関数データ解析の数理的基礎

大阪大学 大学院基礎工学研究科 / 理化学研究所 革新知能統合研究センター 寺田 吉壱
Yoshikazu Terada
Graduate School of Engineering Science, Osaka University/
RIKEN Center for Advanced Intelligence Project (AIP)

1 はじめに

近年，計測技術の進歩に伴い，連続的・断続的に記録されるデータが多くなってきている．例えば，計量化学分野における近赤外線分光法 (NIR) に関連するデータ，運動に関連する軌道データなどが挙げられる．図 1 は，Kalivas (1997) で紹介されている 100 個の小麦サンプルに対する NIR スペクトラルデータを表している．縦軸は各周波数における光の吸収率を表しており，周波数に対して連続的に変化していることが見て取れる．また，図 1 は，Dockery et al. (1983) で紹介されている年齢と肺活量に関するデータである．各被験者に対して肺活量を複数の時点で観測している．このデータは，観測時点が被験者ごとに異なることが特徴である．赤い線は，ある 1 人の被験者に関する経時測定データである．このようなデータに対しては，背後のデータ発生機構として，実数空間上の確率分布を考えるよりも，ある（有界な）領域や区間上のランダムな関数（もしくは，確率過程）を考える方が自然である．ある領域や区間上で連続的・断続的に観測されたデータをランダムな関数や確率過程の実現値として捉えたデータ解析は関数データ解析 (FDA) と呼ばれ，統計科学分野において盛んに研究が進められている (Wang et al., 2016)．関数データ解析という言葉は， Ramsay (1982) によって導入されたものであるが，その考え方の歴史は古く Grenander (1950) や Rao (1958) に遡る．本稿では，関数データ解析に関連する論文ではあまり詳細に述べられることのない，最も基本的な仮定や関数データの特徴について簡単にまとめる．証明やより詳細な議論に関しては，例えば Hsing and Eubank (2015) を参照されたい．

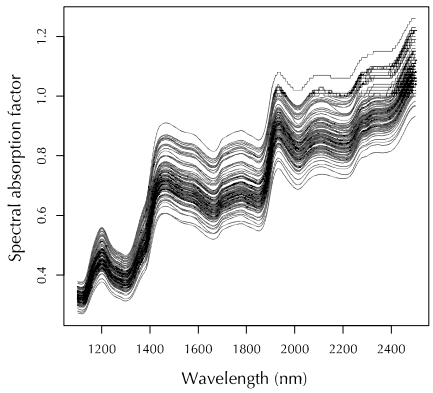

図 1: 小麦サンプルに対するスペクトラルデータ

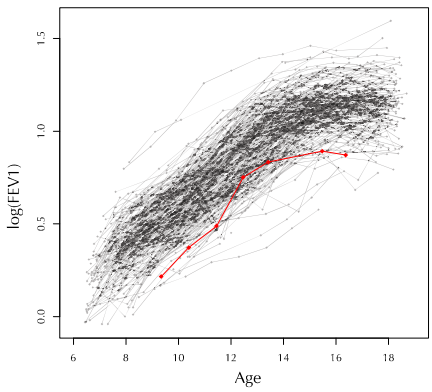

図 2: 肺活量に関する経時測定データ

2 関数データと平均二乗連続な確率過程

実際の観測は離散的なものであるが，関数データ解析ではその背後にランダムな関数を考えている．すなわち，対象 i の時点 $t_{ij} \in \mathcal{I}$（図 1 では周波数）の観測 $X_i(t_j)$ に対して，以下のようなモデルを考えている．

$$X_i(t_{ij}) = \mu(t_{ij}) + Z_i(t_{ij}) + \epsilon_{ij}$$

ここで，$\mathcal{I} \subset \mathbb{R}$ は有界閉区間，$\mu(t_j)$ は平均関数，Z_i は平均 0 のランダムな関数，ϵ_{ij} は Z_i とは独立な観測誤差である．$\mathcal{I}$ は一般のコンパクトな距離空間としても以下の議論は成り立つが，イメージしやすいように有界閉区間のみを扱う．本稿では，関数データの本質的な性質を説明するために，最も理想的な観測誤差のない連続観測の状況を考える．

