を理論的に評価し、判別器が識別能力をもつための条件を、標本数・次元数・グループ間距離の関係式で導出する。また、(1)で開発する高次元データの主成分分析を応用してクラスタリング法を考える。その際、高次元データが未知の分布型の混合分布をもつと仮定して、不均一な混成データを処理できる理論と方法論を確立する。

② 機械学習のカーネルマシンとして知られるサポートベクトルマシンや関連ベクトルマシンに上記①の成果を応用する。サポートベクトルマシンは予測に対する事後確率は

計算できない。一方、関連ベクトルマシンはベイズ理論に基づいて事後確率を計算でき、サポートベクトルマシンよりも疎なモデルが得られやすく、高速な予測が可能になる。しかし、関連ベクトルマシンは逆行列の計算量に問題を抱えているので、①で考案する逆行列の算法を応用する。

研究目的(3)

① マイクロアレイデータにおける相似な機能をもった遺伝子など、高次元データにおいて相関が高い変数をグループにして抽出する変数選択を考える。(1)で開発するクロスデータ行列法を拡張して、母集団分布に依らずに不偏で漸近最小な分散をもつ検定統計量を構成する。その際、計算コストも考慮して、最適なアルゴリズムを構築する。モデル選択には、機械学習の lasso を用いた正則化法を考えるとともに、モデル構築とモデル選択の基準を統一化することを意図して、密度関数のベキ変換によって定義されるパワーダイバージェンスによる情報量規準も考える。適切なバイアス補正を行い、新しい規準を与える。

② 遺伝子ネットワークの推定を考える。計算量が膨大であるため、数千という遺伝子を含むネットワーク推定も、現時点の計算機の性能では、十分な解を得ることはできない。そもそも、数千という遺伝子の依存関係を明らかにするには、標本数が十分ではない。そこで、細胞内における何らかの機能に特化した遺伝子たちや、薬剤に反応する遺伝子たちの部分ネットワークの推定を考える。関連変数群の部分ネットワークを、ベイジアンネットワークとノンパラメトリック回帰で推定する。親ノードが子ノードを制御する非線形関数は、基底関数展開法によって構築する。モデルの選択は、事後確率を最大にするグラフを選択するベイズ型情報量規準(BIC)を考える。グラフィカルモデルの構築において、簡潔さ・計算効率・サンプリングのしやすさを追求して研究を推進する。

４．研究成果

研究課題の研究成果を発表し、情報交換する場として、毎年、国内外でワークショップや特別セッションを企画・開催した。各年度の研究成果は次の通りである。

(1) 2010 年度

高次元小標本データの固有空間の幾何学的表現を世界で初めて発見した。データの球形度と２次の無相関性で収束が決まる球面集中と座標軸集中の２つの表現である。幾何学的表現に基づいて「ノイズ掃き出し法」という高次元空間における主成分分析を考案した。また、母集団分布型に依存しないノンパラメトリックな高次元主成分分析として「クロスデータ行列法」を考案した。これらが、高次元空間における固有空間の次元数推定、固有値・固有ベクトル・主成分スコアの推定に、一致性と漸近分布を与えることを理論的に証明した。また、バンド幅信頼領域、2 標本問題、判別分析、変数選択、回帰分析、パスウエイ解析等について、高次元空間で精度を保証するための推測理論の基礎を構築した。その際、従来の統計学の土台となる大標本漸近理論では高次元データの推測理論を支えきれないので、それに替わる高次元漸近理論を展開した。異常値が混在する高次元空間における潜在分布の推定も行い、パワーダイバージェンスと LASSO に基づくモデル選択基準を考えた。

(2) 2011 年度

高次元空間の様々な統計的推測に特徴量の不偏推定量を統一的に与えるための方法論として、「拡張クロスデータ行列法」を考案した。高次元空間の相関係数ベクトルの検定に応用し、各種特徴量の不偏推定量に基づいて検定統計量を構築し、高次元における漸近正規性を証明した。また、拡張クロスデータ行列法によって与えられる不偏推定量が、母集団分布にガウス性の条件を課して得られる先行研究の不偏推定量と、同等の漸近分散を有することを証明した。なお、先行研究の不偏推定量はガウス性の仮定が崩れると不偏性をもたず、分散は有界にさえならない。それに対して、拡張クロスデータ行列法で与えられる不偏推定量は、分布の仮定に依らず、常に不偏性を有し漸近分散を保証する。さらに、計算コストを格段に削減できる。検定統計量の漸近正規性に基づいて、有意水準と検出力を漸近的に目標値に近づけるための標本数を推定して、検定の精度保証を理論的に証明した。

