関数クラスタリングと関数主要点の関連について

2010年7月9日 統計数理研究所 オープンハウス

清水 信夫　　データ科学研究系 助教

大量の関数データが存在するとき, これらの少数のグループによる分類や特徴的な関数データによる要約が必要となる場合がしばしば生じるが, そのような場合の解析手法として 関数データ解析(Ramsay, 1982 ; Ramsay & Silverman, 1997, 2005)の一手法である関数クラスタリングがよく用いられている.

一方, k-meansクラスタリングと同様の規準に基づいて与えられた確率変数に関し空間をk個の領域に分割したときの重心として主要点(principal points)が定義されている(Flury, 1990). 関数データ解析においても関数データを確率的に扱う概念としてランダム関数(Ibragimov & Rozanov, 1978)が定義されており,それにおける主要点(**関数主要点**)(Tarpey & Kinateder, 2003)が提案されている. ここでは, 正規ランダム関数に従う関数データ集合に対してk-means関数クラスタリングを適用した場合の各クラスターの重心と正規ランダム関数における関数主要点の関連. およびk-means関数クラスタリングを行った場合における局所解の出現回数について示す(清水・水田(2008)).

主要点とは (Flury, 1990, 1993)

p変量確率変数における主要点とは, 確率分布を代表するp次元空間の点の集合として定義される. すなわち, 確率変数Xの分布関数Fについて

$\varsigma_j, y_j \in R^p \ (1 \le j \le k)$, かつ A_j および D_j がそれぞれ

$$A_j = \{x \in R^p ; (x-\varsigma_j)'(x-\varsigma_j) < (x-\varsigma_i)'(x-\varsigma_i), {}^\forall i \ne j\}$$

$$D_j = \{x \in R^p ; (x-y_j)'(x-y_j) < (x-y_i)'(x-y_i), {}^\forall i \ne j\}$$

をみたすとき, 任意のy_j に対して

$$\sum_{j=1}^{k} \Pr(X \in A_j) E\{(X-\varsigma_j)'(X-\varsigma_j) \mid X \in A_j\} \le \sum_{j=1}^{k} \Pr(X \in D_j) E\{(X-y_j)'(X-y_j) \mid X \in D_j\}$$

が成り立つような$\{\varsigma_1, \cdots, \varsigma_k\}$ は, Fに従うp変量確率変数Xの主要点である.

なお, A_j およびD_j はそれぞれς_j およびy_j のDomain of Attractionである.

ランダム関数とは(Ibragimov & Rozanov, 1978)

確率空間(Ω, B, P)において, $L^2(T)$を実数空間T上の2乗可積分かつ可測な全ての関数からなる完備ヒルベルト空間, νをΩからL^2 への可測写像とするとき, νの要素はランダム関数と呼ばれる.

なお, $\xi \in L^2(T)$ を実数空間T上の連続なランダム関数,

$y_j \in L^2(T) \ (1 \le j \le k)$ を非ランダム関数とするとき, y_j のDomain of Attractionは

$$A_j = \{\xi \in L^2[T_L, T_U]; \|\xi - y_j\|^2 < \|\xi - y_i\|^2, {}^\forall i \ne j\}$$

と定義される.

主要点と関数主要点との関係(Tarpey & Kinateder, 2003)

関数データ解析においては, 関数データを正規直交基底を用いて展開する手法が広く利用されている. この手法をランダム関数にも適用することで, ランダム関数の関数主要点を, ランダム関数の正規直交基底の各係数を成分とする空間上における多変量確率変数の主要点に対応する関数として表すことができる.

関数主要点(Tarpey & Kinateder, 2003)

$\varsigma_j, y_j \in L^2(T) \ (1 \le j \le k)$ をそれぞれ非ランダム関数とし, A_j およびD_j をそれぞれ ς_jおよびy$_j$ のDomain of Attraction, zおよびyをそれぞれ $z = \sum_{j=1}^{k} \varsigma_j I_{\{\xi \in A_j\}}$ および $y = \sum_{j=1}^{k} \varsigma_j I_{\{\xi \in D_j\}}$ で表されるランダム関数とする.

ここで$\{\varsigma_1, \cdots, \varsigma_k\}$ がξの関数主要点であるとは, 任意のy_j に対して

$E\|\xi - z\|^2 \le E\|\xi - y\|^2$ が成立することである.

