統計数理(2005)
第 53 巻 第 2 号 189–200

特集「計算推論—モデリング・数理・アルゴリズム—」
[研究詳解]

正定値カーネルによる回帰問題における次元削減法

福水 健次†

(受付 2005 年 4 月 1 日；改訂 2005 年 6 月 30 日)

要　　旨

本論文は，Fukumizu et al. (2004)にしたがって，正定値カーネルを用いた，回帰問題における新しい次元削減法に関する著者らの研究を解説する．説明変数 X を用いて従属変数 Y を説明する回帰の問題設定において，X に含まれる Y の情報をすべて保持するような説明変数空間の低次元部分空間を「有効部分空間」と呼ぶことにし，この部分空間を見つける次元削減の問題を考える．まず，この問題を条件付独立性を用いて定式化し，さらにこの条件付独立性が再生核ヒルベルト空間上の共分散作用素を用いて特徴づけられることを示す．この理論的事実を用いて，与えられた有限サンプルから有効部分空間を推定するための方法を導き，実データに適用した例を示す．本論文で解説する方法は，これまでに提案された同種の方法とは異なり，X や Y の条件付確率，周辺分布および次元などに強い制約を必要としないため，幅広いデータに適用可能である．

キーワード： カーネル法，正定値カーネル，ヒルベルト空間，次元削減，変数選択，回帰．

1. はじめに

さまざまな統計的データ解析において次元削減は重要な手法である．画像，テキスト，遺伝子発現データなど極めて高次元なデータが溢れている今日の状況においては，データの説明や可視化，予測・決定の精度向上のためのノイズ削減，計算量の軽減などさまざまな目的のために次元削減が用いられ，その重要度は高まっている．本論文は，Fukumizu et al. (2004)に従い，m 次元説明変数 X を用いて ℓ 次元従属変数 Y を説明する回帰の問題において，Y に関する情報を保持するような，X の低次元部分空間への射影を見つける次元削減の問題を論じる．

X が与えられたときの Y の条件付確率密度関数を $p_{Y|X}(y|x)$ と書くことにする．本論文では，$\mathbb{R}^m$ の r 次元部分空間 S が存在して，

$$p_{Y|X}(y|x) = p_{Y|\Pi_S X}(y|\Pi_S x) \tag{1.1}$$

が成り立つと仮定する．ここで Π_S は部分空間 S への直交射影である．(1.1)式を満たす部分空間 S のことを Li (1991)にならって有効部分空間と呼ぶ．これは Y の情報を完全に保持する部分空間である．本論文は，与えられた有限サンプルから有効部分空間 S を推定する手法を論じる．

† 統計数理研究所：〒 106-8569 東京都港区南麻布 4–6–7

この問題に対し，分布 $p(X,Y)$ に関するモデルや制約をなるべくおかずに S を推定する，セミパラメトリックなアプローチをとる．回帰問題での次元削減に対する従来法としては，Sliced Inverse Regression（SIR，Li, 1991）や Principal Hessian Directions（pHd, Li, 1992）などの手法が有名であるが，これらは X の周辺分布に楕円型などの強い制約を要する．また，Canonical Correlation Analysis（CCA）や Partial Least Square（PLS）なども用いられることがあるが，これらはもちろん線形モデルを仮定している．また，射影追跡に基づく方法（Friedman and Stuetzle, 1981; Breiman and Friedman, 1985）なども使うことが可能であるが，これも additive model を回帰モデルに仮定している．こういった仮定を置かない本論文のアプローチは，これら従来法よりも一般的である．

セミパラメトリック推定を行う際には，無限自由度を表す関数空間を導入するのが標準的であるが，そのために正定値カーネルの定める再生核ヒルベルト空間を利用する．まず回帰問題における次元削減を条件付独立性の問題として定式化する．さらに，この条件付独立性が，再生核ヒルベルト空間によって特徴づけられることを示す．この条件付独立性の特徴づけは，再生核ヒルベルト空間上の条件付分散作用素という無限次元の線形写像を用いて与えられるが，これは，X を S に射影してできた変数によって Y を推定した際の誤差を表していると考えることもできる．この作用素の最小化によって有効部分空間が特徴付けられる．

本論文は以下のような構成を持つ．まず第2章では，再生核ヒルベルト空間上の条件付共分散作用素を用いて，有効部分空間の特徴づけを行う．次に第3章で，有限サンプルが与えられたときの条件付共分散作用素の推定法を議論し，有効部分空間を数値的に求めるためのカーネル次元削減法のアルゴリズムを説明する．第4章は，このアルゴリズムを実データに用いた結果を示し，第5章でさらに変数選択問題への拡張を論じる．最後に第6章で結論を述べる．

