



高次元漸近理論の統一的研究

著者	矢田 和善
発行年	2018
URL	http://hdl.handle.net/2241/00158981

平成 30 年 6 月 16 日現在

機関番号：12102

研究種目：若手研究(B)

研究期間：2014～2017

課題番号：26800078

研究課題名(和文)高次元漸近理論の統一的研究

研究課題名(英文)Asymptotic Studies for High-Dimensional Data

研究代表者

矢田 和善 (YATA, Kazuyoshi)

筑波大学・数理物質系・准教授

研究者番号：90585803

交付決定額(研究期間全体)：(直接経費) 2,800,000円

研究成果の概要(和文)：高次元ノイズ構造を精密に調査し、ノイズに対する2つの固有値モデル「Strongly spiked eigenvalue (SSE) モデル」と「Non-SSE (NSSE) モデル」を提唱した。NSSEモデルのもとでは各種統計量の高次元漸近正規性が主張できるが、より巨大なノイズを含むSSEモデルにおいてはそれら漸近正規性が主張できず、推測の精度保証が困難であった。それに対し、高次元固有空間の推定に着目し、SSEモデルからNSSEモデルへのデータ変換を行うことで新たな高次元統計的推測法を提案した。さらに、精密な高次元ノイズ解析に基づく新たな高次元判別方式やパスウェイ解析法も提供した。

研究成果の概要(英文)：We provided two disjoint models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. It can be noted that, under the SSE model, asymptotic normality of high-dimensional statistics is not valid because it is heavily influenced by strongly spiked eigenvalues. In order to handle several statistical inferences for the SSE model, we propose a data transformation from the SSE model to the NSSE model by estimating the strongly spiked eigenstructures. We create new statistical test procedures by the data transformation. In addition, we provided several high-dimensional discriminant procedures and pathway analysis methods based on the models.

研究分野：数理統計学

キーワード：高次元統計解析 高次元PCA 強スパイクモデル 高次元判別分析 パスウェイ解析

1. 研究開始当初の背景

ゲノム科学などの現代科学の 1 つの特徴は、データがもつ次元数の膨大さにある。そのような高次元データの理論は、まだ完全には整備されているとは言えず、理論とそれに伴う方法論の構築が急務になっていた。特に、ほぼすべての先行研究において、高次元ノイズ空間が小という実際の高次元モデルとは乖離した仮定のもと統計的推測法が開発されており、統計的な精度保証がしばしば困難であった。

2. 研究の目的

多様な高次元データにも対応でき得る高精度な統計的方法論の構築を目指し、次の 4 つを目的とした。

- (1) 高次元 PCA を用いた漸近理論の構築。
- (2) 高次元パスウェイ解析の新たな推測理論の構築
- (3) 高次元データの判別分析とクラスター分析の最適性指標の構築。
- (4) 高次元固有値モデルにおける理論的体系の構築と新たな統計的方法論の開発。

本研究により、巨大なノイズを包含する高次元データにも頑健かつ高精度な統計的方法論を提供できると考えた。

3. 研究の方法

(1) Aoshima and Yata (2011, SA)において、高次元データの統計量は、ある固有値条件のもとで高次元漸近正規性をもつことを証明した。これは数学的に厳密な結果であるが、実際の高次元データには厳しい条件である。本研究は、Yata and Aoshima (2012, JMA)において考案したノイズ掃き出し法と Yata and Aoshima (2013, JMA)で与えた高次元 PCA の理論を用いて、固有値条件に依存しない精密な高次元漸近理論を構築する。

- (2) パスウェイ解析は変数間の検定が必要

になり、高次元データにおいては、変数の組み合わせが膨大になることに注意する。多重検定の第一種の過誤である Family-wise error rate と検出力は、もはや理論的に制御することができない。本研究は、この問題に対して、冗長な変数の組(サブグループ)を大幅に除去することで対応する。Yata and Aoshima (2013, JMA)で開発した拡張クロスデータ行列法を用いることで、変数のサブグループ間の検定統計量を構築し、要求精度を満足するような変数選択法を提案する。

(3) Aoshima and Yata (2011, SA)において、共分散行列の異質性を母集団間のある距離で測り、2 群判別を提案した。本研究は、Aoshima and Yata (2011)を拡張して多群判別を考える。ただし、母集団間の距離には判別精度を向上させる漸近最適な距離を選ぶ。研究目的(1)の高次元漸近理論から導かれる高精度な推測を利用した判別方式と、既存の判別方式について、母集団間の距離を見積もって判別方式の最適性の指標を構築する。高次元クラスター分析について、Yata and Aoshima (2013, JMA)で提唱した Power Spiked モデルのもと、個体を分類する射影空間の推定の一致性について研究し、理論的な保証を与える。

(4) 多様な高次元固有値モデルに対応でき得る理論的体系を構築する。研究(1)から、高次元統計的推測の鍵となるのが、高次元固有値モデルである。巨大な固有値を包含する場合、それに引きずられ、統計的推測が困難になる。それに対し、巨大な固有値を無効化するような空間を探索し、そこにデータを射影することで、高い精度が保証でき得る統計的方法論を提供する。

