関数データに対するいくつかの解析手法について*

北海道大学　水田 正弘

1　はじめに

関数データ解析は、Ramsay および Silverman などにより 1990 年ころから研究が進められた一連の手法であり、それらの研究をまとめた成書が出版されている (Ramsay & Silverman, 1996)。

通常のデータ解析では、データを多次元空間における点として表現し、点間距離として 1 次元または多次元のユークリッド距離、マハラノビス距離などを適用することが多い。もちろん、扱う量が連続量であったり、離散量であったり、各種のデータ構造を有するなど、多様な状況に対応した多くの解析方法がデータ解析の専門家により、研究されてきたことは事実である。しかし、現在のデータ過多社会におけるデータは非常に多様である。例えば、短い間隔で出現する多くのデータを解析する必要がしばしばある。また、連続的に制御変数を変えることにより、連続的にデータが得られることもある。典型的なデータとしては、時系列データがある。

これらのデータを一般化し、データが「関数」として得られた場合の各種解析法として、関数データ解析がある。通常のデータ解析において多次元データの次元数が無限になったとも解釈できる。従って、多次元データの解析方法の多くのものはとりあえず「関数データ解析対応」に拡張できる。しかし、関数データとしての特殊性を生かす工夫も必要である。

Ramsay & Silverman (1996) は、関数データにおける平均、分散、共分散を定義した後、主成分分析、線形モデル、正準相関分析、判別分析の関数データ対応版を扱っている。さらに、関数の定義域を調整する Registration(見当合わせ)、通常のデータを関数データにする各種平滑化などについても詳細に検討している。また、Nason (1997) は、関数データにおける射影追跡を提案した。下川・水田・佐藤 (2000) は、関数回帰分析を関数重回帰分析に拡張した。さらに、Yamanishi & Tanaka(2001)、山西・田中 (2001) は、関数重回帰分析を拡張し、地理的重み付き関数重回帰分析を提案した。Tokushige, Inada & Yadohisa(2001) は、関数データに関する類似度について検討した。

本報告では、関数データに対する基礎統計量を紹介した後、主として報告者が扱った、関数 (重) 回帰分析 (下川・水田・佐藤, 2000)、関数データにおける主要点 (水田, 1999)、関数主成分分析 (水田, 2000)、関数データに対する多次元尺度構成法 (Mizuta, 2000) について紹介する。

*研究集会「高次元データ解析の研究」2002 年 1 月 10-11 日 (広島大学)

2　関数データに対する基礎統計量

はじめに、関数データに対する基礎統計量を列挙しておく。ただし、データ数を N、2 種類の関数集合 X, Y に属する関数データをそれぞれ $x_i(t) \in X,\ y_i(t) \in Y\ (i = 1, \cdots, N)$ とする。なお、交差共分散関数および交差相関関数は X と Y の関数集合間の共分散や相関を計算したものである。

[平均関数]

$$\bar{x}(t) = N^{-1} \sum_{i=1}^{N} x_i(t)$$

[分散関数]

$$\mathrm{var}_X(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \bar{x}(t)]^2$$

[共分散関数]

$$\mathrm{cov}_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}(t_1)\}\{x_i(t_2) - \bar{x}(t_2)\}$$

[相関関数]

$$\mathrm{corr}_X(t_1, t_2) = \frac{\mathrm{cov}_X(t_1, t_2)}{\sqrt{\mathrm{var}_X(t_1)\mathrm{var}_X(t_2)}}$$

[交差共分散関数]

$$\mathrm{cov}_{X,Y}(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}(t_1)\}\{y_i(t_2) - \bar{y}(t_2)\}$$

[交差相関関数]

$$\mathrm{corr}_{X,Y}(t_1, t_2) = \frac{\mathrm{cov}_{X,Y}(t_1, t_2)}{\sqrt{\mathrm{var}_X(t_1)\mathrm{var}_Y(t_2)}}$$