$$X_i(t) = \mu(t) + Z_i(t) \quad (t \in \mathcal{I})$$

関数データ解析に関する文献においては，このような仮定を下に議論を進めることが多い．

連続な観測 X_i を区間 $\mathcal{I}$ 上の二乗可積分な関数から構成される空間 $L_2(\mathcal{I})$ 上に値をとる確率変数と考えるか，確率過程と考えるかという 2 つの考え方がある．Hilbert 空間上に値をとる確率変数として扱う場合は，議論が抽象的なものとなるため，ここでは，よりイメージしやすいように X_i は確率過程として扱う．これらの 2 つ考え方は，後述するある条件の下では同一なものとなる．

$(\Omega, \mathscr{F}, \mathbb{P})$ を確率空間とし，確率過程 $\{X(t,\omega) \mid t \in \mathcal{I}, \omega \in \Omega\}$ を考える．X の平均関数は，

$$\mu(t) = \mathbb{E}[X(t)] \tag{1}$$

で定義され，共分散関数は，$s, t \in \mathcal{I}$ に対して，

$$K(s,t) := \mathrm{Cov}(X(s), X(t)) \tag{2}$$

で定義される．平均関数と共分散関数が well-defined な確率過程を *second-order process* と呼ぶ．

確率過程の基本的な仮定は，各 t に対して $X(t,\cdot)$ が確率変数（すなわち，$\mathscr{F}$-可測）となることである．しかし，これだけでは，$X(\cdot,\omega)$ が $L^2(\mathcal{I})$ 上の random element とならないことに注意されたい．$X(t,\omega)$ が *jointly measurable*，すなわち，$X(t,\omega)$ が product σ-field $\mathscr{B}(\mathcal{I}) \times \mathscr{F}$ に関して可測となる場合，各 $\omega \in \Omega$ に対して $X(\cdot,\omega)$ は $\mathcal{I}$ 上の可測関数となる．

定理 2.1 (Theorem 7.4.1 in Hsing and Eubank (2015))**.** 確率過程 X が *jointly measurable* であり，各 $\omega \in \Omega$ に対して $X(\cdot,\omega) \in L^2(\mathcal{I})$ が成り立つとする．このとき，X は $L^2(\mathcal{I})$ の *random element* となる．

Jointly measurable は確認することが難しい条件であるが，各 $t \in \mathcal{I}$ に対して $X(t,\cdot)$ が可測であり，各 $\omega \in \Omega$ に対して $X(\cdot,\omega)$ が連続であるとき，$X(t,\omega)$ は jointly measurable となる．

任意の t 及び任意の t に収束する $\mathcal{I}$ 上の列 $\{t_n\}_{n\in\mathbb{N}}$ に対して

$$\lim_{n\to\infty} \mathbb{E}\left[\{X(t_n) - X(t)\}^2\right] = 0 \tag{3}$$

を満たす second-order process に焦点を当てる．このような性質を満たす確率過程を平均二乗連続な確率過程 (*mean-square continuous process*; MSCP) という．確率過程の平均二乗連続性は，平均関数と共分散関数の連続性を特徴づける性質である．

定理 2.2 (Theorem 7.3.1 in Hsing and Eubank (2015))**.** X を *second-order process* とする．このとき，X が平均二乗連続であるための必要十分条件は，平均関数と共分散関数が連続となることである．

もし $\mu(t)$ が連続ならば，共分散関数 $K(s,t)$ が各 (s,t) において連続となるための必要十分条件は，K が “diagonal points” で連続となることである．関数データ解析に関連する多くの文献では，データ X は jointely measurable な平均二乗連続な確率過程と考えている．

3　共分散関数と Mercer の定理

平均二乗連続な確率過程に対してはその共分散関数 K は連続となる．そこで，$L^2(\mathcal{I})$ 上の積分作用素

$$(\mathscr{K} f)(t) = \int K(t,s) f(s)\, ds$$

を考える．これは X の共分散作用素と呼ばれる．また，任意の $a_1, \ldots, a_n \in \mathbb{R}$ に対して

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(t_i, t_j) = \mathrm{Var}\left(\sum_{i=1}^{n} a_j X(t_j)\right)$$