2. 再生核ヒルベルト空間を用いた次元削減

本論文では定理の証明などは省略するので，詳細は Fukumizu et al.（2004）を参照していただきたい．

2.1 次元削減と条件付独立

以降では有効部分空間 S の次元 r は既知とする．S とその直交補空間 $S^{\perp}$ の正規直交基底を並べた行列を，それぞれ B，C とおく．B と C はそれぞれ $m\times r$，$m\times(m-r)$ 行列であり，(B,C) は m 次元直交行列となる．S と $S^{\perp}$ への X の直交射影を $U=B^TX$，$V=C^TX$ であらわすことにする．(B,C) が直交行列であることから，確率密度関数に関して $p_X(x)=p_{U,V}(u,v)$，$p_{X,Y}(x,y)=p_{U,V,Y}(u,v,y)$ が成り立ち，これにより，(1.1)式は

$$p_{Y|U,V}(y|u,v)=p_{Y|U}(y|u) \tag{2.1}$$

と同値である．すなわち，S が有効部分空間であることと，U が与えられたときの Y と V の条件付独立性とは同値である（図1）．

この条件付独立性は，相互情報量を通じて捉えることもできる．2つの確率変数 X，Y の相互情報量 $I(X,Y)$ は

$$I(X,Y)=\int p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}dxdy$$

により定義される．相互情報量に関して

$$I(Y,X)=I(Y,U)+E_U\big[I(Y|U,V|U)\big], \tag{2.2}$$

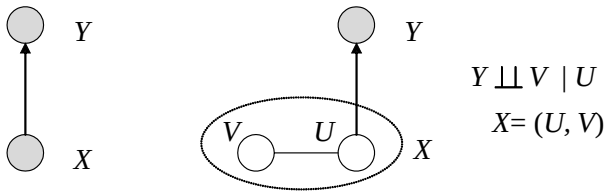

図 1. 回帰問題における次元削減のグラフィカル表現

が成り立つことは容易に示せる．有効部分空間の条件(1.1)式は $I(Y,X)=I(Y,U)$ を意味するので，Y との相互情報量を減じない部分空間が S である．また，$I(Y,X)=I(Y,U)$ は $I(Y|U,V|U)=0$ を意味し，これは再び，U が与えられたときの Y と V の条件付独立性である．

2.2 再生核ヒルベルト空間上の共分散作用素

以下では，条件付独立性を特徴付けるための関数空間として，再生核ヒルベルト空間を用いる．まず正定値カーネルとそれが定める再生核ヒルベルト空間について復習しておこう．詳しくは Aronszajn (1950)や Schölkopf and Smola (2002)を参照していただきたい．

Ω を集合とするとき，$k:\Omega\times\Omega\to\mathbb{R}$ が正定値カーネルであるとは，任意の n 個の点 $x_1,\ldots,x_n\in\Omega$ と 実数 $c_1,\ldots,c_n$ に対し，正定値性

$$\sum_{i,j=1}^{n} c_i c_j k(x_i,x_j)\geq 0 \tag{2.3}$$

が成り立つことをいう．Ω 上の正定値カーネル k に対し，集合 Ω 上の関数からなる(実)ヒルベルト空間 $\mathcal{H}$ が存在し，以下の 2 つの性質を満たす．

(i) 任意の $x\in\Omega$ に対して $k(\cdot,x)\in\mathcal{H}$ であり，$\{k(\cdot,x)\in\mathcal{H}\,|\,x\in\Omega\}$ の張る線形空間は $\mathcal{H}$ の中で稠密である．

(ii) 任意の $f\in\mathcal{H}$ と $x\in\Omega$ に対し，再生性

$$\langle f,k(\cdot,x)\rangle_{\mathcal{H}}=f(x),\qquad(\forall x\in\Omega,\forall f\in\mathcal{H}) \tag{2.4}$$

が成り立つ．ここで $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ は $\mathcal{H}$ の内積を表す．

このようなヒルベルト空間のことを再生核ヒルベルト空間といい，$(\mathcal{H},k)$ であらわす．(ii)の再生性は再生核ヒルベルト空間を用いる上で最も重要な性質であり，ヒルベルト空間内での内積の計算を容易に計算可能にする．例えば $f=\sum_{i=1}^{n}a_i k(\cdot,x_i)$ と $g=\sum_{j=1}^{m}b_j k(\cdot,y_j)$ に対し，(ii)を用いると

$$\langle f,g\rangle=\sum_{i=1}^{n}\sum_{j=1}^{m}a_i b_j k(x_i,y_j)$$

となり，k の値の評価に還元される．また，再生核ヒルベルト空間に属する関数に対しては，その「値」が意味を持つ点が，L^2 のような関数空間とは大きく異なっている．これらの特徴は，有限サンプルからデータ解析を行う際に有利な性質である．

ユークリッド空間 $\mathbb{R}^m$ 上の正定値カーネルの代表的な例は，通常の内積 $k(x_1,x_2)=x_1^T x_2$ のほかに，多項式カーネル

$$k(x_1,x_2)=(x_1^T x_2+c)^d$$

($c \geq 0$，$d \in \mathbb{N}$)や，ガウス RBF (Radial Basis Function)カーネル

$$k(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$$

($\sigma > 0$)などである．本論文では後に述べる理由により，主にガウス RBF カーネルを用いる．

可測空間 $(\Omega_1, \mathcal{B}_1)$，$(\Omega_2, \mathcal{B}_2)$ 上に，有界かつ可測な正定値カーネルを持つ再生核ヒルベルト空間 $(\mathcal{H}_1, k_1)$，$(\mathcal{H}_2, k_2)$ があるとする．$\Omega_1 \times \Omega_2$ に値をとる確率変数 (X, Y) に対し，$\mathcal{H}_1$ から $\mathcal{H}_2$ への相互共分散作用素 Σ_{YX} は，任意の $f \in \mathcal{H}_1$，$g \in \mathcal{H}_2$ に対し

$$\begin{aligned} \langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_2} &= E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \\ &= \mathrm{Cov}[f(X), g(Y)] \end{aligned} \tag{2.5}$$

