

カーネルヒルベルト空間を用いた回帰問題における特徴抽出

Feature Extraction in Regression with Kernel Hilbert Spaces

福水健次*
Kenji Fukumizu

フランシス・バック†
Francis R. Bach

マイケル・ジョーダン‡
Michael I. Jordan

Abstract: We propose a novel method of dimensionality reduction for supervised learning. Given a regression or classification problem in which we wish to predict a variable Y from an explanatory vector X , we treat the problem of dimensionality reduction as that of finding a low-dimensional “effective subspace” of X which retains the statistical relationship between X and Y . We show that this problem can be formulated in terms of conditional independence. To turn this formulation into an optimization problem, we characterize the notion of conditional independence using covariance operators on reproducing kernel Hilbert spaces; this allows us to derive a contrast function for estimation of the effective subspace. Unlike many conventional methods, the proposed method requires neither assumptions on the marginal distribution of X , nor a parametric model of the conditional distribution of Y .

1 はじめに

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においては、変数間の関係の説明や可視化、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、 m 次元入力 X から ℓ 次元出力 Y への確率的関係を考察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に X の低次元部分空間を見つける問題を論じる。

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、

$$p_{Y|X}(y|x) = p_{Y|\Pi_S X}(y|\Pi_S x) \quad (1)$$

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) 式を満たす部分空間 S のことを有効部分空間と呼ぶ。これは Y の情報を完全に保持する

*統計数理研究所, 106-8569 東京都港区南麻布 4-6-7, tel. 03-5421-8730, e-mail fukumizu@ism.ac.jp,
Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569 Japan

†Computer Science Division, University of California, Berkeley, CA 94720-1776, USA. tel. +1-510-642-8277, e-mail fbach@cs.berkeley.edu,

‡Computer Science Division and Dept. Statistics, University of California, Berkeley, CA 94720-1776, USA. tel. +1-510-642-3806, e-mail jordan@cs.berkeley.edu,

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 を推定する手法を導く。

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置かずに 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, [10]) や Principal Hessian Directions (pHd, [11]) 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, [12]) 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 ([6], [5]) も可能であるが、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。

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 [7] 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 ($r < m$) は既知とする。 m 次元直交行列 (B, C) を、 $m \times r$ 行列 B の列ベク

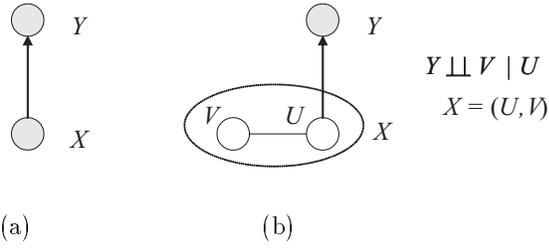


図 1: 回帰問題における次元削減のグラフィカル表現

トルが有効部分空間 S を張るものとして定める。すると、 S と S^\perp への X の直交射影は $U = B^T X$ 、 $V = C^T X$ と表される。 (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) 式は

$$p_{Y|U,V}(y|u, v) = p_{Y|U}(y|u) \quad (2)$$

と同値である。したがって、上式が示しているように、 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 U の条件付独立性とは同値である (図 1)。

この独立性は、相互情報量を通じてもみることが出来る。2つの確率変数 X, Y の相互情報量 $I(X, Y)$ は $I(X, Y) = \int p_{XY}(x, y) \log \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)} dx dy$ により定義される。相互情報量に関して

$$I(Y, X) = I(Y, U) + E_U [I(Y|U, V|U)], \quad (3)$$

が成り立つことは容易に示せる。有効部分空間の条件 (1) 式は $I(Y, X) = I(Y, U)$ を意味しているので、 S は X と Y の相互情報量を保持するような部分空間だと捉えることができる。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 U の条件付独立性であることが確認できる。

2.2 再生核ヒルベルト空間上の共分散作用素

再生核ヒルベルト空間上の相互共分散作用素 [4] を用いて次元削減のための目的関数を導こう。集合 Ω 上の関数からなる (実) 再生核ヒルベルト空間 (\mathcal{H}, k) を考える。ここで $k : \Omega \times \Omega \rightarrow \mathbb{R}$ は正定値核関数であり、 \mathcal{H} の内積を $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ で表す。再生核ヒルベルト空間で最も重要な性質は、再生性

