再生核ヒルベルト空間と統計的学習

情報・システム研究機構　統計数理研究所
（兼）総合研究大学院大学

福水 健次

人工知能学会　DMSM研究会　2006年7月11日

Outline

- はじめに
- 正定値カーネル／再生ヒルベルト空間を用いたデータ処理の方法論　～　カーネル法　～
- 独立性、条件付独立性の特徴づけ
- カーネル次元削減法
- おわりに

イントロダクション

- 非線形データ解析の重要性
  - 古典的な線形データ解析

    データの行列表現

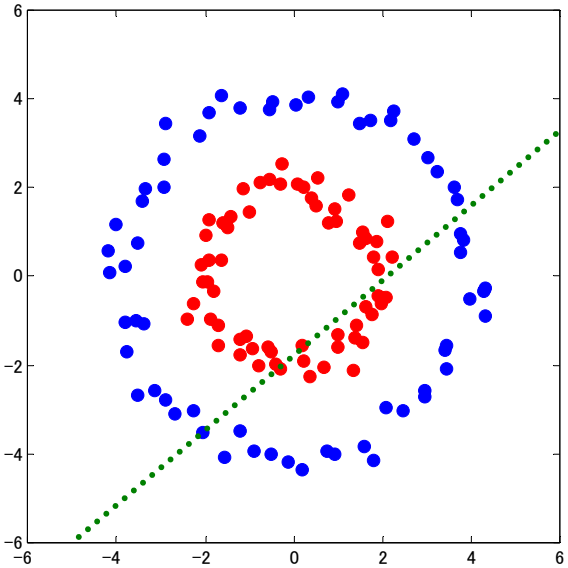
    m 次元　N 点のデータ

$$X = \begin{pmatrix} X_1^1 & \cdots & X_m^1 \\ X_1^2 & \cdots & X_m^2 \\ \vdots & & \vdots \\ X_1^N & \cdots & X_m^N \end{pmatrix}$$

