関数データ解析法とシンボリックデータ解析法
—多様なデータ構造とその解析法—

北海道大学　情報基盤センター
先端データ科学研究室　水田 正弘

１．はじめに

「データ」を解析し、そこから何らかの情報を得ることは、科学一般において普遍的な作業である。ここで、「データ」としてどのようなものを想定するかは、解析における目的や利用できる環境（特にコンピュータ環境）に依存する。コンピュータ環境が不十分な時代には、「データ」をいわゆる１次元の数値として扱うことしかできなかった。しかし、近年では、大量の超高次元の数値として扱うことが可能になってきた。

統計学の初歩的な教科書には、データの型として、量的データ、質的データ、名義尺度、比尺度、間隔尺度、などが説明されることが多い。また、「データ」をベクトルや行列で表現して説明するのが普通である。これらの概念や表記法は有用であるが、近年、統計学に対して期待されている分野では、これらの枠組みで表現できないデータも数多く存在する。例えば、時系列データ、空間データ、時空間データなどがある。S-PLUSも、多くの既存のクラスが提供されており、各クラスに対してメソッドを定義することができる。

さらに、より一般的なアプローチとして、関数データ解析法とシンボリックデータ解析法がある。両者とも、1980年頃に提案され、現在、活発に研究がすすんでいる。この二つのアプローチを中心にして、多様なデータ構造とその解析法について報告する。

２．関数データ解析法

データを関数として扱う関数データ解析法は、1980年台からRamsayにより提案された。その後、多次元データ解析の多くの手法が関数データに対しても適用できるように拡張されてきた。また、関数データ解析法に関する成書も出版された(Ramsay & Silverman, 1997,2005)。さらには、関数データ解析法の応用例をまとめた本も出版されている(Ramsay & Silverman, 2002)。

関数データ解析が有効である場面は、いくつか考えられる。第1の状況は、実際に関数としてデータが入手できる場合である。すなわち、任意の変数を与えることにより数値が得られるようなケースである。ただし、実際の場合にはこのような状況は少ないと思われる。より重要な状況としては、多変量データとして与えられていくデータに対して、平滑化などを施すことにより、データを関数とみなして扱える場合である。一度、データを「関数」に変換することにより、データの「微分」が可能になるなどの利点がある。先に紹介した本(Ramsay & Silverman, 1997,2005)においても、平滑化の説明にかなりのページをさいている。

関数データ解析は、通常の多次元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 (2005)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[20]。さらに、Yamanishi & Tanaka (2001)、山西・田中 (1990)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。

通常のデータと関数データを比較検討する。簡単のために、通常のデータとして n 個の p 変量データ $\mathbf{x}_i \in R^p \quad i=1,2,\cdots,n$ 、関数データとして積分可能な n 個の一変数関数 $x_i(s) \quad i=1,2,\cdots,n$ を考える。確率構造を入れなければ、通常のデータは、 p 次元空間における n 個の点であり、関数データは無限次元空間における n 個の点ととらえることができる。また、 p 変量データではベクトルのノルム、関数データでは L^2 ノルムを利用することで内積が定義でき、データ点間の距離を表現できる。

関数データが得られたとき、コンピュータなどを利用して解析するためには、関数データを有限個の数値で表現しなくてはいけない。関数データ解析における多くの研究では有限個の(正規直交)基底関数を利用して関数データを近似している。これにより関数データは、利用する基底関数の個数と同じ次元をもつ多変量データとなる。すなわち、従来からある多変量データに対する解析法がそのまま使える。

ここで、「関数データが得られたとき」と書いたが、実際の場面では関数ではなく離散的な数値が得られるのが普通である。離散的な数値を関数データにするための平滑化の方法が大きな課題となる。Ramsay & Silverman(2005)においても、関数化・平滑化の説明に多くのページを割いている。このとき、関数データ解析法において利用する基底関数により関数化すると、その後の扱いが容易になる。

離散データの関数化および、関数データを有限個の数値パラメータで表現する方法の両者において、基底関数の選び方、利用する基底関数の個数の選び方が問題となる(荒木・小西,2004など)。

関数データに確率構造を入れる考え方はいくつかある。最も簡単なのは、関数データを基底関数などにより有限個の数値で表現することにより、通常の有限次元空間における確率構造と同様な議論ができる。別のアプローチとして、Random Function (Lifshits, 1995)の考え方を利用することができる。さらに、Sakaori(2002)は、関数データの分布を直接的には利用しないでパーミュテーションテストなどを使う方法を提案している。

