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　元 年　 ６ 月　 ２ 日現在

機関番号：３２６４１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００００５７

研究課題名（和文）スパースモデリングの数理と多変量解析ツールの開発研究

研究課題名（英文）Theoretical developments of sparse modeling and multivariate analysis techniques

研究代表者
小西　貞則（KONISHI, Sadanori）

中央大学・理工学部・教授

研究者番号：４００９０５５０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：複雑な現象を内包する高次元データから有益な情報を高効率に抽出・分析するためのモデリングと多変量データ解析ツールの開発研究に取り組み，次のような研究成果を挙げた．(1) 関数主成分分析と部分空間法を融合した新たな多クラス識別，パターン認識の解析法を開発提案した．(2) モデルの非線形化とスパース性を融合した頑健で汎化能力の優れたスパース回帰モデリングを提唱した．(3) カーネル法による非線形部分空間法の多様なデータへの適用拡大化を図った．(4) 主成分分析，正準相関分析，正準対応分析における次元圧縮法について研究し，情報量，ベイズ理論によるモデル評価基準の提唱と理論・方法論の研究を行った．

研究成果の学術的意義や社会的意義
諸科学，産業界や実社会で日々獲得，蓄積されつつあるデータの多様化と大規模・高次元化の流れの中で，新たなデータ解析技術と効率的な情報処理の必要性が認識されるようになった．本研究で取り組んだ回帰モデリング，識別・判別，パターン認識，分類・クラスタリングなどの多変量解析手法の研究成果は，現象の情報源であるデータを分析，処理し，現象の解明と予測・制御，新たな知識発見や複雑なシステムの理解を促進するツールとして役立つと考えられる．また，大規模データの高速処理を可能とする高度なアルゴリズムの開発研究と相俟って，柔軟で汎化能力の優れた機械学習の新たな解析法として寄与することが期待される．

研究成果の概要（英文）：Huge amount of data with complex structure and/or high-dimensional data have been accumulating from diverse sources. Through this research we have investigated the problem of analyzing such datasets to extract useful information and pattern, and proposed various modeling and multivariate analysis techniques: (1) Multi-class classification methods for high-dimensional longitudinal data are proposed based on class-featuring information compression with the help of multivariate functional principal component. (2) Sparse kernel subspace methods are proposed to learn the complex structure of high-dimensional data. (3) Model selection criteria are provided for Bayesian probabilistic dimensionality reduction in principal component and canonical correlation analyses. (4) With the development of modeling techniques such as sparse and Bayes modeling, we investigate a general theory for constructing model selection criteria to evaluate models constructed by various estimation procedures.

研究分野：統計科学

キーワード：線形・非線形スパースモデリング　関数データ解析　部分空間法　多クラス識別・パターン認識　確率的次元圧縮　カーネル非線形モデリング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

計算機関連技術と計測・測定技術の高度な進展は，大規模・高次元データの獲得，蓄積を可能とし，多様な形式でデータベースとして組織化されつつあった．現象を解明，理解し，新たな知識発見に繋げるためには，情報を内包するデータ集合に基づく適切なモデル化が必要である．しかし，超高次元データに基づくモデルは必然的に複雑かつ大規模となり，従来のモデルの推定，評価のプロセスは有効に機能しないだけでなく，不必要な情報をモデルの中に過剰に取り込むことから汎化能力が低下し，現象の解明と予測には繋がらない．

このような状況の中で，大規模・高次元データの背後に潜む有益な情報やパターンを高効率に抽出・処理するためのモデリングの開発と理論・方法論の研究，新たな多変量データ解析法の開発研究，そして諸科学への応用を目的として研究を推進することを計画するに至った．

２．研究の目的

複雑な現象を内包する超高次元のデータを分析するためのモデルは必然的に大規模となり，従来のモデリングの推定，評価プロセスは有効に機能しないだけでなく，不必要な情報を過剰に取り込むことから，現象の解明と予測・制御には繋がらない．

本研究では，高次元データ集合に基づく線形・非線形スパースモデリングについて理論的・数値的両側面から研究し，回帰，識別・判別，パターン認識，次元圧縮等に関して，柔軟で汎化能力の優れたモデリングと新たな多変量解析ツールの提唱，そして諸科学・産業界への応用研究を推進し，従来の多変量データ解析手法では捉えきれないさまざまな現象解析に役立つ手法を開発し研究することを目的とした．