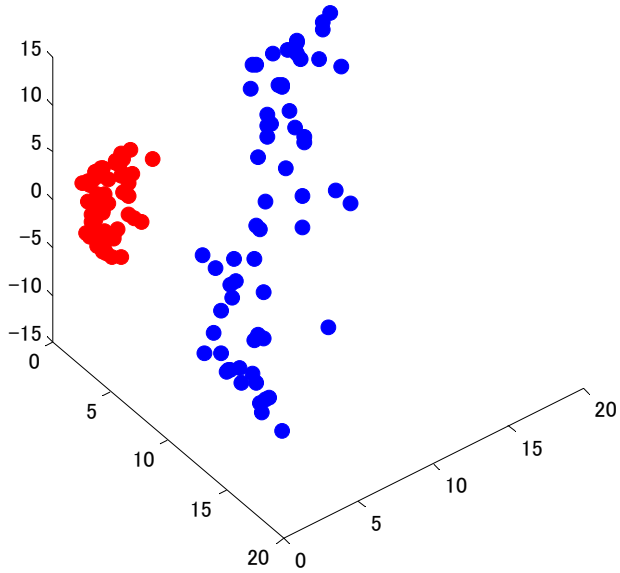
    ⇒ 線形の処理（主成分分析，正準相関分析，線形回帰．．．）

  - 線形で十分か？

線形識別不能



線形識別可能



$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

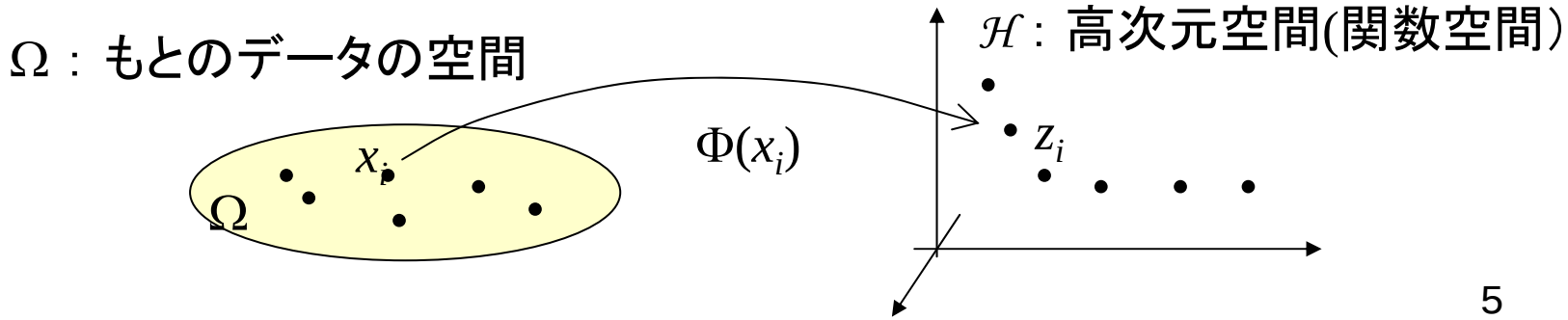
カーネル法の概略

- 線形手法の非線形化
  - 線形パラメータの推定 ⇒ 非線形関数の推定

線形の関数　$w^T X$　⇒　非線形関数　$f(X)$

  - 非線形関数の処理をうまく扱えるようにしたのがカーネル法

- もうひとつの見方：　高次元空間での線形データ処理
  - データ x を高次元（無限次元）空間のベクトル $\Phi(x)$ に変換して処理.
  - もとのデータ空間 Ω は、ベクトル空間でなくてもよい



正定値カーネル／再生核ヒルベルト空間を用いたデータの処理

正定値カーネル

- 正定値カーネル

  Ω: 集合. $k : \Omega \times \Omega \to \mathbf{R}$ $k(x,y)$ が Ω 上の正定値カーネル:

  1. （対称性） $k(x,y) = k(y,x)$
  2. （正定値性） 任意の自然数 n と，任意のΩ の点 $x_1, \ldots, x_n$ に対し，

  $n \times n$ 行列 $\left(k(x_i, x_j)\right)_{i,j=1}^{n} = \begin{pmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{pmatrix}$ （グラム行列）

  が（半）正定値. i.e. 任意の実数 $c_1, \ldots, c_n$ に対し，$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$

  - 例（$\mathbf{R}^m$上）
    - ユークリッド内積 $k(x,y) = x^T y$
    - 多項式カーネル $k(x,y) = (x^T y + c)^d$ （d: 自然数，$c \geq 0$ ）
    - ガウスカーネル $k(x,y) = \exp\left(-\|y-x\|^2 / \sigma^2\right)$

