Dimension Reduction for Regression with Reproducing Kernels

Kenji Fukumizu
Institute of Statistical Mathematics, Japan
Visiting UC Berkeley

Statistical Colloquium. March 18, 2003

Joint work with Michael Jordan and Francis Bach in Berkeley

Outline

- Introduction
  - Dimension reduction for regression
- Conditional Independence and RKHS
  - Dimension reduction and conditional independence
  - Reproducing kernel Hilbert space
  - Conditional covariance operator
- Kernel Dimension Reduction for Regression
  - Algorithm and experimental results
- Extension to Variable Selection
- Summary

Introduction

Dimension reduction for regression

- Regression

$$Y \sim f(X, Z) \qquad \text{or} \qquad p(Y \mid X)$$

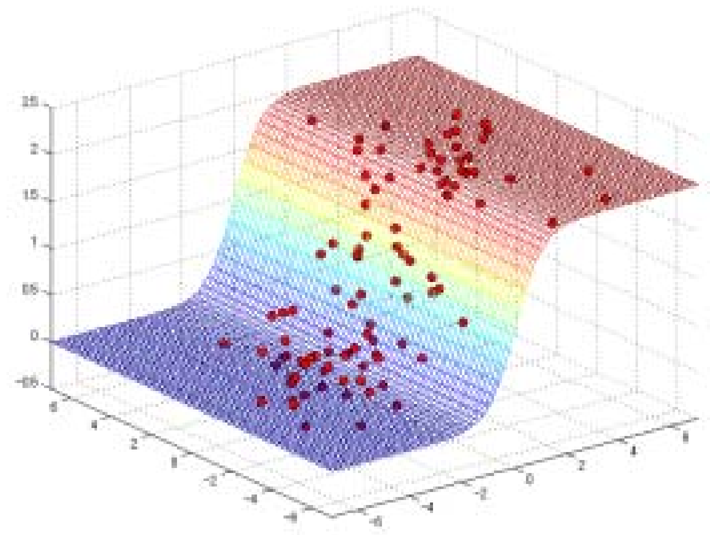
Y: response variable, X: m-dim. explanatory variable, Z: noise

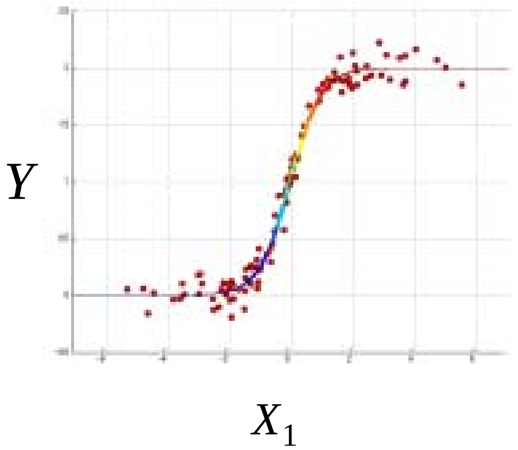
- Goal: Find effective subspace defined by B.

$$\tilde{p}(Y \mid B^T X) = p(Y \mid X) \qquad B\text{: } m \times d \text{ matrix} \qquad d \text{ is fixed.}$$

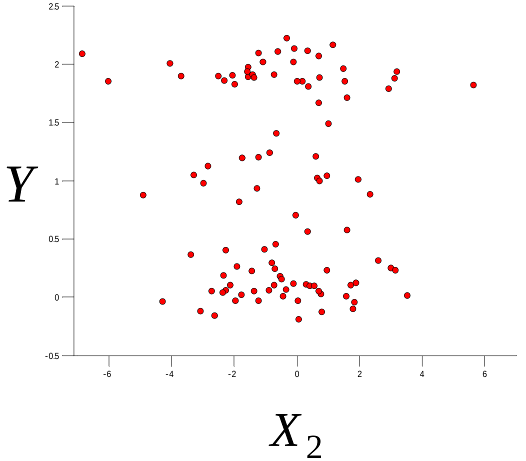
- Effective subspace to explain Y.
- Compact representation of the statistical relation.
  - data analysis : what determines Y?.
  - preprocessing of regression:
    accuracy of regression, computational efficiency.

– Example





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



Semi-parametric problem

Assume $p_{Y|X}(Y \mid X) = \tilde{p}(Y \mid B_0^T X)$ B_0: $m \times d$ matrix

i.i.d. sample $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$ given.

