関数データの主成分分析と判別分析について

山西芳裕（岡山大学大学院自然科学研究科）
田中　豊（岡山大学環境理工学部）

1　はじめに

関数データ解析は曲線のデータ解析法の一つであり、データはベクトルとしてではなく関数として構成される。つまり、適当な観測時刻 $t_1, t_2, \cdots$ で観測されたデータ $\{x(t_j), j = 1, 2, \cdots\}$ を、時刻を変数とする多変量データではなく、関数 $x(t)$ として取り扱う。この分野の回帰分析や主成分分析、正準相関分析などが、Ramsay and Silverman (1997) によって提案されている。本報告では、関数データ解析における主成分分析、判別分析をとりあげ、データ間における時間的な変動を解析する手法として、ペナルティ付き関数主成分分析の有効性を示す。また、正準判別分析を関数データ解析の文脈において定義し、群間平方和、郡内平方和の観点からペナルティを用いた正則化の導入を考える。また、各個体が複数の関数データを持つ場合においても適用を可能にするため、ペナルティ付き関数正準判別分析を拡張し、基底展開を用いたアルゴリズムを導出する。さらに、クロスバリデーションを利用した平滑化パラメータの選択法を提案する。

2　関数主成分分析

2.1　関数主成分分析

N 個の関数データ $x_i(s)$ $(i = 1, \cdots, N)$ に対して、重み関数 $\xi(s)$ を用い、合成変量 $f_i = \int \xi(s)x_i(s)ds$ の分散の最大化を考える。つまり、重み関数 ξ は分散

$$PCASV = \int\int \xi(s)v(s,t)\xi(t)dtds \tag{1}$$

を最大化するように選ばれる。ここで、$v(s,t)$ は分散共分散関数である。制約

$$\int \xi_l(t)^2 dt = 1, \int \xi_l(t)\xi_m(t)dt = 0 \ (l < m) \tag{2}$$

に従いながら、$PCASV$ を最大化することは次のような固有値問題に置き換えられる。

$$\int v(s,t)\xi(t)dt = \rho\xi(t). \tag{3}$$

2.2　ペナルティ付き関数主成分分析

正則化を導入するため、重み関数 $\xi(s)$ の曲率である二乗積分をペナルティとして $PEN(\xi) = \int \left\{\frac{D^2\xi(s)}{Ds^2}\right\}^2 ds$ と定義する。ペナルティ付き標本分散は平滑化パラメータ λ を用いて、

$$PCAPSV = \frac{PCASV}{\| \xi \|^2 + \lambda \times PEN(\xi)} \tag{4}$$

と書ける。つまり、主成分の分散を最大化するということと、$\xi(s)$ を平滑化するというトレードオフを、平滑化パラメータ λ が制御することを意味する。すると、次のようなペナルティ付き固有方程式が得られる。

$$\int v(s,t)\xi(t)dt = \rho(I + \lambda D^4)\xi(s). \tag{5}$$

2.3　アルゴリズム

関数データ $x_i(s)$ と主成分重み関数 $\xi(s)$ が、基底関数の集合 $\phi_p(s) = (\phi_{1p}(s), \cdots, \phi_{Hp}(s))^T$ の線形結合によって近似されていると仮定すると、それぞれ以下のように基底展開される。

$$x_i(s) = \sum_{h=1}^{H} C_{ih}\phi_h(s) = \mathbf{C}_i\phi(s), \quad \xi(s) = \sum_{h=1}^{H} y_h\phi_h(s) = \mathbf{y}^T\phi(s). \tag{6}$$

ここで、$\mathbf{C}$ は $N \times H$ 係数行列、 $\mathbf{y}$ は $H \times 1$ 係数行列、 H は基底関数の数を表す。このとき、$\mathbf{V}$ をベクトル C_i の分散共分散行列とし、$\mathbf{J}_\phi = \int \phi(s)\phi(s)^T ds, \mathbf{K}_\phi = \int (D^2\phi(s))(D^2\phi(s))^T ds$ と置くと、ペナルティ付き関数方程式は、次のような一般化固有値問題に帰着する。

$$(\mathbf{J}_\phi \mathbf{V} \mathbf{J}_\phi) y = \rho(\mathbf{J}_\phi + \lambda\mathbf{K}_\phi) y. \tag{7}$$

