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　２９　年　　６　月　１２　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３００１６

研究課題名（和文）時間・空間依存性を考慮した超多変量関数データ解析法の開発と生命科学への応用

研究課題名（英文）Multivariate functional data analysis for temporally and spatially dependent data and its application to life science

研究代表者

山本　倫生（Yamamoto, Michio）

京都大学・医学研究科・講師

研究者番号：５０７２１３９６

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：データ収集技術・環境の飛躍的な向上に伴って、統計科学で扱うデータは複雑・膨大化してきている。そのようなデータに対する解析のうち、例えば経時的に測定されるデータをある種の関数とみなして解析を行う方法を関数データ解析と呼ぶ。本研究では多変量関数データに対して、次元縮小とクラスタリングを同時に達成する新たな方法を開発した。また、関連する研究として、高次元二値データに対する次元縮小を伴うクラスタリング法を開発した。さらに、アウトカムのクラスタリングだけでなく、説明変数によるクラスター構造の予測を考慮した新たな分析手法を開発した。

研究成果の概要（英文）：Due to the recent advances in data collection and storage, data sets for statistical analysis have become complex and enormous. In the analysis of repeated measures data, for example, the data are often considered as a certain function, and such an analysis is called functional data analysis. In this study, I developed a new clustering method that conducted clustering and dimension reduction of multivariate functional objects simultaneously. Related to the method, I developed another clustering method with dimension reduction for multivariate binary data. In addition, I developed a new clustering method that identified a cluster structure of outcome variables and predicted cluster memberships of future individuals based on explanatory variables.

研究分野：　統計科学

キーワード：　クラスタリング　関数データ　次元縮小

１．研究開始当初の背景

データ収集技術・環境の飛躍的な向上に伴って、統計科学で扱うデータはこれまでにない程、複雑・膨大化している。例えば、脳認知科学分野では、脳全体の血流の変化を把握可能な fMRI（functional magnetic resonance imaging）技術が、ヒトの認知行動・疾病と脳との関連についての研究に利用されている。fMRI では、生物の脳や脊髄の活動に関連した血流動態反応（BOLD 信号）を視覚化する方法で、認知科学研究の中心的なデータ形式の一つである。一つの脳に対して数万単位の変数（各ボクセル上での BOLD 信号、一般に量的変数）を時系列で得ることが可能である。特に、疾病の予測に利用される臨床バイオマーカーの探索を目的として、fMRI と疾病等との関連を調べる際には、患者の分類（クラスタリング）と関連のあるボクセルの探索（特徴抽出）が重要な目的となる。

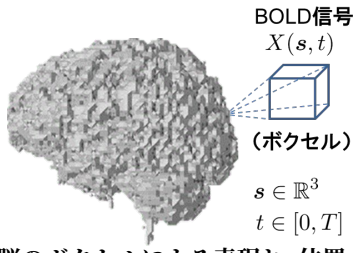

図 1:脳のボクセルによる表現と、位置 s、時点 t におけるボクセルの特徴量（BOLD 信号）の関数データ表現

このような fMRI データに対して、関数データ解析によるアプローチが試みられていた。関数データとは、通常得られる離散・連続データを一種の「関数」として扱われるデータを指す。関数データ解析では、一つの変数（例えば体重の経時測定データ）のみを扱うことが多い。しかし、fMRI データの解析においては、多数のボクセル上で得られる時系列データを扱うことから、多変量の関数データを扱う必要がある。また、一般に、関数データ全体をそのままクラスタリングに利用する場合、クラスター構造に関係しない誤差を多量に含んでしまい、クラスター構造の推定が難しくなることが判明していた。そこで、多変量関数データからクラスター構造に寄与する特徴抽出を行うとともに、データに内在するクラスター構造を推定する方法が、研究代表者により複数開発されていた。

しかし、それらの方法を用いて臨床バイオマーカー探索のための fMRI データの解析を行う場合、次の 3 つの問題が生じる。
(1) 各ボクセルは空間的に依存しており、それを無視して単なる多変量関数データとして扱うと、誤った結論を導きかねない。
(2) fMRI データのような膨大な数の変数からなる関数データに対しては、一般的な高次元データの問題と同様に、パラメータの推定が不可能になるか、少なくとも推定が不安定になるなど、多くの問題が生じる。
(3) 臨床バイオマーカーとして、一般に、性別や体重といった背景情報および遺伝子情報などのいわゆる予後因子を用いることが多い。しかし、既存の方法では、予後因子などの通常のベクトルデータと関数データのハイブリッドなデータに対しては適用できない。