を満たす有界作用素として定義される．存在および一意性は Riesz の表現定理による．共役作用素に関して，$\Sigma_{YX}^* = \Sigma_{XY}$ が成り立つ．特に Σ_{XX} は自己共役である．相互共分散作用素の一般的な理論は Baker (1973)に詳しい．

相互共分散作用素は，以下のように確率変数の独立性を特徴付ける．

定理 1. $(\mathcal{H}_1, k_1)$，$(\mathcal{H}_2, k_2)$ を，それぞれ $\mathbb{R}^m$，$\mathbb{R}^\ell$ 上のガウス RBF カーネルを持つ再生核ヒルベルト空間とし，X，Y をそれぞれ $\mathbb{R}^m$，$\mathbb{R}^\ell$ 上の確率ベクトルとする．このとき，

$$X \perp\!\!\!\perp Y \quad \Longleftrightarrow \quad \Sigma_{XY} = O \quad \text{(零作用素)} \tag{2.6}$$

が成り立つ．ここで $X \perp\!\!\!\perp Y$ は X と Y が独立であることを表す．

この定理は，特性関数を用いた確率変数の特徴づけ

$$X \perp\!\!\!\perp Y \quad \Longleftrightarrow \quad E_{XY}[e^{\sqrt{-1}u^T X} e^{\sqrt{-1}v^T Y}] = E_X[e^{\sqrt{-1}u^T X}]E_Y[e^{\sqrt{-1}v^T Y}] \tag{2.7}$$

の類似とみなすことができる．実際，$\mathbb{R}^m$ 上の関数 $k(x, y) = e^{\sqrt{-1}x^T y}$ は(複素数値)正定値カーネルとなっており，(2.6)(2.7)式の右辺は，これらカーネルの定義する再生核ヒルベルト空間に属する関数に対して X と Y の非線形共分散がゼロであることを意味する．$e^{\sqrt{-1}x^T y}$ やガウス RBF カーネルを持つヒルベルト空間は十分豊かな非線形関数を含んでおり，(2.6)(2.7)式の右辺は，それらの定める非線形共分散がすべて消えることを意味していることから，上の定理の主張は容易に納得できると思う．

(2.5)から，条件付期待値に関する以下の事実が示される．

定理 2. 可測空間 $(\Omega_1, \mathcal{B}_1)$，$(\Omega_2, \mathcal{B}_2)$ 上に，有界かつ可測な正定値カーネルを持つ再生核ヒルベルト空間 $(\mathcal{H}_1, k_1)$，$(\mathcal{H}_2, k_2)$ があるとし，(X, Y) を $\Omega_1 \times \Omega_2$ に値をとる確率変数とする．さらに，任意の $g \in \mathcal{H}_2$ に対し条件付期待値 $E_{Y|X}[g(Y) \mid X = \cdot]$ が Ω_1 上の関数として $\mathcal{H}_1$ に属すると仮定する．このとき

$$\Sigma_{XX} E_{Y|X}[g(Y) \mid X = \cdot] = \Sigma_{XY} g, \qquad (\forall g \in \mathcal{H}_2) \tag{2.8}$$

が成立する．

系 1. 定理 2 の仮定のもと，$\tilde{\Sigma}_{XX}^{-1}$ を Σ_{XX} の $(\mathrm{Ker}\Sigma_{XX})^\perp$ 上の右逆作用素とすると，任意の $f \in (\mathrm{Ker}\Sigma_{XX})^\perp$，$g \in \mathcal{H}_2$ に対し

$$\langle f, \tilde{\Sigma}_{XX}^{-1} \Sigma_{XY} g \rangle_{\mathcal{H}_1} = \langle f, E_{Y|X}[g(Y) \mid X = \cdot] \rangle_{\mathcal{H}_1} \tag{2.9}$$

が成り立つ.

Σ_{XX} が可逆であると(2.9)式 は

$$E_{Y|X}[g(Y)\,|\,X=\cdot]=\Sigma_{XX}^{-1}\Sigma_{XY}g \tag{2.10}$$

を意味している．よく知られているように，X，Y が有限次元ガウス確率ベクトルであるとき，任意のベクトル a に対し

$$E_{Y|X}[a^TY\,|\,X=x]=x^TV_{XX}^{-1}V_{XY}a$$

(ここでは V_{XX}，V_{XY} は通常の分散共分散行列)が成り立つので,(2.10)式はガウス分布の条件付平均の一般化とみなすこともできる.

2.3 共分散作用素による条件付独立性の特徴づけ

ここで条件付共分散作用素を定義する．可測空間 $(\Omega_1,\mathcal{B}_1)$，$(\Omega_2,\mathcal{B}_2)$ 上に，有界かつ可測な正定値カーネルを持つ再生核ヒルベルト空間 $(\mathcal{H}_1,k_1)$，$(\mathcal{H}_2,k_2)$ が与えられており，(X,Y) は $\Omega_1\times\Omega_2$ に値をとる確率変数とする．このとき，X が与えられたときの Y の条件付共分散作用素 $\Sigma_{YY|X}$ とは

$$\Sigma_{YY|X}:=\Sigma_{YY}-\Sigma_{YX}\tilde{\Sigma}_{XX}^{-1}\Sigma_{XY} \tag{2.11}$$

により定まる $\mathcal{H}_2$ 上の半正定値自己共役作用素のことである.