ここで得られる固有ベクトル y を用い、式 (6) から主成分重み関数 ξ を求めることができる。

2.4　クロスバリデーションによる平滑化パラメータ λ の選択

もとのデータが主成分によってどれほど説明されているかに注目し、クロスバリデーション得点を次のように定義する。

$$CV(\lambda) = \sum_{i=1}^{N} \parallel x_i(s) - \sum_{m=1}^{M}\sum_{l=1}^{M}(\mathbf{G}^{-1})_{ml}\left(\int \xi_m^{[-i]}(t)x_i(t)dt\right)\xi_l^{[-i]}(s) \parallel^2 . \tag{8}$$

ここで、$\mathbf{G}$ は、(m, l) 要素が $\int \xi_m(t)\xi_l(t)dt$ である $M \times M$ 行列、M はとりあげる主成分の数、$\xi^{[-i]}$ は i 番目の個体を落として計算したことを表す。この $CV(\lambda)$ を最小にする λ を、平滑化パラメータとして決定する。

3　数値例

ペナルティ付き関数主成分分析を、日本の 1999 年の気象のデータに適用する。観測地点として 60 都市を選び、それぞれの観測所で観測された毎日の平均気温を関数データと見なす。図 1 は、関数データの例として、気温と日照時間のデータを示している。図 2 では、第 1、第 2 主成分の重み関数（左図)、1 軸と 2 軸の主成分得点の散布図（右図）を示している。それぞれの重み関数より、第 1 主成分は年間を通してすべて正なので総合的な尺度、第 2 主成分は夏と冬のコントラストを表すものとして解釈できる。

4　関数正準判別分析

4.1　関数正準判別分析

関数データ $x(s)$ に対して、任意の重み関数 $a(s)$ を用いて、$z = \int a(s)x(s)ds$ のような線形結合 z をつくり、この値によって判別することを考える。$N(= n_1 + \cdots + n_G)$ 個の個体の各々に対して、合成変量

$$z_i^{(g)} = \int a(s)x_i(s)ds \qquad (g = 1, \cdots, G; \quad i = 1, \cdots, n_g) \tag{9}$$

を構成する。ここで、G は群の数を表す。群間と郡内の平方積和関数を、それぞれ

$$b(s,t) = \sum_{g=1}^{G} n_g(\bar{x}^{(g)}(s) - \bar{x}(s))(\bar{x}^{(g)}(t) - \bar{x}(t)), \quad w(s,t) = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\bar{x}_i^{(g)}(s) - \bar{x}^{(g)}(s))(\bar{x}_i^{(g)}(t) - \bar{x}^{(g)}(t)) \tag{10}$$