Find the subspace B_0 without knowing anything about $p_{Y|X}$ (or $\tilde{p}$).

There is the infinite degree of freedom on unestimated p.

→ Semiparametric problem.

Approach

- Formulate the problem by conditional independence.
- Use reproducing kernel Hilbert spaces as functional spaces for the infinite degree of freedom.

Existing Methods

- Sliced Inverse Regression (SIR, Li 1991)
  - PCA of $\mathrm{E}[X|Y]$ → use slice of Y.
  - Semiparametric method: no assumption on $p(Y|X)$.
  - Elliptic assumption on the distribution of X is necessary.
- Principle Hessian Direction (pHd, Li 1992)
  - Average Hessian $\Sigma_{yxx} \equiv E[(Y-\bar{Y})(X-\bar{X})(X-\bar{X})^T]$ is used.
  - If X is Gaussian, eigenvectors gives the effective directions.
  - Gaussian assumption on X. Y must be one-dimensional.
- Projection pursuit approach (e.g. Friedman et al. 1981)
  - Additive model is used for regressor.
- Canonical Correlation Analysis (CCA) / Partial Least Square (PLS)
  - Linear assumption on the regression.

Conditional Independence

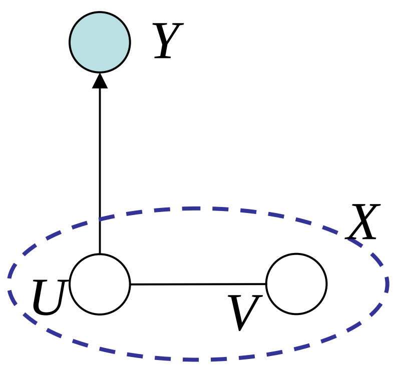
- Dimension reduction and conditional independence

$(U, V) = (B^T X, C^T X)$ for $(B, C) \in O(m)$

B gives the effective subspace $\Leftrightarrow$ $p_{Y|X}(y \mid x) = p_{Y|U}(y \mid B^T x)$

$\Leftrightarrow$ $p_{Y|U,V}(y \mid u, v) = p_{Y|U}(y \mid u)$ for all y, u, v

$\Leftrightarrow$ Conditional independence $Y \perp V \mid U$



- Characterization of conditional independence

⇨ Reproducing kernel Hilbert space (RKHS)

Reproducing Kernel Hilbert Space

Definition

Ω : set. H : Hilbert space $\subset \{ f : \Omega \to \mathbf{R} \}$

H : reproducing kernel Hilbert space (RKHS)

$\underset{def}{\Leftrightarrow}$ $\exists\, k : \Omega \times \Omega \to \mathbf{R}$ symmetric function (reproducing kernel) s.t.

1) $k(\cdot\,, x) \in H$ for all $x \in \Omega$.

2) $\langle k(\cdot\,, x), f \rangle_H = f(x)$ for $\forall f \in H$, $x \in \Omega$. reproducing property

Reproducing property makes computation easy and feasible.

e.g.) For $f = \sum_{i=1}^{n} a_i k(\cdot\,, X_i)$, $g = \sum_{j=1}^{m} b_j k(\cdot\,, X_j)$

$$\langle f, g \rangle_H = \sum_{ij} a_i b_j k(X_i, X_j)$$

– Example: Gaussian kernel

$k : \mathbf{R}^m \times \mathbf{R}^m \to \mathbf{R}$, $\quad k(x, y) = \exp\left(- \| x - y \|^2 / \sigma^2 \right)$

⇨ There is a RKHS on $\mathbf{R}^m$ with reproducing kernel k.

RKHS and Independence

- Independence and characteristic functions

Random variables X and Y are independent

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

$e^{\sqrt{-1}\omega^T x}$ and $e^{\sqrt{-1}\eta^T y}$ work as test functions which account for the infinite degree of freedom (L^2).

- RKHS characterization

H_X and H_Y are RKHS on Ω_X and Ω_Y, respectively.

Random variables $X \in \Omega_X$ and $Y \in \Omega_Y$ are independent

$$\Leftrightarrow \quad E_{XY}[f(X)g(Y)] = E_X[f(X)]E_Y[g(Y)] \qquad \text{for all } f \in H_X,\ g \in H_Y$$

This is true if H_X and H_Y are RKHS for Gaussian kernels.