系 1 を用いると次の定理は容易に示される.

定理 3. 定理 2 の仮定のもと，任意の $f,g\in\mathcal{H}_2$ に対し,

$$\begin{aligned}\langle g,\Sigma_{YY|X}f\rangle_{\mathcal{H}_2}&=E_Y[f(Y)g(Y)]-E_X\big[E_{Y|X}[f(Y)|X]E_{Y|X}[g(Y)|X]\big]\\&=E_X\big[\mathrm{Cov}_{Y|X}\big[f(Y),g(Y)\,|\,X\big]\big]\end{aligned} \tag{2.12}$$

が成り立つ.

(2.10)式の場合と同様に,(2.11),(2.12)式はガウス確率変数に関するよく知られた関係式

$$\mathrm{Cov}[a^TY,b^TY|X]=a^T(V_{YY}-V_{YX}V_{XX}^{-1}V_{XY})b$$

の拡張と考えることができる.

定理 3 より，$\Sigma_{YY|U}$ が自己共役作用素の半正定値性で定まる半順序に関して小さいほど，条件付分散 $\mathrm{Var}_{Y|U}[f(Y)|U]$ は小さくなり，U は Y をよりよく説明することができる．この事実を有効部分空間 S の特徴づけに用いることを考えるのは自然である．このアイデアを正当化するために次の定義をしよう．可測集合 $(\Omega,\mathcal{B})$ 上に，有界かつ可測な正定値カーネルを持つ再生核ヒルベルト空間 $(\mathcal{H},k)$ があるとする．$(\Omega,\mathcal{B})$ 上のすべての確率分布からなる集合を $\mathcal{M}$ で表すとき，再生核ヒルベルト空間 $\mathcal{H}$ が確率決定性を持つとは，写像

$$\mathcal{M}\ni P\quad\mapsto\quad(f\mapsto E_{X\sim P}[f(X)])\in\mathcal{H}^* \tag{2.13}$$

が単写であることをいう．ここで $\mathcal{H}^*$ は $\mathcal{H}$ の双対空間を表す．このとき次の事実が成り立つ.

定理 4. 任意の $\sigma>0$ に対し，ガウス RBF 関数 $k(x,y)=\exp(-\|x-y\|^2/2\sigma^2)$ を正定値カーネルに持つ再生核ヒルベルト空間は確率決定性を持つ.

ガウス RBF カーネルの定める再生核ヒルベルト空間は十分に豊富な関数族を含むので，そこに属する関数に対する期待値をすべて調べれば，確率がひとつに決まるというのが定理の主張である．

集合 Ω_1，Ω_2 上の再生核ヒルベルト空間 $(\mathcal{H}_1, k_1)$，$(\mathcal{H}_2, k_2)$ の直積 $\mathcal{H}_1 \otimes \mathcal{H}_2$ とは，値の積で定まる正定値カーネル $k_1 k_2$ を持つ $\Omega_1 \times \Omega_2$ 上の再生核ヒルベルト空間のことであった(Aronszajn, 1950)．以上の準備のもと条件付独立性は次のように特徴付けられる．

定理 5. $(\mathcal{H}_{11}, k_{11})$，$(\mathcal{H}_{12}, k_{12})$，$(\mathcal{H}_2, k_2)$ をそれぞれ可測集合 Ω_{11}，Ω_{12}，Ω_2 上の再生核ヒルベルト空間とし，正定値カーネルはすべて連続かつ有界と仮定する．(U, V, Y) を $\Omega_{11} \times \Omega_{12} \times \Omega_2$ に値をとる確率変数とし，$X = (U, V)$ および $\mathcal{H}_1 = \mathcal{H}_{11} \otimes \mathcal{H}_{12}$ と表すことにする．また，任意の $g \in \mathcal{H}_2$ に対し $E_{Y|U}[g(Y)|U = \cdot] \in \mathcal{H}_{11}$ と $E_{Y|X}[g(Y)|X = \cdot] \in \mathcal{H}_1$ を仮定する．このとき，自己共役作用素の半順序に関して

$$\Sigma_{YY|U} \geq \Sigma_{YY|X} \tag{2.14}$$

が成立する．さらに $\mathcal{H}_2$ が確率決定性を持つとすると，

$$\Sigma_{YY|X} = \Sigma_{YY|U} \quad \Longleftrightarrow \quad Y \perp\!\!\!\perp V \,|\, U \tag{2.15}$$

の同値性が成立する．

証明の概略． 条件付分散に関するよく知られた関係式

$$\mathrm{Var}_{Y|U}[g(Y)|U] = E_{V|U}\left[\mathrm{Var}_{Y|U,V}[g(Y)|U,V]\right] + \mathrm{Var}_{V|U}\left[E_{Y|U,V}[g(Y)|U,V]\right]$$