より，共分散関数 K は非負定値 (nonnegative definite) となる．K は nonnegative definite であるため，K を再生核とする再生核 Hilbert 空間 (RKHS) $\mathcal{H}_K$ を考えることができる．関数データ解析においても多変量解析の手法を関数データへと拡張する際などに，RKHS $\mathcal{H}_K$ は重要な役割を担う．ただし，関数データ X は $\mathcal{H}_K$ の元とはならないことに注意されたい．

定理 3.1 (Theorem 7.5.4 in Hsing and Eubank (2015))**.** X を $E[\|X\|^2] < \infty$ を満たす $\mathcal{H}_K$ の *random element* とする．ここで，K は X の共分散関数である．このとき，$\mathcal{H}_K$ は有限次元でなければならない．

共分散作用素 $\mathscr{K}$（もしくは，その kernel K）は，従来の多変量データ解析における共分散行列に対応している．以下の定理の後半部分は，Mercer の定理と呼ばれ，共分散関数の固有値分解に対応している．

定理 3.2. K を平均二乗連続な確率過程 X の共分散関数とする．このとき，ある非負単調減少列 $\{\lambda_j\}_{j\in\mathbb{N}}$ とある正規直交系 $\{e_j\}_{j\in\mathbb{N}}$ が存在して，

$$\int K(s,t) e_j(s)\, ds = \lambda_j e_j(t) \quad (t \in \mathcal{I},\ j \in \mathbb{N})$$

が成り立つ．λ_j と e_j は，それぞれ共分散作用素 $\mathscr{K}$ の固有値と固有関数と呼ばれる．さらに，

$$K(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t) \quad (s,t \in \mathcal{I})$$

が成り立つ．ここで，この収束は $\mathcal{I}\times\mathcal{I}$ 上の一様かつ絶対な収束である．

連続観測を想定している場合，平均関数や共分散関数は従来の多変量解析と同様に，標本平均や標本共分散を考えることで簡単に推定できる．また，固有値 λ_j や固有関数 e_j は，関数主成分分析 (FPCA) によって推定することができる．離散的な観測や観測誤差がある場合の平均・共分散関数の推定については，Zhang and Wang (2016) などを参考にされたい．

4　Karhunen-Loéve (KL) 展開と関数データの無限次元性

前節では，共分散関数の固有値分解について述べた．本節では，関数データ解析において最も重要となる平均二乗確率過程の展開を考える．$L^2(\mathcal{I})$ 上の関数 f は，$L^2(\mathcal{I})$ の完全正規直交系 (CONS) $\{e_j\}_{j\in\mathbb{N}}$ を用いて，

$$f(t) = \sum_{j=1}^{\infty} \langle f, e_j\rangle e_j(t) \quad (t \in \mathcal{I})$$

と展開できる．ここで，$\langle f, e_j \rangle := \int f(t)e_j(t)\,dt$ である．平均二乗連続な確率過程 X に対しては，以下の Karhunen-Lòeve 展開が成り立つ．

定理 4.1 (e.g., Theorem 7.3.5 in Hsing and Eubank (2015))．X を平均 0 の平均二乗連続な確率過程とする．$\{(\lambda_j, e_j)\}$ を X の共分散関数から定義される共分散作用素 $\mathscr{K}$ の固有値・固有関数とする．このとき，以下の展開

$$X(t) = \sum_{j=1}^{\infty} I_X(e_j)e_j(t) \quad (t \in \mathcal{I})$$

が成り立つ．ここで，$I_X(e_j)$ $(j \in \mathbb{N})$ は以下を満たす実数値確率変数である．

$$\mathbb{E}[I_X(e_j)] = 0, \quad \mathbb{E}[I_X(e_i)I_X(e_j)] = \lambda_j \delta_{ij}$$

この展開は，以下の意味で成り立つ．

$$\lim_{n\to\infty} \sup_{t\in E} \mathbb{E}\left[\left(X(t) - \sum_{j=1}^{n} I_X(e_j)e_j(t)\right)^2\right] = 0$$