(Bach & Jordan 2002)

Cross-covariance Operator

Definition

X and Y: random variable on Ω_X and Ω_Y, respectively.

H_X and H_Y : RKHS on Ω_X and Ω_Y, respectively, with bounded kernels.

We can define a bounded operator $\Sigma_{YX} : H_X \to H_Y$ by

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

for all $f \in H_X,\ g \in H_Y$

Σ_{YX} is called cross-covariance operator.

Cross-covariance operator and Independence

Theorem

H_X and H_Y : RKHS with Gaussian kernel.

X and Y are independent $\Leftrightarrow$ $\Sigma_{YX} = O$

RKHS and Conditional Independence

Conditional covariance

X and Y are random vectors. H_X , H_Y : RKHS with kernel k_X, k_Y, resp.

Assumption: $\exists\, \Sigma_{XX}^{-1}$, $E_{Y|X}[g(Y) \mid X] \in H_X$ for all $g \in H_Y$.

$$\boxed{\left\langle f, \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} g \right\rangle = E_X\left[\mathrm{Cov}_{Y|X}[f(Y), g(Y) \mid X]\right]}$$

Def. $\Sigma_{YY|X} \equiv \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$: conditional covariance operator

c.f. For Gaussian $\mathrm{Cov}_{Y|X}[a^T Y, b^T Y \mid X = x] = a^T\left(V_{YY} - V_{YX}V_{XX}^{-1}V_{XY}\right)b$

– Monotonicity of conditional covariance operators

Y, $X = (U,V)$: random vectors

$$\boxed{\Sigma_{YY|U} \geq \Sigma_{YY|X}}$$

$\geq$: in the sense of self-adjoint operators

RKHS and Conditional Independence

Conditional independence

Theorem

$X = (U,V)$ and Y are random vectors.

H_X , H_U , H_Y : RKHS with Gaussian kernel k_X, k_U, k_Y, resp.

$E_{Y|X}[g(Y)\,|\,X] \in H_X$ and $E_{Y|U}[g(Y)\,|\,U] \in H_U$ for all $g \in H_Y$.

$\Rightarrow$ $$Y \perp V \,|\, U \quad \Leftrightarrow \quad \Sigma_{YY|U} = \Sigma_{YY|X}$$

Minimization of conditional covariance operator

$$\min_{B : U = B^T X} \Sigma_{YY|U}$$ $\Rightarrow$ B gives the effective subspace

– Evaluation
  - Operator norm -- maximum eigenvalue.
  - Trace norm -- sum of eigenvalues
  - Determinant -- product of eigenvalues

Kernel Dimension Reduction

- Estimation of conditional covariance operator

$(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$: i.i.d. sample from the true joint probability.

The space is restricted in the linear hull of $\{k(\cdot, X^{(i)}) \mid 1 \le i \le n\}$ and $\{k(\cdot, Y^{(i)}) \mid 1 \le i \le n\}$

Replace $\Sigma_{YY|U}$ by n x n matrix

$$\hat{\Sigma}_{YY|U} \equiv \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}{}^{-1} \hat{\Sigma}_{UY}$$

where

$$\hat{\Sigma}_{UU} = (G_U + \varepsilon I_n)^2, \quad \hat{\Sigma}_{YY} = (G_{YY} + \varepsilon I_n)^2, \quad \hat{\Sigma}_{UY} = G_U G_Y$$

ε : regularization coefficient

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

reproducing property and empirical average

Kernel dimension reduction (KDR)

$$\min_B \quad \hat{\Sigma}_{YY|U} \equiv \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY} \qquad U = B^T X$$

$$\Leftrightarrow \min_B \quad \det\left[I_n - \hat{\Sigma}_{YY}^{-1/2}\hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY}\hat{\Sigma}_{YY}^{-1/2}\right]$$

$$\Leftrightarrow \min_B \quad \boxed{\frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}} \qquad \text{where} \quad \hat{\Sigma}_{[YU][YU]} = \begin{pmatrix} \hat{\Sigma}_{YY} & \hat{\Sigma}_{YU} \\ \hat{\Sigma}_{UY} & \hat{\Sigma}_{UU} \end{pmatrix}$$

Kernel generalized variance (KGV, Bach & Jordan 2002)

Kernel Dimension Reduction (KDR) = minimization of KGV

Minimization method – gradient-based method.