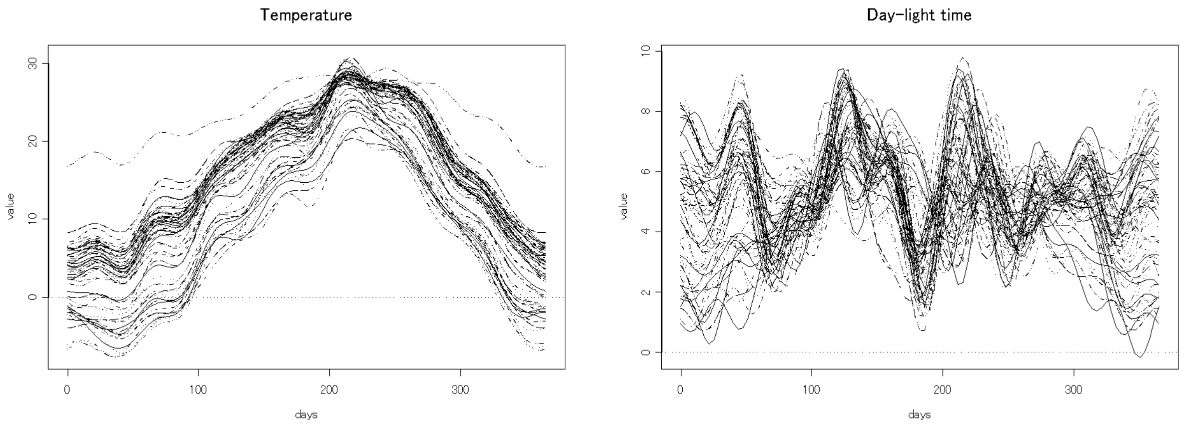

図 1: 関数データの例：気温、日照時間

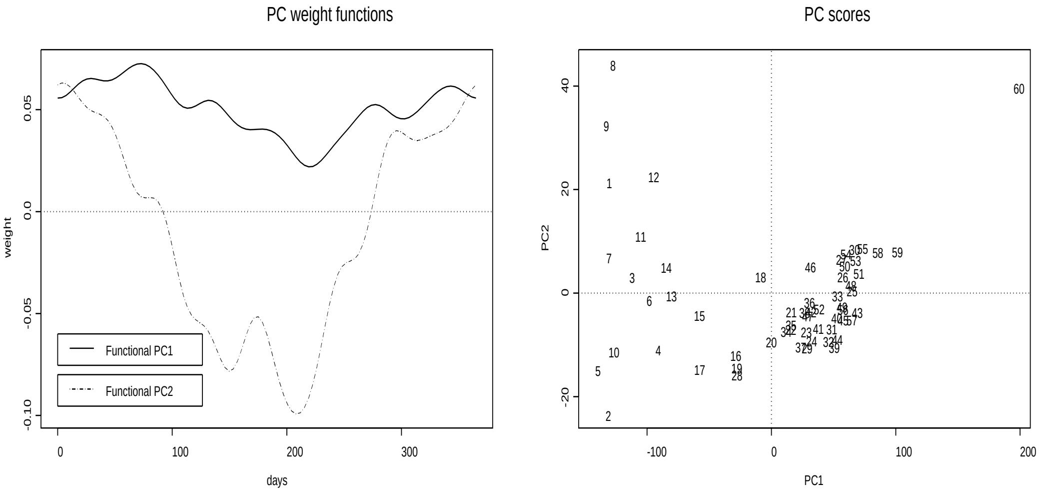

図 2: 主成分重み関数（実線：第 1、点線：第 2）と主成分得点

と定義する。z により G 個の群がよく判別されるということを、群間平方和 S_B が郡内平方和 S_W に対して大きくなると考えると、以下のような比を最大化するような重み関数 $a(s)$ を推定するという問題になる。

$$\rho = S_B/S_W = \int\int a(s)b(s,t)a(t)dsdt \Big/ \int\int a(s)w(s,t)a(t)dsdt. \tag{11}$$

4.2 ペナルティ付き関数正準判別分析

正則化を導入するため、重み関数 $a(s)$ の曲率である二乗積分をペナルティとして $PEN(a) = \int \left\{ \frac{D^2 a(s)}{Ds^2} \right\}^2 ds$ と定義する。ここで、群間平方和 S_B の郡内平方和 S_W に対する比の分母に付加することを考える。

$$\rho = \frac{S_B}{S_W + \lambda PEN(a)} = \frac{\int\int a(s)b(s,t)a(t)dsdt}{\int\int a(s)w(s,t)a(t)dsdt + \lambda PEN(a)}. \tag{12}$$

これは、群がうまく判別されるということと、$a(s)$ を平滑化するというトレード・オフを、平滑化パラメータ λ が制御することを意味する。

5　拡張関数正準判別分析

5.1　拡張関数正準判別分析

実際の適用においては、各個体が複数の種類の関数データを持つ場合が考えられる。ここでは、P 個の関数データが各個体に対し得られていると仮定する。関数データ $x_p(s)$ に対して、任意の重み関数 $a_p(s)$ を用いて、$z = \sum_{p=1}^{P} \int a_p(s)x_p(s)ds$ のような線形結合 z をつくり、この値によって判別することを考える。$N(= n_1 + \cdots + n_G)$ 個の個体の各々に対して、合成変量

$$z_i^{(g)} = \sum_{p=1}^{P} \int a_p(s)x_{ip}(s)ds \qquad (g = 1, \cdots, G; \quad i = 1, \cdots, n_g) \tag{13}$$