正定値カーネルと再生核ヒルベルト空間

- 定理（再生核ヒルベルト空間, Reproducing kernel Hilbert space, RKHS）

  $k(x,y)$: 集合 Ω 上の正定値カーネル

  ⇨

  Ω 上の関数からなるヒルベルト空間 $\mathcal{H}_k$ が一意に存在して，次の3つが成立

  (1) $k(\cdot, x) \in \mathcal{H}_k$ （$x \in \Omega$ は任意に固定）

  (2) 有限和 $f = \sum_{i=1}^{n} c_i k(\cdot, x_i)$ の形の関数全体は $\mathcal{H}_k$ の中で稠密

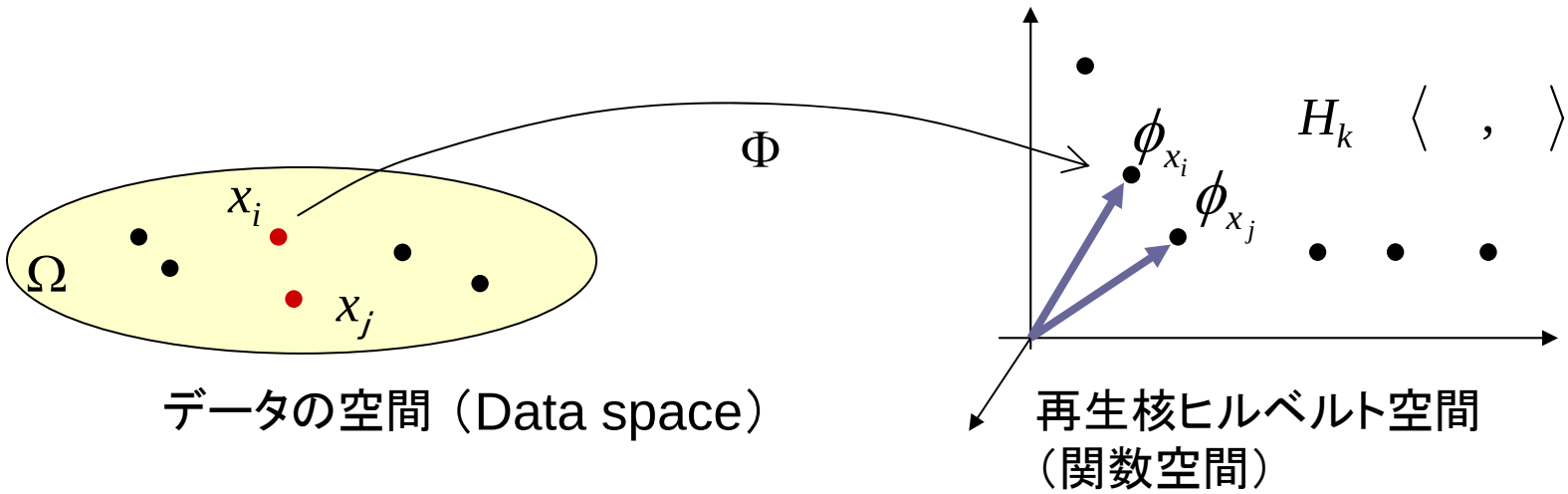
  (3)（再生性） $f(x) = \langle f, k(\cdot, x) \rangle$ $\forall f \in \mathcal{H}_k, x \in \Omega$

  関数の値が内積によって与えられる

  注） $k(\cdot, x)$ ・・・ x を固定した1変数関数

カーネルによるデータ変換

- データの変換　　$x \mapsto \Phi(x) = k(\cdot, x) = \phi_x$



- 内積計算：　$\langle \Phi(x), \Phi(y) \rangle = \langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y)$　・・・　カーネルトリック

- 特徴空間における線形アルゴリズム
  → データの空間での非線形アルゴリズム
  サポートベクターマシン, カーネルPCA,
  カーネルCCA　　etc

カーネル法の例： カーネルPCA

- PCAの非線形化

$$\max_{\|a\|=1} \quad \mathrm{Var}[a^T X] = \frac{1}{N}\sum_{i=1}^{N}\left(a^T X_i - \frac{1}{N}\sum_j a^T X_j\right)^2$$

非線形化

$$\max_{\|f\|=1} \quad \mathrm{Var}[f(X)] = \frac{1}{N}\sum_{i=1}^{N}\left(f(X_i) - \frac{1}{N}\sum_j f(X_j)\right)^2$$

関数 f を探す空間として
「再生核ヒルベルト空間」をとると

$$\max_{f\in H_k, \|f\|=1} \quad \mathrm{Var}[\langle f, \Phi(X)\rangle] = \frac{1}{N}\sum_{i=1}^{N}\left(\langle f, \Phi(X_i)\rangle - \frac{1}{N}\sum_j \langle f, \Phi(X_j)\rangle\right)^2$$

（再生性）

$f, \Phi(X_i)$ は H_k のベクトル　⇒　H_k における「線形」な問題
ベクトル $\Phi(X_1), \ldots, \Phi(X_N)$ に対する H_k 内でのPCA

- カーネルPCA（Schölkopf et al 98）

カーネル k を設定

特徴ベクトル $\Phi(X_1), ..., \Phi(X_N)$ に対するPCA

$$\max_{f \in H_k, \|f\|=1} \mathrm{Var}[\langle f, \Phi(X) \rangle] = \frac{1}{N} \sum_{i=1}^{N} \left(\langle f, \tilde{\phi}_i \rangle \right)^2$$

ただし $\tilde{\phi}_i = \Phi(X_i) - \frac{1}{N} \sum_{j=1}^{N} \Phi(X_j)$

$f = \sum_{i=1}^{N} \alpha_i \tilde{\phi}_i$ としてよい （∵ 直交する方向は分散に寄与しない）

⇒ 分散 $= \frac{1}{N} \sum_{a=1}^{N} \left\langle \sum_{j=1}^{N} \alpha_j \tilde{\phi}_j, \tilde{\phi}_a \right\rangle^2 = \frac{1}{N} \alpha^T \tilde{K}^2 \alpha$ ただし $\tilde{K}_{ij} = \langle \tilde{\phi}_i, \tilde{\phi}_j \rangle$

