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

滋賀大学 データサイエンス学部 / JST さきがけ　松井 秀俊
アステラス製薬株式会社 データサイエンス部　三角 俊裕
大正製薬株式会社 データサイエンス部　横溝 孝明
中央大学 理工学部　小西 貞則

1. 概要

複数の個体に対して経時的に観測・測定されたデータを関数化処理することで得られる関数化データ集合に基づく解析を行う方法は関数データ解析とよばれ，多くの分野でその有用性が報告されている (例えば，Ramsay & Silverman, 2005)．本報告では，基底関数展開に基づく非線形混合効果モデルを適用することで，経時観測データを平均効果関数および個体ごとの変動を表したランダム効果関数を用いて関数データとして表現する．そして，ランダム効果関数集合を対象としたさまざまなクラスタリング手法を適用する．提案した手法を気象学などの分野のデータ解析に適用し，その有効性を検証する．なお，本報告は松井他 (2016) の結果に基づくものである．

2. 非線形混合効果モデル

いま，N 個の個体のある特性について，それぞれが経時的に観測されたデータを $\{(t_{ij}, y_{ij})\ ;\ j = 1, \ldots, n_i\}$ $(i = 1, \ldots, N)$ とする．ここで，t_{ij} は区間 $\mathcal{T} \subset \mathbb{R}$ で観測された第 i 個体の j 番目の観測時点であり，y_{ij} は時点 t_{ij} における観測値，n_i は第 i 個体の観測時点数である．t_{ij} や y_{ij} については，個体ごとにその値や数が異なっていたり，欠損があったりしてもよい．このとき，基底関数展開に基づく非線形混合効果モデルを用いることで，t_{ij} と y_{ij} の関係は次で表される（Rice & Wu, 2001）．

$$y_{ij} = \sum_{k=1}^{m_f} \beta_k b_k^{(f)}(t_{ij}) + \sum_{l=1}^{m_r} \gamma_{il} b_l^{(r)}(t_{ij}) + \varepsilon_{ij} = \boldsymbol{\beta}' \boldsymbol{b}^{(f)}(t_{ij}) + \boldsymbol{\gamma}_i' \boldsymbol{b}^{(r)}(t_{ij}) + \varepsilon_{ij}. \tag{1}$$

ここで，$\boldsymbol{b}^{(f)}(t) = (b_1^{(f)}(t), \ldots, b_{m_f}^{(f)}(t))'$, $\boldsymbol{b}^{(r)}(t) = (b_1^{(r)}(t), \ldots, b_{m_r}^{(r)}(t))'$ はそれぞれ平均効果項，ランダム効果項に対する基底関数からなるベクトル，$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m_f})'$ は平均効果項の係数パラメータベクトル，$\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{im_r})'$ はランダム効果項の係数で，平均ベクトル $\mathbf{0}$，分散共分散行列 Γ の m_r 変量正規分布に従う確率変数ベクトルとする．また，$\varepsilon_{i1}, \ldots, \varepsilon_{in_i}$ は互いに独立に平均 0，分散 σ^2 の正規分布に従う観測誤差とする．平均効果係数 $\boldsymbol{\beta}$ やランダム効果係数 $\boldsymbol{\gamma}_i$ については，EM アルゴリズムを適用することでそれぞれ推定値と予測値を得る．

3. 関数データクラスタリング

混合効果モデルに基づき関数化することで得られるランダム効果関数 $\{\boldsymbol{\gamma}_i' \boldsymbol{b}^{(r)}(t);\, i = 1, \ldots, N\}$ を対象とした関数データクラスタリングを行う．いま，$\xi_i(t) = \boldsymbol{\gamma}_i' \boldsymbol{b}^{(r)}(t)$ とし，基底関数の積の積分を要素にもつ行列を $W = \int_{\mathcal{T}} \boldsymbol{b}^{(r)}(t) \boldsymbol{b}^{(r)}(t)' dt$ とおくと，関数 $\xi_i(t)$ の関数空間上での L_2 ノルムは $\|\xi_i\|_{L_2}^2 = \boldsymbol{\gamma}_i' W \boldsymbol{\gamma}$ となる．行列 W は正値対称行列であるため，ある上三角行列 U に対してコレスキー分解 $W = UU'$ が適用でき，ゆえに $\xi_i(t)$ の L_2 ノルムは，$\|\xi_i\|_{L_2}^2 = \|U\boldsymbol{\gamma}_i\|_2^2$ のようにユークリッドノルムを用いて表される．このようにして得られたベクトル集合 $\{\tilde{\boldsymbol{\gamma}}_i = U\boldsymbol{\gamma}_i;\, i = 1, \ldots, N\}$ を対象としたクラスタリングを行うことで，関数空間上の距離を保存した分析が可能となる (Kayano *et al.*, 2010)．

参考文献

Kayano, M., Dozono, K., and Konishi, S. (2010). Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. *J. Classification* **230**, 211–230.

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 **44**, 25–45.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis* 2nd ed. Springer.

Rice, J. and Wu, C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259.