基底展開法による関数回帰モデリング

Functional regression modeling via basis expansions

数学専攻 大屋 拓磨
OHYA, Takuma

1　はじめに

関数データ解析 (functional data analysis) は, 離散点で経時的に観測された複数のデータ系列それぞれを関数化処理したデータ集合を解析の対象とする手法であり, Ramsay and Silverman (1997) によって提案された. これまでに, 回帰分析, 主成分分析, 正準相関分析, 判別分析などの多変量解析手法に対応した関数データ解析手法が提案されている (Ramsay and Silverman, 1997, 2002).

離散点で観測されたデータを関数化することによって, 複数の経時的データが, 観測時点に依存しないことや欠測値を含む場合にも対応できるという利点がある. さらに, データを関数化処理することによって, 関数の導関数を求め, 速度や加速度といった新たな情報が得られ, これらの情報を有効に活用できることが関数データ解析の特徴といえる.

本稿では, Ramsay and Silverman (1997) によって提唱された関数データ解析を研究の対象とし, 特に, 基底展開法に基づく関数回帰モデリングについて検討した. さらに, 関数データ解析における問題点を踏まえ, バーンスタイン基底関数に基づく関数回帰モデリングを提案した.

2　データの関数化

関数データとは, 例えば時間 $\{t_i \in \tau \subset \mathbb{R};\ i = 1, 2, ..., n\}$ の経過に伴って観測されたデータ $\{(x_1, t_1), (x_2, t_2), ..., (x_n, t_n)\}$ に対して, 平滑化や補間を行って一つの関数 $x(t)$ $(t \in \tau \subset \mathbb{R})$ として処理されたデータである. データの関数化は, 主として基底展開法によって行われ, これまでにフーリエ級数や B-スプラインなどを用いられているが, ここでは, 一般の基底関数を ϕ として統一的に述べる.

いま, データ $\{(x_i, t_i); t_i \in \tau, i = 1, 2, ..., n\}$ は,

$$x_i = u(t_i) + \varepsilon_i, \qquad i = 1, 2, ..., n \tag{1}$$

に従って観測されたとする. ただし, $u(t)$ は未知の滑らかな関数であり, 誤差 ε_i は互いに独立で平均 0, 分散 σ^2 の正規分布に従うとする. 関数 $u(t)$ は基底関数の線形和として

$$u(t) = w_0 + \sum_{k=1}^{m} w_k \phi_k(t) \tag{2}$$

と表す. この未知の関数である $u(t)$ をデータから推定して関数データとして用いる.

実際には (1) 式と (2) 式より

$$x_i = \boldsymbol{w}^T \boldsymbol{\phi}(t_i) + \varepsilon_i, \qquad i = 1, 2, ..., n \tag{3}$$

を得る. ただし, $\boldsymbol{w} = (w_0, w_1, ..., w_m)^T, \boldsymbol{\phi}(t_i) = (1, \phi_1(t_i), ..., \phi_m(t_i))^T$ とする. 従って最尤法を用いることに

より, パラメータ $\boldsymbol{w}$ と σ^2 の推定量は

$$\hat{\boldsymbol{w}} = (B^T B)^{-1} B^T \boldsymbol{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \{x_i - \hat{\boldsymbol{w}} \boldsymbol{\phi}(t_i)\}^2 \tag{4}$$

で与えられる. ただし, $B = (\boldsymbol{\phi}(t_1), ..., \boldsymbol{\phi}(t_n))^T, \boldsymbol{x} = (x_1, ..., x_n)$ である.

しかし, 関数回帰モデリングでは, これら関数化した関数データを同じ基底関数で構成する必要がある. このため, 基底関数の個数の決定が重要であり, 適切に設定することが求められる. そこで, 基底関数を統一し, なおかつ個々にモデルの複雑さを調整する方法として正則化法が用いられている. 正則化最尤推定法によるパラメータ $\boldsymbol{w}$ と σ^2 の推定量は

$$\hat{\boldsymbol{w}} = (B^T B + \gamma \hat{\sigma}^2 K)^{-1} B^T \boldsymbol{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \{x_i - \hat{\boldsymbol{w}} \boldsymbol{\phi}(t_i)\}^2 \tag{5}$$

で与えられる. ただし, γ は正則化パラメータ, K は $(m+1) \times (m+1)$ 次非負値定符号行列である. 調整パラメータの推定法としては, 一般化情報量規準や平滑化行列に基づく方法などが提案されている (詳しくは, 小西・北川 (2004), Konishi and Kitagawa (2008) を参照されたい).