を U に関して期待値をとると

$$E_U\left[\mathrm{Var}_{Y|U}[g(Y)|U]\right] - E_X\left[\mathrm{Var}_{Y|X}[g(Y)|X]\right] = E_U\left[\mathrm{Var}_{V|U}[E_{Y|X}[g(Y)|X]]\right] \geq 0$$

が得られ，(2.14)式が成り立つ．等号成立は，ほとんどすべての X に対して $E_{Y|X}[g(Y)|X] = E_{Y|U}[g(Y)|U]$ となる場合であるが，$\mathcal{H}_2$ の確率決定性より(2.15)式を得る．□

線形回帰の場合から類推できるように，条件付分散 $E_X[\mathrm{Var}_{Y|X}[g(Y)|X]]$ は，X を用いて $g(Y)$ を推定したときの推定誤差を表すものと考えることができる．すると，(2.14)式は，情報が部分的になれば，Y を推定した際の誤差が増加するという当然の事実を表している．(2.15)式は，推定誤差が増加しなければ，X と U は Y に関して同じだけの情報量を持つことを意味しており，非常に自然な結果である．

定理 5 より，確率決定性を持つ再生核ヒルベルト空間を用いると，有効部分空間 S は次の最小化問題の解として与えられる．

$$\min_S \Sigma_{YY|U}, \quad \text{subject to} \quad U = \Pi_S X \tag{2.16}$$

次章では，これに基づいて有効部分空間を推定するため目的関数を導く．

3. カーネル次元削減法

(2.16)式から有限サンプルによる目的関数を導くためには，サンプルを用いて条件付共分散作用素を推定する必要がある．以降では，定理 4 にもとづいて，正定値カーネルとしてガウス RBF 関数のみを考えることにする．

Bach and Jordan (2002)に従って(相互)共分散作用素を以下のように推定する．n 個のサンプル $(X_1,Y_1),\dots,(X_n,Y_n)$ が与えられているとする．$\tilde{k}_1(\cdot,X_i)$，$\tilde{k}_2(\cdot,Y_i)$ をそれぞれ $\tilde{k}_1(\cdot,X_i)=k_1(\cdot,X_i)-\frac{1}{n}\sum_{j=1}^n k_1(\cdot,X_j)$，$\tilde{k}_2(\cdot,Y_i)=k_2(\cdot,Y_i)-\frac{1}{n}\sum_{j=1}^n k_2(\cdot,Y_j)$ と定めよう．(2.5)式の期待値をサンプル平均に置き換えると

$$\frac{1}{n}\sum_{i=1}^n \langle f,\tilde{k}_1(\cdot,X_i)\rangle_{\mathcal{H}_1}\langle \tilde{k}_2(\cdot,Y_i),g\rangle_{\mathcal{H}_2}$$

に一致する．さらに，ヒルベルト空間 $\mathcal{H}_1$，$\mathcal{H}_2$ をそれぞれ $\{\tilde{k}_1(\cdot,X_i)\}_{i=1}^n$，$\{\tilde{k}_2(\cdot,Y_i)\}_{i=1}^n$ の張る $n-1$ 次元空間に制限し，これらを冗長な基底として，作用素 Σ_{YX} の制限を行列表示すると，再生性を用いて，

$$P_nG_XP_nG_YP_n$$

が得られる．ここで射影行列 P_n は，$\mathbf{1}_n=(1,\dots,1)^T$ として $P_n=I_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ により定義され，$G_{X,ij}=k_1(X_i,X_j)$，$G_{Y,ij}=k_2(Y_i,Y_j)$ はグラム行列と呼ばれる $n\times n$ 行列である．以上により，

$$\widehat{K}_X=P_nG_XP_n,\qquad \widehat{K}_Y=P_nG_YP_n$$

と書くことにすると，

$$\widehat{\Sigma}_{YX}=\widehat{K}_Y\widehat{K}_X \tag{3.1}$$

を共分散作用素の推定量として用いることができる．

条件付共分散作用素の推定量を得るためには，逆作用素を考える必要があるが，一般に $\widehat{\Sigma}_{ZZ}$ は P_n を含むために非可逆である．そこで，自己共分散作用素 Σ_{ZZ} を推定する際には，正則化を用い，

$$\widehat{\Sigma}_{ZZ}=(\widehat{K}_Z+\varepsilon_nI_n)^2$$

($\varepsilon_n>0$)を推定量として使うことにする．ここで ε_n は正則化のための正定数で $n\to\infty$ のときに $\varepsilon_n\to 0$ となるように定める．以上により，条件付共分散作用素の推定量 $\widehat{\Sigma}_{YY|U}$ を

$$\widehat{\Sigma}_{YY|U}:=\widehat{\Sigma}_{YY}-\widehat{\Sigma}_{YU}\widehat{\Sigma}_{UU}^{-1}\widehat{\Sigma}_{UY} \tag{3.2}$$

により定め，この正定値行列を最小化する．

正定値対称行列の半順序をはかるには，トレース，行列式，最大固有値などいろいろなものが考えられるが，まず行列式を考えよう．行列式の Schur 分解を用いると，

$$\widehat{\Sigma}_{[YU][YU]}=\begin{pmatrix}\widehat{\Sigma}_{YY} & \widehat{\Sigma}_{YU}\\ \widehat{\Sigma}_{UY} & \widehat{\Sigma}_{UU}\end{pmatrix}=\begin{pmatrix}(\widehat{K}_Y+\varepsilon_nI_n)^2 & \widehat{K}_Y\widehat{K}_U\\ \widehat{K}_U\widehat{K}_Y & (\widehat{K}_U+\varepsilon_nI_n)^2\end{pmatrix}$$