Extension of Kernel ICA

- Kernel ICA (Bach & Jordan 02): kernel method for independence.
  → KDR: kernel method for conditional independence.

Wide applicability of KDR

- Semiparametric method: no assumptions on $p(Y|X)$.
- KDR needs no strong assumption on the distribution of X, Y and dimensionality of Y.

c.f. other method; SIR, pHd, CCA, PLS, etc.

Computational cost

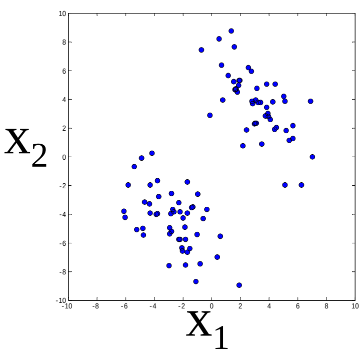
- Multiplication of n x n matrices is computationally hard.
  → Incomplete Cholesky decomposition
- Local minimum → annealing is used in gradient method.

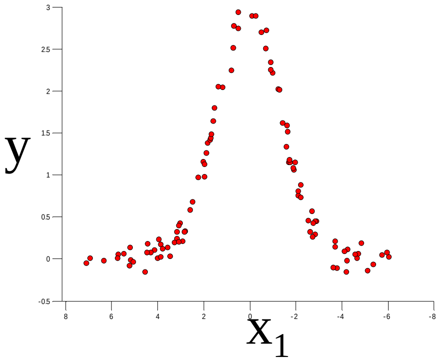
Experiments

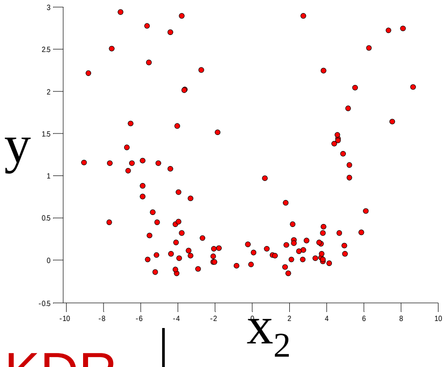
- Synthesized data
  - Data

    X: 2 dim, Y: 1 dim
    100 data

    $$Y \sim 2\exp(-X_1^2) + N(0;\, 0.1^2)$$





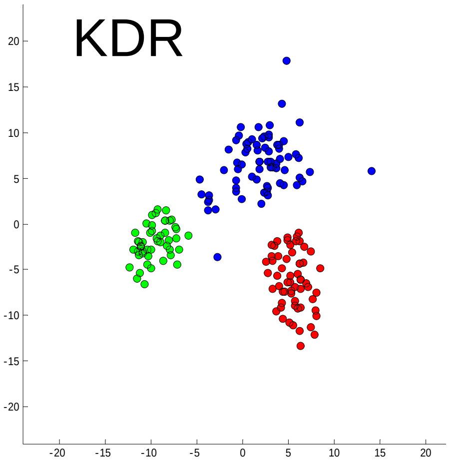


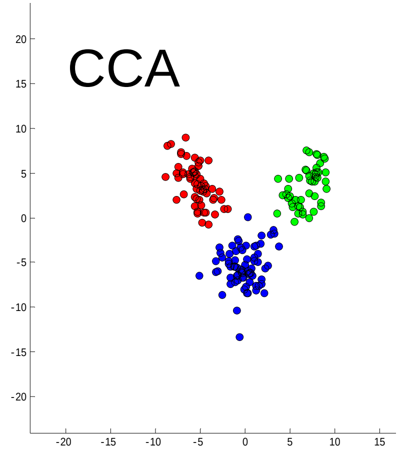
  - Results

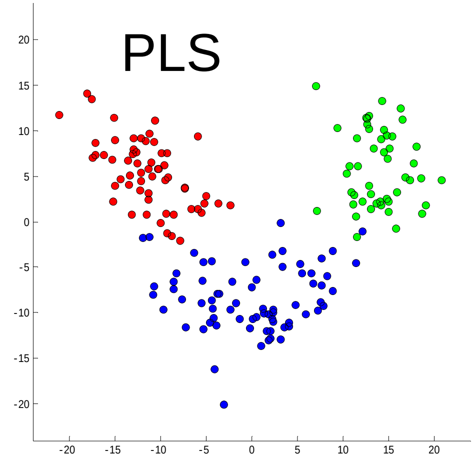
	SIR	pHd	CCA	PLS	KDR
Angle (deg.)	-86.522	57.015	-10.416	-26.093	0.298