3　関数重回帰分析

はじめに、Ramsay らによる関数回帰分析について紹介する。

例えばある都市で観測された 1 年間の気温の関数が与えられたとして、そこから降水量の関数を予測することを考える。関数回帰分析の目的は通常の回帰と同様、基本的には予測であり、このように関数を予測する場合のほかに、予測値がスカラーである場合も考えられる。しかし後者は前者の特別な場合と考えることができるので、以下関数予測の場合のみを考える。

データ数を N とする。以下では説明変量に対応する関数データを $x_i(s),\ s \in \mathcal{T}_X \subseteq \boldsymbol{R}$, 目的変量に対応する関数データを $y_i(t),\ t \in \mathcal{T}_Y \subseteq \boldsymbol{R}\ (i = 1, \cdots, N)$ と表す。ただし $\mathcal{T}_X, \mathcal{T}_Y$ はそれぞれの定義域とし、閉区間とする。

Ramsay らの関数回帰モデルは通常の回帰モデルを説明変量および目的変量が関数の場合に拡張したもので、

$$y_i(t) = \alpha(t) + \int_{\mathcal{T}_X} x_i(s)\beta(s,t)ds + \epsilon_i(t)$$

と与えられる。ただし $\alpha(t)$ は平均関数、$\beta(s,t)$ は回帰の重み関数、$\epsilon_i(t)$ は誤差関数である。このとき、積分 2 乗誤差

$$\texttt{LMISE} = \sum_{i=1}^{N} \int_{\mathcal{T}_Y} [y_i(t) - \alpha(t) - \int_{\mathcal{T}_X} x_i(s)\beta(s,t)ds]^2 dt$$

を最小にするような 2 次元の重み関数 $\beta(s,t)$ を求める。

関数回帰分析を用いた実際の解析で、Ramsay らは気温から降水量を予測している。しかし降水量を決定づける要因が気温だけであるとは考えにくく、他の種類のデータも得られるならばそれも予測に利用した方が、一般的にはよい予測が得られる。例えば気温のほかに湿度や日照時間などの関数データ、また標高や緯度などのベクトルデータが与えられた場合に、これら複数の説明変量から予測を行なうモデルを考える。

以下では Ramsay らの関数回帰分析を、説明変量にあたる関数が 1 つではない場合に拡張する提案を行なう。

拡張した関数回帰モデルを定義する。目的変量に対応する関数データ $y_i(t)$ が、G 個の関数データ $x_{gi}(s),\ s \in \mathcal{T}_{X_g} \subseteq \boldsymbol{R}\ (g = 1, \cdots, G)$ および H 個の関数データではない変数 $\boldsymbol{w}_i' = (w_{1i}, \cdots, w_{Hi})$ によって表現できるモデル

$$y_i(t) = \alpha(t) + \sum_{g=1}^{G} \int_{\mathcal{T}_{X_g}} x_{gi}(s)\beta_g(s,t)ds + \boldsymbol{w}_i'\boldsymbol{\gamma}(t) + \epsilon_i(t)$$

を考える。ただし $\beta_g(s,t)$ は $x_{gi}(s)$ に対する重み関数、$\boldsymbol{\gamma}(t)$ は $\boldsymbol{w}_i$ の各要素に対する重み関数を要素にもつ H 次元の関数ベクトルとする。

簡単のために $x^*_{gi}(s_g) = x_{gi}(s_g) - \bar{x}_g(s_g),\ y^*_i(t) = y_i(t) - \bar{y}(t),\ \boldsymbol{w}^*_i = \boldsymbol{w}_i - \bar{\boldsymbol{w}}$ として $\alpha(t)$ を消去する。

$$y^*_i(t) = \sum_{g=1}^{G} \int_{\mathcal{T}_{X_g}} x^*_{gi}(s_g)\beta_g(s_g,t)ds_g + \boldsymbol{w}'^*_i\boldsymbol{\gamma}(t) + \epsilon_i(t)$$