の記法のもと，$\det\widehat{\Sigma}_{YY|U}=\det\widehat{\Sigma}_{[YU][YU]}/\det\widehat{\Sigma}_{UU}$ となる．これにより，有効部分空間 S を推定するための目的関数が

$$\min_B \frac{\det\widehat{\Sigma}_{[YU][YU]}}{\det\widehat{\Sigma}_{YY}\det\widehat{\Sigma}_{UU}},\quad \text{ただし } U=B^TX \tag{3.3}$$

により得られる．ここで $\det\widehat{\Sigma}_{YY}$ は定数であるが，目的関数の対称性のために加えた．

またトレースを用いると，(3.2)式の第 2 項のトレースの最大化を考えることにより

$$\max_B \mathrm{Tr}\left[\left(\widehat{K}_Y+\varepsilon_nI_n\right)^{-1}\widehat{K}_Y\widehat{K}_U\left(\widehat{K}_U+\varepsilon_nI_n\right)^{-2}\widehat{K}_U\widehat{K}_Y\left(\widehat{K}_Y+\varepsilon_nI_n\right)^{-1}\right] \tag{3.4}$$

を用いることが可能である．

(3.3)，(3.4)式を用いて，部分空間 S ないし行列 B を求める最適化問題を，カーネル次元削減法(Kernel dimensionality reduction, KDR)と呼ぶことにする．

(3.3)式は，ガウス確率変数の相互情報量(のマイナス)の一種の拡張とみなせる．Bach and Jordan (2002)では，これを一般の確率変数の相互情報量の代用として提案し独立成分分析に用いたが，本論文では代用ではなく理論的な導出を行っている．

カーネル次元削減法を実行するためには，目的関数の最小化／最大化を行う必要があるが，この目的関数は非線形かつ非凸であり，非線形最適化手法が必要となる．以下では，直線探索を併用した最急勾配法を用いる．さらに局所解の問題を避けるために，ガウス RBF カーネルの分散パラメータを徐々に小さくしていく，一種のアニーリング手法を用いている．また，(3.3)式からわかるように，最適化には $n \times n$ 行列の演算を数多く行う必要があり，サンプル数 n が大きいと計算量が増大する．これに対し，不完全 Cholesky 分解によって $\widehat{K}_Y$ などを低ランク行列で近似すると演算量を削減することが可能である(Bach and Jordan, 2002)．

また，(3.4)式の最大化を行う代わりに $\mathrm{Tr}[\widehat{K}_Y \widehat{K}_U (\widehat{K}_U + \varepsilon_n I_n)^{-1}]$ の最大化を考え，さらに $(A+\varepsilon I)^{-1}A = I - \varepsilon(A+\varepsilon I)^{-1}$ であることを用いると，近似的に

$$\min_B \mathrm{Tr}\left[\widehat{K}_Y \left(\widehat{K}_U + \varepsilon_n I_n\right)^{-1}\right] \tag{3.5}$$

の最小化を達成する B を見つける問題に変換される．この目的関数を用いると計算量は大幅に削減される．

4. カーネル次元削減法の実データへの応用

カーネル次元削減法(KDR)を実データに応用し，結果を SIR，pHd，CCA，PLS といった従来法と比較した．予備的な実験では (3.4)式と(3.3)式の結果にあまり違いが見られなかったため，以下の実験では，目的関数として(3.3)式を用いた．

まずデータ可視化の能力を見る目的で，UCI machine learning repository (Murphy and Aha, 1994)の Wine データを用いた．このデータは 3 種類のワインに対する 13 次元の属性を 178 サンプル集めたデータである．クラスの情報をなるべく保持するように，各手法で 2 次元部分空間を求めた結果が図 2 である．KDR が 3 クラスを最もよく判別しており，2 次元空間で完全な識別が可能なことがわかる．CCA も 3 クラスを完全に分けているが，他の手法の結果では判別は不完全である．

第二の実験では，推定された部分空間の中に，クラス判別に必要な情報がどれぐらいよく残されているかを調べる目的で，UCI レポジトリの 3 種類の実データに対し，次元削減を行った後，その部分空間へ射影したデータを用いてサポートベクターマシンによる識別器を構成し，訓練データとは別に用意されたテストデータに関する正答率を調べた．ところで，多くの次元削減の従来法は，判別問題，特に 2 クラス判別の問題に適用が難しいものが多い．SIR は，Y の空間をスライスに切り，各スライス内で X のサンプル平均を取るので，クラス数が小さいと適用するのが困難になる．また，線形手法である CCA や PLS では，クラス数以上の部分空間を見つけることはできない．この実験では，2 クラス識別にも適用可能な pHd との比較を行った．図 3 にさまざまな次元の部分空間における正答率を示した．KDR は pHd に比べて低次元でも高い正答率を保っていることが見て取れる．特に Ionosphere データに対しては，5，10，20 次元の正答率は全次元を用いた場合の正答率を上回っている．これは KDR が判別に不要な成分を有効に取り除き，ノイズ除去の役割を果たしたためだと考えられる．