を構成し、このときの群間と郡内の平方積和関数を、それぞれ

$$b_{lm}(s,t) = \sum_{g=1}^{G} n_g(\bar{x}_l^{(g)}(s) - \bar{x}_l(s))(\bar{x}_m^{(g)}(t) - \bar{x}_m(t)), \quad w_{lm}(s,t) = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\bar{x}_{il}^{(g)}(s) - \bar{x}_l^{(g)}(s))(\bar{x}_{im}^{(g)}(t) - \bar{x}_m^{(g)}(t)) \tag{14}$$

と定義する。すると、以下のような比を最大にするような重み関数 $a_p(s)$ を推定するという問題になる。

$$\rho = S_B/S_W = \sum_{l=1}^{P}\sum_{m=1}^{P} \int\int a_l(s)b_{lm}(s,t)a_m(t)dsdt / \sum_{l=1}^{P}\sum_{m=1}^{P} \int\int a_l(s)w_{lm}(s,t)a_m(t)dsdt. \tag{15}$$

5.2　ペナルティ付き拡張関数正準判別分析

正則化を導入するため、ペナルティを $PEN(a_p) = \int \left\{ \frac{D^2 a_p(s)}{Ds^2} \right\}^2 ds$ と定義すると、群間平方和 S_B の郡内平方和 S_W に対するペナルティ付きの比を最大化する問題となる。

$$\rho = \frac{S_B}{S_W + \sum_{p=1}^{P}\lambda_p PEN(a_p)} = \frac{\sum_{l=1}^{P}\sum_{m=1}^{P}\int\int a_l(s)b_{lm}(s,t)a_m(t)dsdt}{\sum_{l=1}^{P}\sum_{m=1}^{P}\int\int a_l(s)w_{lm}(s,t)a_m(t)dsdt + \sum_{p=1}^{P}\lambda_p PEN(a_p)}. \tag{16}$$

5.3　アルゴリズム

$x_p(s)$、$a_p(s)$ が、基底関数の集合 $\phi_p(s) = (\phi_{1p}(s), \cdots, \phi_{Hp}(s))^T$ の線形結合によって近似されていると仮定すると、それぞれ以下のように基底展開される。

$$x_p(s) = \sum_{h=1}^{H} C_{hp}\phi_{hp}(s) = \mathbf{C}_p\phi_p(s), \quad a_p(s) = \sum_{h=1}^{H} y_{hp}\phi_{hp}(s) = \mathbf{y}_p^T\phi_p(s). \tag{17}$$

ここで、$\mathbf{C}_p$ は $N \times H$ 係数行列、 $\mathbf{y}$ は $H \times 1$ 係数行列、 H は基底関数の数を表す。このとき、eq.(16) は、

$$\rho = \frac{S_B}{S_W + \sum_{p=1}^{P}\lambda_p PEN(a_p)} = \frac{\sum_{l=1}^{P}\sum_{m=1}^{P}\mathbf{y}_l^T\mathbf{J}_{\phi_l}\mathbf{B}_{lm}\mathbf{J}_{\phi_m}\mathbf{y}_m}{\mathbf{y}_l^T\mathbf{J}_{\phi_l}\mathbf{W}_{lm}\mathbf{J}_{\phi_m}\mathbf{y}_m + \lambda_p\sum_{p=1}^{P}\mathbf{y}_p^T\mathbf{K}_{\phi_p}\mathbf{y}_p} \tag{18}$$

と書き直せる。ここで、 $\mathbf{J}_{\phi_p} = \int \phi_p(s)\phi_p(s)^T ds$ 、$\mathbf{K}_{\phi_p} = \int \left(\frac{D^2\phi_p(s)}{Ds^2}\right)\left(\frac{D^2\phi_p(s)}{Ds^2}\right)^T ds$ である。ρ を最大にするような係数ベクトル $\mathbf{y}_p$ を選ぶということは、以下のような一般化固有値問題に帰着する。

$$\mathbf{J}_{\phi_l}\left(\sum_{m=1}^{P}\mathbf{B}_{lm}\mathbf{J}_{\phi_m}\mathbf{y}_m\right) = \rho\left(\mathbf{J}_{\phi_l}(\sum_{m=1}^{P}\mathbf{W}_{lm}\mathbf{J}_{\phi_m}\mathbf{y}_m) + \lambda_p\sum_{p=1}^{P}\mathbf{K}_{\phi_p}\mathbf{y}_p\right). \tag{19}$$