3 関数回帰モデリング

関数化されたデータにおける関数回帰モデリングについて述べる. 関数データを目的変数か説明変数, あるいは両方に用いるかによってモデルは異なるが, ここでは荒木・小西 (2004) をもとに, 説明変数が関数の場合における関数回帰モデリングについて述べる. 通常の回帰モデリングと同様に, 基本的な目的は予測である.

目的変数 Y と説明変数 X に対し, n 個の観測データ $\{(y_i, x_i(t)); t \in \tau, i = 1, 2, ..., n\}$ が得られたとする. ここで $x_i(t)$ は関数化によって得られた関数データ

$$\begin{aligned} x_i(t) &= \sum_{j=1}^{m} w_{ij} \phi_j(t) \\ &= \boldsymbol{w}_i^T \boldsymbol{\phi}(t) \qquad i = 1, 2, ..., n \end{aligned} \tag{6}$$

とする. ただし, $\boldsymbol{w}_i = (w_{i1}, ..., w_{im})^T, \boldsymbol{\phi}(t) = (\phi_1(t), ..., \phi_m(t))^T$ とおく. 通常の回帰モデリングでは, 離散点で観測されたデータの線形結合としてモデル化されるのに対して, 関数回帰モデリングでは変数間の関係を

$$y_i = \alpha + \int_{\tau} x_i(t) \beta(t) dt + \varepsilon_i, \qquad i = 1, 2, ..., n \tag{7}$$

とモデル化する. ただし, $\beta(t)$ は τ 上の滑らかな関数, 誤差 ε_i は, 互いに独立で平均 0, 分散 σ^2 の正規分布に従うと仮定して, n 個のデータに基づいてパラメータ $\alpha, \beta(t)$ を推定する.

(6) 式の関数データに含まれる基底関数 $\boldsymbol{\phi}(t)$ を用いて, (7) 式の関数パラメータ $\beta(t)$ を

$$\begin{aligned} \beta(t) &= \sum_{k=1}^{m} \beta_k \phi_k(t) \\ &= \boldsymbol{\beta}_x^T \boldsymbol{\phi}(t) \end{aligned} \tag{8}$$

とする. ただし, $\boldsymbol{\beta}_x = (\beta_1, ..., \beta_m)^T$ とおく. このとき, 基底展開法によって得られた (6) 式と (8) 式を (7) 式に

代入すると

$$\begin{aligned} y_i &= \alpha + \int_\tau \boldsymbol{w}_i^T \boldsymbol{\phi}(t) \boldsymbol{\beta}_x^T \boldsymbol{\phi}(t) dt + \varepsilon_i \\ &= \alpha + \boldsymbol{w}_i^T \Phi \boldsymbol{\beta}_x + \varepsilon_i \end{aligned} \tag{9}$$

と表すことができる. ただし, Φ は $m \times m$ 行列で, その (j,k) 要素は

$$\phi_{jk} = \int_\tau \phi_j(t)\phi_k(t)dt, \qquad j,k = 1,2,...,m \tag{10}$$

で与えられる. (9) 式より, n 個のデータに対してベクトルと行列を用いて表記すると

$$\boldsymbol{y} = Z\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n) \tag{11}$$

を得る. ただし, $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T, \boldsymbol{\beta} = (\alpha, \boldsymbol{\beta}_x^T)^T, \boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$,

$$Z^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \Phi^T \boldsymbol{w}_1 & \Phi^T \boldsymbol{w}_2 & \cdots & \Phi^T \boldsymbol{w}_n \end{bmatrix}$$

とする. 従って, (11) 式より最尤法によって, モデルのパラメータ $\boldsymbol{\beta}, \sigma^2$ を推定することが可能となり, その最尤推定量は

$$\hat{\boldsymbol{\beta}} = (Z^T Z)^{-1} Z^T \boldsymbol{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - Z\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - Z\hat{\boldsymbol{\beta}}) \tag{12}$$

である.