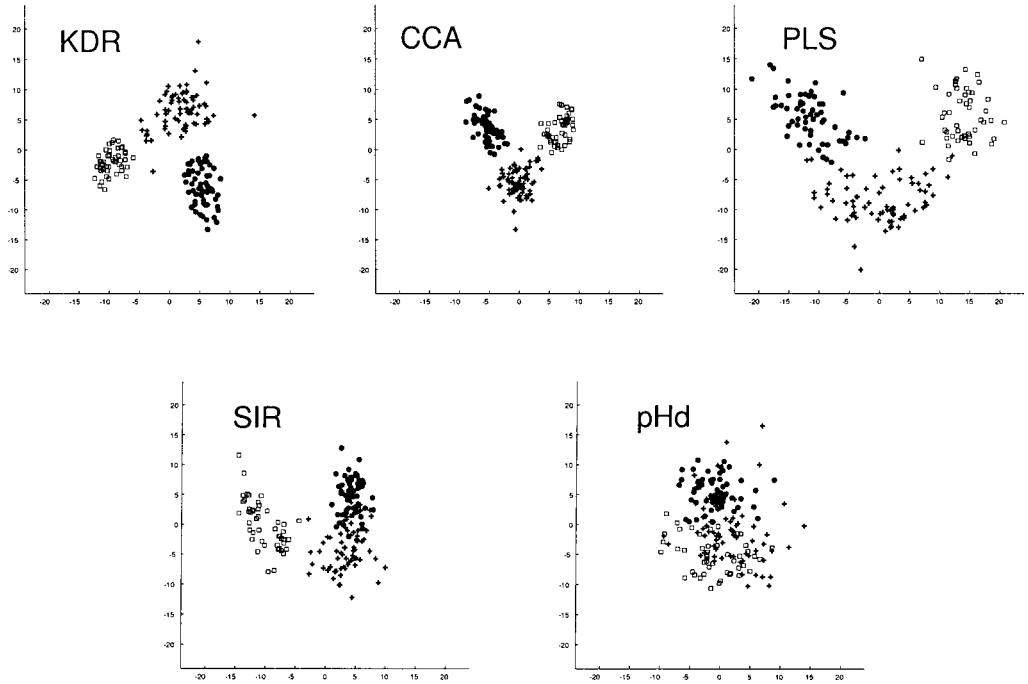

図 2. Wine **データの** 2 **次元射影.** “+”，“•”，“□” **が** 3 **クラスに対応.**

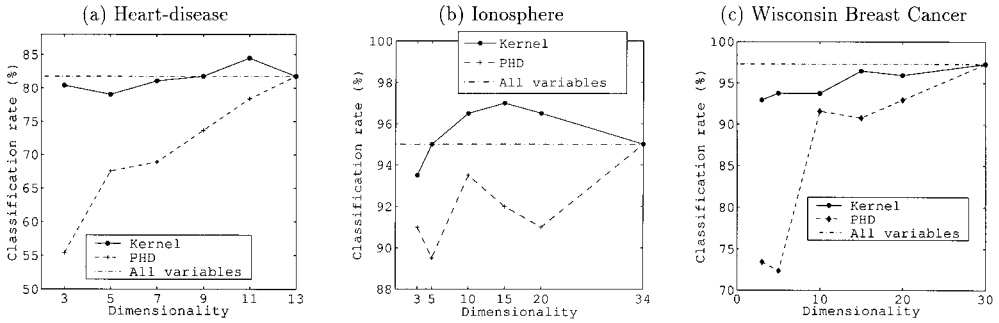

図 3. **次元削減後のテストデータに対する** SVM **の判別正解率.** (a)$m=13$, $N=149$, $T=148$, (b)$m=34$, $N=151$, $T=200$, (c)$m=30$, $N=200$, $T=369$ (m: **説明変数の次元**, N: **訓練データ数**, T: **テストデータ数**).

5. 変数選択への応用

ここまで次元削減の方法として説明変数の線形和を求める方法を考えてきたが，KDR の手法は説明変数の部分集合を求める「変数選択」にも応用可能である．そのためには，(3.3)式の最小化問題の探索空間を，部分空間全体ではなく，説明変数の部分集合(の張る部分空間)全体に置き換えればよい．

この変数選択法を Boston Housing データ(Harrison and Rubinfeld, 1978)に適用した．このデータは 13 変数を用いて各地域の住宅価格の平均値を説明するもので，506 サンプルからなる．4 個の説明変数を選んだところ，RM，LSTAT，PTRATIO，TAX が最も有効な変数として選ばれた．これは Breiman and Friedman (1985)が ACE という手法を用いて選んだものと同一である．