２．研究の目的

上記を背景に、fMRI データによる臨床バイオマーカーの探索を行うための統計解析法という観点から、(1) 時間と空間の両方に依存するデータ、(2) 変数の数が膨大な多変量関数データ、(3) ハイブリッドなデータに適用可能な方法を開発することが目的である。
(1) まず、時間と空間両方の依存性を表現可能なデータ構造を導入する。そして、その新たなデータ構造を用いて既存の方法を再定式化し、モデルの推定アルゴリズムを開発する。また、提案モデルの良さを示す性質として、損失関数の一致性を導出する。さらに、シミュレーションによって、データが時間・空間依存性をもつ場合に、提案モデルによる市のクラスター構造の再限度を評価する。
(2) 制約付き最適化手法を用いて、上記の(1)で開発した手法を変数の数が非常に多い多変量関数データに適用可能なモデルへと拡張する。その際に、関数データに対するスパース制約を用いてモデルを定式化し、パラメータの推定アルゴリズムを開発する。また、人工データおよび実際の fMRI データを利用して、提案モデルの性能を評価する。
(3) ハイブリッドデータに適用するために、データが存在する空間を拡張し、関数空間と実数空間の直積を新たなデータ空間と定義する。そして、上記の(2)で開発した手法を新たなデータ空間上で数学的に定式化し、モデルの推定アルゴリズムを開発する。また、実際の fMRI データと予後因子のデータに提案モデルを適用し、臨床バイオマーカーの探索を行う。

３．研究の方法

(1) 時間・空間依存性を考慮した関数データのクラスタリングを行うために、誤差成分が空間依存性を持つシンプルなモデルを導入した。しかし、提案方法のような次元縮小を伴うクラスタリング法では、データにある種の構造（disturbing structure と呼ぶ）が存在する場合、クラスター構造を正しく推定することができないことが判明した。そこで、まずは空間依存性をもたない時間のみに依存するデータを対象に、上記の disturbing structure を数学的に定式化し、データの構造に合わせて柔軟なモデリングが可能なク

ラスタリング法（FGRC 法）を開発した。FGRC 法では、その最適化問題を無限次元空間上のデータに対する低次元空間上のクラスター内／クラスター間分散の大きさに対する制約付きの次元縮小問題として定式化した。このような制約付き最適化問題として定式化することにより、上記の disturbing structure を適切に回避できることが予想された。
(2) スパース制約を利用した次元縮小を伴うクラスタリング法の考え方を発展させ、画像データや SNP データなど、高次元二値データを特徴量にもつ対象のクラスタリング法を開発した。具体的には、いわゆる潜在クラス分析モデルの拡張として、低次元空間上にクラスター中心の布置を仮定し、各データが、これらクラスター中心によって生成されているという因子分析モデルを導入した。このような低次元空間を仮定するモデルは近年いくつか提案されているが、それらは低次元空間上の点（因子）に対してパラメトリックなモデルを仮定するものがほとんどである。その場合、パラメータ推定の際に積分計算を必要とするなど、実際の利用に際して問題が生じる。そこで、本研究では低次元空間上の点に対して分布の仮定を置かないノンパラメトリックなモデルを採用した。さらに、低次元空間が一部の変数によって規定されると仮定し、これをスパース推定によって表現した。このような低次元空間を仮定することにより、データに内在するクラスター構造に強い影響を与えている変数を特定することが容易となると考えられた。
(3) 予備的な実験により、上記で考案した関数データのクラスタリング手法は、データに内在するクラスター構造を把握するためには有用と考えられるが、ヒトの認知行動や疾病と関連のあるクラスター構造を推定するための方法としては不十分であることがわかった。そこで、まずは通常の有限次元データを対象として、認知行動や疾病などのアウトカムと関連のあるクラスター構造を推定するための方法を開発した。具体的には、最適化のための目的関数を、アウトカムのクラスター構造を表現する部分と、説明変数によるクラスター構造の予測精度を表現する部分の凸結合で表現した。このような定式化により、分析者が現象やデータに合わせてアウトカムのクラスターの良さや予測の精度を柔軟に検討できると考えられた。