4. 研究成果

- (1) PCAを用いた高精度な高次元共分散行列

の推定法の開発し，固有値と固有ベクトルの推測に精密な漸近分布を導出した．さらに，それらを用いて固有値の条件に依存しない高次元共分散構造に関するパラメータの推定・検定法を提案した．

(2) Yata and Aoshima (2013, JMA)で提案した拡張クロスデータ行列法を相関行列の推測に適用することで，高次元パスウェイ解析における推定・検定方式を構築し，高速で頑健かつ高精度なパスウェイ解析を可能にした．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．

(3) Aoshima and Yata (2011, SA)で提案した幾何学的表現に基づく2群判別を多群に拡張し，誤判別確率に関して精度を保証するための判別アルゴリズムを構築した．さらに，高次元判別方式における漸近正規性を導出することで，最適性に関する理論を構築した．その結果，従来のマハラノビス距離に基づく最適性理論が高次元の枠組みで崩壊することを示した．一方で，高次元混合モデルにおける幾何学的一致性を示し，クラスタリングにおいて分類可能な空間を特定した．

(4) 研究(3)により，最適性の指標がノイズのスパイク度に大きく依存することを突き止め，ノイズに対する2つの固有値モデル「Strongly spiked eigenvalue (SSE) モデル」と「Non-SSE (NSSE) モデル」を提唱した．NSSEモデルのもとでは各種統計量の高次元漸近正規性が主張できるが，より巨大なノイズを含むSSEモデルにおいてはそれら漸近正規性が主張できず，推測の精度保証が困難であった．そこで，研究(1)で導出した固有空間に関する精密な漸近理論と高次元統計解析法を融合し，新たな統計的推測法を考案した．

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 19件)

M. Aoshima, D. Shen, H. Shen, K. Yata, Y.-H. Zhou, J.S. Marron:

A survey of high dimension low sample size asymptotics (査読あり, 国際共著論文). Aust. N. Z. J. Stat.

(Special Issue in Honour of Peter Gavin Hall), 60, 4-19, 2018.

DOI: 10.1111/anzs.12212

M. Aoshima, K. Yata: Two-sample tests for high-dimension, strongly spiked eigenvalue models (査読あり). Stat. Sinica, 28, 43-62, 2018.

DOI: 10.5705/ss.202016.0063

K. Yata, M. Aoshima:

High-dimensional inference on covariance structures via the extended cross-data-matrix methodology (査読あり). J. Mult. Anal., 151, 151-166, 2016. DOI:

10.1016/j.jmva.2016.07.011

K. Yata, M. Aoshima:

Reconstruction of a high-dimensional low-rank matrix (査読あり). Electron. J. Stat, 10, 895-917, 2016.

DOI: 10.1214/16-EJS1128

A. Ishii, K. Yata, M. Aoshima: Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context (査読あり). J. Statist. Plan. Infer., 170, 186-199, 2016.

DOI: 10.1016/j.jspi.2015.10.007

M. Aoshima, K. Yata: Geometric classifier for multiclass, high-dimensional data (査読あり).

Sequential Anal., 34, 279-294, 2015.

DOI: 10.1080/07474946.2015.1063256

M. Aoshima, K. Yata: Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions (査読あり). Methodol. Comput. Appl. Probab., 17, 419-439, 2015.

DOI: 10.1007/s11009-013-9370-7

〔学会発表〕(計 53 件)

K. Yata, M. Aoshima:

Regularized noise-reduction methodology for high-dimensional data (Invited Talk).

10th Conference of the IASC-ARS/68th Annual NZSA Conference, Auckland (New Zealand), Dec. 13, 2017.

K. Yata, M. Aoshima:

High-dimensional correlation tests with sample size determination (Invited talk).

Sixth International Workshop in Sequential Methodologies, Rouen (France), June 22, 2017.

K. Yata: Estimation of low-rank matrices in high-dimensional settings (Invited talk).

A Symposium on Complex Data Analysis 2017, Hsinchu (Taiwan), May 26, 2017.

K. Yata, M. Aoshima: Inference on high-dimensional covariance structures via the extended cross-data-matrix methodology (Invited talk).

Eighth International Workshop on Applied Probability, Toronto (Canada), June 23, 2016.

K. Yata, M. Aoshima: PCA consistency for high-dimensional multiclass mixture models and its applications (Invited talk).

ISNPS Meeting “Biosciences, Medicine, and novel Non-Parametric Methods”, Graz (Austria), July 13, 2015.

K. Yata, M. Aoshima:

Quadratic-type classification for

high-dimensional data (Invited talk). Third Institute of Mathematical Statistics Asia Pacific Rim Meeting, Taipei (Taiwan), July 2, 2014.

〔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・准教授

研究者番号: 90585803