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

平成　３１　年　　４　月　１３　日現在

機関番号：　１４２０１
研究種目：　若手研究(B)
研究期間：　2016～2018
課題番号：　１６Ｋ１６０２０

研究課題名（和文）経時測定データに対する統計モデルの構築およびスパース正則化法に基づく推定

研究課題名（英文）Formulation of statistical models for longitudinal data and estimation by the sparse regularization

研究代表者

松井　秀俊（Matsui, Hidetoshi）

滋賀大学・データサイエンス学部・准教授

研究者番号：９０６３３３０５

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：観測、測定された大規模なデータを集約し、有効な情報を抽出するための統計的モデリング手法の開発に取り組んだ。経時測定データに滑らかな関数を当てはめて、関数化データ集合を分析対象にする方法は関数データ解析とよばれている。本研究では、関数データとして与えられた説明変数と目的変数との関係を表現する新たな関数回帰モデルと、関数データを判別するためのロジスティック回帰モデルの推定方法を提案した。特に、関数回帰モデルとしては、非線形な関係性を捉えるための２次回帰モデルを、関数ロジスティック回帰モデルに対しては、スパース推定を用いて決定境界を選択するための方法を提案した。

研究成果の学術的意義や社会的意義
計測機器の発達やデータサイエンス分野の学問の発展に伴い、より多くのデータが測定、保有されるようになってきた。このようなデータの中には、時間の経過に伴い繰り返して計測される形式のデータも多く含まれる。関数データ解析は、このような形式のデータから解釈の容易な情報を抽出するための有効な手法である。本研究の開発により、説明変数、目的変数間の非線形な関係を捉え、かつその関係性を定量化するための方法を得ることができた。また、経時測定データの判別問題において、どの変数がどの群の判別に寄与しているかといった情報を、スパース推定を用いて浮かび上がらせることができた。

研究成果の概要（英文）：We developed statistical modeling procedures to extract useful information from the collection of repeatedly measured data. The basic idea for functional data analysis is to transform repeatedly measured individuals into smooth functions and then to analyze a set of functional data.
In this work, we proposed a new functional regression model that represents the relationship between predictor and response variables given as functional data, and a method for estimating a logistic regression model to classify functional data.
In particular, we introduced a quadratic regression model for a functional predictor and a functional response to capture the nonlinear relationships between variables. We also introduced a method for selecting decision boundaries for a functional logistic regression model using a sparse regularization.

研究分野：　統計科学

キーワード：　関数データ解析　スパース推定　回帰分析　モデル選択

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

近年の計算機の発展とそれに伴う計測規模の拡大により、データの構造が多種多様化・複雑化の傾向を見せている。特に、図 1 に点で示すデータのように、時間の経過に伴い繰り返し測定される形式のデータが、さまざまな分野で計測されるようになってきた。繰り返し測定データに対しては、主に次の 2 点の性質から、しばしば古典的な多変量解析手法を適用することが困難となる。

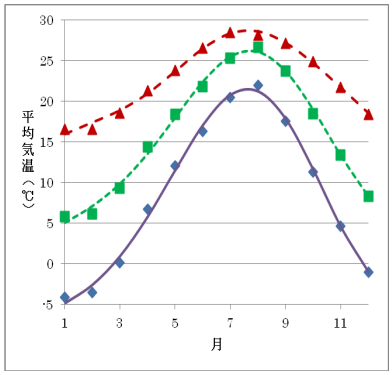

図 1: 3 都市の月別平均気温

- 観測時点数の増加によりデータの次元が増大する
- 観測の欠損や観測時点または時点数の不一致が生じる

繰り返し測定データの分析のためのアプローチの１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、図1の曲線のように、得られた関数集合をデータとして扱う解析方法である。これにより、上記2点の問題点を解消できる。関数データ解析に基づく統計的モデリング手法の開発は発展途上であるため、予測や解釈可能性の観点からより良いモデルや推定法の開発が求められている。

