科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１８　日現在

機関番号：　１２１０２
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７４００６６

研究課題名（和文）高次元小標本の理論的体系の構築

研究課題名（英文）Constructing theoretical system for high-dimension, low-sample-size data

研究代表者
矢田　和善（YATA, KAZUYOSHI）

筑波大学・数理物質系・助教

研究者番号：９０５８５８０３

交付決定額（研究期間全体）：（直接経費）　　2,500,000円、（間接経費）　　750,000円

研究成果の概要（和文）：　高次元小標本における新しい理論と方法論を構築した．高次元小標本ならではの特徴的な2つの幾何学的構造を発見し，これらの幾何学的構造に基づいて，新しいPCA手法である「ノイズ掃き出し法」を提案した．高次元の推測に現れる各種パラメータに対して，計算コストを著しく削減し，漸近分散が小さな不偏推定量を与える「拡張クロスデータ行列法」という推定法を提案した．高次元データの判別分析について，母集団間のユークリッド距離に基づく判別方式と幾何学的差異を利用した判別方式を提案し，それらが誤判別確率に関して高い精度を保証することを示した．さらに，誤判別確率に対して要求される精度を保証するような判別方式も与えた．

研究成果の概要（英文）：　We proposed statistical theories and methodologies for high-dimension, low-sample-size (HDLSS) data. We showed that HDLSS data have two distinct geometric representations. We proposed the noise-reduction methodology that was brought from the geometric representations. We proposed the extended Cross-data-matrix methodology, which offers an unbiased estimator having small asymptotic variance and low computational cost, for parameters appearing in high-dimensional data analysis. We provided two effective discriminant procedures: a distance-based classifier and a geometric classifier, which can ensure high accuracy in misclassification rates and hold misclassification rates less than a threshold.

研究分野：　数物系科学

科研費の分科・細目：　数学一般（含確率論・統計数学）

キーワード：　高次元小標本データ　高次元漸近理論　PCA　判別分析　クラスター分析　マイクロアレイデータ

１．研究開始当初の背景

近年，情報化の進展に伴い，データの次元数が標本数より遥かに大きな高次元小標本(HDLSS)データが解析対象になる場面が増えてきている．しかしながら，研究開始当初，HDLSS データの理論は完全に整備されているとはいえず，理論の構築が急務になっていた．さらに，従来の統計的方法論が HDLSS に適用できないため，HDLSS 特有の統計的方法論の開発も急務であった．

２．研究の目的

HDLSS 漸近理論を構築し，HDLSS における種々の統計的推測を構築することを目指し，次の３つを研究目的とした．

(1) HDLSS の各種特徴量に対して HDLSS 漸近理論と新たな推測理論の構築．
(2) HDLSS の判別分析とクラスター分析に関して統計的推測の構築．
(3) HDLSS におけるパスウェイ解析と変数択法の理論的構築．

本研究により，HDLSS の統計学に新たに HLDSS 漸近理論を構築し，HDLSS データならではの特色ある統計的推測に，精密な理論と実用的な方法論を体系的に提供できると考えた．

３．研究の方法

(1) まず，Hall et al.(2005, JRSS-B)が発見した HDLSS データの球面集中現象に着目する．この幾何学的構造に着目すれば，平均ベクトルに関する特徴量の存在領域をよりシャープに特定できると考える．この HDLSS 信頼領域を，半径が異なる２つの超球に挟まれた領域として定義し，球面に集中する推定量の挙動を漸近的に捉える．具体的には，球面上での漸近正規性を証明し，平均ベクトルの推定・検定に要求精度を保証するような方法論を与える．さらに，球面上での高次元ノイズを特定することで，主成分分析(PCA)の各種特徴量の推定も考える．

(2) 既存の先行研究では，各母集団の共分散行列に共通性など HDLSS にとって大変に厳しい制約条件に基づいていた．そこで本研究では，研究目的(1)で扱った幾何学的構造に基づき，母集団間での幾何学的差異を利用した柔軟な判別関数を考え，判別精度を向上させる．さらに，共分散行列の逆行列の推定量を単位行列等で代用したユークリッド距離に基づく判別関数も考えて，HDLSS 漸近正規性を証明し，判別精度に関する検討を行う．

一方で，クラスター分析については，分類にとって有効な空間を探索する手法を PCA の HDLSS 漸近理論に基づき提案する．

(3) パスウェイ解析は変数間の偏相関について検定が必要であるが，HDLSS データに対するパスウェイ解析では，変数の組み合わせが膨大になることに注意する．本研究では，この問題に対して，逐次推定論を用いてアプローチする．数回の前処理で有意な変数を大幅に絞り込みができる変数選択法を提案する．その後に多重検定を行うことで，HDLSS のもと第一種の過誤である FWER と検出力を同時に制御する方法論を提案する．

４．研究成果

(1) Hall 等の先行研究で見つけることができなかった特徴的な２つの幾何学的構造を発見し，HDLSS データが元来もつこれらの幾何学的構造に基づいて「ノイズ掃き出し法」を与え，ノイズを除去して固有空間を有効に推定するための PCA 手法を提案した．さらに，その手法に基づく固有値・固有ベクトルと主成分スコアの推定量を与え，HDLSS のもと一致性をもつことを示した．