主成分は

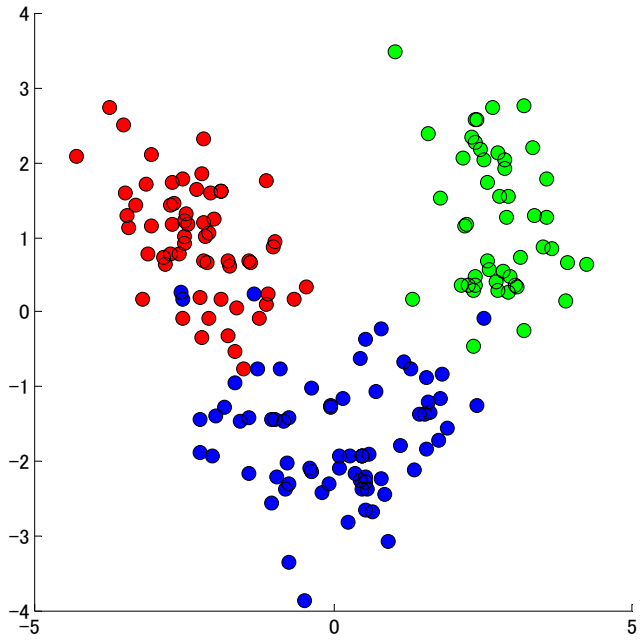
$$\begin{cases} \max_{\alpha} \alpha^T \tilde{K}^2 \alpha \\ \text{制約条件 } \alpha^T \tilde{K} \alpha = 1 \end{cases} \quad \Leftarrow \quad \| f \|_{H_k}^2 = \left\langle \sum_i \alpha_i \tilde{\phi}_i, \sum_i \alpha_i \tilde{\phi}_i \right\rangle = \alpha^T \tilde{K} \alpha$$

カーネルPCAの実験例

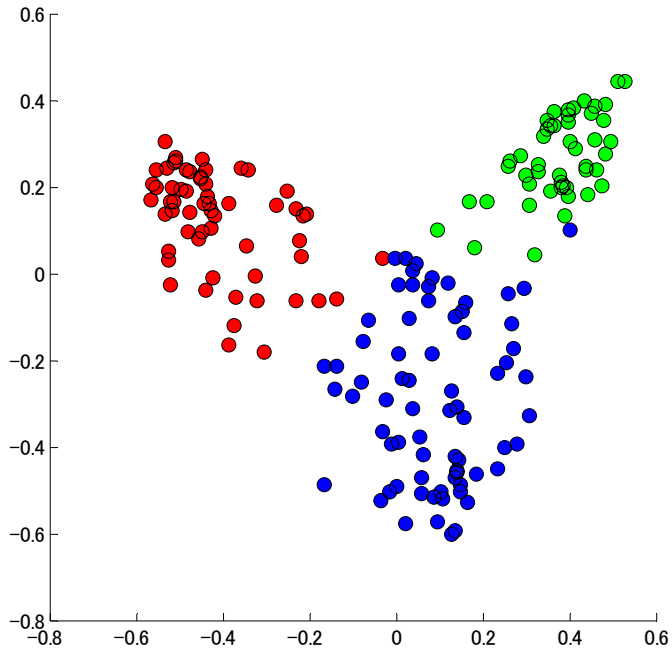
‘Wine’ データ（UCI Machine Learning Repository）

13次元，178データ，3種類のワインの属性データ

2つの主成分を取った（3クラスの色は参考に付けたもの)



PCA（線形)



KPCA（Gauss，$\sigma = 3$)