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x), \quad (\forall x \in \Omega, \forall f \in \mathcal{H}) \quad (4)$$

である。以降では、主にガウス核関数

$$k(x_1, x_2) = \exp(-\|x_1 - x_2\|^2 / \sigma^2)$$

を用いる。再生核ヒルベルト空間に関する詳細は [2] や [1] を参照していただきたい。

可測空間 $(\Omega_1, \mathcal{B}_1)$ 、 $(\Omega_2, \mathcal{B}_2)$ 上に、有界かつ可測な核関数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

$$\begin{aligned} \langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_2} &= E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \\ &= \text{Cov}[f(X), g(Y)] \end{aligned} \quad (5)$$

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。

相互共分散作用素は、以下のように確率変数の独立性と関係する。

定理 1 ([3]). (\mathcal{H}_1, k_1) 、 (\mathcal{H}_2, k_2) を、それぞれ $\mathbb{R}^m, \mathbb{R}^\ell$ 上のガウス核関数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と、

$$X \perp\!\!\!\perp Y \iff \Sigma_{XY} = 0 \quad (\text{零作用素})$$

が成り立つ。 $(X \perp\!\!\!\perp Y)$ は X と Y が独立であることを表す。

ガウス核関数を持つヒルベルト空間が十分豊かな非線形関数を含んでおり、相互共分散作用素がその非線形相関を表していることから、この定理の主張は容易に納得できると思う。

(5) 式の定義を用いると、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、有界かつ可測な核関数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 $E_{Y|X}[g(Y) | X = \cdot]$ 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

$$\Sigma_{XX} E_{Y|X}[g(Y) | X] = \Sigma_{XY} g, \quad (\forall g \in \mathcal{H}_2) \quad (6)$$

が成立する。

系 3. 定理 2 の仮定のもと、 $\tilde{\Sigma}_{XX}^{-1}$ を Σ_{XX} の $(\text{Ker} \Sigma_{XX})^\perp$ 上の右逆作用素とすると、任意の $f \in (\text{Ker} \Sigma_{XX})^\perp, g \in \mathcal{H}_2$ に対し

$$\langle f, \tilde{\Sigma}_{XX}^{-1} \Sigma_{XY} g \rangle_{\mathcal{H}_1} = \langle f, E_{Y|X}[g(Y) | X] \rangle_{\mathcal{H}_1} \quad (7)$$

が成り立つ。

証明の概略. Σ_{XY} は、ある有界作用素 W を用いて $\Sigma_{XY} = \Sigma_{XX}^{1/2} W \Sigma_{YY}^{1/2}$ と分解できるので (Theorem 1, [4])、

$\overline{\text{Range}\Sigma_{XY}} \subset \overline{\text{Range}\Sigma_{XX}} = (\text{Ker}\Sigma_{XX})^\perp$ により $\tilde{\Sigma}_{XX}^{-1}\Sigma_{XY}$ は well-defined である。よって系の主張は定理 2 からしたがう。 \square

特に Σ_{XX} が可逆であると (7) 式は

$$E_{Y|X}[g(Y) | X] = \Sigma_{XX}^{-1}\Sigma_{XY}g \quad (8)$$

を意味している。よく知られているように、ガウス確率変数 X, Y に対しては、任意のベクトル a に対し

$$E_{Y|X}[a^T Y | X = x] = x^T V_{XX}^{-1} V_{XY} a$$

(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、(8) 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。

次に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、有界かつ可測な核関数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、 (X, Y) を $\Omega_1 \times \Omega_2$ 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

$$\Sigma_{YY|X} := \Sigma_{YY} - \Sigma_{YX} \tilde{\Sigma}_{XX}^{-1} \Sigma_{XY} \quad (9)$$