３．研究の方法

(1) 多数のパラメータを含む大規模モデルに対しては，モデルの想定，想定したモデルの推定と評価・選択という従来のモデリングのプロセスは有効に機能しない．この問題の解決法として一つの方向性を示したのが，パラメータの L_1 ノルム正則化項を損失関数に課すという lasso であった．この研究に端を発し，生命科学，システム工学，医用画像工学など科学のさまざまな分野で応用が試みられてきた．適用分野の拡がりは問題提起へと繋がり，lasso の基本的な考え方を拡張，発展させ，各分野のデータの特性，分析目的に合わせて様々な L_1 型正則化回帰モデリング（スパースモデリング）の提唱と推定アルゴリズムの開発研究が集中的に行われてきた．これらの研究は，主として線形回帰，一般化線形モデルの枠組みで行われてきたが，本研究では，諸科学・産業界の多様なデータの分析に役立つ非線形回帰，主成分，正準相関，識別・判別，パターン認識，次元圧縮等に対するスパースモデリングと新たな解析手法の開発研究に取り組んだ．

(2) 時間の経過にともなって経時的に観測・測定されたデータや時空間データは，極めて次元の高いベクトルデータと捉えることができる．このようなデータは，気象学，地球環境科学，経済学，マーケティングなど様々な分野で日々獲得・蓄積されつつある．そこで，観測時点の不均一性や欠測，さらに個体差や内在するノイズを考慮して，高次元ベクトルデータを関数化処理し，処理した関数化データ集合に基づく非線形現象解明のためのモデリングと理論・方法論，新たな多変量解析手法の開発研究を推進した．

(3) 大量かつ超高次元データ，少数かつ高次元データなど多様なデータ集合に基づくモデリングの過程で重要な役割を果たすモデルの評価・選択に対して，適用上の問題点の克服とモデルの汎化能力の向上を目指して，情報量理論，ベイズアプローチによって新たなモデル評価基準の提唱に取り組んだ．

４．研究成果

(1) 識別・判別，パターン認識は，文字・画像認識，医用画像工学，生物科学，気象学など科学の様々な分野で適用される重要な研究課題である．本研究では，多クラスパターン認識の解析手法の研究に取り組み，主成分分析法によって高次元空間に散らばる複数のデータ集合を低次元の部分空間へと次元圧縮し，所属未知のデータと圧縮した空間との類似度を基準として分類する部分空間法について研究し，以下の成果を挙げた．

① モデルの推定と予測に本質的な変数の選択を同時に実行する fused lasso と呼ばれる L_1 ノルム正則化法を融合したスパース部分空間法を提案した．提案手法は，文字画像データの解析を通して，従来手法と比べて汎化能力が高いことを立証した．

② 複雑な非線形構造を内包する高次元データの分類を目的として，カーネル法による非線形部分空間法とモデルの推定と変数選択を同時に実行する新たなスパース非線形部分空間法を提案した．

③ 部分空間法に本質的な圧縮次元数の決定法に対して，ベイズアプローチによる確率的主成分によって情報量およびベイズの観点からモデル評価基準を導出し，データに基づく圧縮次元の推定法を提唱した．

本研究から得た知見を一般化することを目的として，解析の対象とする現象の情報を確率分布モデルで捉えて，ベイズアプローチによって融合したモデリングの理論・方法論の研究を推

進中である．

(2) 現象過程や動作過程の連続的な実験・計測データ，高頻度で観測される経済データなどのように離散時点で経時的に観測・測定された多数のデータ系列（高次元ベクトルデータ集合）に非線形回帰モデルを当てはめて関数化処理して，関数化したデータ集合を対象とした解析手法の開発研究に取り組み，以下の成果を挙げた．

① 高次元ベクトルデータの関数化に関して，バーンスタイン基底に基づく基底展開法によって関数化データ集合を構成し，関数主成分分析と部分空間法を融合した新たな判別・識別，パターン認識の解析手法を提案し，手法の有効性を実データ，シミュレーションを通して立証した．この関数データ解析手法と部分空間法を融合した手法は，新たな解析手法へと繋がることが期待され，現在，手法の適用拡大化と理論研究を推進中である．

② スプライン基底，バーンスタイン基底などに基づく基底展開法によって非線形関数化する方法を研究し，関数化データ集合に基づく分類，クラスタリング手法を提唱した．提唱した関数クラスタリング解析手法を運動・体幹機能データの解析に適用し，早期機能異常の検出等に応用する研究を推進中である．

