3 カーネルの実際
-3.2　カーネル正準相関分析

10NM733X　林華

3.2.1　正準相関分析の復習-1

- 正準相関分析(CCA)：2種類の多変量データX,Yに依存関係を調べる線形手法である。

•X は p 次元のデータをn 個集めた n×p のデータ行列．
•Y は q 次元のデータをn 個集めた n×q のデータ行列．p≤q としておく．
共分散行列：
•S_{XX} =1/ n $\tilde{X}^T \tilde{X}$
•S_{YY} =1/n $\tilde{Y}^T \tilde{Y}$
•S_{XY} =1/n $\tilde{X}^T \tilde{Y}$
（ただし，$\tilde{X}$ と $\tilde{Y}$ は，それぞれ X とY から，それぞれの列方向の平均値を，各要素から引いた平均偏差行列．）
p 次元ののベクトル a とq 次元ののベクトル b を導入．これらのベクトルと，$\tilde{X}$ や$\tilde{Y}$とを一次結合したベクトルの相関を考える．

$$\rho = \max_{a \in R^m, b \in R^l} \frac{a^T S_{XY} b}{\sqrt{a^T S_{XX} a}\sqrt{b^T S_{YY} b}}$$

この式を最大化するa とb を，第1正準相関ベクトルと呼び，a 1 とb 1 で表す．

第k 正準相関ベクトル(k=1,...,p)は， a_i S_{XX} a_i =1, b_i S_{YY} b_i =1を満たすものの中で，ρを最大化するように求める．

3.2.1　正準相関分析の復習-2

ここでV_{XY},V_{XX},V_{YY}は標本分散共分散行列で、上の最大化問題は

$$\max a^T V_{XY} b \qquad subject \quad to \qquad a^T V_{XX} a = b^T V_{YY} b = 1$$

という形に書き直せる．さらにLagrange乗数法を適用することにより、以下の一般化固有値問題の最大固有値に一致

$$\begin{pmatrix} O & V_{XY} \\ V_{YX} & O \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \rho \begin{pmatrix} V_{XX} & O \\ O & V_{YY} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

3.2.2　カーネルCCA-1

- 独立同分布データX,Y
- 対応する再生核ヒルベルト空間H_X,H_Y
- 特徴写像Φ_X,Φ_Y、 $\tilde{\Phi}_X(X_i)$,は標本平均を差し引いた特徴ベクトル

正規相関を与える方向ベクトルf　H_X, g　H_Yと正準関数ρは次と定義

$$\rho = \max_{f \in H_x, g \in H_y} \frac{\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_x(X_i) \right\rangle_{H_x} \left\langle g, \tilde{\Phi}_y(Y_i) \right\rangle_{H_y}}{\sqrt{\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_x(X_i) \right\rangle_{H_x}^2} \sqrt{\sum_{i=1}^{N} \left\langle g, \tilde{\Phi}_y(Y_i) \right\rangle_{H_y}^2}}$$

ここで上式を最大にするf　H_X, g　H_Yは下式の形と考えればよい。

$$f = \sum_{i=1}^{N} \alpha_i \tilde{\Phi}_x(X_i), \qquad g = \sum_{i=1}^{N} \beta_i \tilde{\Phi}_y(Y_i)$$

$$\rho = \max_{\alpha \in R^N, \beta \in R^N} \frac{\alpha^T \tilde{K}_X \tilde{K}_Y \beta}{\sqrt{\alpha^T \tilde{K}_X^2 \alpha} \sqrt{\beta^T \tilde{K}_Y^2 \beta}}$$

3.2.2　カーネルCCA-2

問題：　　$u = \tilde{K}_X \alpha , \upsilon = \tilde{K}_Y \beta$ とすれば、

$$\rho = \max_{u \in R(K_X), \upsilon \in R(K_Y)} \frac{u^T \upsilon}{\|u\| \|\upsilon\|}$$

対策：正則化係数ε_Nを導入して正則化

$$\|f\|^2 = \alpha^T \tilde{K}_X \alpha , \|g\|^2 = \beta^T \tilde{K}_Y \beta \quad \text{により}$$

$$\rho = \max_{f \in H_x, g \in H_y} \frac{\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_x(X_i) \right\rangle_{H_x} \left\langle g, \tilde{\Phi}_y(Y_i) \right\rangle_{H_y}}{\sqrt{\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_x(X_i) \right\rangle^2_{H_x} + N\varepsilon_N \|f\|^2} \sqrt{\sum_{i=1}^{N} \left\langle g, \tilde{\Phi}_y(Y_i) \right\rangle^2_{H_y} + N\varepsilon_N \|g\|^2}}$$

$$\rho = \max_{\alpha \in R^N, \beta \in R^N} \frac{\alpha^T \tilde{K}_X \tilde{K}_Y \beta}{\sqrt{\alpha^T \tilde{K}_X^2 \alpha + N\varepsilon_N \tilde{K}_X \alpha} \sqrt{\beta^T \tilde{K}_Y^2 \beta + N\varepsilon_N \tilde{K}_Y \beta}}$$