４．研究成果

(1) disturbing structure を数学的に定式化し、データがある種の相関構造を持つ場合、既存の関数データクラスタリング手法である FPCK 法や FFKM 法がクラスター構造の推定に失敗することを示した。シミュレーションおよび実際のデータを用いて、開発した手法（FGRC 法）が既存の方法に比べて真のクラスター構造の推定精度が良いことを示した。また、disturbing structure が存在する場合に、既存手法と FGRC 法とで推定される低次元空間が異なることを確認し、その結果得られる異なるクラスター構造の解釈の方法を提示した。
(2) 高次元二値データのクラスタリングに関して、シミュレーションや実際のデータを用いて既存のクラスタリング手法と比較し、クラスター構造の推定精度は既存手法の中での最も良い方法と同程度であることがわかった。さらに、提案方法では、推定された低次元空間上に布置された対象を利用したクラスター構造の解釈や、クラスター構造に影響を与える変数の解釈を行うことができる点で、応用上有用な方法であることがわかった。なお、提案手法を実装したソフトウェア R のパッケージ cbird を開発し、CRAN 上に公開した。
(3) 説明変数によるクラスター構造の予測精度を考慮したクラスタリング法について、交互最小二乗法による最適化アルゴリズムを開発した。また、人工データを用いたシミュレーションにより、アウトカムのクラスター構造推定精度と説明変数によるクラスター構造の予測精度について、通常のクラスタリング手法と比較して性能がよいことを示した。さらに、既存の次元縮小を伴うクラスタリング手法と同様に、推定されたクラスター中心の一致性、および、損失関数の平均 2 乗誤差に対する非漸近的な上界を導出した。

当初の計画では空間的に依存性をもつデータに対するモデルを提案する予定であったが、一般に、クラスター構造の推定に悪影響を与える構造が存在することが判明したため、まずはその解決策となる方法を開発した。今後は上記の研究成果をベースとして、空間的に依存性をもつデータに対するモデルを開発する予定である。

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 9 件）

①Yamamoto, M. Hirose, K. and Nagata, H. Graphical tool of sparse factor analysis. Behaviormetrika, 44, 229-250, 2017.（査読あり）
DOI: 10.1007/s41237-016-0007-3

②Yamamoto, M. and Hwang, H. Dimension-reduced clustering of functional data via subspace separation. Journal of Classification, 2017.（査読あり）
DOI: 10.1007/s00357-017-9232-z

③Choi, J.Y., Hwang, H., Yamamoto, M.,

Jung, K., and Woodward, T.S. A unified approach to functional principal component analysis and functional multiple-set canonical correlation. Psychometrika, 2016.（査読あり）
DOI: 10.1007/s11336-015-9478-5

④Yamamoto, M. and Hayashi, K. Clustering of multivariate binary data with dimension reduction via L1-regularized likelihood maximization. Pattern Recognition, 48, 3959-3968, 2015.（査読あり）
DOI: 10.1016/j.patcog.2015.05.026

⑤Yamamoto, M. and Terada, Y. Functional factorial K-means analysis. Computational Statistics and Data Analysis, 79, 133-148, 2014.（査読あり）
DOI: 10.1016/j.csda.2014.05.010

〔学会発表〕（計 22 件）
① Yamamoto, M. Dimension-reduced clustering of functional data via variance-penalized optimization. The 9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016). 2016 年 12 月 10 日. Seville (Spain).

②Yamamoto, M., Kawaguchi, A., and Hwang, H. Predictive clustering using a component- based approach. The 22nd International Conference on Computational Statistics (COMPSTAT 2016). 2016 年 8 月 23 日. Oviedo (Spain).

③Yamamoto, M. and Terada, Y. Canonical correlation analysis for multivariate functional data. The 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics). 2015 年 12 月 13 日. London (the United Kingdom).

④ Yamamoto, M. and Kawaguchi, A. A component-based approach to find outcome-related clusters. The 80th Annual meeting of the Psychometric Society (IMPS 2015). 2015 年 7 月 13 日. Beijing (China).

⑤Yamamoto, M. A simultaneous analysis of dimension reduction and clustering with correlated error variables. The 2015 conference of the International Federation of Classification Societies (IFCS 2015). 2015 年 7 月 8 日. Bologna (Italy).

⑥ Yamamoto, M. and Hayashi, K. Simultaneous analysis of clustering and dimension reduction for binary variables with application to biomedical data. The 27th International Biometric Conference. 2014 年 7 月 11 日. Florence (Italy).

〔その他〕
ソフトウェアパッケージ cbird
https://cran.r-project.org/web/packages/cbird/index.html

６．研究組織
(1)研究代表者
山本　倫生（YAMAMOTO, Michio）
京都大学・大学院医学研究科・講師
研究者番号：50721396