一方，平均ベクトルに関して，幾何学的表現に基づき高次元球面上の与えられたバンド幅の信頼領域や二標本問題，ノルムの信頼区間について，与えた各推定量・統計量の漸近正規性を示して，要求精度を満足するような推定・検定方式を与えた．さらに，高次元共分散行列の推定・検定，高次元回帰分析，変数選択問題など，HDLSS の統計的推測において推測問題を提示し，それらのオリジナルの理論と方法論を与えた．さらに，マイクロアレイデータを用いた実解析例において，提案手法が有効に機能することも確認できた．

(2) HDLSS における判別分析に関して，研究成果(1)で構築した球面上での HDLSS 漸近理論を拡張し，多母集団にも適用できるようなユークリッド距離に基づく判別超平面を考え，逐次解析の理論と融合させ「Misclassification rate adjusted classifier (MRAC)」を提案した．MRAC により，多母集団の HDLSS データにおける判別分析に，精度保証を与える判別方式を提案した．さらに，共分散行列の幾何学的差異を利用した判別関数も提案し，平均ベクトルだけでなく共分散行列の差異も考慮することで，判別精度の高い二次判別ルールを構築した．

一方で，クラスター分析について，研究成果(1)で導出した HDLSS における幾何学的構造を，混合分布を含むクラスに拡張し，データを幾何的に分類するためのクラスタリング手法を提案した．

(3) Yata and Aoshima (2010, JMVA)で提案したクロスデータ行列法を漸近最適な組み合わせに基づき拡張し，HDLSS における各種パラメータの推定と検定に，計算コストを著しく削減し，漸近分散が小さい不偏推定量を与えるための「拡張クロスデータ行列法(ECDM)」を開発し，先行研究よりも高速かつ高精度に推定量を構築できることを示し

た.

一方で，ECDM に基づき，HDLSS の枠組みで重相関係数の多重検定を考え，事前に設定された精度を保証するような，逐次解析を用いた新しい多重検定法も提案した．さらに，この多重検定法を変数選択に応用することで，パスウェイ解析において，精度を保証した有意な遺伝子群の抽出が可能となり，実際のマイクロアレイデータを用いた実解析例において，提案手法が有効に機能することも確認できた.

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　16 件）

① K. Yata, M. Aoshima : PCA consistency for the power spiked model in high-dimensional settings(査読あり).
J. Mult. Anal., 122, 334-354, 2013.
DOI: 10.1016/j.jmva.2013.08.003

② K. Yata, M. Aoshima : Correlation tests for high-dimensional data using extended cross-data-matrix methodology (査読あり).
J. Mult. Anal., 117, 313-331, 2013.
DOI: 10.1016/j.jmva.2013.03.007

③ M. Aoshima, K. Yata: A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data (査読あり).
Ann. Inst. Statist. Math., in press, 2014.
DOI: 10.1007/s10463-013-0435-8

④ 青嶋　誠, 矢田和善:　高次元データの統計的方法論 (査読あり).
日本統計学会誌, 43, 123-150, 2013.
DOI: なし

⑤ 青嶋　誠, 矢田和善: 高次元小標本における統計的推測 (査読あり).
数学, 65, 225-247, 2013
DOI: なし

⑥ K. Yata, M. Aoshima : Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations (査読あり).
J. Mult. Anal., 105, 193-215, 2012.
DOI: 10.1016/j.jmva.2011.09.002

⑦ M. Aoshima, K. Yata : Two-stage procedures for high-dimensional data (査読あり).
Seq. Anal. (Editor ' s special invited paper), 30, 356-399, 2011.
DOI: 10.1080/07474946.2011.619088

〔学会発表〕（計　34 件）

① K. Yata, M. Aoshima: PCA consistency for high-dimensional data under the power spiked model (Invited talk).
KSS/JSS/CSA International Session in KSS Semi-Annual Meeting, Seoul (Korea), Nov. 2, 2013.

② K. Yata: Asymptotic normality for inference on multi-sample, high-dimensional mean vectors under mild conditions (Invited talk).
Fourth International Workshop in Sequential Methodologies, Georgia (U.S.A.), July 18, 2013.

③ 青嶋　誠, 矢田和善: 高次元小標本データの統計学 (日本統計学会各賞受賞者講演).
統計関連学会連合大会, 北海道大学, 2012 年 9 月 10 日.

④ K. Yata: Effective PCA for large p, small n scenario under generalized models (Invited talk).
Sixth International Workshop on Applied Probability, Jerusalem (Israel), June 14, 2012.

⑤ 矢田和善: 高次元小標本における統計的推測 (特別講演).
日本数学会秋季総合分科会特別講演, 信州大学, 2011 年 9 月 30 日.

⑥ K. Yata: Effective PCA for large p, small n context with sample size

determination (Invited talk).
Third International Workshop in Sequential Methodologies, Stanford (U.S.A.), June 15, 2011.

〔図書〕（計　1 件）

① M. Aoshima, K. Yata: Effective methodologies for statistical inference on microarray studies. In P.E. Spiess (Ed.), Prostate Cancer - From Bench to Bedside, InTech, 2011, pp. 13-32.

〔その他〕
ホームページ等
研究者総覧（筑波大学）：
http://www.trios.tsukuba.ac.jp/researcher/0000000526

６．研究組織
(1)研究代表者
矢田　和善　　（YATA KAZUYOSHI）
筑波大学・数理物質系・助教
研究者番号：90585803