Dissimilarity for functional data clustering based on smoothing parameter commutation

研究方法（一）期末報告

統碩一 112354031
賴真萍

Contents

介紹

- We propose a novel and easy method to implement **dissimilarity measure for functional data clustering based on smoothing splines and smoothing parameter commutation.**
- Our method takes into account the estimation uncertainty using smoothing parameter commutation and is not strongly affected by outliers. It can also be used for **outlier detection.**

Functional Data Clustering Methods

Functional data clustering :

1. dissimilarity-based methods
2. decomposition-based methods
3. model-based methods

Dissimilarity-based methods : using pointwise dissimilarities between pairs of subjects, are the most straightforward approach.

Smoothing Spline

Assume that the curve of the i-th subject is observed as a set of measurements $\{y_{i1}, y_{i2}, \ldots, y_{ik_i}\}$ contaminated by noises at distinct finite time points $\{t_{i1}, t_{i2}, \ldots, t_{iK_i}\}$in an interval $[T_L, T_U]$ according to the model

$$y_{ik} = f_i(t_{ik}) + \varepsilon_{ik} \, , k = 1, 2, \ldots, K_i \, , i = 1, 2, \ldots, n$$

A reasonable estimation of f_i (a smoothing spline $\hat{f}_i(\cdot; \lambda)$) is to

$$min \, \frac{1}{K_i} (y_i - f_i)^T (y_i - f_i) + \lambda \int_{T_L}^{T_U} (f_i''(t))^2 \, dt$$

Find λ

Mixed-effects model to find λ:

$$y_i = X_i\beta_i + u_i + \varepsilon_i$$

Under the **Gaussian assumption** for ε_i and u_i, the two variance components $\sigma_u^2$and σ^2can be determined based on the restricted maximum likelihood method (REML), so that $\boldsymbol{\lambda}$ **is also determined**.

Smoothing Parameter Commutation dissimilarity

<u>Concept</u> **:** If the 'true' f_i and f_j are similar, it is expected that $\hat{f}_i$ and $\hat{f}_j$ should be close, given an identical smoothing parameter λ.

The dissimilarity between subjects i and j:

$$d_{i,j} = \frac{1}{2}\left\{ \left[\int_{T_L}^{T_U} \left(\hat{f}_i(t;\hat{\lambda}_i) - \hat{f}_j(t;\hat{\lambda}_i)\right)^2 dt\right]^{\frac{1}{2}} + \left[\int_{T_L}^{T_U} \left(\hat{f}_i(t;\hat{\lambda}_j) - \hat{f}_j(t;\hat{\lambda}_j)\right)^2 dt\right]^{\frac{1}{2}} \right\}$$

we call it a Smoothing Parameter Commutation dissimilarity

結論

1. data observed at irregular time points can be handled directly.
2. The concept of the proposed dissimilarity measure is simple and easy to implement.
3. The dissimilarity also serves as a useful tool for outlier detection.

Thank you