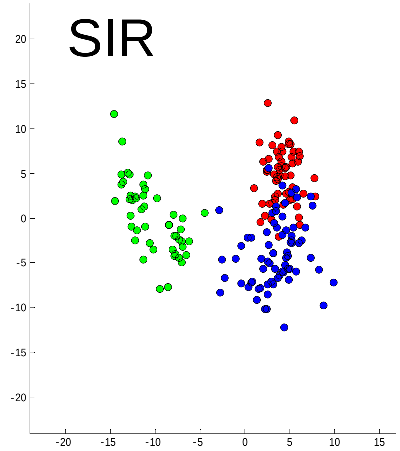
Wine data

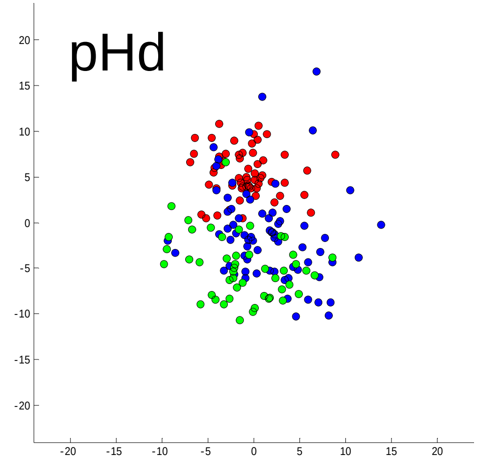
- Data
  13 dim. 178 data.
  3 classes
  2 dim. projection











Classification accuracy

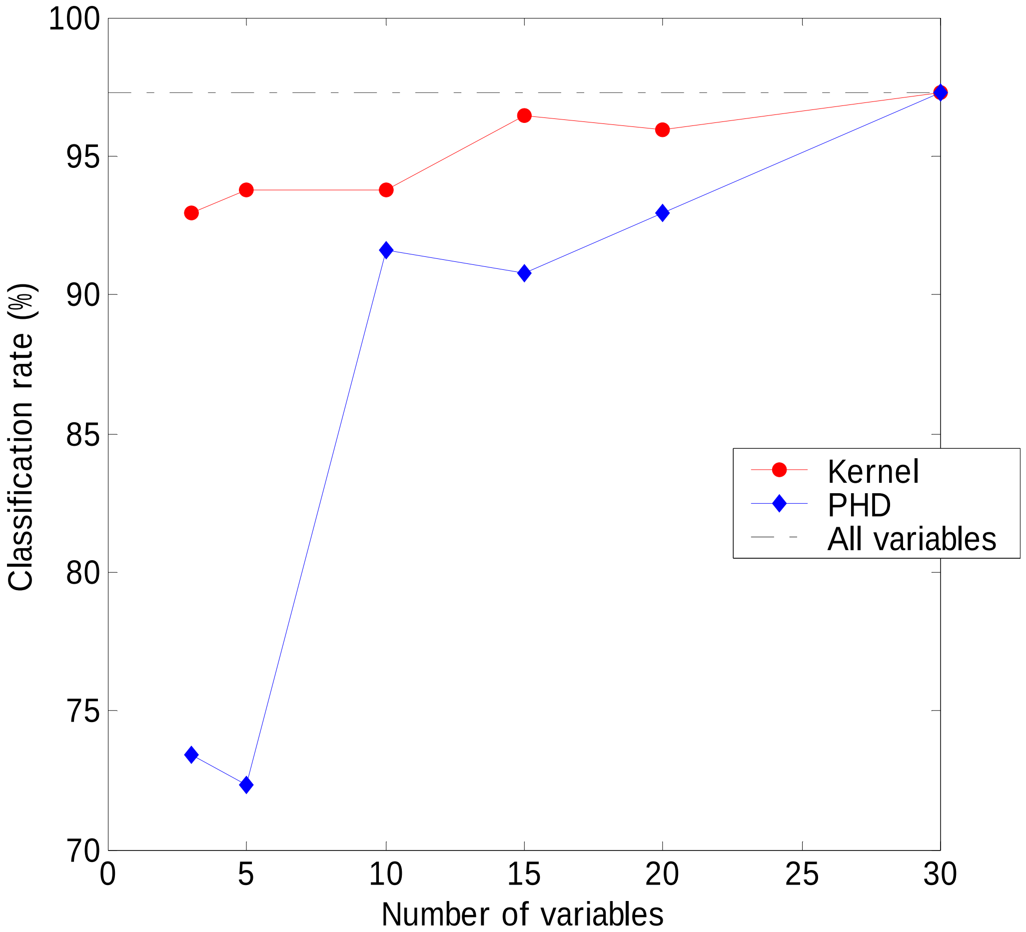
- Purpose:
  to see how much information on Y is maintained in the low-dimensional subspace of X.
- Test classification accuracy of Support Vector Machine after reducing dimensionality.
- Data sets for binary classification from UCI repository.
- Comparison with pHd.
  Many methods are NOT applicable for binary classification tasks.