さらに $\boldsymbol{w}^*_i$ の各要素を定数関数と見て $x^*_{gi}(s)$ に, $\boldsymbol{\gamma}$ の各要素を s に関して定数関数と見て $\beta_g(s,t)\ (g = G+1, \cdots, G+H)$ に含めれば、

$$y^*_i(t) = \sum_{g=1}^{G+H} \int_{\mathcal{T}_{X_g}} x^*_{gi}(s)\beta_g(s,t)ds + \epsilon_i(t)$$

と表現することができる。このとき

$$\texttt{LMISE} = \sum_{i=1}^{N} \int_{\mathcal{T}_Y} [y^*_i(t) - \sum_{g=1}^{G+H} \int_{\mathcal{T}_{X_g}} x^*_{gi}(s)\beta_g(s,t)ds]^2 dt$$

が最小となるような $\beta_g(s,t),\ g = 1, \cdots, G+H$ を求めることになる。

4　関数データにおける主要点

Flury(1990) は分布に関する主要点 (Principal Points) を提案した。これは、分布を代表する p 次元空間の点の集合である。

$f(x)$ を確率変数 X の密度関数、 $F(x)$ を分布関数とする。p 次元空間における点 $\boldsymbol{x} \in R^p$ と点の集合 $\{\boldsymbol{y}_j\}, \boldsymbol{y}_j \in R^p$ との距離を

$$d(\boldsymbol{x}|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k) = \min_{1 \le h \le k} \{(\boldsymbol{x} - \boldsymbol{y}_h)^T (\boldsymbol{x} - \boldsymbol{y}_h)\}^{1/2}. \tag{1}$$

によって定義する。このとき、

$$E_F\{d^2(X|\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_k)\} = \min_{\boldsymbol{y}_j \in R^p} E_F\{d^2(X|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k)\}.$$

が成立する $\boldsymbol{\xi}_j \in R^p$ $(1 \le j \le k)$ を分布 F における k-主要点 (k-principal points) と定義する。この定義は、クラスター分析における k-means 法の基準と同じである。

上述の主要点の考え方を拡張し、関数データに適用することを試みる。ただし、今回は関数データの分布が与えられているのではなく、N 個の関数データ $x_i(t), t \in \mathcal{T} \subseteq \boldsymbol{R}, (i = 1, \cdots, N)$ を考える。

このとき

$$H = \sum_{i \in S_1} \int_{\mathcal{T}} (x_i(t) - \mu_1(t))^2 dt + \cdots + \sum_{i \in S_k} \int_{\mathcal{T}} (x_i(t) - \mu_k(t))^2 dt$$

を最小とする $\mu_1(t), \cdots, \mu_k(t)$ を k-主要点と定義する。ただし、S_j は $\{1, 2, \cdots, N\}$ の分割で、

$$S_j = \{i| \int_{\mathcal{T}} (x_i(t) - \mu_j(t))^2 dt \le \int_{\mathcal{T}} (x_i(t) - \mu_h(t))^2 dt \text{ for all } h \ne j\}$$

を満たすものとする。

$k = 1$ のときは、$\mu_1(t) = \frac{1}{N} \sum_{i=1}^N x_i(t)$ であることは初等的にも証明できるが、変分法による証明の概略を示す。$\mu(0, t) = \mu_1(t)$ となる連続微分可能な任意の 2 変数関数 $\mu(s, t)$ を考える。

$$H(s) := \sum_{i=1}^N \int_{\mathcal{T}} (x_i(t) - \mu(s, t))^2 dt$$

とおくと、

$$\frac{d}{ds} H(s)|_{s=0} = -2 \sum_{i=1}^N \int_{\mathcal{T}} (x_i(t) - \mu(s, t)) \mu_s(0, t) dt = -2 \int_{\mathcal{T}} (\sum_{i=1}^N x_i(t) - N\mu(s, t)) \mu_s(0, t) dt$$

となる。ここで、任意の $\mu(s, t)$ について、$\frac{d}{ds} H(s)|_{s=0} = 0$ であるためには、$\mu(t) = \frac{1}{N} \sum_{i=1}^N x_i(t)$ でなくてはならない。