２．研究の目的

観測、測定された大規模なデータを集約し、有効な情報を抽出するための統計的モデリング手法を開発するために、2 点のテーマを軸に研究を進めた。

(1) 関数データに基づく多群判別モデルに対して**スパース正則化**を適用することで、繰り返し測定データの判別における情報縮約法を確立する。

(2) 関数回帰モデルにおいて交互作用を考慮に入れたモデリング手法について研究することで、これまであまり研究されて来なかった、関数データの説明変数同士の関連を考慮に入れたモデルの推定・選択方法を構築する。

本研究を遂行することで、従来の方法では発見できなかった、経時測定データを含んだデータ間の複雑な関係を浮かび上がらせる。

３．研究の方法

(1) 関数データに基づく多群判別モデルに対する推定に、スパース正則化の一種である bi-level selection を用いる。Bi-level selection は本来、カテゴリ変数のようにグループ化されて扱うべき説明変数群において、グループとしてだけでなく、個々の変数も同時に選択するために導出された制約であり、このような性質を持った制約は既にいくつか提案されている。本研究では bi-level selection の制約を関数データに対する多群ロジスティック回帰モデルに応用し、関数データとして与えられた変数と判別境界の選択を同時に行うことができるような制約を構築する。

(2) 説明変数と目的変数が共に関数データとして与えられた関数回帰モデルにおいて、交互作用項を導入したモデルを構築する。関数回帰モデル推定の過程では、説明変数およびその係数関数を基底関数展開によって表現する方法が広く知られている。従って、交互作用項に対してもスプラインなどを用いた基底関数展開を導入する。そして、正則化法に基づく推定法を導出することで、安定した推定量の導出をめざす。

４．研究成果

(1) 関数データに基づくロジスティック回帰モデルにおいて、スパース推定法の一つである

sparse group lasso を用いて推定することで、関数データの変数選択と決定境界の選択を同時に行う方法を提案した。推定アルゴリズムとして、座標降下法を提案手法に応用したものを導出した。また、モデル推定に伴う調整パラメータの値を、モデル評価基準を用いて決定した。提案手法を、人工データおよび実データ（遺伝子発現データ）の解析に適用することで、有効性を検証した。成果は論文として採択済である。また、関数データに基づく線形回帰モデルにおいて、変数選択を効率的に行うためのアルゴリズムを新たに提案した。多変数の説明変数において、変数を複数のグループに分割し、各サブグループで変数選択を行う処理を繰り返すことで、全変数で変数選択を行う場合よりも計算コストを大幅に抑えることができる。本手法は、スパース推定に基づく変数選択などに比べて優れた変数選択の精度を与えた。この研究成果も論文として採択済である。

(2) 前年度から遂行していた、関数データに基づく二次回帰モデリングについては、推定法を最尤法の枠組みから正則化法へ拡張した。そして、数値実験を行うことで既存手法に比べて優れた予測結果を得た。研究成果は論文として執筆中である。このモデルに対して、線形回帰モデルの枠組みを拡張させ、2 次の項を導入した関数応答 2 次回帰モデルを新たに提案した。また、提案したモデルを正則化法に基づき推定する方法およびアルゴリズムを導出し、さらに、推定されたモデルを評価するための基準として、情報量規準などを、既存のモデルに対するものを応用することで導出した。以上の手法を人工データおよび実データの分析に適用することで、既存のモデルよりも優れた結果を得た。関数回帰モデルに関しては、これまでに非線形モデルへの発展も多く報告されているが、得られるモデルの解釈が困難であることが多い。これに対して 2 次回帰モデルは、推定結果から、説明変数と目的変数の関係性として比較的解釈が容易な結果を与えることができる。

(3) 変量間の相関を考慮に入れた多変量関数データのクラスタリング手法を提案した。従来の関数データクラスタリングでは、変量間で独立に関数化を行っていた。これに対して本研究では、非線形混合効果ジョイントモデルおよび pairwise fitting を適用することで、変量間の相関を考慮に入れつつ、なおかつパラメータ数を削減することで、効率的にモデルを推定できる。本研究ではさらに、得られた関数データ集合に対して自己組織化マップを適用することで関数データに対するクラスタリングを行った。そして、提案手法を台風データの分析の適用し、強い台風と弱い台風の性質を分類した。