ここで, データの関数化と同様, 関数回帰モデリングにおいても複雑な構造をもつ現象に対してモデルのあてはめを行う際には, 基底関数の個数の決定が重要となるが, 推定するパラメータ $\boldsymbol{\beta}$ の次元はデータの関数化における基底関数の個数 m によって $(m+1)$ 個に決まってしまうために, 正則化法を用いてデータに合わせてパラメータを調整することが本質的である. $\boldsymbol{\beta}$ と σ^2 の正則化最尤推定量は

$$\hat{\boldsymbol{\beta}} = (Z^T Z + \lambda \hat{\sigma}^2 K)^{-1} Z^T \boldsymbol{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - Z\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - Z\hat{\boldsymbol{\beta}}) \tag{13}$$

で与えられる. ここでも, パラメータの推定には調整パラメータ λ の決定の課題が残る. 調整パラメータの推定法としては, 同様に, 一般化情報量規準や平滑化行列に基づく方法が用いられる.

4　バーンスタイン基底関数による関数回帰モデリング

関数回帰モデリングの特徴として, 基底関数の取り方に依存せず全て同じプロセスで推定ができ, 統一的な理論構成が可能であることがわかる. この特徴を生かすためには, 様々な基底関数の中から当該のデータを解析するために最適な基底関数の選択が必要であると考える. これまでに基底関数はフーリエ級数や B-スプライン, 動径基底関数などが提案されている.

そこで基底関数として, バーンスタイン基底関数を関数回帰モデルに導入することを提案する. m 次のバーンスタイン基底関数は $t \in [0,1]$ に対して

$$b_{k,m}(t) = {}_mC_k \; t^k (1-t)^{m-k}, \qquad k = 0,1,...,m \tag{14}$$

として与えられ, バースタイン基底関数の線形結合による

$$B(t) = \sum_{k=0}^{m} \beta_k b_{k,m}(t) \tag{15}$$

は m 次のバーンスタイン多項式である. また, β_k はバーンスタイン係数である.

バーンスタイン基底関数を適用するにあたっては, $[0,1]$ 区間にデータを変換する必要がある. また, 基底関数の個数が変化することによって ${}_mC_k$ の値が変化し, 基底の形が変化するため, 基底関数の個数の設定が課題である.

また, 関数回帰モデリングでの課題として (10) 式の計算があるが, バーンスタイン基底関数は積分区間が $[0,1]$ なので (10) 式の第 $(j+1,k+1)$ 要素は

$$\begin{aligned}\phi_{(j+1,k+1)} &= {}_mC_j\,{}_mC_k \int_0^1 t^{j+k}(1-t)^{2m-j-k}dt \\ &= {}_mC_j\,{}_mC_k \frac{(j+k)!(2m-j-k)!}{(2m+1)!}, \qquad j,k=0,1,...,m \end{aligned} \tag{16}$$

となる. したがって, 理論的にバーンスタイン基底関数を関数回帰モデリングに適用することが可能である.

関数回帰モデルは, 実際の現象に対応して, 目的変数と説明変数のいずれかを関数化するモデルや説明変数および目的変数を共に関数化して, 変数間の関係をモデル化する方法がある (Matsui, 2009). バーンスタイン基底関数に基づくこれらの関数回帰モデリングについては, 修士論文を参照されたい.

5 おわりに

本研究を通して, バーンスタイン基底関数が, 経時的に観測・測定されたデータ系列を関数化するのに有効な基底の一つであることがわかった. 今後の研究課題として, バーンスタイン基底関数を用いてデータの関数化を行う際に, データを [0, 1] 区間の中にどのように変換するのが最適であるか検討することが挙げられる. また, バーンスタイン基底関数を改良し, データの関数化と関数回帰モデルの構築に適切に機能する手法の研究などを行いたい.

参考文献

[1] 荒木由布子, 小西貞則, (2004). 動径基底関数展開に基づく関数回帰モデリング. 応用統計学 **33-3**, 243-256.

[2] 小西貞則, (2010). 『多変量解析入門-線形から非線形へ-』. 岩波書店.

[3] 小西貞則, 北川源四郎, (2004). 『情報量規準』. 朝倉書店

[4] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

[5] Matsui, H. (2009). Regularized Functional Regression Modeling and its Applications. Ph. D. Thesis, Kyushu University.

[6] Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag, New York.

[7] Ramsay, J.O. and Silverman, B.W. (2002). *Applied Functional Data Analysis.* Springer-Verlag, New York.

[8] Joy, K.I. Bernstein Polynomials. *On-Line Geometric Modeling Notes.* URL:http://www.idav.ucdavis.edu/education/CAGDNotes/Bernstein-Polynomials.pdf.