上述の定理で，$I_X(e_j)$ は $\langle X, e_j \rangle$ に対応する確率積分である．一般に，$X(\cdot, \omega)$ は $L^2(\mathcal{I})$ の元とは限らないため，X の確率積分を考える必要がある．実際に，jointly measurable の仮定の下では，確率変数 $I_X(e_j)$ は

$$I_X(e_j) = \langle X, e_j \rangle = \int X(t)e_j(t)\,dt$$

と表現できる．

共分散作用素 $\mathscr{K}$ が strictly positive であれば $\{e_j\}$ は $L^2(\mathcal{I})$ の完全正規直交系となる．また，この様な状況でない場合でも，零核の完全正規直交系を $\{e_j\}$ に適当な順序で付け足すことで固有関数に基づく完全正規直交系を構成することができる．そのため，平均二乗連続な確率変数 X の平均関数 μ は以下のように展開できる．

$$\mu = \sum_{j=1}^{\infty} \langle \mu, e_j \rangle e_j$$

したがって，平均二乗連続な確率変数 X は以下のように展開できる．

$$X(t) = \sum_{j=1}^{\infty} W_j e_j(t) \quad (t \in \mathcal{I})$$

ここで，W_j は以下を満たす確率変数である．

$$\mathbb{E}[W_j] = \langle \mu, e_j \rangle, \quad \mathrm{Cov}(W_i, W_j) = \begin{cases} \lambda_i & (i = j) \\ 0 & (i \neq j) \end{cases}$$

これにより，関数データ X は無限個の実数値確率変数によって構成されていることがわかる．この無限次元性は，関数データ解析を考えていく上で重要な性質となる．

5　おわりに

本稿では，統計科学の分野で研究が進んでいる関数データ解析について，関連する研究論文ではあまり詳細に触れられることのない基本的な仮定や重要な性質などの数学的な基礎を述べた．関数データ解析の各手法の性質は，これらの仮定を下に議論される．

最後に，関数データ解析に関するいくつかの書籍について言及しておく．Ramsay and Silverman (2005) では，各対象の関数データが基底関数展開によって表現できる場合を中心に，様々な多変量解析を関数データへと拡張した方法が詳細に述べられている．Ramsay and Silverman (2002) では，Ramsay and Silverman (2005) で紹介されている解析方法の実データへの応用が詳細に述べられている．Ramsay et al. (2009) では，Ramsay and Silverman (2005) や Ramsay and Silverman (2002) で紹介されている各手法の統計ソフトウェアである R や matlab での実行方法について詳細に解説されている．これらの書籍は，少し古典的なものであるが，観測時点が密であるような状況において，関数データ解析を具体的な実行方法を知るために有用である．理論的な内容を含む最近の書籍は以下の通りである．Horváth and Kokoszka (2012) では，関数データ X が連続的に観測され，$L^2(\mathcal{I})$ の元であることを仮定して関数データ解析の基本的な手法とその理論的な性質を紹介している．また，R での実行方法についても述べられている．そして，Kokoszka and Reimherr (2017) では，連続観測でない場合も含めて関数データ解析の様々な手法，それらの理論的な性質，R での実行方法を紹介している．Hsing and Eubank (2015) では，関数データ解析の理論を考える上で必要となる関数解析の基礎からはじめ，関数データ解析の理論的な側面について詳細に述べられている．また，共分散関数を再生核とする RKHS がどの様に関数データ解析において重要な役割を担うかについても言及されている．関数データ解析の重要性は応用分野でも徐々に認識されており，今後のさらなる発展が期待される．

参考文献

[1] Dockery, D.W., Berkey, C.S., Ware, J.H., Speizer, F.E. and Ferris, B.G. (1983). Distribution of FVC and FEV1 in children 6 to 11 years old. *American Review of Respiratory Disease*, **128**, 405–412.

[2] Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv for Matematik*, **1**, 195–277.

[3] Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, Springer.

[4] Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, Wiley.

[5] Kalivas, J. H. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory System*, **37**, 255–259.

[6] Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*, CRC press.

[7] Ramsay J.O. (1982). When the data are functions. *Psychometrika*, **47**, 379–396.

[8] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*, Springer.

[9] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Springer.

[10] Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*, Springer.

[11] Rao, C.R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, **14**, 1–17.

[12] Rice, J. and Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B*, **53**, 233–243.

[13] Wang, J. L., Chiou, J. M., and Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

[14] Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Annals of Statistics*, **44**, 2281–2321.