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　５ 年　 ６ 月 ２６ 日現在

機関番号：１４３０１
研究種目：研究活動スタート支援
研究期間：2020～2022
課題番号：２０Ｋ２２３０５

研究課題名（和文）非線形特徴量選択に関する高次元小標本漸近論

研究課題名（英文）High-dimension, low-sample-size asymptotic theory for nonlinear feature selection

研究代表者
中山　優吾（Nakayama, Yugo）

京都大学・情報学研究科・助教

研究者番号：４０８８４１６９

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：高次元データの非線形性を調査するために，カーネル関数を用いた主成分分析を高次元漸近理論の枠組みで調査した．主成分分析を用いたクラスタリングと外れ値検出の手法を提案し，特に，経験的に度々使用されるガウシアンカーネルのチューニングパラメータに関する理論評価を与えることで，その最適性を議論した．外れ値の有無に関して，主成分分析を用いた検定方式を提案し，複数の外れ値を特定できるような手法も考案した．また，外れ値に関連し，高次元データのロバスト性についてもSpatial Signに着目し，研究を進めた．

研究成果の学術的意義や社会的意義
近年観測されるデータの次元数は非常に多くなっており，例えば，遺伝子発現データでは数万の遺伝子を観測することができる．しかし，実験にかかるコストなどの問題から，解析に十分なサンプル数を確保することができない．このようなデータは解析が難しいため，本研究では非線形な特徴量に注目し，カーネル関数を用いた主成分分析を用いた解析手法を提案した．これにより，高次元データのクラスタリングや外れ値検出が可能となった．提案手法は標本数が少ない高次元データでも機能し，計算コストが問題となる高次元データ解析において効果的である．

研究成果の概要（英文）：We investigated principal component analysis (PCA) with kernel functions in the framework of high-dimensional asymptotic theory to reveal non-linearity in high-dimensional data. We proposed clustering and outlier detection methods by using PCA and discuss their optimality, in particular, by providing a theoretical evaluation for the tuning parameters of the Gaussian kernel, which is often used empirically. In the presence of outliers, we proposed a test method using principal component scores, and devised a method that can identify multiple outliers. With respect to outliers, we also studied the robustness of high-dimensional data, focusing on spatial signs.

研究分野：数理統計学

キーワード：高次元データ　機械学習　非線形　高次元小標本　外れ値検出　クラスタリング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年観測されるデータの規模は非常に膨大であり，ビッグデータ利活用が重要である．ビッグデータの 1 つである遺伝子発現データに見られる特徴は，遺伝子数が患者数に比べて非常に大きいことである．さらに近年では，次世代シーケンサーのコストが減少することでゲノムワイドなデータの取得が可能になり，数千万にも渡る特徴量を持つ超高次元データも登場している．このような背景で，高次元統計解析がますます重要になってきている．

膨大な特徴の中から解釈するために，特徴量選択は高次元小標本データを解析する強力な手法として知られている．高次元データは高次元球面上に集中するという非線形な構造を持っている．そこで，高次元小標本データの解析を大成させるためには，非線形な特徴量を解明することが鍵だと思われる．

このような背景のもと，研究開始当初，研究代表者の所属するグループにおいて，高次元小標本データに対する統計手法を開発していた．その研究結果の 1 つである，サポートベクターマシン(SVM)の高次元小標本における漸近的性質を調査し，高次元小標本で爆発するノイズとして過学習を起こしている原因だと数学的に特定でき，ノイズ除去まで踏み込むことができた．このように一見問題を持つ手法も高次元小標本における漸近的性質を駆使することで改善できる可能性を持っている．また，SVM の代表的な非線形の 1 つであるガウシアンカーネルを用いた SVM についても研究を進め，高次元データに対して非線形性の解析が重要であるという示唆が得られた．しかし，高次元統計解析において非線形性に関する報告は少なかった．

２．研究の目的

本研究は，これまで議論されてこなかった高次元空間の非線形性について調査し，その構造を把握することで，従来の線形手法に対する優位性を理論的に評価することを目指し，次の 2 つを研究目的とした．

