UNIVERSITY OF LEEDS

Classification of disease evolution based on longitudinal metabolomics data

A new k-means clustering algorithm for functional data

Supervisors: Jeanine Houwing-Duistermaat, Emanuele Paci and Haiyan Liu

Yusheng Liang

201325291

Aim and Data Description

UNIVERSITY OF LEEDS

Aim

This project bases on a new k-means clustering algorithm with functional samples' derivative information(Meng et al., 2018) to assign 18 patients into 3 clusters, “Primary function”, “delay graft function” and “acute rejection” (Krivov et al., 2014).

Data Description

The data is the evolution of ^{1}H NMR spectra from 18 patients' erythrocyte extracts of blood who were undergoing kidney transplantation.

Each patient has up to 9 samples which were taken before surgery and daily up to one week after.

Functional Data

UNIVERSITY OF LEEDS

data pre-processing

I normalize the original data to the total sum of the spectral intensities and then calculate the average with a interval size of 0.16 ppm. After that, these averages within each interval were logarithmically transformed as $I_k = \log(10^6 I_k + 1)$(Krivov et al., 2014).

ppm	[0, 0.16)	[0.16, 0.32)	[0.32, 0.48)	[0.48, 0.64)	[0.64, 0.80)	[0.80, 0.96)
k1 pre	0.017201875	0.049864461	0.090524084	0.168729507	0.310384468	0.533306779
k1 int	-0.008570977	-0.017500336	-0.008777328	0.010900726	0.072171404	0.440929964
k1 post	0.024259156	0.046388461	0.09048075	0.163218022	0.293168031	0.58440538
k1 01	-0.000817262	-0.003775021	0.00732207	0.038675323	0.124647038	0.572603933
k1 03	0.040649264	0.100401186	0.176809451	0.290840931	0.450802113	0.735228078
k1 04	0.031376128	0.039564006	0.061431381	0.128506092	0.244638398	0.581555431
k1 07	0.004352273	-0.007994192	-0.005684271	0.017745552	0.115300386	0.615833703

From Discrete to Functional data

UNIVERSITY OF LEEDS

A function x(t) defined with basis systems is expressed in mathematical notation as: $x(t) = \sum_{i=1}^{K} \phi_i(t) c_i$

- Φ_k is basis systems , and then combine them linearly.Because of the data is aperiodic, so Bspline basis was choosen. Splines are piecewise polynomials. Order four splines are often used, consisting of cubic polynomial segments (degree three), The number of basis functions K = order + number o f interior knots.
- The parameters c_i are the coefficients of the function x(t) (Ramsay et al., 2009).

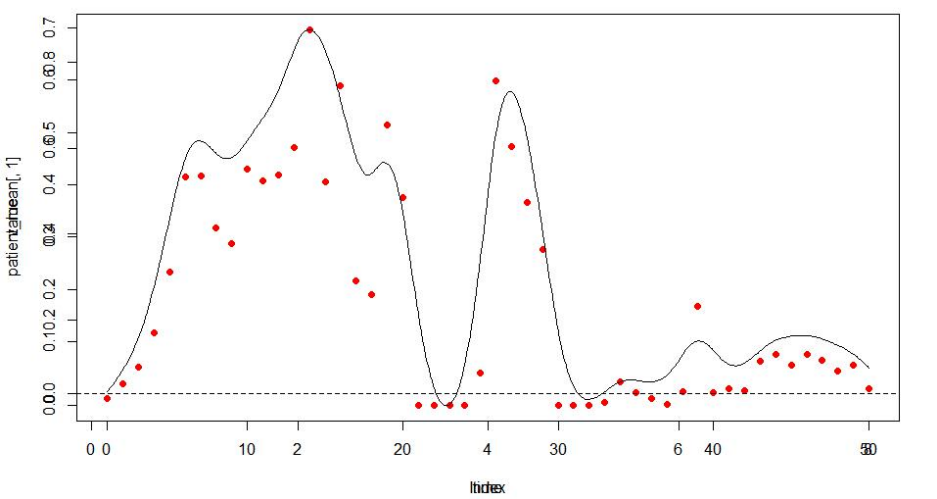

Fig 1. The point and functional object for patient 1.