関数データ解析法の例として、関数主成分分析法を簡単に紹介する。一次元関数データ $\{x_i(t), i=1,\ldots,n\}$ に対する関数主成分分析とは、 $\int \zeta_1(s)^2 ds = 1$ の制約条件において $n^{-1}\sum_{i=1}^{n}(\int \zeta_1(s)x_i(s)ds)^2$ が最大となる第一主成分 $\zeta_1(s)$ を求めることからはじめる。実際の計算においては、関数データおよび $\zeta_1(s)$ を正規直交関数系で展開することで、固有値問題に帰着できる。

本堂・南・白土・水田(2004)は、関数主成分分析を用いて動体追跡照射データの解析結

果を報告している。動体追跡放射線照射とは、体内にある腫瘍に対し、その位置を追跡しながら適切なタイミングで放射線を照射する方法である。位置を知るために直径2mm程度の金マーカを利用している。金マーカの動きを、3次元空間を値域とする関数データとみなして解析した。関数主成分分析による第1主成分から第5主成分まで寄与率は順に、84.7%, 5.8%, 3.5%, 1.9%, 1.5%である。第1主成分は、Size factorと解釈できる。第2主成分は、呼吸における呼気の特徴を示している。第3主成分は、吸気の特徴を示している。第4主成分および第5主成分は、呼吸と心臓の鼓動とのタイミングを表していると思われる。

これ以外にも、関数データ解析法の応用例は、数多く報告されている。特に、Ramsay & Silverman (2002)では、骨の形状、人間の成長、手書き文字、など数多くの応用例が掲載されている。

３．シンボリックデータ解析法

関数データ解析と独立に提案され、発展してきた新しいデータの捉え方としてDiday(1987)らによるシンボリックデータがある。これは、従来のデータ構造の枠組みを一般化し、多様なタイプのデータを許容するシンボリックデータを定義し、それを解析する方法である。シンボリックデータのオブジェクトは、通常の質的データ、量的データ以外に、区間データ、分布などを同時に含むことができ、これらに重みをつけることもできる。さらに、シンボリックデータ自体を含むことができる。例えば、通常のクラスター分析により、クラスター(個体の集合)が得られるが、各クラスター自体をオブジェクトとして扱うことができる。すなわち、「クラスのクラス」、さらには、「クラスのクラスのクラス」のような階層構造を有するデータも扱うことができる。また、1980 年代に盛んであった人工知能(あるいは知識工学)との接点もあり、属性の継承などの概念も考慮されている。シンボリックデータ解析法のためのソフトウェアとしては、SODAS (Symbolic Official Data Analysis System)が公開されている。シンボリックデータ解析に関しては 2 つの成書、Billard & Diday (2006), Bock & Diday eds.(2000)および、電子ジャーナル http://www.jsda.unina2.it/newjsda/index.htm が参考になる。

このような多様なシンボリックデータのうち、特に、区間データに対する解析法が活発に研究されている。シンボリックデータ解析法の例として区間データに対する主成分分析法を紹介する。 p 次元区間データを $\mathbf{X}^I = \{\mathbf{x}_i^I, i = 1, \ldots, n\}$ とおく。ただし、 $\mathbf{x}_i^I = [\underline{\mathbf{x}_i}, \overline{\mathbf{x}_i}]$, $\underline{\mathbf{x}_i}, \overline{\mathbf{x}_i} \in \mathbf{R}^p$, $\underline{\mathbf{x}_i} \le \overline{\mathbf{x}_i}$ とする。この区間データに対する主成分分析法として、 p 次元空間における区間の全ての頂点に着目した頂点法と、区間データの重心 $\mathbf{c}_i = (\underline{\mathbf{x}_i} + \overline{\mathbf{x}_i}) \in \mathbf{R}^p$ に着目した中心法が提案されている。ともに、分散共分散行列の固有値問題に帰着できる。また、区間値をとる固有ベクトル $\mathbf{u}^I$ と固有値 λ^I による区間固有値問題 $\mathbf{X}^I\mathbf{u}^I = \lambda^I\mathbf{u}^I$ とする方法もある。