Breast-cancer-Wisconsin

X: 30 dim.

training data=200

test data=369

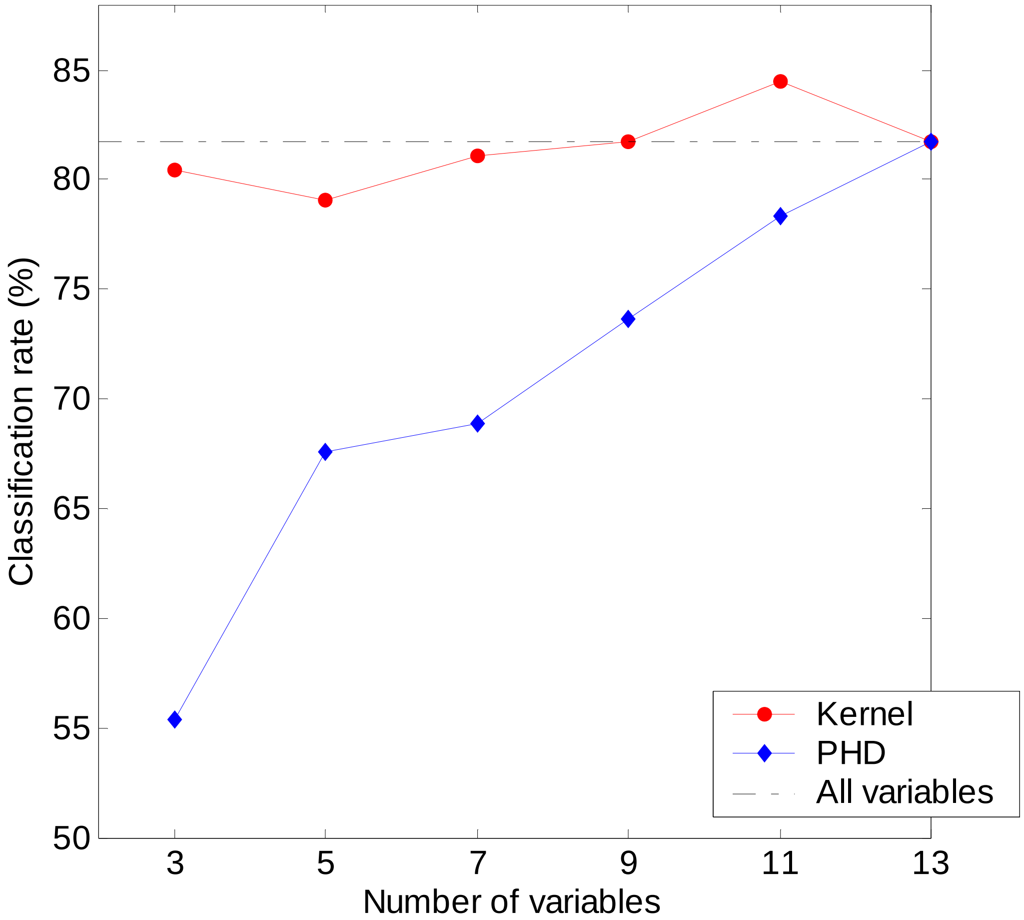


Heart-disease

X: 13 dim.