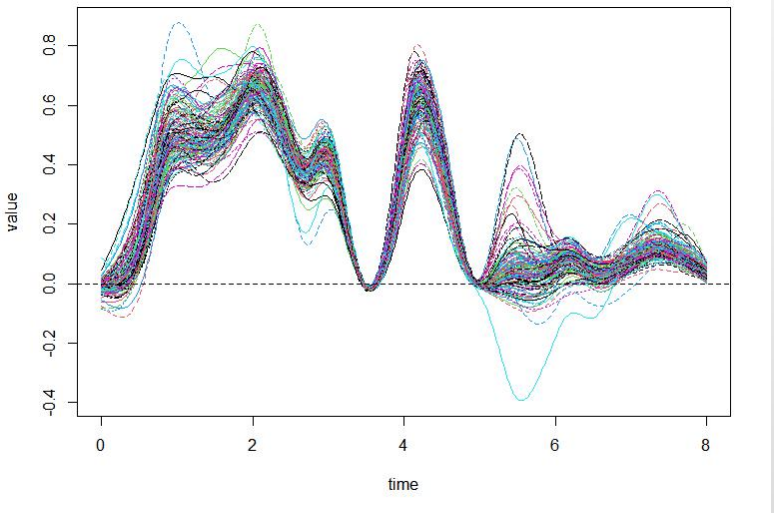

Fig 2. The functional objects for 18 patients.

K-means Algorithm with Derivative

UNIVERSITY OF LEEDS

The new distance for the k-means algorithm:

$$\mathrm{d}(x_i(t), x_j(t)) = \sqrt{\sum_{l=1}^{p}\left[\int_T (x_i^l(t) - x_j^l(t))^2\,dt + \int_T (Dx_i^l(t) - Dx_j^l(t))^2\,dt\right]} \qquad (1)$$

Update cluster centroids:

$$\mu_{\mathrm{s}}^l(t) = \frac{\sum_{i:C_i^{(m)}=S} x_i^l(t)}{\#\left\{i : C_i^{(m)} = S\right\}} \qquad (2)$$

Cluster assignment:

$$C_{\mathrm{i}}^m = \arg\min_{s=1\cdots k} d(x_i(t), \mu_s^{(m-1)}(t)), i = 1, \cdots, N \qquad (3)$$

K-means Algorithm(P =1):

1. Choose K function samples as the start point randomly.
2. Calculate the distance between function samples and three initial cluster centroids according to Eqs(1), and get the cluster assignment with Eqs(3). S is the index of cluster centroids, S = 1,…,K.
3. Update the cluster centroids according to Eqs(2).
4. Calculate the distance between function samples and three new cluster centroids according to Eqs(1) , and get the cluster assignment.
5. If two cluster assignments are same, then stop; otherwise go to step 3 (Meng et al., 2018).

Result

UNIVERSITY OF LEEDS

start points	6	15	1
clusters	2, 4, 6, 7, 8, 11, 14	9, 10, 15, 16	1, 3, 5, 12, 13, 17, 18
start points	3	11	17
clusters	1， 3, 5, 6, 12, 13, 14, 18	2， 4, 7, 8, 9, 10, 11	15, 16, 17
start points	3	12	8
clusters	1， 3, 5, 6, 13, 18	10, 12, 14, 16, 17	2， 4, 7, 8, 9, 11, 15
start points	6	18	14
clusters	2， 4, 6, 7, 8, 9, 10, 11, 15, 16	1， 3, 5, 12, 13, 17, 18	14
start points	10	14	4
clusters	9, 10, 12, 15, 16, 17, 18	14	1, 2, 3, 4, 5, 6, 7, 8, 11, 13

1. For different start points, there are different cluster assignments.

2. The data of ^{1}H NMR spectra are similar to each other, so even if I update the cluster centroid after one iteration, the distance does not change enough to assign one patient to a new cluster.

3. Patients' erythrocyte extracts of blood samples were taken at different times after surgery, which means patients' data has differnet dimensions. So I will use the FPA score of each patient to do the multivariate case ($p \neq 1$).

Reference

Meng, Y., Liang, J., Cao F. and He Y. 2018. A new distance with derivative information for functional k-means clustering algorithm. Information Sciences. Vol.463-464, pp.166-185.

Krivov, S.V., Fenton, H., Goldsmith, P.J., Prasad, R.K., Fisher, J. and Paci, E. 2014. Optimal Reaction Coordinate as a Biomarker for the Dynamics of Recovery from Kidney Transplant. PLoS Computational Biology. 10(6), p.e1003685.

Ramsay, J.O., Hooker, G. and Graves, S. 2009. Functional Data Analysis with R and MATLAB. Gentleman, R., Hornik, K. and Parmigiani, G. eds. New York : Springer.