(4) 関数データに基づく線形回帰モデルにおいて、変数選択を効率的に行うためのアルゴリズムを新たに提案した。多変数の説明変数において、変数を複数のグループに分割し、各サブグループで変数選択を行う処理を繰り返すことで、全変数で変数選択を行う場合よりも計算コストを大幅に抑えることができる。本手法は、スパース推定に基づく変数選択などに比べて優れた変数選択の精度を与えた。

(5) これまでに身に着けた関数データ解析に関する知識や研究成果を総合報告論文として報告し、和文誌へ採択された。これまでに関数データ解析に関する総合報告の和文論文は無かったため、国内における関数データ解析の普及に貢献できるのではと考えている。

５．主な発表論文等

〔雑誌論文〕（計 5 件）

① Matsui, H. (2019). Quadratic regression for functional response models. Econometrics and Statistics. **採択済．査読あり．**
doi:10.1016/j.ecosta.2018.12.003

② 松井秀俊 (2019). **関数データに基づく統計的モデリング. 統計数理, 採択済. 査読あり．**

③ Misumi, T., Matsui, H., and Konishi, S. (2018). Multivariate functional clustering and its application to typhoon data. Behaviormetrika 46, 163-175. **査読あり.**
doi:10.1007/s41237-018-0066-8

④ Matsui, H. (2018). Sparse group lasso for multiclass functional logistic regression models. Communications in Statistics - Simulation and Computation, **採択済. 査読あり.**
doi:10.1080/03610918.2018.1423693

⑤ Smaga, Ł., Matsui, H. (2018). A note on variable selection in functional regression via random subspace method. Statistical Methods & Applications 27(3), 455-477. **査読あり.**
doi:10.1007/s10260-018-0421-7

〔学会発表〕（計 24 件）

① **松井秀俊. 関数データに基づく交互作用モデルとその推定. 科研費シンポジウム「予測モデリングとその周辺 -機械学習・統計科学・情報理論からのアプローチ-」, 成蹊大学, 2018 年 11 月.**

② **三角俊裕, 松井秀俊, 小西貞則. 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, 2018 年 9 月.**

③ Matsui, H. Quadratic regression for function-on-function models. The 5th Institute of Mathematical Statistics Asia Pacific Rim Meeting (IMS-APRM 2018), National University of Singapore. Jun. 2018. (Invited talk: TCP02)

④ Matsui, H. Selection of variables and decision boundaries in functional logistic regression model. Conference of the International Federation of Classification Societies (IFCS-2017), Tokai University, Japan. Aug. 2017. (Invited talk: SP10)

⑤ **松井秀俊. 関数データに基づく経時測定データの分析. 科研費シンポジウム「複雑な生命現象を読み解くための大規模データとモデリング」, 久留米シティプラザ（福岡県久留米市）, 2016 年 11 月.**

〔図書〕（計 5 件）

① **竹村彰通, 姫野哲人, 高田聖治 (編), 和泉志津恵, 市川治, 梅津高朗, 北廣和雄, 齋藤邦彦, 佐藤智和, 白井剛, 高田聖治, 竹村彰通, 田中琢真, 姫野哲人, 松井秀俊 (著) (2019). データサイエンス入門（データサイエンス大系）. 学術図書出版社. 207 ページ.**

② Miodrag Lovric (編), **日本統計学会（翻訳）(2018). 統計科学百科事典. 丸善出版. 2130 ページ.**

③ **滋賀大学データサイエンス学部（編）(2018). 大学生のためのデータサイエンス(I) ーオフィシャルスタディノート. 日本統計協会. 130 ページ.**

④ **川野秀一, 松井秀俊, 廣瀬慧 (2018). スパース推定法による統計モデリング（統計学 One Point）. 共立出版. 155 ページ.**

⑤ Peter Flach (著), **竹村 彰通 (監修, 翻訳), 田中 研太郎, 小林 景, 兵頭 昌, 片山 翔太, 山本 倫生, 吉田 拓真, 林 賢一, 松井 秀俊, 小泉 和之, 永井 勇 (翻訳) (2017). 機械学習—データを読み解くアルゴリズムの技法. 朝倉書店. 392 ページ.**

〔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/home

６．研究組織

(1)研究分担者　なし

(2)研究協力者
研究協力者氏名：三角　俊裕
ローマ字氏名：Toshihiro Misumi

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。