非線形化としてのカーネル法

- 線形のアルゴリズム

  データが $\mathbf{R}^m$ のベクトル → 線形回帰、主成分分析、正準相関分析 etc

  - $a^{\mathrm{T}}X$ として，パラメータ a を求める
  - 相関、分散共分散行列による計算

- カーネル法の2つの見方

  - 非線形化

    $a^T X \quad \Longrightarrow \quad f(X)$

  - 特徴ベクトルへの線形アルゴリズム

    特徴ベクトル $\Phi(X_1), \ldots, \Phi(X_N)$

- カーネルによる非線形化
  - $a^{\mathrm{T}}X$ の代わりに $f(X)$ （ f はRKHS内の関数）
  - $f(X_i) = \langle f, \Phi(X_i) \rangle_{H_k}$ （再生性）
  - 多くの場合 $f = \sum_{i=1}^{N} \alpha_i \Phi(X_i) = \sum_{i=1}^{N} \alpha_i k(\cdot, X_i)$ でOK （representer定理）

    $$\langle f, \Phi(X_i) \rangle_{H_k} = \sum_j \alpha_j \langle \Phi(X_j), \Phi(X_i) \rangle = \sum_j \alpha_j k(X_j, X_i)$$

    （カーネルトリック）
  - グラム行列（データ数のサイズ）による計算
- 特徴ベクトルへの線形アルゴリズムとしての見方
  - データ $X_1, \ldots, X_N$ ⇒ 特徴ベクトル $\Phi(X_1), \ldots, \Phi(X_N)$
  - RKHS 内での線形アルゴリズム

    $\langle f, \Phi(X_i) \rangle$ として f を求める
  - 以下同様

RKHSを使う利点

- 数理的な観点から
  - k がC^r 級ならば $\mathcal{H}_k$ に属する関数は C^r 級
  - $E_P[k(X,X)] < \infty$ ならば $\mathcal{H}_k \subseteq L^2(P)$ かつ包含写像は連続

- 情報処理的観点から
  - 関数の「値」が定まる （c.f. L^2 空間）
  - 関数空間での内積計算が容易

$$f = \sum_i a_i k(\cdot, X_i), g = \sum_j b_j k(\cdot, X_j) \Rightarrow \langle f, g \rangle = \sum_{ij} a_i b_j k(X_i, X_j)$$

  - 基底による展開をする必要が無い（Ωの次元によらない）
  - 非ベクトルデータの処理

    もとのデータ空間はベクトル空間でなくてもよい

    → グラフデータ、ツリーデータ、ヒストグラムなどの処理

独立性、条件付独立性と再生核ヒルベルト空間

独立性とRKHS

- 特性関数による独立性の特徴づけ（確率論の復習）

確率ベクトル X と Y が独立

$$\Leftrightarrow \quad E_{XY}\left[e^{\sqrt{-1}\omega^T X} e^{\sqrt{-1}\eta^T Y}\right] = E_X\left[e^{\sqrt{-1}\omega^T X}\right] E_Y\left[e^{\sqrt{-1}\eta^T Y}\right] \quad \text{for all } \omega \text{ and } \eta.$$

$$\Leftrightarrow \quad \mathrm{Cov}\left[e^{\sqrt{-1}\omega^T X}, e^{\sqrt{-1}\eta^T Y}\right] = 0 \qquad \text{for all } \omega \text{ and } \eta.$$

$e^{\sqrt{-1}\omega^T x}$, $e^{\sqrt{-1}\eta^T y}$ は様々な関数による非線形相関を調べるテスト関数

RKHSによる独立性の特徴づけ

$\mathcal{H}_X, \mathcal{H}_Y$: Ω_X, Ω_Y 上のRKHS

X: Ω_X 上に値を持つ確率変数, Y: Ω_Y 上に値を持つ確率変数

X と Y が独立

$$\Leftrightarrow \quad E_{XY}\left[f(X)g(Y)\right] = E_X\left[f(X)\right]E_Y\left[g(Y)\right] \quad \text{for all } f \in \mathcal{H}_X,\ g \in \mathcal{H}_Y$$

という特徴づけは可能か？

定理（Bach and Jordan, 2002）

X, Y に対して、ガウスカーネル $k(z,\tilde{z}) = \exp\left(-\|z-\tilde{z}\|^2 / \sigma^2\right)$ を使うと

X, Y が独立 $\Leftrightarrow$ $E_{XY}\left[f(X)g(Y)\right] = E_X\left[f(X)\right]E_Y\left[g(Y)\right]$

for all $f \in \mathcal{H}_X,\ g \in \mathcal{H}_Y$

- 独立成分分析への応用

条件付独立性とRKHS

- 条件付分散
  - ガウスの場合（X, Y: ガウス）

$$\mathrm{Var}[a^T Y \mid X] = a^T (V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}) a$$

