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　２８　年　　４　月　１５　日現在

機関番号：　１７１０２
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５７３００１７

研究課題名（和文）大規模計測データに対する正則化統計モデリング手法の開発

研究課題名（英文）Development of statistical modeling with regularization for large scale data

研究代表者

松井　秀俊（Matsui, Hidetoshi）

九州大学・数理（科）学研究科（研究院）・助教

研究者番号：９０６３３３０５

交付決定額（研究期間全体）：（直接経費）　　1,600,000円

研究成果の概要（和文）：各個体が時間の経過などに応じて繰り返し計測値を得た経時測定データから，有効な情報を集約して抽出するための統計的モデリング手法の開発を行った．本研究では特に，経時測定データを関数として処理し解析対象とする関数データ解析において，モデル推定と同時に変数選択を行うことができるスパース正則化を適用するための推定，評価の手順を構築した．さらに，提案手法を実際のデータへ適用し，手法の有効性の検証や新たな知見獲得を行った．

研究成果の概要（英文）：I have developed some statistical modeling strategies for repeated measurement or longitudinal data in order to extract important information from data. In this work I focused on two topics; functional data analysis and sparse regularization. The basic idea behind functional data analysis is to express longitudinal data in the form of a function and then draw information from a set of functional data. I applied sparse regularization techniques to several statistical models for functional data in order to estimate model and select functional variables simultaneously. I also derived algorithms for estimating parameters and model selection criteria for evaluating the estimated models. The proposed methods were applied to the analyses of real data sets and then examined the effectiveness of the methods.

研究分野：　統計科学

キーワード：　関数データ解析　スパース正則化　モデル選択

１．研究開始当初の背景

近年の計算機の発展に伴い、様々な分野で可能な限り多くの情報を保存しようと大規模なデータが取得されるようになってきた。これに伴い、データの構造が多種多様化・複雑化の傾向を見せており、従来から用いられている統計手法を直接適用することが困難になりつつある。このようなデータを解析するために、「関数データ解析に基づく繰り返し測定データ解析」と「スパース正則化による次元圧縮」という 2 つのテーマを軸にして研究を進めた。

２．研究の目的

１つの個体が時点の経過や位置の変化などに応じて複数の観測値を得た形式のデータは「繰り返し測定データ 」と呼ばれている。現在の計算機の発展に伴う計測規模の拡大により、このような形式のデータの観測は増加しつつある。繰り返し測定データは、その特性からしばしば従来の多変量解析手法を直接適用することが困難となる。繰り返し測定データに対するアプローチの一つである関数データ解析は、離散時点で観測されたデータを関数化処理し、得られた関数をデータとして扱う解析方法である。この手法を適用することでデータの情報を効率的に集約し、統計解析手法が容易に適用できるようになる。

また、スパース正則化は、回帰モデルの係数パラメータに L1 ノルム等の制約を課した推定法で、制約の性質からモデル推定と変数選択を同時に行える手法として近年の統計科学の分野で最も多く利用されている手法の一つである。

本研究では、繰り返し測定データに対する統計的モデリング手法として、次の研究を構想した。

（１）多群、すなわち３群以上の関数データ判別モデルに対してスパース正則化を適用することで、適切に変数選択を行うための制約の形状について検討する。加えて、各変数がどの群の判別に寄与しているかの選択、つまり決定境界の選択のための制約についても研究する。

（２）説明変数および目的変数が共に関数データとして与えられた回帰モデルを拡張し、目的変数が多変量の関数データで与えられた関数多変量回帰モデルに対する推定、評価法を導出する。

３．研究の方法

（１）多群判別モデルに対するスパース正則化に関する先行研究では、スパース制約の効果を十分に発揮することができず、不要な変数をモデルから除去できる保証がない。これに対して本研究では、group lasso の考え方を適用してパラメータをグループ化することで、群間のパラメータを同時に 0 に縮小する制約を構築する。さらに、構築した制約に基づく推定アルゴリズムや、推定されたモデルを予測の観点から評価するための評価基準を導出する。推定および変数選択に関する精度や安定性を既存の制約と比較するため、数値実験を行い、さらに、医学データの解析を通して複数種類の疾病の判別を試みる。

（２）説明変数と目的変数が共に多変量の関数データとして与えられた場合、変量間の関係を表現するための回帰モデルを構築する。関数データとして与えられた変数は基底関数展開によって表されるとし、パラメータの推定問題を簡略化させる。推定されたモデルを評価するための基準として、情報量およびベイズアプローチの観点から導出されたモデル評価基準を、多変量回帰モデルの枠組みで導出する。