(3) 2012 年度

高次元データのパターン認識について、多クラスの分類問題を考えた。ユークリッド距離に基づく分類器を考案し、高次元において誤判別率がゼロになることを理論的に証明した。さらに、高次元小標本の幾何学的表現に基づく分類器も考案し、クラス間の共分散行列の差異を利用することで、高次元において誤判別率がゼロになることを理論的に証明した。また、相似な機能をもった遺伝子等の相関が高い変数をグループにして抽出する変数選択を考え、相関係数ベクトルの多重検定を構築した。拡張クロスデータ行列法を応用して、ノンパラメトリックな設定で推測の精度を保証するための標本数の決定法を与えた。

研究成果は国内外で高く評価され、次の 3 つの受賞に結びついた。

① 青嶋　誠：筑波大学 Best Faculty Member Award (2013 年 3 月)

② 青嶋　誠・矢田和善：日本統計学会研究業績賞 (2012 年 9 月)

③ 青嶋　誠・矢田和善：Abraham Wald Prize in Sequential Analysis (2012 年 6 月)

(4) 2013 年度

高次元データの固有空間を正しく表現するためにパワースパイクモデルを提唱して、

固有空間の推測を精密に研究した。本研究課題で開発したクロスデータ行列法が従来型主成分分析の性能を遥かに凌ぐことを理論的に証明し且つ数値的に明らかにした。また、遺伝子ネットワークの推定に、計算量の問題を克服するための拡張クロスデータ行列法の応用を研究した。

青嶋は 7 月に米国で開催された Fourth International Workshop in Sequential Methodologies で基調講演を行い、満員の会場から多数の質問を受け、研究成果に高い関心を集めた。また、青嶋は 8 月に香港で開催された The 59th World Statistics Congress に招待され招待講演を行い、矢田は 11 月に開催された韓国統計学会で招待講演を行った。青嶋と矢田の一連の研究成果は、日本統計学会研究業績賞受賞者特別寄稿論文や日本数学会の論説として出版され、さらに特徴ある研究として日米研究インスティチュート(USJI)リサーチレポートにも紹介された。

(5) 2014 年度

高次元データの効率的ネットワーク推定法を確立するために、遺伝子ネットワークを構築するための遺伝子のグループ間検定を考えた。従来のネットワーク推定法は計算量が膨大となり、数千という遺伝子でも現時点の計算機の性能では十分な解が得られず、また、遺伝子の依存関係を明らかにするには標本数が不足する。本研究課題で開発した拡張クロスデータ行列法を用いて、高次元データに対して低い計算コストで検定統計量を作り、有意水準と検出力の両方に精度を保証するネットワーク推定法を考えた。また、高次元空間に高次モーメントを利用したパターン認識法を確立するために、高次元データの非ガウス性と非線形性を高次モーメントによる幾何学的表現で捉え、分類器の族を考えた。これにより、高次元データの分類器を統一的に扱う理論を構築し、最適性を論じることができる。分類器の推定、精度に関する一致性・漸近正規性、変数選択に至るまで、統一的な理論と方法論を考えた。

青嶋は、6 月にスペインで開催された 2nd International Society of NonParametric Statistics に招待され招待講演を行い、3 月に台湾で開催された Workshop on Statistical Methods for Large Complex Data で基調講演を行った。

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　45 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. 66 (2014) 983-1010. DOI: 10.1007/s10463-013-0435-8

② Aoshima, M., Yata, K. Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. Methodology and Computing in Applied Probability. 査読有. 印刷中. 2013. DOI: 10.1007/s11009-013-9370-7

③ 青嶋　誠, 矢田和善. 日本統計学会研究業績賞受賞者特別寄稿論文：高次元データの統計的方法論. 日本統計学会誌. 査読有. 43 (2013)123-150. URL: http://www.terrapub.co.jp/journals/jjssj/pdf/4301/43010123.pdf

④ 青嶋　誠, 矢田和善. 論説：高次元小標本における統計的推測. 数学. 査読有. 65 (2013) 225-247. URL:http://mathsoc.jp/publication/sugaku/index65.pdf

⑤ Yata, K., Aoshima, M. PCA consistency for the power spiked model in high-dimensional settings. Journal of Multivariate Analysis. 査読有. 122 (2013) 334-354. DOI: 10.1016/j.jmva.2013.08.003

⑥ Yata, K., Aoshima, M. Correlation tests for high-dimensional data using extended cross-data-matrix methodology. Journal of Multivariate Analysis. 査読有. 117 (2013) 313-331. DOI: 10.1016/j.jmva.2013.03.007

⑦ Yata, K., Aoshima, M. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. Journal of Multivariate Analysis. 査読有. 105 (2012) 193-215. DOI: 10.1016/j.jmva.2011.09.002

⑧ Yata, K., Aoshima, M. Inference on high-dimensional mean vectors with fewer observations than the dimension. Methodology and Computing in Applied Probability. 査読有. 14 (2012) 459-476. DOI: 10.1007/s11009-011-9233-z