  - 再生核ヒルベルト空間

$\mathcal{H}_X, \mathcal{H}_Y$ ： Ω_X, Ω_Y 上のRKHS

X: Ω_X 上に値を持つ確率変数，　Y: Ω_Y 上に値を持つ確率変数

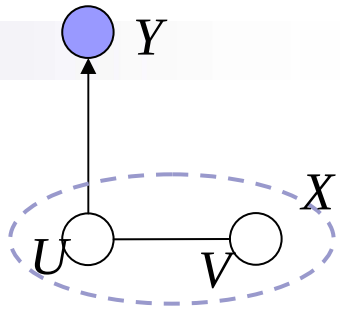
$\mathcal{H}_X, \mathcal{H}_Y$ が十分豊かなクラスだと仮定すると

$$\mathrm{Var}[g(Y) \mid X] = \left\langle g, (\Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}) g \right\rangle_{H_Y} \quad \text{for all} \quad g \in \mathcal{H}_Y$$

Σ_{YX} などは、無限次元の分散共分散行列

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_Y} = E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \quad (= \mathrm{Cov}[f(X), g(Y)])$$

$$\textit{c.f.} \quad a^T V_{YX} b = \mathrm{Cov}[a^T X, b^T Y]$$



条件付独立性の特徴付け

確率変数 Y, U, V に対し、

$Y \perp V \mid U$　　あるいは同値な条件　$Y \perp X \mid U$　　$(X = (U,V))$

の特徴づけを考える。

定理6（Fukumizu et al. 2004）

U, V, Y: Ω_U , Ω_V , Ω_Y に値をとる確率変数。

$\mathcal{H}_U, \mathcal{H}_V, \mathcal{H}_Y$: Ω_U , Ω_V , Ω_Y 上のRKHS

$X = (U,V)$,　　$\mathcal{H}_X = \mathcal{H}_U \otimes \mathcal{H}_V$（直積）

$\mathcal{H}_X, \mathcal{H}_U$ は十分豊かな関数族を含むことを仮定。

⇒　　$\Sigma_{YY|U} \geq \Sigma_{YY|X}$　　i.e. $\mathrm{Var}[g(Y)|U] \geq \mathrm{Var}[g(Y)|X]$

