関数データに対する部分空間クラスタリング法とその性質

大阪大学 大学院基礎工学研究科，理化学研究所 革新知能統合研究センター 寺田 吉壱
岡山大学 大学院環境生命科学研究科 山本 倫生

計量化学分野における近赤外線分光法に関連するデータ，運動に関連する軌道データ，経時測定データのように連続的に変化するデータや連続的・断続的に記録されるデータは，従来の多変量データとは異なる構造をもっている．以下の図は，Kalivas (1997) で紹介されている 100 個の小麦サンプルに対する近赤外線分光法 (NIR) によって得られたスペクトルデータである．このように変数の並びに意味があるデータに対しては，背後のデータ発生機構として，実数空間上の確率分布を考えるよりも，ある（有界な）領域や区間上のランダムな関数を考える方が自然である．そして，ある領域や区間上で連続的・断続的に観測されたデータをランダムな関数や確率過程の実現値として捉えたデータ解析は関数データ解析 (FDA) と呼ばれ，統計科学分野を中心に盛んに研究が進められている．

本稿では，関数データ解析の中でも，部分空間を利用したクラスタリング法について考える．関数データに対するクラスタリング法では，関数主成分分析 (FPCA) に基づくものが多く挙げられる．最も単純な方法として，関数データを FPCA により有限次元の部分空間へ射影した後，多変量解析におけるクラスタリング法を適用する 2-step approach が挙げられる．Yamamoto (2012) や Yamamoto and Terada (2014) では，FPCA によって得られる部分空間が必ずしもクラスタリングに適していないことを指摘し，クラスタリングに適した部分空間を見つける方法を提案している．また，部分空間を用いた異なるアプローチとして，Chiou and Li (2007) では，クラスタごとに特有の部分空間を考え，射影距離の意味で最も近いクラスタ部分空間へ対象を割り振るというクラスタリング法が提案されている．

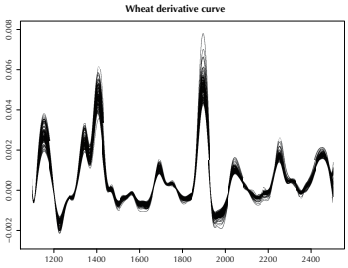

図 1: 関数データの例

Delaigle and Hall (2012) によって，教師あり判別問題に対しては関数データの実数空間 $\mathbb{R}$ への射影が有効であることが明らかになっているため，これらの部分空間への射影を考えたクラスタリング法は関数データに対して有効であると考えられる．しかし，これらの方法は平均構造に大きな違いがある場合以外では上手く機能しないことが多い．図 1 で示されるデータには，一見クラスタ構造を確認することはできないが，別途測定された水分量の観点からは明確に 2 つのクラスタに分離することができる．教師あり判別問題ではこれらのクラスを非常に高い精度で分類することができるが，既存のクラスタリング法ではこのデータの背後にあるクラスタ構造を捉えることができない．

本発表では，関数データの背後に隠れたクラスタ構造を反映した部分空間を求める新しい部分空間クラスタリング法を提案し，その理論的性質を明らかにする．また，数値実験や実データ解析を通じて，既存手法では捉えられないクラスタ構造が提案手法を用いて発見できることを示す．

参考文献

[1] Chiou, J.M. and Li, P.L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B*, **69**, 679–699.

[2] Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B*, **74**, 267–286.

[3] Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory System*, **37**, 255–259.

[4] Yamamoto, M. (2012). Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification*, **6**, 219–247.

[5] Yamamoto, M. and Terada, Y. (2014). Functional factorial k-means analysis. *Computational Statistics & Data Analysis*, **79**, 133–148.