４．研究成果

（１）関数データに対する多群ロジスティック回帰モデルに対して、適切に変数選択または決定境界の選択を行うための制約として、多変量回帰モデルの推定のために提案されていたものを拡張してそれぞれ導出した。提案した２種類の制約は、解析目的（変数選択、決定境界選択）に応じて使い分けることが良いと考えられる。これらの制約を、経時測定された遺伝子発現データの解析に適用することで、変数選択または決定境界の選択が適切に行われていることを検証した。今後の展望として、２種類の制約を包括し一般化した新たな制約の構築が考えられる。

（２）繰り返し測定データとして与えられた説明変数と目的変数との因果関係を明らかにする回帰モデルの一つである変化係数モデルに対して、スパース正則化を適用することでモデル推定と変数選択を同時に行う方法を提案した。一般的に、変化係数モデルやスパース正則化に基づく推定では、推定量を解析的に導出することは困難である。そこで本研究では、推定アルゴリズムとして coordinate descent algorithm を適用した。これにより、変化係数モデルに対して逐次的に各説明変数の係数を推定することができる。そして、提案手法を医学のデータに適用した結果、疾病の進行に関連のある情報を適切に選択することができた。

（３）説明変数と目的変数が共に時間の関数データとして与えられた場合、これらの関係をモデル化する関数回帰モデルにおいて、特に時間の因果関係を考慮に入れたモデルに対する推定方法について研究した。ここでは、モデルに含まれるパラメータを、スパース正則化を用いて推定する方法を提案した。提案手法を気象学のデータへ適用し、台風の進路

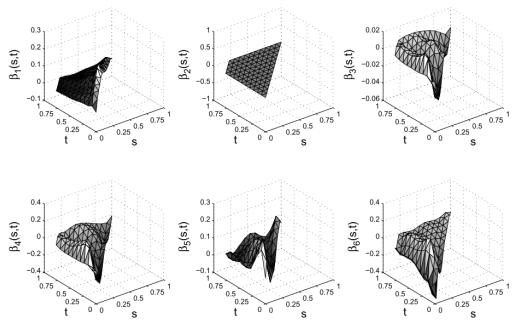

図1　スパース正則化に基づき推定された係数曲面

に影響を与えていると考えられる情報の選択を行った。図１は、提案手法を適用することで推定された、６個の説明変数に対する係数曲面である。本手法を適用することで、２番目（中央上）の係数が零関数として推定されている。これは、対応する説明変数が目的変数に影響を与えていないことを示している。

（４）関数データとして与えられた説明変数と、スカラーとして与えられた目的変数がいずれも多変量データのとき、これらの関係性を包括的にモデル化する関数多変量回帰モデルに対して、スパース正則化を適用することで説明変数を選択する方法について研究した。関数多変量回帰モデルでは、一つの説明変数に関連するパラメータが複数与えられるため、スパース正則化を用いて変数選択を行う場合、その制約の形状が重要となる。本研究では、その形状を構築するとともに、パラメータの推定量を導出するためのアルゴリズムを、既存のアルゴリズムを拡張して導出した。そして、提案手法を近赤外線スペクトルデータの解析に適用した。肉標本に対して照射された近赤外線スペクトルの吸収率は、その肉の成分含有量と関係があることが知られている（図２）。本研究では、スペクトルデータを関数データとして扱い、さらにこの微分データを含めて説明変数とし、３種類の成分含有量を目的変数として関数回帰モデルを構築した。そして、提案手法を適用することで、何次の微分情報が成分含有量に関連しているかを調査した。

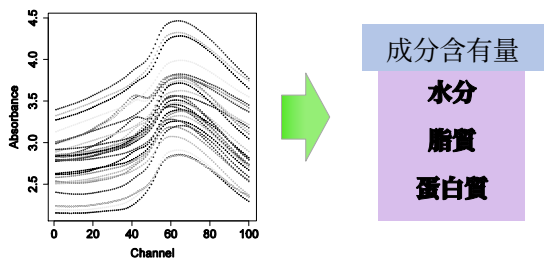

図2　スペクトルデータ（左）と成分含有量（右）とのモデル化

（５）スパース正則化に関する最近の研究に、bi-level selection とよばれる手法がある。これは、説明変数がいくつかのグループに分割される場合、説明変数をグループとして一括して選択し、かつ個々の変数も同時に選択するための方法であり、遺伝子データ解析などに応用されている。本研究では、これまでに提唱されてきた複数の bi-level selection の手法に対して、予測精度および正しい変数を選択する精度を数値的に比較、検証した。その結果、制約の種類に応じて予測、変数選択の精度の観点から特徴に違いがあることが明らかになった。今後の課題として、bi-level selection を研究成果（１）で述べた研究に応用することで、変数と決定境界を同時に選択するための手法を構築することが考えられる。