以上の議論を一般の k で行なうことにより、

$$\mu_m(t) = \frac{1}{N} \sum_{i \in S_m}^N x_i(t), \quad (m = 1, 2, \cdots, k)$$

が成立する。従って、N 個の関数データに対する k-主要点 $\mu_1(t), \cdots, \mu_k(t)$ は、クラスター分析における k-means 法と同様に求めることができる。

また、Ramsay & Silverman(1996) や下川・水田・佐藤 (2000) にあるように関数データを p 個の基底関数で展開して近似すると、近似誤差を除いて、通常の p 次元データにおける主要点と同様に扱うことができる。

5 関数主成分分析

関数データのための主成分分析は、関数データ解析における比較的早い時期に研究がなされた。しかし、一般化主成分分析など非線形な主成分分析の検討をはじめ、多くの問題が残っている。

通常の主成分分析では、N 個の p 変量データを $x_{ij}, (i = 1, \cdots, N; j = 1, \cdots, p)$ として、$f = \boldsymbol{a}_1^T \boldsymbol{x}$ の分散が最大となる $\boldsymbol{a}_1$ を求めることから始める。

τ を定義域とする N 個の関数データを $x_i(s), s \in \tau, (i = 1, \cdots, N)$ とする。関数主成分分析では、データの変動を表す関数 $\xi_k(s), (k = 1, \cdots, K)$ を順に求めることが目的となる。正確には、以下の 3 つの同値な方法により説明できる。ただし、以下では、各関数データから平均関数 $N^{-1}\sum_{i=1}^N x_i(s)$ を引くことにより、関数データの平均関数は 0 関数であると仮定する。また、関数 $\alpha(s), \beta(s)$ のたたみ込みを $\langle \alpha, \beta \rangle = \int_\tau \alpha(s)\beta(s)ds$ と書く。

分散を用いた方法

$$f_{ik} = \int_\tau \xi_k(s) x_i(s) ds, \quad \langle \xi_k, \xi_k \rangle = 1, \ \ (k = 1, \cdots, K)$$

とし、$N^{-1}\sum_{i=1}^N f_{i1}^2$ を最大とする $\xi_1(s)$ を求める。次に $\langle \xi_1(s), \xi_2(s) \rangle = 0$ の制約条件のもとで、$N^{-1}\sum_{i=1}^N f_{i2}^2$ を最大とする $\xi_2(s)$ を求める。以下、同様に $\xi_K(s)$ まで求める。

経験的正規直交基底による方法

$$\hat{x}_i(s) = \sum_{k=1}^K \langle x_i, \xi_k \rangle \xi_k(s)$$

とおき、

$$\sum_{i=1}^N \int_\tau (x_i(s) - \hat{x}_i(s))^2 ds$$

が最小となる正規直交基底 $\{\xi_k(s)\}$ を求める。

固有関数による方法

$$N^{-1}\sum_{i=1}^N x_i(s)\langle x_i, \xi \rangle = \rho \xi(s), \quad \langle \xi, \xi \rangle = 1$$

を満たす $\xi(s)$ のうち ρ の値が大きいものから順に K 個選ぶ。

通常のデータに対する一般化主成分分析は、p 変量データ $\boldsymbol{x}$ を事前に決めた写像 ϕ により q 変量データ $(q \geq p)$ に拡張した後、主成分分析を適用する。最も簡単な 2 変量 2 次の一般化主成分分析では、データ (x, y) を $\phi(x, y) = (x, y, x^2, xy, y^2)$ により 5 変量データに拡張