変数選択においては，m 個の説明変数の中から r 個選ぶ組み合わせは ${}_mC_r$ だけあり，m が大きいとすべての場合を調べ尽くすのは困難になる．その場合には何らかの最適化手法が必要となる．詳細は省くが，ある種のランダムサーチを用いて，遺伝子発現データ Leukimia (Golub et al., 1999)からの遺伝子選択を行った．Leukimia は 2 種類の急性白血病を 7129 次元の遺伝子発現データから判別するためのマイクロアレイデータである．38 個の訓練用サンプルを用いて 50 個の有効な変数(遺伝子)を選択した．その中の 25 遺伝子が， Golub et al. (1999)によって別の方法で選ばれたものと一致しており，意味のある遺伝子が選ばれていることがわかる．また，選ばれた 50 遺伝子を用いてサポートベクターマシンによる識別子を作ったところ，訓練サンプルとは別に取られた 34 個のテストサンプルに対する正答数は 32 であった．Golub et al. (1999)では，識別結果の信頼度に閾値を設けて判定不能(リジェクト)を許した場合，ある閾値において 5 個のリジェクトのほか 29 個すべてが正答であったと報告されているので，それと比較しても識別に対して有効な遺伝子が選択されていると言える．

6. おわりに

本論文は，正定値カーネルで定まる再生核ヒルベルト空間を用いた，回帰問題における次元削減の方法を紹介した．この方法は，有効部分空間を求める問題を条件付独立性として捉え，それをヒルベルト空間上の共分散作用素を使って特徴付けることにより導かれた．

このカーネル次元削減法 KDR は，条件付確率や周辺分布にモデルや強い条件をおかずに導かれているため適用範囲が非常に広い．回帰における次元削減の従来法である SIR，pHd，CCA，PPR などの方法は，条件付確率や周辺分布に強い制約があり，その適用範囲は KDR よりも限定されている．本論文では，KDR を実データに適用してその有効性を確認した結果や，変数選択問題への拡張も述べた．

KDR はモデルを仮定しない高い汎用性のかわりに，数値的な最適化にともなう困難さを持っており，次元やデータ数が大きくなると実用的な時間で解を求めることが難しくなる．本稿では数十次元程度のデータからの次元削減の例を示したが，数百を越えるような次元，1000 を越えるようなデータ数を持つようなデータに対しては，計算機の性能にもよるが，本稿で述べた方法をそのまま適用するのは難しいと思われる．この問題に対しては，(3.5)式の方法をさらに詳しく検討したり，もっと本質的な最適化の改良を考察したりする必要がある．また，十分に問題の構造を反映したモデルがある場合，そのモデルに依存した次元削減法がより有効なのは言うまでもないが，特に高次元のデータの場合には，データを生成する真の構造が明確にわかる場合は少ないと考えられるため，本稿のようなセミパラメトリックな方法論が有望であると考える．

KDR は理論的な背景に基づく手法であるが，その有効性の確認は実験的に行っており，得

られた推定量の統計的性質などの理論解析は今後の課題である．特に，本論文では有効部分空間の次元 r を固定して議論したが，言うまでもなくその次元の選択は重要な問題である．この問題に対しては，最終的な目的が予測精度で測られるのであれば，クロスバリデーションなどの方法を適用することも可能であるが，その正当性を理論検証するためにも推定量の性質を詳しく知ることは重要である．

本論文では，回帰問題における次元削減だけを述べたが，共分散作用素による条件付独立性の特徴づけは，もっと広い問題に適用することが可能であろう．特に，条件付独立性はグラフィカルモデルを定義する際の基本的な道具であり，本論文の方法論をもっと一般のグラフィカルモデルへ拡張することは興味深い問題である．

謝　辞

本研究の一部は科研費 15700241 により行われた．

参 考 文 献

Aronszajn, N. (1950). Theory of reproducing kernels, *Transaction of the American Mathematical Society*, **69**(3), 337–404.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis, *Journal of Machine Learning Research*, **3**, 1–48.

Baker, C. R. (1973). Joint measures and cross-covariance operators, *Transaction of the American Mathematical Society*, **186**, 273–289.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, **80**, 580–598.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association*, **76**, 817–823.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, *Journal of Machine Learning Research*, **5**(Jan), 73–99.

Golub, T. R., Slonim, D. and Tamayo, P. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531–537.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics Management*, **5**, 81–102.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association*, **86**, 316–342.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma, *Journal of the American Statistical Association*, **87**, 1025–1039.

Murphy, P. M. and Aha, D. W. (1994). UCI repository of machine learning databases, Tech. Report, Department of Information and Computer Science, University of California, Irvine, http://www.ics.uci.edu/~learn/MLRepository.html

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*, MIT Press, Cambridge, Massachusetts.

200 Proceedings of the Institute of Statistical Mathematics Vol. 53, No. 2, 189–200 (2005)

Dimensionality Reduction in Regression with Positive Definite Kernels

Kenji Fukumizu

The Institute of Statistical Mathematics

This paper explains our recent research on dimensionality reduction in regression with positive definite kernels, following Fukumizu et al. (2004). In regression problems, where the response variable Y is explained by explanatory variables X, the best subspace within the explanatory space to explain Y is called *effective subspace*. Our goal of dimensionality reduction in regression is to find the effective subspace. We formulate this problem as conditional independence of variables, and show that this conditional independence can be characterized by using covariance operators on the reproducing Hilbert spaces, which are given by positive definite kernels. Based on this theoretical result, we derive a method of estimating the effective subspace, given a finite number of samples. This dimensionality reduction method is very general and applicable to a wide class of data, because we do not need any strong assumptions on the conditional and marginal distributions, or dimension of variables. The experimental results on real data are shown to demonstrate the effectiveness of the method.

Key words: Kernel method, positive definite kernel, Hilbert space, dimensionality reduction, variable selection, regression.