training data=149,

test data=148

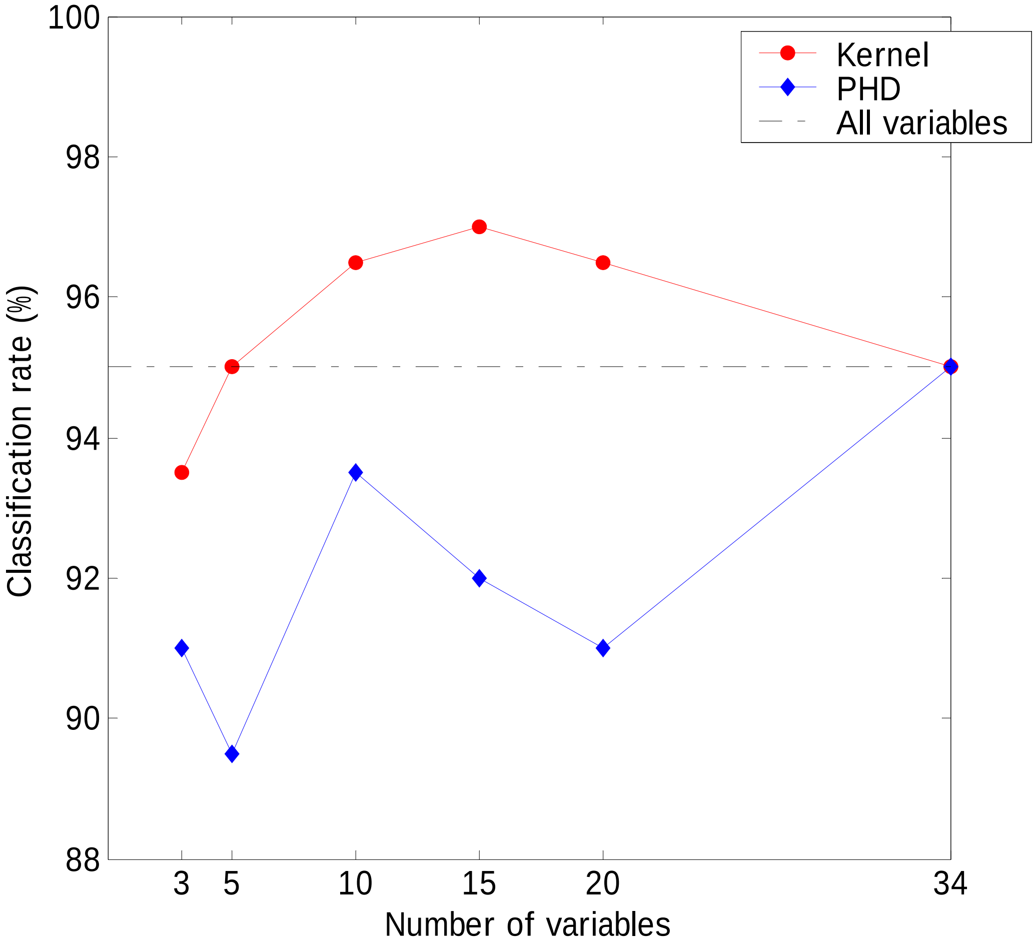


Ionosphere

X: 34 dim.

training data=151

test data=200



Extension to Variable Selection

- Variable selection by KGV
  - Select subset $(X_{i_1}, \ldots, X_{i_d})$ from $\{X_1, \ldots, X_m\}$.
  - Principle

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

  - KGV gives an objective function for variable selection.

$$\boxed{\min_U \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}}$$

$\min$ is taken over subsets $U = (X_{i_1}, \ldots, X_{i_d})$ where $1 \le i_1 < \cdots < i_d \le m$

  - Problem: combinatorial explosion
    - ${}_m\mathrm{C}_d$ evaluations are needed.
    - Calculation of all the combinations is possible only for small m and d.

Experiments of Variable Selection

- Small data set
  - *Boston Housing*:
    X :13 dim.,
    Y = house price,
    506 data.
  - 4 variables are selected.
    ${}_{13}C_4 = 715$.

	1st	2nd	3rd	ACE
CRIM		O		
ZN				
INDUS				
CHAS				
NOX				
RM	O	O	O	O
AGE				
DIS			O	
RAD				
TAX	O		O	O
PTRATIO	O	O		O
B				
LSTAT	O	O	O	O

ACE: Breiman & Friedman (1985)

Variable Selection for Large Data Sets

- Computational issue
  - Combinatorial explosion
    If m and d are large, e.g. m=1000, d=20, evaluation of all the subsets is intractable.