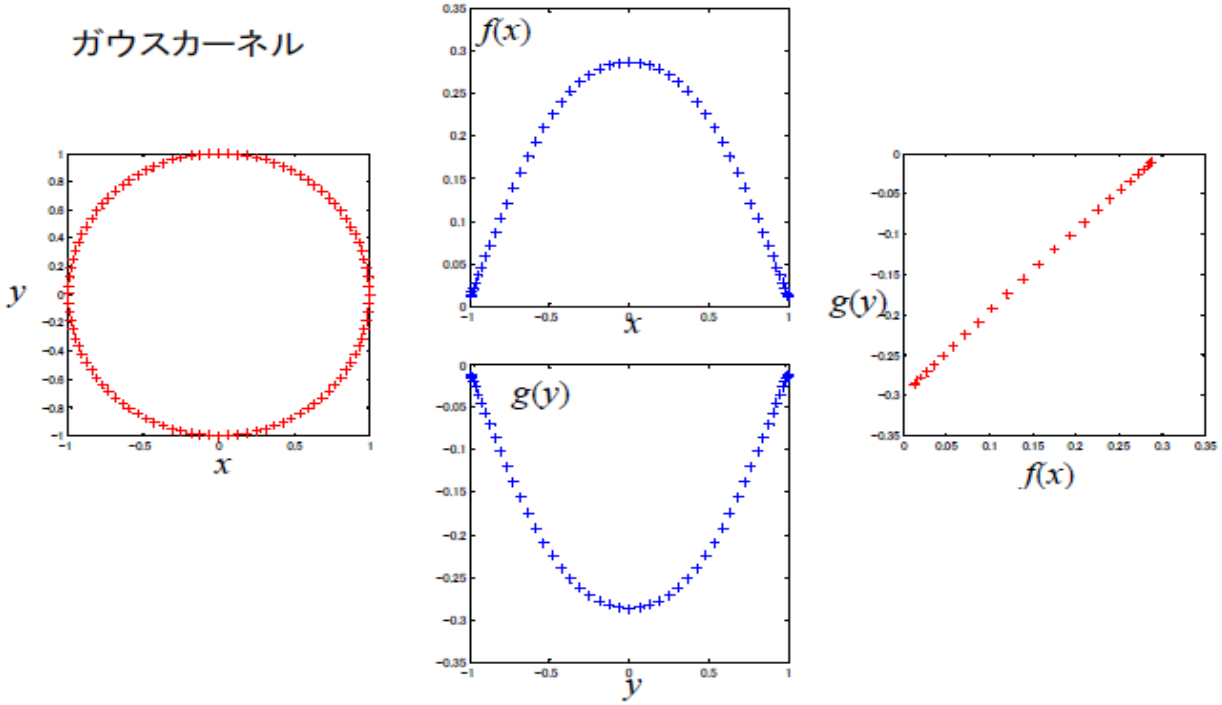
$$\begin{pmatrix} O & \tilde{K}_X \tilde{K}_Y \\ \tilde{K}_Y \tilde{K}_X & O \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} \tilde{K}_X^2 + N\varepsilon_N K_X & O \\ O & \tilde{K}_Y^2 + N\varepsilon_N K_Y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

3.2.3　カーネルCCAの応用例-1

- X,Y：円上にある100点からなる1次元データ
- k(x,y)=exp(-1/2(x-y)^2)
- ε＝1/10^5

- カーネルCCAの実験例

ガウスカーネル



3.2.3　カーネルCCAの応用例-2

- X: image　色情報データ, Y: text　単語の頻度データ
- それぞれ400データ、中の200データがテスト用

- データにカーネルＣＣＡを実施し、大きいd個の固有値に対応する固有ベクトル$f_1,\ldots,f_d,g_1,\ldots,g_d$を求める
- 各画像Xiに対するd次元部分空間への正写像 $\xi_i = \left(\left\langle \tilde{\Phi}_x(X_i), f_a \right\rangle_{H_x} \right)_{a=1}^{d} \in R^d$ を計算
- 新たなテキストデータY_{new}が与えられたとき、その特徴を計算

$$\zeta = \left(\left\langle \tilde{\Phi}_y(Y_{new}), g_a \right\rangle_{H_x} \right)_{a=1}^{d} \in R^d$$

- テキストに最も適した画像データとして $\arg\max_i = \xi^T \zeta$ を与えるXiを出力する