により定まる \mathcal{H}_2 上の自己共役作用素のことである。

系 3 により次の定理は容易に示される。

定理 4. 定理 2 の仮定のもと、任意の $f, g \in \mathcal{H}_2$ に対し、

$$\begin{aligned} & \langle g, \Sigma_{YY|X} f \rangle_{\mathcal{H}_2} \\ &= E_Y[f(Y)g(Y)] - E_X[E_{Y|X}[f(Y)|X]E_{Y|X}[g(Y)|X]] \\ &= E_X[\text{Cov}_{Y|X}[f(Y), g(Y) | X]] \end{aligned} \quad (10)$$

が成り立つ。

(8) 式の場合と同様に、(9), (10) 式はガウス確率変数に関するよく知られた関係式 $\text{Cov}[a^T Y, b^T Y | X] = a^T (V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}) b$ の拡張と考えることができる。

2.3 共分散作用素による有効部分空間の特徴づけ

定理 4 より、 $\Sigma_{YY|U}$ が自己共役作用素の意味で小さいほど、条件付分散 $\text{Var}_{Y|U}[f(Y)|U]$ は小さくなり、 U は Y をよりよく説明する。この事実を S の特徴づけに用いるのは自然な発想である。これを正当化するために次の定義をしよう。可測集合 (Ω, \mathcal{B}) 上に、有界かつ可測な核関数を持つ再生核ヒルベルト空間 (\mathcal{H}, k) があると、 (Ω, \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

$$\mathcal{M} \ni P \mapsto (f \mapsto E_{X \sim P}[f(X)]) \in \mathcal{H}^* \quad (11)$$

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。[3] の Theorem 2 は次の事実を示している。

定理 5 ([3]). 任意の $\sigma > 0$ に対し、ガウス関数 $k(x, y) = \exp(-\|x - y\|^2/2\sigma^2)$ を核関数に持つ再生核ヒルベルト空間は確率決定性を持つ。

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、核関数 $k_1 k_2$ 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 [2]。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 6. $(\mathcal{H}_{11}, k_{11}), (\mathcal{H}_{12}, k_{12}), (\mathcal{H}_2, k_2)$ をそれぞれ可測集合 $\Omega_{11}, \Omega_{12}, \Omega_2$ 上の再生核ヒルベルト空間とし、核関数はすべて連続かつ有界であると仮定する。 (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 $E_{Y|U}[g(Y)|U] \in \mathcal{H}_{11}$ と $E_{Y|X}[g(Y)|X] \in \mathcal{H}_1$ を仮定する。このとき、自己共役作用素の順序に関して

$$\Sigma_{YY|U} \geq \Sigma_{YY|X} \quad (12)$$

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

$$\Sigma_{YY|X} - \Sigma_{YY|U} = 0 \iff Y \perp\!\!\!\perp V | U \quad (13)$$

の同値性が成立する。

略証. 条件付分散に関するよく知られた関係式

$$\begin{aligned} \text{Var}_{Y|U}[g(Y)|U] &= E_{V|U}[\text{Var}_{Y|U,V}[g(Y)|U, V]] \\ &\quad + \text{Var}_{V|U}[E_{Y|U,V}[g(Y)|U, V]] \end{aligned}$$

を U に関して期待値をとると $E_U[\text{Var}_{Y|U}[g(Y)|U]] - E_X[\text{Var}_{Y|X}[g(Y)|X]] = E_U[\text{Var}_{V|U}[E_{Y|U,V}[g(Y)|U, V]]] \geq 0$ により (12) 式を得る。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 (13) 式を得る。 \square

定理 6 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。

$$\min_S \Sigma_{YY|U}, \quad \text{subject to } U = \Pi_S X \quad (14)$$

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次章で導く。

3 カーネル次元削減法 (KDR)

(14) 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、核関数としてガウス関数のみを考えることにする。