⑨ Aoshima, M., Yata, K. Two-stage procedures for high-dimensional data (Editor's special invited paper). Sequential Analysis. 査読有. 30 (2011) 356-399. DOI: 10.1080/07474946.2011.619088

⑩ Aoshima, M., Yata, K. Asymptotic second-order consistency for two-stage estimation methodologies and its applications. Ann. Inst. Statist. Math. 査読有. 62 (2010) 571-600. DOI: 10.1007/s10463-008-0188-y

⑪ Yata, K., Aoshima, M. Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. Journal of Multivariate Analysis. 査読有. 101 (2010) 2060-2077. DOI: 10.1016/j.jmva.2010.04.006

〔学会発表〕（計 67 件）

① Aoshima, M. High-Dimensional Quadratic Classifiers in Non-Sparse

Settings. Workshop on Statistical Methods for Large Complex Data. March 13, 2015. Kaohsiung (Taiwan) **基調講演**

② 矢田和善. Principal component analysis based clustering for high-dimension, low-sample-size data. **第 9 回日本統計学会春季集会**. 2015 **年** 3 **月** 8 **日**. **明治大学（東京都）招待講演**

③ 青嶋　誠. **高次元データの分類－判別分析とクラスター分析の諸問題と高次元現象**. The Applied Statistics Workshop. 2014 **年** 12 **月** 19 **日**. **東京大学（東京都）招待講演**

④ Yata, K. Quadratic-Type Classifications for High-Dimensional Data. The 3rd IMS Asia Pacific Rim Meeting. July 2, 2014. Taipei (Taiwan) **招待講演**

⑤ Aoshima, M. Quadratic-Type Classifications for Non-Gaussian, High-Dimensional Data. Second Conference of the International Society of NonParametric Statistics. June 13, 2014. Cadiz (Spain) **招待講演**

⑥ Aoshima, M. New PCAs for High-Dimensional Data. Workshop on Statistics for High-Dimensional and Dependent Data. March 21, 2014. Taipei (Taiwan) **招待講演**

⑦ Yata, K. PCA Consistency for High-Dimensional Data under the Power Spiked Model. Korea Statistical Society Semi-Annual Meeting. Nov. 2, 2013. Seoul (Korea) **招待講演**

⑧ Aoshima, M. Effective PCA for High-Dimensional Data and Its Applications. 59th ISI World Statistics Congress. Aug. 27, 2013. Hong Kong Convention and Exhibition Centre (Hong Kong) **招待講演**

⑨ Aoshima, M. Effective Methodologies for High-Dimensional Data. Fourth International Workshop in Sequential Methodologies. July 20, 2013. Georgia (U.S.A.) **基調講演**

⑩ 矢田和善. PCA Consistency for the Power Spiked Model in High-Dimensional Settings. **第 7 回日本統計学会春季集会**. 2013 **年** 3 **月** 3 **日**. **学習院大学（東京都）招待講演**

⑪ 青嶋　誠. **高次元小標本データの統計学(日本統計学会研究業績賞受賞者講演)**. 2012 **年度統計関連学会連合大会**. 2012 **年** 9 **月** 10 **日**. **北海道大学（北海道）招待講演**

⑫ Aoshima, M. Misclassification Rate Adjusted Classifier for Multiclass, High-Dimensional Data. The Sixth International Workshop on Applied Probability. June 14, 2012. Jerusalem (Israel) **招待講演**

⑬ Yata, K. Effective PCA for Large p, Small n Scenario under Generalized Models. The Sixth International Workshop on Applied Probability. June 14, 2012. Jerusalem (Israel) **招待講演**

⑭ Aoshima, M. Effective Methodologies for High-Dimensional Statistical Inference. Joint Meeting of the 2011 Taipei International Statistical Symposium and 7th Conference of the Asian Regional Section of the IASC. Dec. 17, 2011. Taipei (Taiwan) **招待講演**

⑮ Aoshima, M. Effective Classification for High-Dimension, Non-Gaussian Data and Sample Size Determination. Third International Workshop in Sequential Methodologies 2011. June 15, 2011. California (U.S.A.) **招待講演**

⑯ Yata, K. Effective PCA for Large p, Small n Context with Sample Size Determination. Third International Workshop in Sequential Methodologies 2011. June 15, 2011. California (U.S.A.) **招待講演**

⑰ Aoshima, M. Two-stage inference methods for large p, small n scenarios. The Fifth International Workshop in Applied Probability. July 5, 2010. Madrid (Spain) **招待講演**

〔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

イリチュ美佳（SATO-ILIC, Mika）
筑波大学・システム情報系・教授
研究者番号：60269214

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

小池健一（KOIKE, Ken-ichi）
筑波大学・数理物質系・准教授
研究者番号：90260471

大谷内奈穂（OHYAUCHI, Nao）
筑波大学・数理物質系・助教
研究者番号：40375374