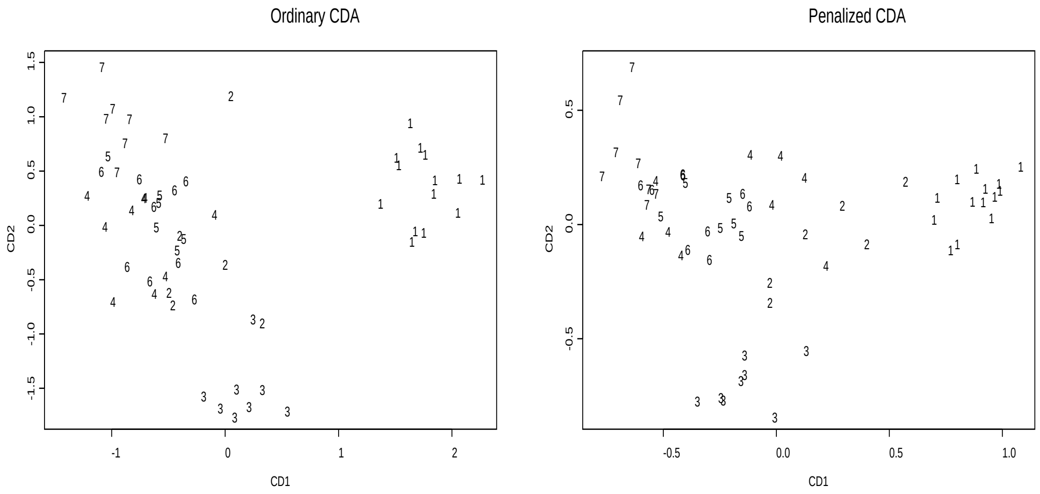

図 3: 正準判別得点（1 軸と 2 軸の散布図）: ペナルティ無し（左図）とペナルティ付き（右図）

5.4　クロスバリデーションによる平滑化パラメータの選択

適切な正則化を導入するため、最適な平滑化パラメータ λ を選ぶ必要があるが、ここではクロスバリデーションの利用した選択法を提案する。i 番目の個体を落としたときの重み関数 $a_p^{[-i]}(s)$ に基づく正準得点を

$$z_i^{(g)*} = \sum_{p=1}^{P} \int a_p^{[-i]}(s) x_{ip}(s) ds \qquad (g = 1, \cdots, G; \quad i = 1, \cdots, n_g) \tag{20}$$

とおく。ここで、$[-i]$ は i 番目の個体を落とすことを意味する。正準得点の空間上において、個体の属する群の平均 $\bar{z}^{(g)*}$ への距離 d_i^W が、全体の平均 $\bar{z}^*$ への距離 d_i^T に対して小さくなればよいという観点から、以下のようなクロスバリデーション得点を構成でき、これが最小になるような $\lambda_p (p = 1, \cdots, P)$ を選ぶ。

$$CV(\lambda_1, \cdots, \lambda_P) = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \frac{d_i^W}{d_i^T} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \frac{d(z_i^{(g)*}, \bar{z}^{(g)*})}{d(z_i^{(g)*}, \bar{z}^*)}. \tag{21}$$

6　数値例

ペナルティ付き拡張関数正準判別分析を、日本の 1999 年の毎日の気象データに適用する。観測地点として 60 都市を選び、それぞれ、気温、降水量、日照時間から構成される関数データであるとみなす。判別グループは、1:北海道、2:東北、3:関東、4:中部、5:関西、6:中国四国、7:九州沖縄とした。図 3 では、1 軸と 2 軸のグループ別の正準判別得点の散布図を示している。左側がペナルティをつけないとき、右側がつけたときのものである。それぞれの地方間における気候の特色の相対的な違いを理解することができる。数値的には、ペナルティを導入すると、グループ間の違いが少しぼやけるのが分かる。図 4 で、気温、降水量、日照時間それぞれに対する、第一正準判別重み関数（実線）と第二正準判別重み関数（点線）を示している。左側がペナルティをつけないとき、右側がつけたときのものである。正則化の導入によって、重み関数の時間的なコントラストが解釈しやすくなることが分かる。