[3] にならって (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。\$a = 1, 2\$ に対し、\$\tilde{k}_a(\cdot, X_i)\$ を \$\tilde{k}_a(\cdot, X_i) = k_a(\cdot, X_i) - \frac{1}{n} \sum_{j=1}^n k_a(\cdot, X_j)\$ により定めよう。(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、これらを基底にして、作用素 \$\Sigma_{YX}\$ の制限を行列表示すると、再生性を用いることにより、

$$P_n G_X P_n G_Y P_n$$

が得られる。ここで射影行列 \$P_n\$ は、\$\mathbf{1}_n = (1, \dots, 1)^T\$ の記法のもと、\$P_n = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T\$ により定義され、また、\$n \times n\$ 行列 \$G_{X,ij}, G_{Y,ij}\$ はグラム行列と呼ばれ、\$(i, j)\$ 成分がそれぞれ \$k_1(X_i, X_j), k_2(Y_i, Y_j)\$ により定義される。以上により、

$$\hat{K}_X = P_n G_X P_n, \quad \hat{K}_Y = P_n G_Y P_n$$

と書くことにすると、

$$\hat{\Sigma}_{YX} = \hat{K}_Y \hat{K}_X \quad (15)$$

が作用素の推定量として使える。

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 \$\hat{\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 \$\Sigma_{ZZ}\$ を推定する際には、正則化を用い、

$$\hat{\Sigma}_{ZZ} = (\hat{K}_Z + \varepsilon I_n)^2$$

(\$\varepsilon > 0\$) を推定量として使うことにする。以上により、条件付共分散作用素の推定量 \$\hat{\Sigma}_{Y|U}\$ を

$$\hat{\Sigma}_{Y|U} := \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}^{-1} \hat{\Sigma}_{UY} \quad (16)$$

により定める。有効部分空間を見つけるためには、この正定値行列を最小化するように、\$U\$ を決める部分空間を探せばよい。

正定値対称行列としての大きさをはかるには、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、本論文では \$\hat{\Sigma}_{Y|U}\$ の行列式を考える。行列式の Schur 分解を用いると、

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

の記法のもと、\$\det \hat{\Sigma}_{Y|U} = \det \hat{\Sigma}_{[YU][YU]} / \det \hat{\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

$$\min_{B \in \mathbb{R}^{m \times r}} \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}, \quad \text{ただし } U = B^T X \quad (17)$$

により得られる。ここで \$\det \hat{\Sigma}_{YY}\$ は定数であるが、目的関数の対称性のために加えた。\$S\$ ないし行列 \$B\$ を求めるこの最小化問題を、カーネル次元削減法 (Kernel dimensionality reduction, KDR) と呼ぶことにする。

(17) 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。[3] では、これを一般の確率変数の相互情報量の近似として提案し独立成分分析に用いたが、本論文の導出は近似ではなく理論的特徴づけに基づく。

カーネル次元削減法を実行するためには、目的関数の最小化を行う必要があるが、これは非線形かつ非凸な関数の最小化であり、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た。さらに局所解の問題を避けるために、ガウス核関数の分散パラメータを徐々に小さくする、一種のアニーリング法を用いている。また、(17) 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 \$\hat{K}_Y\$ などを低ランク行列で近似すると演算量を大幅に削減することが可能である [3]。

4 KDR 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。

まずデータ可視化の能力を見る目的で、UCI レポジトリ (<http://www.ics.uci.edu/~mllearn/MLRepository.html>) の Wine データを用いた。これは 3 種類のワインに対する 17 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のが見える。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く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 つの実データに対し、