し、5 変量データとして分散共分散行列を求め固有値問題を解く。最小固有値に対応する固有ベクトルにより「データを当てはめる」2 次曲線が定義できる。

ただし、「当てはめ」の解釈や、写像 ϕ の決定法など、いくつか検討しなくてはならない問題が残されている (Mizuta, 1984)。

一般化主成分分析の考え方等を利用して、関数主成分分析の拡張を検討する。

$$f_{ik} = \int \beta_k(s) x_i(s) ds + \int \gamma_k(s) x_i(s)^2 ds$$

とおき、$\int \beta_k(s)^2 ds + \int \gamma_k(s)^2 ds = 1$ の制約条件のもと、

$$N^{-1} \sum_{i=1}^{N} f_{ik}^2 = N^{-1} \sum_{i=1}^{N} (\langle \beta_k, x_i \rangle^2 + (\int \gamma_k(s,t) x_i(s)^2 ds)^2 + 2\langle \beta_k, x_i \rangle \int \gamma_k(s) x_i(s)^2 ds)$$

を最大にする $\beta_k(s), \gamma_k(s)$ を順次求めることが考えられる。実際の計算では、関数主成分分析と同様に、定義域を均等に分割したり、関数の基底展開により、通常のデータにおける最適化問題に帰着することができる。

関数主成分分析の 1 つの拡張を検討した。これ以外にもいくつかの拡張方法は考えられる。実際の関数データに対する有効性の評価とともに今後の課題としたい。さらに、関数主成分分析および拡張した関数主成分分析において、関数の個数 K の決定法、寄与率の定義とその解釈については検討が必要である。

6 関数多次元尺度構成法

通常の多次元尺度構成法では、n 個のオブジェクトの (非) 類似度 $S = \{s_{ij}\}(i, j = 1, 2, \cdots, n)$ から、それらの (非) 類似度を適切に表現する p 次元空間における n 個の点 $X = \{\boldsymbol{x}_i\}(i = 1, 2, \cdots, n)$ を構成する。(非) 類似度に関して、$s_{ij} \geq 0$, $s_{ij} = s_{ji}$, $s_{ii} = 0$ を仮定する場合も多い。2 点 $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ のユークリッド距離を d_{ij}; $d_{ij} := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ とする。多次元尺度構成法とは、$d_{ij} \simeq s_{ij}$ となる付置 X を求める手法であると言える。ここで、d_{ij} と s_{ij} との当てはまりの良さの規準により多くの手法が提案されている。

ここで、(非) 類似度データが関数データとして与えられた場合における多次元尺度構成法を検討する。すなわち、n 個のオブジェクト間の (非) 類似度が変数 t に依存しており、$S(t) = \{s_{ij}(t)\}(i, j = 1, 2, \cdots, n), t \in [a, b]$ と表現されるとする。以下では、説明の都合上、付置するユークリッド空間の次元は 2 とする。

以下では、このようなデータに対する多次元尺度構成法を報告する。

はじめに、変数 t について、変数を固定して通常の多次元尺度構成法 (2 次元) を適用する。従って、各 t について、n 個のオブジェクトの 2 次元における付置が得られる。これらを、$X(t) = \{\boldsymbol{x}_i(t)\}(i = 1, 2, \cdots, n)$ とする。ここで、$X(t)$ が t に関して連続である保証はないが、とりあえず連続性と微分可能性を仮定する。ここで、直交行列 $Q(t)$ を利用して、$X(t)$ を t $Q(t)X(t)$ により回転させることを考える。多次元尺度構成法の解を直交変換させることの妥当性については、利用する手法に依存するが、付置におけるオブジェクト間の距離は不変なので、大部分の手法について実用上の問題はないと思われる。

2 次元空間における曲線 $\boldsymbol{x}_i(t)$ の距離は、

$$l = \int_a^b \sqrt{\left\| \frac{d\boldsymbol{x}(t)}{dt} \right\|^2}\, dt.$$

と定義される。そこで、

$$l(Q) = \int_a^b \sum_{i=1}^n \left\| \frac{dQ(t)\boldsymbol{x}_i(t)}{dt} \right\|^2 dt.$$