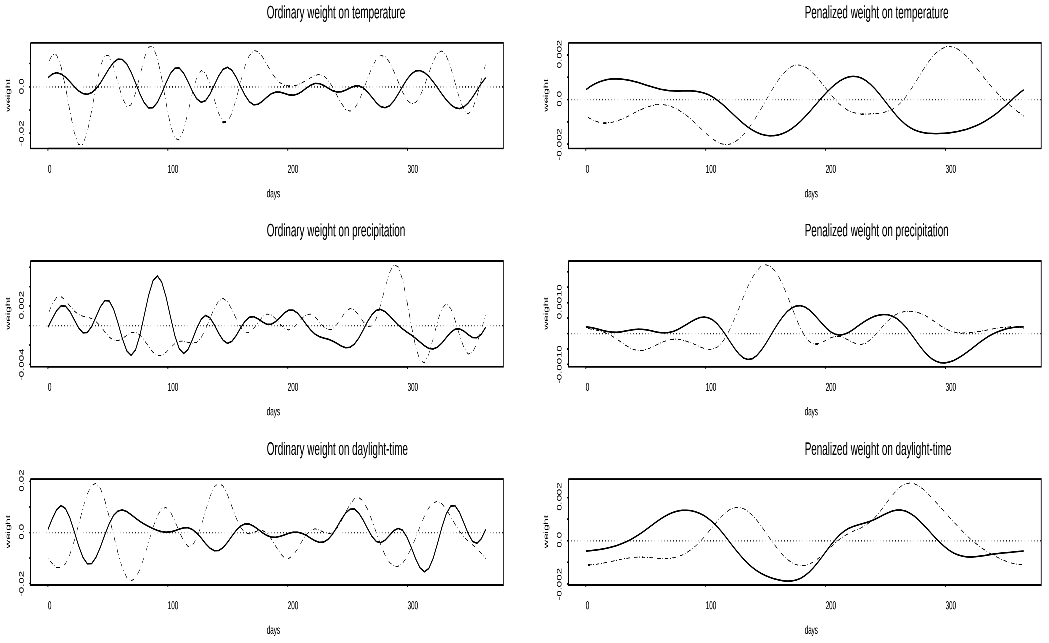

図 4: 正準判別重み関数（実線：第 1、点線：第 2）：ペナルティ無し（左図）とペナルティ付き（右図）

参考文献

[1] 気象庁監修 (1999). 平成 11 年アメダス年報,（財）気象業務支援センター.

[2] 山西芳裕，田中豊 (2000). 関数データの主成分分析：感度分析と数値的検討, 第 14 回日本計算機統計学会大会, 論文集 92-95

[3] 山西芳裕，田中豊 (2001). 関数データのペナルティ付き正準判別分析, 第 15 回日本計算機統計学会シンポジウム, 論文集 97-100

[4] Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions, Psychometrika, 51, 285-311.

[5] Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *Annals of Statistics*, 23, 73-102.

[6] Leurgans, S., Moyeed, R. and Silverman, B. (1993). Canonical correlation analysis when the data are curves, *J. Royal Statist. Soc.*, B 55, 725-740.

[7] Ramsay, J. and Dalzell, C. (1991). Some tools for functional data analysis (with discussion), *J. Royal Statist. Soc.*, B 53, 539-572.

[8] Ramsay, J. O. and Silverman, B.W. (1997). Functional Data Analysis, *Springer*.

[9] Yamanishi, Y. and Tanaka, Y. (2000). Sensitivity Analysis in Functional Principal Component Analysis, *Proceedings of the Seventh Japan-China Symposium on Statistics*, 37-40

[10] Yamanishi, Y., Tanaka, Y. (2000). Principal Component Analysis for Functional Data, *Journal of Faculty of Environmental Science and Technology, Okayama University*, Vol.6, 25-34.