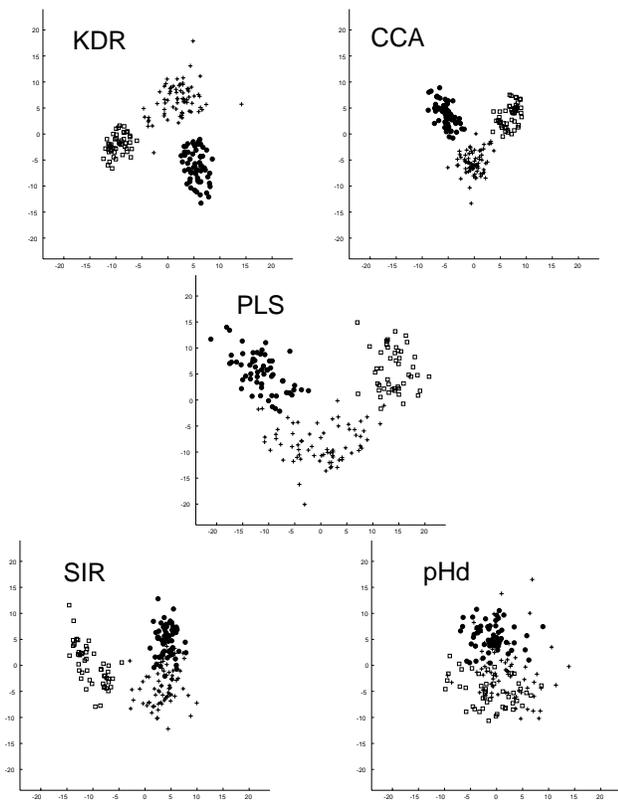


図 2: Wine データの 2 次元射影。" + ", " • ", " □ " が 3 クラスに対応。

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 SVM 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ことで、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。

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、(17) 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 [9] 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 [5] 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 ${}_m C_r$ 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 greedy サーチを用いて、遺伝子発現データ Leukimia [8] 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、その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率は 32 であった。[8] では、5 個リジェクトした場合に 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、再生核ヒルベルト空間を用いて回帰問題における次元削減を論じた。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、新しい次元削減法を提案した。

この次元削減法 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するとともに、変数選択問題への拡張も述べた。

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, ONR MURI N00014-00-1-0637, NSF grant IIS-9988642, および Intel Corporation からの研究資金により行われた。

参考文献

- [1] D. Alpay. *The Schur Algorithm, Reproducing Kernel Spaces and System Theory*. American Mathematical Society, 2001.
- [2] N. Aronszajn. Theory of reproducing kernels. *Trnas. Amer. Math. Soc.*, 69(3):337–404, 1950.
- [3] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Machine Learning Research*, 3:1–48, 2002.
- [4] C. R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973.
- [5] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *J. Amer. Stat. Assoc.*, 80:580–598, 1985.
- [6] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Stat. Assoc.*, 76:817–823, 1981.
- [7] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. TechReport 641, UC Berkeley, Dept. of Statistics, 2003.
- [8] T. R. Golub, D. Slonim, P. tamayo, and C. H. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
- [9] D. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.*, 5:81–102, 1978.
- [10] K.-C. Li. Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Stat. Assoc.*, 86:316–342, 1991.
- [11] K.-C. Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein’s lemma. *J. Amer. Stat. Assoc.*, 87:1025–1039, 1992.
- [12] H. Wold. Partial least squares. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. Wiley, 1985.

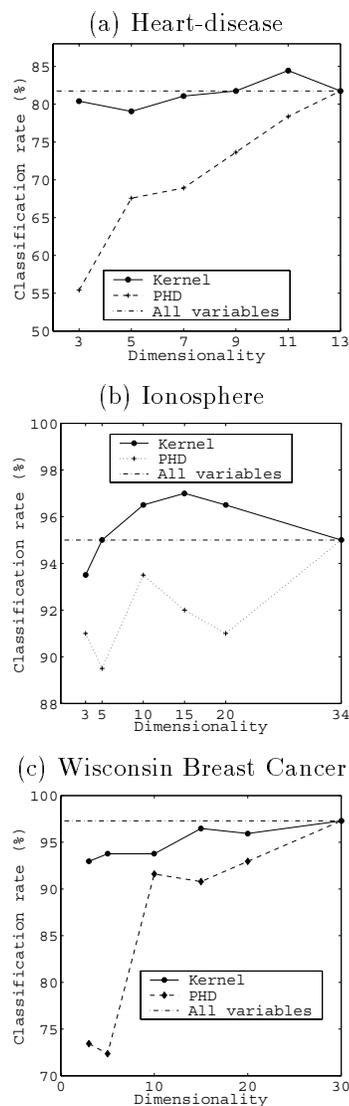


図 3: 次元削減後のテストデータに対する SVM の判別正解率。(a) $m = 13$, $N = 149$, $T = 148$ (b) $m = 34$, $N = 151$, $T = 200$ (c) $m = 30$, $N = 200$, $T = 369$ (m : 説明変数の次元, N : 訓練データ数, T : テストデータ数)