を最小とする直交行列の関数 $Q(t)$ を考える。$Q(t)$ は、$Q(t) = \begin{pmatrix} \sin\phi(t) & \cos\phi(t) \\ -\sin\phi(t) & \sin\phi(t) \end{pmatrix}$,
と表現することができる。ここで、$\phi(t)$ は t の関数である。そこで、

$$l(Q) = \int_a^b \left(\sum_{i=1}^n \left\| \frac{d\boldsymbol{x}_i(t)}{dt} \right\|^2 + 2 \left(\sum_{i=1}^n \frac{d\boldsymbol{x}_i(t)^T}{dt} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \boldsymbol{x}_i(t) \right) \phi'(t) + \left(\sum_{i=1}^n \|\boldsymbol{x}_i(t)\|^2 \right) \phi'(t)^2 \right) dt.$$

となる。従って、

$$\phi'(t) = \frac{-\sum_{i=1}^n \frac{d\boldsymbol{x}_i(t)^T}{dt} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \boldsymbol{x}_i(t)}{\sum_{i=1}^n \|\boldsymbol{x}_i(t)\|^2},$$

のとき、$l(Q)$ は最小値となる。

この関数による付置 $Q(t)X(t)$ は関数多次元尺度構成法の解とみなすことができる。関数多次元尺度構成法の解は、n 個の点の動きを動的なグラフィックスで表示したり、$Q(t)X(t)$ を n 本の軌跡として表現することも有効である。

7 おわりに

本報告では、関数データ解析の手法をいくつか紹介した。関数データを解析するための新たな手法の開発は重要な課題である。また、関数データ解析の研究において不足していると思われるものに、(1) 確率論的な扱い、(2) 変分法をはじめとする関数解析の技法の適用、がある。関数データ解析を実用的レベルにするために、これらの課題を含め、数学的検討および計算機的検討の両面から関数データ解析に関する研究をすすめていきたいと思う。

参考文献

[1] Flury, B. A.(1990). Principal Points. *Biometrika*, **77**, 1, 33–41.

[2] Mizuta,M. (1984). Generalized Principal Components Analysis Invariant under Rotations of a Coordinate System. *Journal of the Japan Statistical Society*, **14**, 1–9.

[3] Mizuta, M.(2000). Functional Multidimensional Scaling. *Proceedings of the Tenth Japan and Korea Joint Conference of Statistics*, 77–82.

[4] Nason, G. P. (1997). Functional Projection Pursuit. *Computing Science and Statistics*, **23**, 579–582. http://www.stats.bris.ac.uk/~ guy/Research/PP/PP.html

[5] Ramsay, J. O. and Silverman, B. W. (1996). *Functional Data Analysis*, Springer.

[6] Tokushige, S., Inada, K. and Yadohisa, H.(2001). Dissimilarity and Related Methods for Functional Data. *Proceedings of the International Conference on New Trends in Computational Statistics with Biomedical Applications*, 295–302.

[7] Yamanishi, Y. and Tanaka, Y.(2001). Geographically Weighted unctional Multiple Regression Analysis: A Numerical Investigation. *Proceedings of the International Conference on New Trends in Computational Statistics with Biomedical Applications*, 287–294.

[8] 下川真由子・水田正弘・佐藤義治 (2000). 関数データ解析における回帰分析の拡張, 応用統計学, **29(1)**, 27–39.

[9] 水田正弘 (1999). 関数データ解析における主要点について, 第 67 回日本統計学会講演報告集、355–356.

[10] 水田正弘 (2000). 関数データに対する主成分分析について, 第 68 回日本統計学会講演報告集、195–196.

[11] 山西芳裕・田中 豊 (1990). 関数データの主成分分析：感度分析と数値的検討, 日本計算機統計学会第 14 回大会論文集, 92–95.

著者連絡先

〒 060-0811 札幌市北区北 11 西 5
北海道大学 情報メディア教育研究総合センター情報メディア科学基礎分野
(大学院工学研究科システム情報工学専攻 担当)
水田正弘　e-mail: mizuta@main.eng.hokudai.ac.jp