さらに$\mathcal{H}_Y$ がガウスカーネルならば

$$\Sigma_{YY|U} = \Sigma_{YY|X} \quad \Leftrightarrow \quad Y \perp X \mid U$$

＊　$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$

カーネル次元削減法

回帰問題における次元削減

- 回帰問題における有効な部分空間
  - 回帰問題 ・・・ Y を X で説明する

    $p(Y \mid X)$ の推定

  - 次元削減

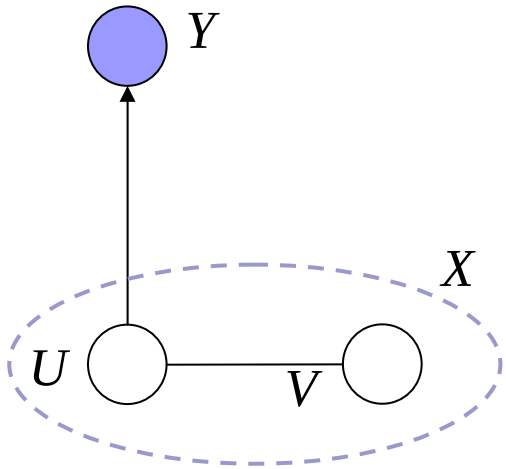
    X : m 次元ベクトル

    $B = (b_1, \ldots, b_d)$ m x d 行列

$$p(Y \mid X) = p(Y \mid B^T X)$$

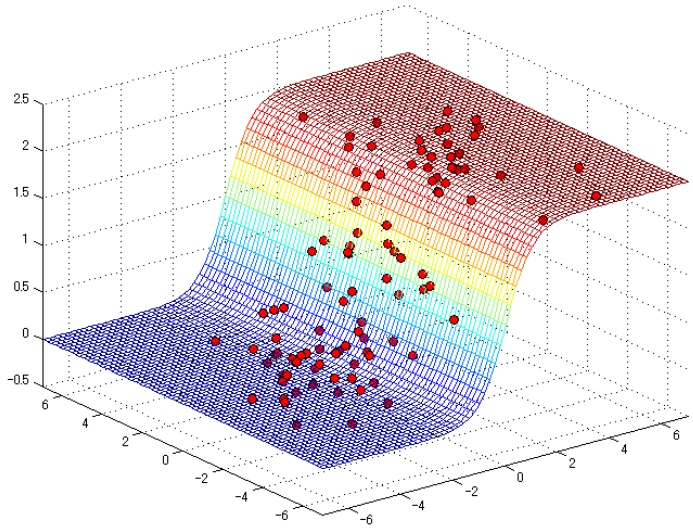
$$\Leftrightarrow \quad Y \perp X \mid B^T X$$

となる B を探す.

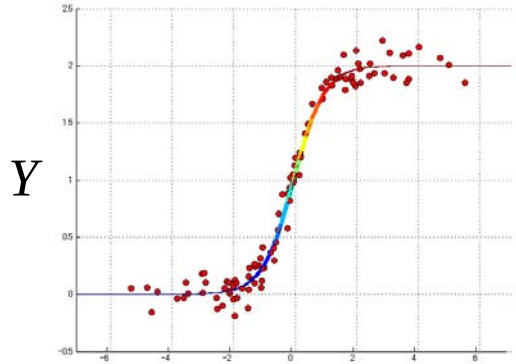


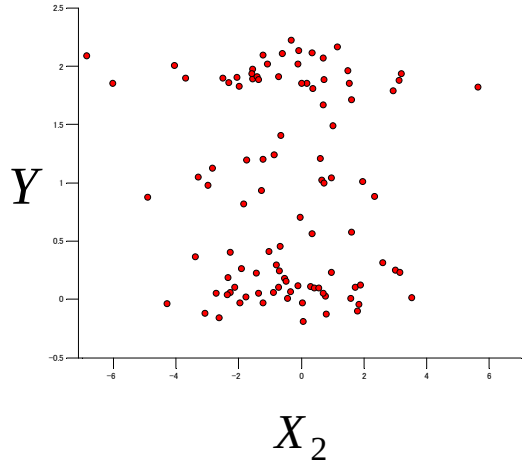
$B^T X = (b_1{}^T X, .., b_d{}^T X)$ は，Y を説明する目的では，X と同じ情報を持つ

有効部分空間（特徴ベクトル）



$$Y = \frac{2}{1+\exp(-2X_1)} + N(0; 0.1^2)$$





カーネル次元削減法

有限個のデータ $X_1, \ldots, X_N,\ Y_1, \ldots, Y_N$ から条件付共分散作用素を推定

$$\min_{B:B^T B=I_d} \mathrm{Tr}\left[G_Y \left(G_X^B + N\varepsilon_N I_N \right)^{-1} \right]$$

ただし

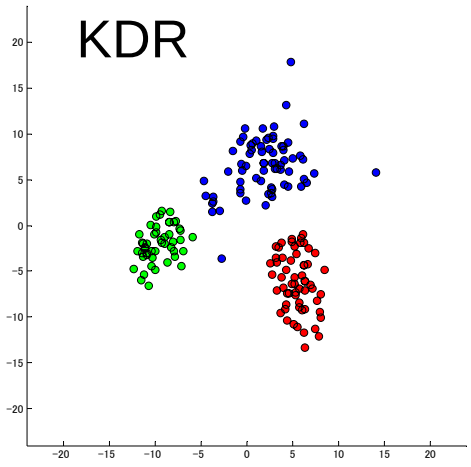
$$G_X{}^B = (I_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n{}^T)\left(k_d(B^T X^{(i)}, B^T X^{(j)})\right)(I_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n{}^T)$$