(1) 非線形性を用いることで高次元データに関する関係性と計算効率化および予測性能向上

(2) 非線形性を持つ統計手法の高次元における数学的な評価による理論保証

本研究により、高次元データの持つ非線形構造を浮き彫りにし，高精度で柔軟な解析法を提案できると考えた．

３．研究の方法

先行研究で非線形判別手法であるガウシアンカーネルを用いた SVM を一般のカーネル関数に拡張することで，判別分析の枠組みで非線形性について考察する．高次元データの持つ非線形構造を把握すべく固有解析を行う．固有構造の解析のために線形主成分分析 (PCA)を用いる．上位のスペクトルだけに興味があることを踏まえると，線形の PCA では標本共分散行列の固有値分解を考えるが，高次元小標本の場合，その双対関係にあたるグラム行列を考える方が計算コストを低く抑えつつも目的のスペクトルを解析することができる．この方法によって得られたスペクトルに対する高次元小標本漸近評価から始める．研究課題である非線形の場合への拡張は SVM に対する考察を踏まえ，カーネル関数を考えることで達成できる．また，カーネル関数でもグラム行列から同様にスペクトルを解析できる．実データ解析で得られたスペクトルの振る舞いから，それらに対して適切なモデルを提案し，固有解析を進めていく．最終的には，先行研究では高次元で混入するバイアスや過学習が報告されているため，同様に除去・緩和したスペクトルを用いて非線形特徴量選択や固有射影を構成する．機械学習タスクに固有解析に基づく固有射影を適応させる．これらの結果は数値実験と遺伝子発現データを用いて，その有効性を提示し，国内外の研究集会及び国際雑誌において報告する．

４．研究成果

下記の研究成果が得られた．各研究成果について，図 1 に関係性をまとめる．

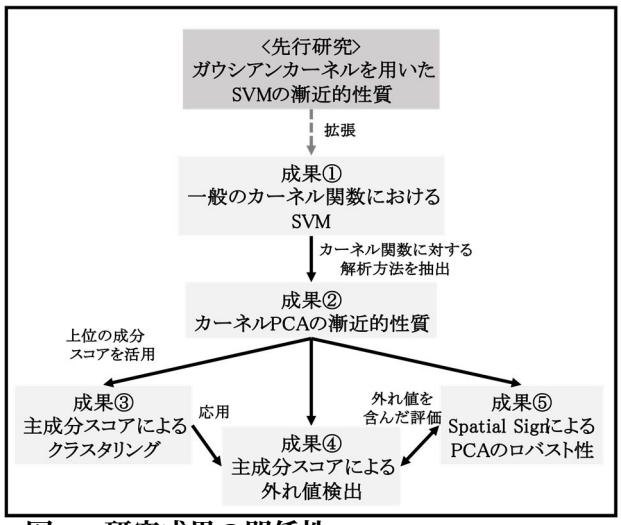

図 1．研究成果の関係性．

(1) SVM の漸近的性質についてカーネル関数のある条件を満たす部分集合上での一般論を展開し，バイアス補正とチューニング方法を提案した．この 2 つの提案手法は先行研究における結果を包含している．特に，チューニングに関してガウシアンカーネルの結果に加え，多項式カーネルとラプラスカーネルに対しても有効であることを数値実験から確認できた．この結果は査読付き国際雑誌に掲載されている．

(2) 高次元統計解析では，PCA が解析されるが，高次元空間の非線形性を解き明かすために，一般のカーネル関数に対する PCA の漸近的性質を調査した．

(3) 研究成果(2)の解析結果を用いることで，カーネル主成分スコアを用いたクラスタリング手法を考案した．第 1 主成分スコアの正負で 2 つのクラスターを分離することができる．カーネル関数を用いて解析できたことで先行研究の線形カーネルの場合を包含できた．研究成果(1)と同様にカーネル関数に対する一般論も議論し，提案手法がある族上のカーネル関数にも有用であることを示した．統一的な枠組みで解析できたことから先行研究よりも仮定を緩めることに成功した．特に，ガウシアンカーネルの性質について注目し，線形カーネルとの数学的な比較をすることで，非線形性が 2 次モーメントの差分という意味でクラスタリングに生かされていることを明らかにした．また，ガウシアンカーネルの場合に 3 つのクラスターの場合のクラスタリング方法も与えた．カーネル関数に含まれるハイパーパラメータの選択も高次元では計算コストが膨大になるため，ガウシアンカーネルにおいて反復法を必要としない高速な選択法を導出した．この結果は査読付き国際雑誌に掲載されている．