関数k-meansクラスタリングアルゴリズム

$y(t) \in L^2[T_L, T_U]$: ランダム関数

$y_i(t) \in L^2[T_L, T_U]$: $y(t)$の標本関数(非ランダム関数) $(1 \le i \le n)$

$m_j(t) \ (1 \le j \le k)$: クラスターjにおける中心関数

n_j: クラスターjに含まれる標本関数の個数

$A_j = \{\xi \in L^2[T_L, T_U]; \|\xi - y_j\|^2 < \|\xi - y_i\|^2, {}^\forall i \ne j\}$: $y_j(t)$ のDomain of Attraction

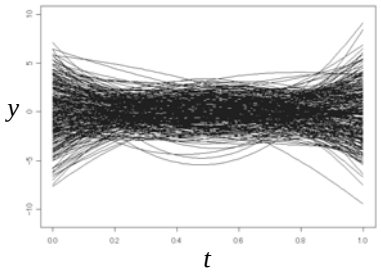
(1) $m_j(t)$の初期値をN組与える.

(2)各$y_i(t)$について$\int_{T_L}^{T_U} (m_j(t) - y_i(t))^2 dt$ が最小となるjを求め, $y_i(t)$ をA_j に振り分ける.

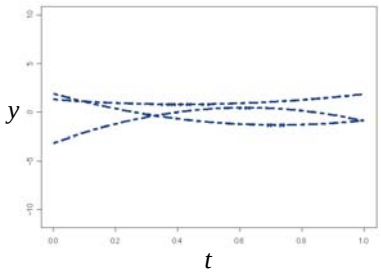
(3) $m_j(t) = \sum_{y_i(t) \in A_j} y_i(t)/n_j$ を計算する.

(4) (2) ～(3)を繰り返し, $m_j(t)$ が変化しなくなったときの値およびA_j に含まれる関数データ集合を中心関数および各クラスターとする.

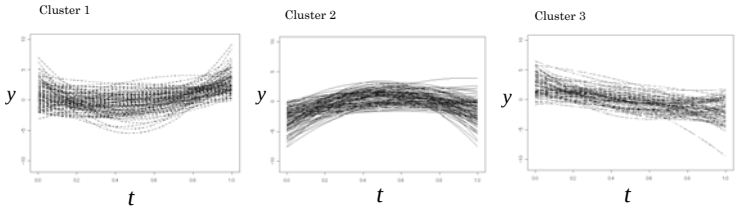
$y(t) = a_1 + a_2 t + a_3 t^2 = b_1 + \sqrt{12}(t - 1/2) b_2 + \sqrt{12}(t^2 - t + 1/6) b_3 \ (0 < t < 1)$ における
関数クラスタリング$((b_1, b_2, b_3)' \sim N_3((0,0,0)', diag(1,1,1)) \quad (n = 250, k = 3))$



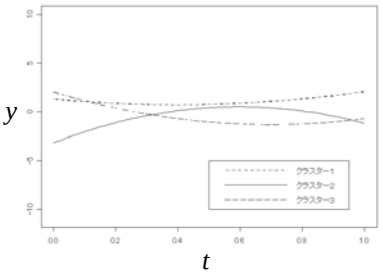
$y(t)$ のk=3個の関数主要点



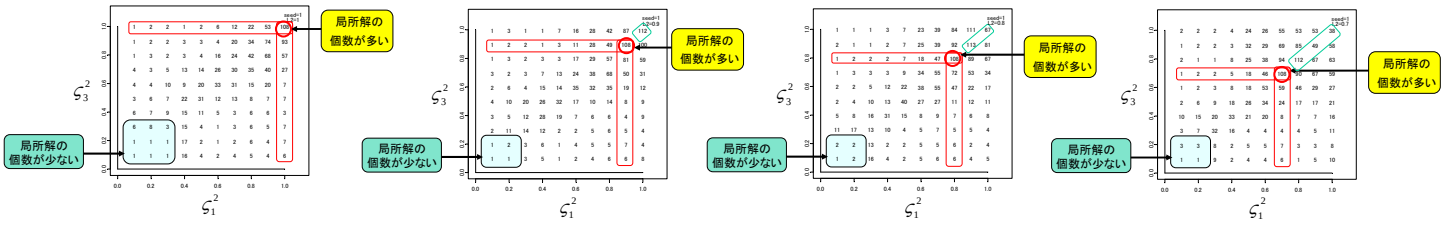
各クラスターに含まれる関数データ



各クラスターの中心関数



$y(t) = a_1 + a_2 t + a_3 t^2 = b_1 + \sqrt{12}(t - 1/2) b_2 + \sqrt{12}(t^2 - t + 1/6) b_3 \ (0 < t < 1)$ の関数クラスタリングにおける局所解の個数
$((b_1, b_2, b_3)' \sim N_3((0,0,0)', diag(\varsigma_1^2, \varsigma_2^2, \varsigma_2^2)) \quad (n = 250, k = 3, N = 1000))$



$\varsigma_2^2 = 1$　　$\varsigma_2^2 = 0.9$　　$\varsigma_2^2 = 0.8$　　$\varsigma_2^2 = 0.7$