（６）難病に指定されている多発性硬化症に対する治療の効果に関連がある遺伝子の選択を行うための判別手法を提案した。データとして、治療直後から経時的に測定された遺伝子発現量が得られている。本研究ではこれらを関数データとして扱い、さらに、治療の効果の有無を判別するためにロジスティック回帰モデルを適用した。そして、判別に影響を与えている変数（遺伝子）を、スパース正則化を用いて選択した。提案した手法を上記データの解析に適用した結果、生物学的に重要であると考えられている遺伝子を、統計モデルに基づく解析によって重要なものとして抽出することに成功した。図３上がその遺伝子の発現量の、各患者における経時変化を図示したものである。黒は治療により改善が見られた群、赤は改善が見られなかった群である。また、図３下は関数ロジスティック回帰モデルを適用することで得られた係数曲線の推定量である。この曲線が絶対値の意味で大きい時点が、判別への寄与が特に大きい期間を示している。
本研究で扱ったデータは、あらかじめ遺伝子の数が限定されていたが、全ての遺伝子に対する網羅的解析を行う場合、今回のモデルを直接適用することは困難となる。そこで、変数のスクリーニングを導入した二段階推定を行うことが今後の課題である。

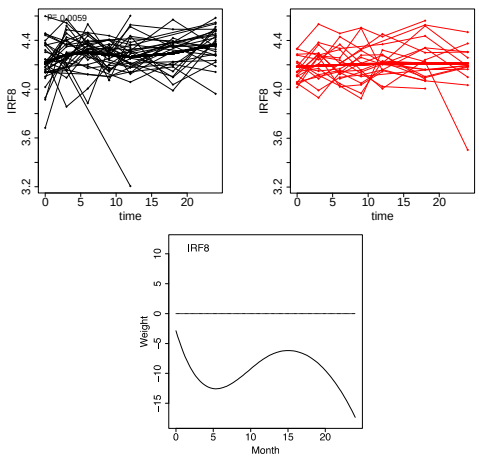

図3　多発性硬化症患者に対する遺伝子発現量（上）と推定された判別モデルの係数曲線（下）

（７）繰り返し測定データを関数データ化するための方法の一つに、混合効果モデルを用いるものがある。本研究では、混合効果モデルに基づく関数データ化の過程で得られるランダム効果関数を対象とした、関数データのクラスタリング方法を提唱した。ランダム効果関数に限定することで、各個体に特有の変動のみに着目したクラスタリングを行うことができる。提案した手法を、環境学や気象学などのデータに適用し、安定したクラスタリング結果が得られていることを確認した。

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１２件）

① Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S. and Miyano, S. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection. Biostatistics, 査読有, Vol. 17, 2016, 235-248. doi:10.1093/biostatistics/kxv037

② Koseki, J., Matsui, H., Konno, M., 他 6 名 . A trans-omics mathematical analysis reveals novel functions of the ornithine metabolic pathway in cancer stem cells. Scientific Reports, 査読有, Vol. 6, 2016, Article number: 20726. doi:10.1038/srep20726

③ Matsui, H. and Misumi, T. Variable selection for varying coefficient models with the sparse regularization. Computational Statistics, 査読有, Vol. 30, 2015, 43-55. doi:10.1007/s00180-014-0520-3

④ Matsui, H. Variable and boundary selection for functional data via multiclass logistic regression modeling. Computational Statistics & Data Analysis, 査読有, Vol. 78, 2014, 176-185. doi:10.1016/j.csda.2014.04.015

⑤ Matsui, H., Misumi, T. and Kawano, S. Model selection criteria for the varying-coefficient modeling via regularized basis expansions. Journal of Statistical Computation and Simulation, 査読有, Vol. 84, 2014, 2156-2165. doi:10.1080/00949655.2013.785548

〔学会発表〕（計２８件）

① Matsui, H. Classification of functional data using bi-level selection. 8th International Conference of the ERCIM WG on Computational and Methodological Statistics, London (United Kingdom), 2015 年 12 月 13 日.

② Matsui, H. Selection of variable and classification boundary by functional logistic regression. 7th International Conference of the ERCIM WG on Computational and Methodological Statistics, Pisa (Itary), 2014 年 12 月 7 日.

③ 松井秀俊. Bi-level selection を用いた関数ロジスティック回帰モデルの推定. 2015 年度統計関連学会連合大会，岡山大学（岡山県岡山市）, 2015 年 9 月 8 日.

④ 松井秀俊. スパース正則化に基づく関数多変量回帰モデルの推定. 2014 年度統計関連学会連合大会，東京大学（東京都文京区），2014 年 9 月 14 日.

⑤ 松井秀俊. 関数データに基づく回帰モデルの構築と正則化. 2013 年度統計関連学会連合大会, 大阪大学（大阪府豊中市）, 2013 年 9 月 9 日.

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　０件）

○取得状況（計　０件）

〔その他〕
ホームページ等
https://sites.google.com/site/hidetoshimatsui/

６．研究組織
(1)研究代表者

松井　秀俊（MATSUI, Hidetoshi）
九州大学・大学院数理学研究院・助教
研究者番号：90633305

(2)研究分担者　なし

(3)連携研究者　なし