(4) 研究成果(2), (3)を用いた応用として，主成分スコアを用いた外れ値混入に関する高次元データの検定問題を考えた．検定方式に対して第 1 種の過誤と検出力について理論的な結果を与えた．この検定方式を応用することで複数の外れ値を連続的に検出できる外れ値検出手法を提案した．この手法は，外れ値を正しく検出できる Sure Screening という性質を持つことを示した．高次元小標本の枠組みにおける他の手法に比べ，偽陽性と真陰性を低く抑えながらも外れ値を検出できるという優位性を数値実験から示した．

(5) 研究成果(4)の外れ値に関連して，ロバスト性についても Spatial Sign に着目した PCA の漸近的性質を調査した．通常の PCA に比べて，外れ値に対す影響を緩和できることを示した．

これらの結果は数値実験と遺伝子発現データを用いて，その有効性を提示し，国内外の研究集会及び国際雑誌において報告した．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Nakayama Yugo、Yata Kazuyoshi、Aoshima Makoto	４．巻 185
２．論文標題 Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings	５．発行年 2021年
３．雑誌名 Journal of Multivariate Analysis	６．最初と最後の頁 104779～104779
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jmva.2021.104779	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 Nakayama Yugo	４．巻 1
２．論文標題 Support vector machine and optimal parameter selection for high-dimensional imbalanced data	５．発行年 2020年
３．雑誌名 Communications in Statistics - Simulation and Computation	６．最初と最後の頁 1～16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/03610918.2020.1813300	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計12件（うち招待講演　2件／うち国際学会　2件）

１．発表者名 中山優吾，矢田和善，青嶋誠
２．発表標題 高次元主成分スコアに基づく異常値の検出法
３．学会等名 日本数学会2022年度年会
４．発表年 2022年

１．発表者名 中山優吾，矢田和善，青嶋誠
２．発表標題 高次元におけるカーネル主成分分析の漸近的性質とその応用
３．学会等名 多様な高次元モデルの理論と方法論：最前線の動向
４．発表年 2022年

１．発表者名
中山優吾，矢田和善，青嶋誠

２．発表標題
Asymptotic properties of high-dimensional kernel PCA and its applications

３．学会等名
International Symposium on New Developments of Theories and Methodologies for Large Complex Data（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
中山優吾

２．発表標題
ガウシアンカーネルに基づく高次元データの分類問題

３．学会等名
2021年度秋季総合分科会

４．発表年
2021年

１．発表者名
中山優吾，矢田和善，青嶋誠

２．発表標題
高次元における重み付き判別分析とデータ変換法について

３．学会等名
2021年度統計関連学会連合大会

４．発表年
2021年

１．発表者名
Yugo Nakayama, Kazuyoshi Yata, Makoto Aoshima

２．発表標題
Clustering by kernel PCA with Gaussian kernel and tuning for high-dimensional data

３．学会等名
The 4th International Conference on Econometrics and Statistics（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
中山優吾，矢田和善，青嶋誠

２．発表標題
高次元におけるカーネル主成分分析の漸近的性質と異常値の検出への応用

３．学会等名
日本数学会2021年度年会

４．発表年
2021年

１．発表者名
中山優吾，矢田和善，青嶋誠

２．発表標題
高次元データにおける異常値の検出について

３．学会等名
科研費シンポジウム「機械学習・統計学・最適化の数理とＡＩ技術への展開 」

４．発表年
2020年

１．発表者名
中山優吾，矢田和善，青嶋誠

２．発表標題
高次元カーネル主成分分析に基づく異常値の検出

３．学会等名
科研費シンポジウム「大規模複雑データの理論と方法論：最前線の動向と新たな展開」

４．発表年
2020年

１．発表者名
中山優吾，矢田和善，青嶋誠

２．発表標題
Clustering by kernel principal component analysis for high-dimensional data

３．学会等名
日本数学会 2020年度秋季総合分科会

４．発表年
2020年

１．発表者名 中山優吾，矢田和善，青嶋誠
２．発表標題 高次元小標本における異常値の検出
３．学会等名 2020年度統計関連学会連合大会
４．発表年 2020年

１．発表者名 中山優吾，矢田和善，青嶋誠
２．発表標題 高次元主成分分析における頑健性について
３．学会等名 日本数学会 2022年度秋季総合分科会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関