- Efficient optimization
  - Greedy algorithm
    1. Start from one variable.
    2. For already chosen t variables $S_t = \{X_{i_1}, \ldots, X_{i_t}\}$, evaluate KGV of $S_t \cup \{X_j\}$ for all j, and select the best one.
    3. Repeat this to d variables.
  - Random optimization
    Genetic algorithm

Application: Gene Selection

- AML/ALL classification (Golub et al. 1999)
  - Microarray data: 6817 dim. 38 data.
  - Class label:
    AML (acute myeloid leukemia) / ALL (acute lymphoblastic leukemia).
  - Golub et al (1999 Science) show 50 effective genes using nearest neighborhood analysis.

- Results
  - 50 genes are selected by the kernel method and compared with previous works.

Application: Gene Selection

Kernel	Golub99	Lee03	Szabo02	Li02	Fuj
1 Leukotriene C4 synthase (LTC4S)	o			o	o
2 Zyxin	o	o		o	o
3 FAH Fumarylacetoacetate	o			o	o
4 LYN V-yes-1 Yamaguchi sarcoma	o	o		o	o
5 LEPR Leptin receptor	o			o	o
6 CD33 CD33 antigen (differentiati	o	o		o	o
7 Liver mRNA for interferon-gamma					o
8 "PRG1 Proteoglycan 1, secretory	o				o
9 GB DEF = Homeodomain protein Hox	o				
10 DF D component of complement (ad	o	o	o		o
11 INTERLEUKIN-8 PRECURSOR	o	o			o
12 INDUCED MYELOID LEUKEMIA	o				o
13 "PEPTIDYL-PROLYL CIS-TRANS	o				o
14 Phosphotyrosine independent liga	o				o
15 ATP6C Vacuolar H+ ATPase proton	o				
16 CST3 Cystatin C (amyloid angiopa	o	o	o	o	o
17 Interleukin 8 (IL8) gene	o	o	o		o
18 CTSD Cathepsin D (lysosomal aspa	o				o
19 "ITGAX Integrin, alpha X (antige	o				o
20 "LGALS3 Lectin, galactoside-bind	o				o
21 Epb72 gene exon 1	o				o
22 MAJOR HISTOCOMPATIBILITY	o				
23 LYZ Lysozyme	o				o
24 Azurocidin gene	o				o
25 "PFC Properdin P factor, complem	o				o
26 Lysophospholipase homolog (HU-K5					
27 PPGB Protective protein for beta		o			o
28 "Catalase (EC 1.11.1.6) 5'flank	o				
29 FTH1 Ferritin heavy chain					o
30 "CD36 CD36 antigen (collagen typ					o
31 EUKARYOTIC PEPTIDE CHAIN					
32 GB DEF = CD36 gene exon 15					
33 CSF1 Colony-stimulating factor 1					
34 CA2 Carbonic anhydrase II					o
35 Hepatocyte growth factor-like pr					
36 MPO Myeloperoxidase		o			o
37 "CHRNA7 Cholinergic receptor, ni					o
38 AFFX-HUMTFRR/M11507_M_at					
39 "C1NH Complement component 1 inh					
40 "GB DEF = Glycophorin Sta (type					
41 GYPE Glycophorin E					
42 AFFX-HUMTFRR/M11507_3_at					
43 Metabotropic glutamate receptor					
44 "GB DEF = Neutrophil elastase ge			o		
45 "ELA2 Elastatse 2, neutrophil"		o		o	o
46 GB DEF = Kazal-type serine prote					
47 LCAT Lecithin-cholesterol acyltr					
48 "ALDH2 Aldehyde dehydrogenase 2,					
49 ANX8 Annexin VIII					
50 "PRSS3 Protease, serine, 3 (tryp					
#agree/#selected	25/50	10/28	4/9	8/10	29/50

Summary

- Kernel method for dimension reduction in regression
  – Dimension reduction for regression = conditional independence.
  – Conditional covariance operators gives the criterion for the conditional independence.
- Kernel dimension reduction / variable selection
  – have wide applicability to dimension reduction / variable selection. *c.f*. other methods have some restrictions.
  – find effective subspaces / variables in practical problems.
- Future/ongoing studies
  – Theoretical analysis of the estimator: consistency etc.
  – How to choose the number of dimensions.
  – More efficient optimization techniques for variable selection.
  – Mixture of effective subspaces.