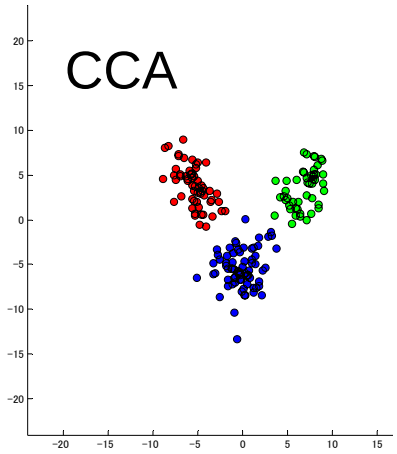
$$G_Y = (I_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n{}^T)\left(k_Y(Y^{(i)}, Y^{(j)})\right)(I_n - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n{}^T)$$

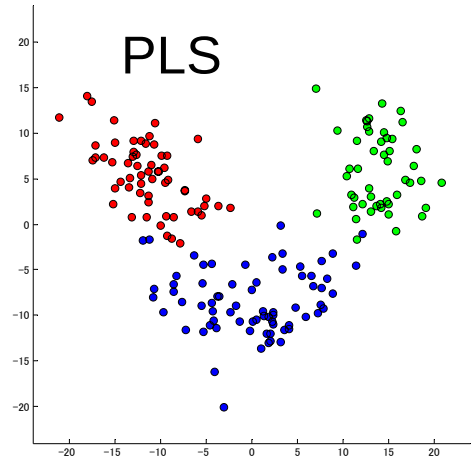
ε_N: 正則化パラメータ

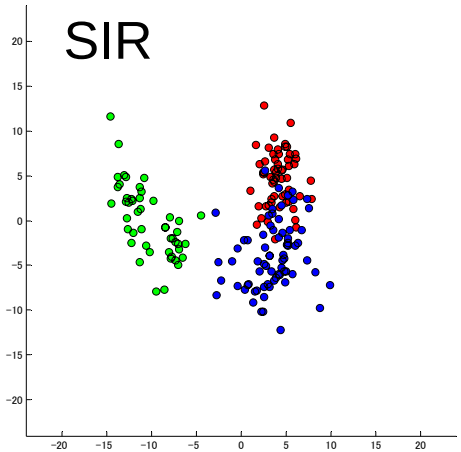
非凸な目的関数 → 勾配法による最適化が必要

- Wine data （他の次元削減法との比較)
  - Data
    13 dim. 178 data.
    3 classes
    2 dim. projection









カーネル次元削減法の特徴

- どんなデータでも扱える
  - 回帰問題の次元削減としては最も一般的
    $p(Y|X)$ のモデル（線形など)を使わない.
  - X, Y の分布に条件がいらない. Y が離散値や高次元でもOK.
    従来法（SIR, pHd, CCA, PLS, etc）ではさまざまな制約
- 計算量の問題
  - N x N 行列を用いた演算.
    → Incomplete Cholesky decomposition の利用
  - 非線形最適化に伴う局所解／計算時間の問題

カーネル次元削減法の一致性

- 定理（Fukumizu et al. 2006）

  最適なパラメータを

$$S_0 = \left\{ B \mid B^T B = I_d,\ B = \arg\min \mathrm{Tr}[\Sigma_{YY|B^T X}] \right\}$$

  とおく。正則化パラメータが

$$\varepsilon_N \to 0, \quad \sqrt{N}\varepsilon_N \to \infty \quad (N \to \infty)$$

  を満たすとき、ある種の正則条件のもと、任意の $\delta > 0$ に対し

$$\Pr\left(\mathrm{dist}\left(\widehat{B}, S_0 \right) > \delta \right) \to 0 \quad (N \to \infty)$$

  証明の核）

$$\sup_{B:B^T B = I_d} \left| \mathrm{Tr}\left[\widehat{\Sigma}_{YY|X}^{B(N)} \right] - \mathrm{Tr}\left[\Sigma_{YY|X}^{B} \right] \right| \to 0$$

  が成り立つ

おわりに

- カーネル法：再生核ヒルベルト空間を用いたデータ処理の方法論
  - RKHS：　関数の値が意味を持つヒルベルト空間
  - 内積計算が容易
  - 線形手法の非線形化が可能
- 独立性、条件付独立性の特徴づけ
  - 相互共分散作用素を用いた線形（ガウス）手法の非線形化
- カーネル次元削減法
  - 回帰における最も一般的な次元削減法
  - 条件付独立性の特徴づけにもとづく