(3) 正準相関分析，正準対応分析は，2 つの確率変数ベクトル間の関連性の程度を基準とする多変量解析法で，前者は，自然言語処理，システム工学，経済学などにおいて，後者は，生態学，水質科学，生物科学，社会学など幅広い分野で現象解析の有用な手法として用いられている．理論的にはある種の固有値問題へ帰着され，固有ベクトルを係数とする射影軸を逐次求めて次元圧縮を行い，データ行列に内包される情報とパターン抽出を行う．近年，極めて次元の高いデータや複雑な非線形構造を内包する多次元データが観測されるようになり，正準相関分析の新たな推定法や非線形化の研究が必要となってきた．また，正準対応分析については，誤差を内包する観測度数データ行列に対して，推測論の研究の必要性を認識した．

これらの問題に対処するための研究を行い，高次元化に伴って起因する従来手法の限界を克服するスパース正準相関分析，非線形構造を内包する多次元データの分析に有効に機能するカーネル法による非線形正準相関分析，モデルの非線形化とスパース性を融合したスパース非線形正準相関分析について理論的・数値的両側面から研究し，実際上有用な解析手法の提案を目指して研究を行った．特に，正準相関分析に対して，確率的正準相関の概念を定式化して情報量，ベイズの観点からモデル評価基準を導出した．正準対応分析に関しては，解析的・代数的操作をアルゴリズム化する統計的計算法であるブートストラップ法を適用して，新たな次元圧縮法を提案した．現在，深層学習化に向けた研究を推進中である．

(4) 円周上のデータ，球面上のデータ，より一般に超球面上のデータは，方向性（角度）を有しており，方向性データあるいは角度データと呼ばれ，その解析手法は，気象学，地球環境科学，地質学などの分野で広く用いられている．本研究では，方向性データの解析に関して，連続的でなくジャンプが起る時系列データの場合における自己回帰型のモデル化の提案，説明変数に角度を含む非対称回帰モデルにおいて影響診断を行う方法と太陽光発電のための基礎データの解析など，理論研究と実際問題への適用研究を行った．

５．主な発表論文等

〔雑誌論文〕（計 17 件）

① 小西貞則 (2019). 情報量規準 AIC の統計科学に果たしてきた役割, 統計数理, 査読有, 掲載決定.

② Misumi, T., Matsui, H. and Konishi, S. (2019). Multivariate functional clustering and its application to typhoon data, Behaviormetrika, 46, 163-175, 査読有.
DOI: 10.1007/s41237-018-0066-8

③ Park, H. and Konishi, S. (2018). Sparse common component analysis for multiple high-dimensional datasets via non-centered principal component analysis, Statistical Papers, 査読有，掲載決定.
DOI: 10.1007/s00362-018-1045-6.

④ Shimamura, K., Ueki, M., Kawano, S. and Konishi, S. (2018). Bayesian generalized fused lasso modeling via NEG distribution, Communications in Statistics - Theory and Methods, 査読有，掲載決定.
doi/abs/10.1080/03610926.2018.1489056?journalCode=lsta20.

⑤ Matsuda, K., Kawano, S. and Konishi, S. (2018). Predictive information criteria for robust relevance vector regression models, Bulletin of Informatics and Cybernetics, 50, 65-80, 査読有.

⑥ Imoto, T., Shimizu, K. and Abe, T. (2018). A cylindrical distribution with heavy-tailed linear part, Japanese Journal of Statistics and Data Science, 査読有，掲載決定.
DOI: 10.1007/s42081-019-00031-5.

⑦ Zhan, X., Ma, T., Liu, S. and Shimizu, K. (2018). Markov-switching linked autoregressive model for non-continuous wind direction data, Journal of Agricultural,