４．関数区間データに対する主成分分析法

関数データ解析とシンボリックデータ解析は、比較的独立して発展してきた。形式的には、関数をシンボリックデータとして扱うことができるので、関数データはシンボリックデータの一種と解釈することができる。しかし、「関数」を広く定義するとシンボリ

ックデータの大部分を記述できる。

しかし、両者の重要な差異は、その解析手法にあると思われる。関数データ解析では、関数を有限個の基底関数により近似することで、有限次元の多次元データに対する解析法が利用できる。シンボリックデータ、特に区間データでは、その「集合」全体について成立する解を求めるか、または簡易法として区間の頂点を利用する。

関数データ解析法とシンボリックデータ解析法を融合させることにより、データ解析の適用範囲を広げることが可能になると思われる。例えば、一次元関数区間データ $\{x_i^I(t), i=1,\ldots,n\}$(ただし、 $x_i^I(t)=[\underline{x_i(t)},\overline{x_i(t)}]$, $\underline{x_i(t)},\overline{x_i(t)}\in R^1$, $\underline{x_i(t)}\le\overline{x_i(t)}$)に対して、前節の関数データに対する主成分分析法を併用すれば、シンボリックデータ解析における頂点法または中心法に基づく主成分分析法を本データに適用するのは容易である。また、$\int\zeta_1(s)^2ds=1$の制約条件において $n^{-1}\sum_{i=1}^n(\iint_{x\in x_i^I(s)}\zeta_1(s)x\,dxds)^2$ が最大となる $\zeta_1(s)$ を求める問題とすることもできる。

５．おわりに

多様なデータ構造に対応した解析法として、関数データ解析法とシンボリックデータ解析法を紹介した。さらに、両者を組み合わせた手法の可能性について、主成分分析法を例にとって紹介した。S-PLUS で実行可能な関数データ解析用のソフトウェアは公開されている。シンボリックデータ解析については、（少なくとも報告者は）S-PLUS で直接実行できるソフトウェアで著名なものはないと思われる。しかし、S-PLUS が有するクラスやメソッドに対する柔軟な記述能力(垂水他,2002)はシンボリックデータ解析に適している。

今回、報告した内容は「多様なデータ構造に対応した解析法」の開発のための第一歩に過ぎない。統計科学の分野はもとより、知識工学を含む情報科学の分野との研究交流により、この目標が達成できると思われる。

参考文献

Billard,L., Diday,E. (2006), *Symbolic Data Analysis* - Conceptual Statistics and Data Mining -, Wiley.

Bock,H.H., Diday,E. eds. (2000), *Analysis of Symbolic Data*. - Exploratory Methods for Extracting Statistical Information from Complex Data-, Series: Studies in Classification, Data, and Knowledge Organization, 15. Springer-Verlag, Berlin.

Diday E. (1987). The symbolic approach in clustering and related methods of Data Analysis, *Classification and Related Methods of Data Analysis", Proc. IFCS, Aachen, Germany.*

Lifshits, M. A. (1995). *Gaussian Random Functions*. Kluwer Academic Publishers.

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

Mizuta, M. (2003a). Hierarchical clustering for functional dissimilarity data. *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics* 2003, Vol.V, 223-227.

Mizuta, M. (2003b). K-means method for functional data. *Bulletin of the International Statistical Institute, 54th Session*, Book 2, 69-71.

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

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

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

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd Edition, New York: Springer-Verlag.

Tokushige,S., Inada,K. and Yadohisa,H. (2001). Dissimilarity and related methods for functional data. *Proceeding of the International Conference on New Trends in Computational Statistics with Biomedical Applications*, 295-302.

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

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.

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.

垂水共之・越智義道・水田正弘・森 裕一・山本義郎 訳 John Chambers **著**(2002). **データによるプログラミング**, **森北出版**.

本堂義行・南 弘征・白土 博樹・水田 正弘(2004). **関数主成分分析による動体追跡照射データの解析、日本計算機統計学会第** 18 **回シンポジウム論文集** 5-8.

水田正弘(2005). **関数データとその解析法**, **知識と情報(日本ファジィ学会誌)**、17(4), 413-417.

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.

関数データ解析に関する重要な Web Page: **http://ego.psych.mcgill.ca/misc/fda/**
シンボリックデータ解析に関する重要な Web Page: **http://www.jsda.unina2.it/**