Biological and Environmental Statistics, **査読有**, **掲載決定.**
DOI 10.1007/s13253-018-0331-z.
⑧ Nurul Hidayah Sadikon, Adriana I. N. Ibrahim, Ibrahim Mohamed and Kunio Shimizu (2018). A new test of discordancy in cylindrical data, Communications in Statistics - Simulation and Computation, **査読有**, **掲載決定.**
doi.org/10.1080/03610918.2018.1458131.
⑨ Jan Dolinsky, D., Hirose, K. and Konishi, S. (2017). Readouts for echo-state networks built using locally regularized orthogonal forward regression, Journal of Applied Statistics (Published online).
DOI: 10.1080/02664763.2017.1305331.
⑩ Park, H. and Konishi, S. (2017). Principal component selection via adaptive regularization method and generalized information criterion, Statistical Papers, 58, 147-160**, 査読有.**
⑪ **清水邦夫** (2017). **方向統計学における確率分布の最近の話題, 日本統計学会誌, 第** 47 **巻**, **第** 2 **号,** 103-140, **査読有.**
⑫ Liu, S., Ma, T., SenGupta, A., Shimizu, K., Wang, M.-Z. (2017). Influence diagnostics in possibly asymmetric circular-linear multivariate regression models, Sankhya B, 79(1), 76-93, **査読有.**
doi: 10.1007/s13571-016-0116-8.
⑬ Misumi, T. and Konishi, S. (2016). Mixed effects historical varying coefficient model for evaluating dose-response in flexible-dose trials, Journal of the Royal Statistical Society: Series C (Applied Statistics), 65, 331-344, **査読有.**
⑭ **松井秀俊・三角俊裕・横溝孝明・小西貞則** (2016). **非線形混合効果モデルに基づく関数データクラスタリング**, **応用統計学** 45-1&2, 25--45, **査読有.**
⑮ Park, H. and Konishi, S. (2016). Robust nonlinear regression modeling via L1-type regularization, Bulletin of Informatics and Cybernetics, 48, 47-61, **査読有.**
⑯ Park, H. and Konishi, S. (2016). Robust logistic regression modeling via the elastic net-type regularization and tuning parameter selection, Journal of Statistical Computation and Simulation, 86-7, 1450 - 1461, **査読有.**
DOI:10.1080/00949655.2015.1073290.
⑰ Park, H. and Konishi, S. (2016). Robust coordinate descent algorithm robust solution path for high dimensional sparse regression modeling, Communications in Statistics - Simulation and Computation, 45-1, 115-129, **査読有.**
DOI:10.1080/03610918.2013.854910.

〔学会発表〕(計 11 **件)**
① 三角俊裕, 松井秀俊, 小西貞則 (2018). ジョイントモデリングに基づく多変量関数クラスタリングと気象データへの応用, 2018 年度統計関連学会連合大会, 中央大学.
② 松田 和己, 川野 秀一, 小西貞則 (2018). ロバスト関連ベクター回帰モデルにおける予測情報量規準, 2018 年度統計関連学会連合大会, 中央大学.
③ 松川 達也, 三角 俊裕, 小西 貞則 (2018). 多変量経時データの共通主成分による次元圧縮と推移分析, 第 23 回情報・統計科学シンポジウム, 九州大学.
④ 新井 仁智, 福田 竜也, 三角 俊裕, 小西 貞則 (2018). 正則化基底展開法による多変量関数データクラスタリング, 第 23 回情報・統計科学シンポジウム, 九州大学.
⑤ 入江 敦子, 三角 俊裕, 小西 貞則 (2018). 正準対応分析による次元圧縮とブートストラップ推測, 第 23 回情報・統計科学シンポジウム, 九州大学.
⑥ Fukuda, T., Misumi, T., Matsui, H. and Konishi, S. (2018). Multivariate functional subspace methods for classifying high-dimensional longitudinal data, 11th International Conference of the ERCIM WG on Computational and Methodological Statistics, Italy.
⑦ Shimizu,K. (2018). A distribution for observations on a hyper-cylinder, The 4TH ISM International Statistical Conference (ISM-IV), Sunway University, Malaysia.
⑧ 清水邦夫, 井本智明, 阿部俊弘 (2017). A Pareto-type distribution on the cylinder**, 2017 年度統計関連学会連合大会, 南山大学.**
⑨ Shimizu, K, Imoto, T. and Abe, T. (2017). Probability distributions for cylindrical data, ADISTA17, Roma, Italy.
⑩ Shimizu, K, Imoto, T. and Grace S. Shieh (2017). Discrete distributions on the circle, Keio International Symposium: Statistical Analysis for High-Dimensional, Circular or Time Series Data, Yokohama.
⑪ Shimizu, K, Imoto, T. and Grace S. Shieh (2016). A new method of obtaining probability mass functions on the circle, The 3rd ISM International Statistical Conference, Kuala Lumpur, Malaysia.

６．研究組織
(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名： 清水　邦夫
ローマ字氏名： SHIMIZU, kunio