

# 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法

福水 健次

統計数理研究所

2003年9月26日 日本数学会

M.I. Jordan, F.R. Bach (UC Berkeley) との共同研究

# Outline

## ■ Introduction

- Dimensionality reduction for regression

## ■ Conditional Independence and RKHS

- Dimensionality reduction and conditional independence
- Reproducing kernel Hilbert space
- Conditional covariance operator

## ■ Kernel Dimensionality Reduction for Regression

- Algorithm and experimental results

## ■ Extension to Variable Selection

## ■ Summary

# Introduction

## ■ Dimensionality reduction

- Dimensionality reduction is important for high-dimensional data such as gene expression, text, image, etc.
- Dimensionality reduction
  - **feature selection** – linear or nonlinear combination of variables.
  - **variable selection** – subset of variables.
- Purposes of dimensionality reduction
  - Compact and readable explanation of statistical relationship
  - Computational efficiency
  - Accuracy of estimation

## ■ Dimensionality reduction for regression

- Regression: analysis of statistical dependence of  $Y$  on  $X$ ,

Conditional probability density  $p(Y | X)$

$Y$  : response variable,

$X = (X_1, \dots, X_m) \in \mathbf{R}^m$  explanatory variables

- Goal of dimensionality reduction  
= Find an **effective subspace** to select feature vector for regression.

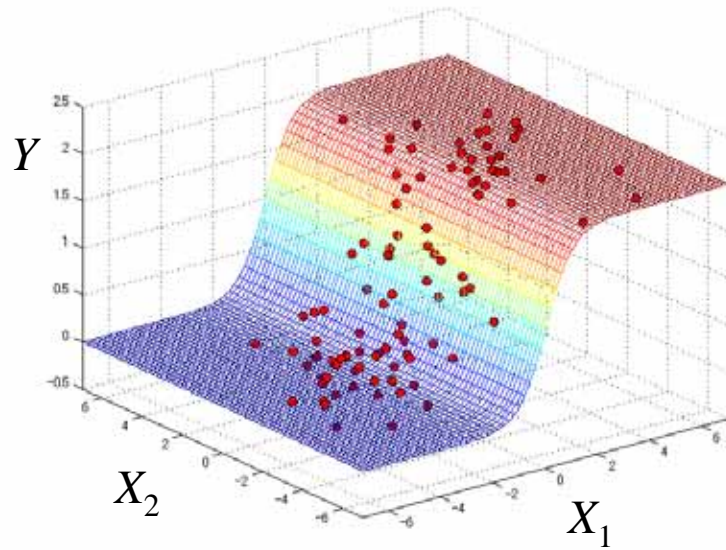
$$p(Y | X) = \tilde{p}(Y | b_1^T X, \dots, b_d^T X) \quad \left( = \tilde{p}(Y | B^T X) \right)$$

$B = (b_1, \dots, b_d)$ :  $m \times d$  matrix  $d (< m)$  is fixed.

Feature vector as a linear combination of  $X_1, \dots, X_m$ .

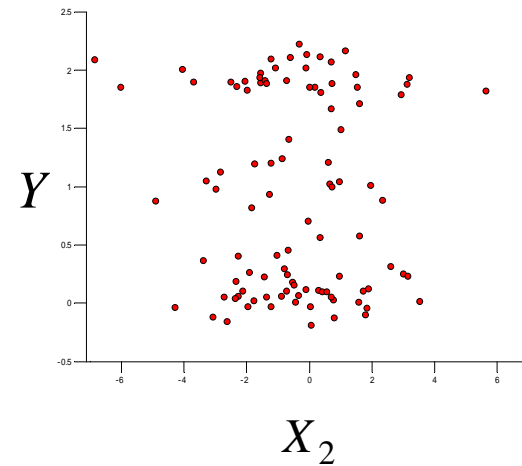
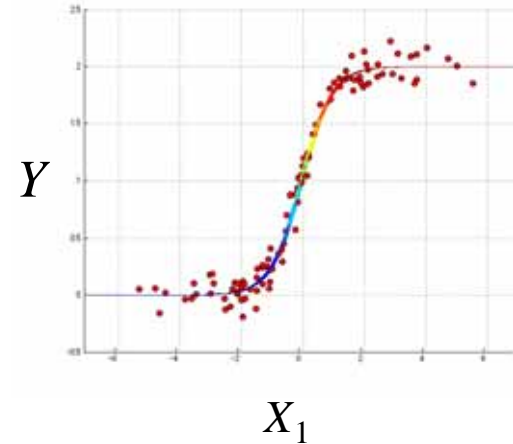
Effective subspace contains all the information in  $X$  to explain  $Y$ .

## – Example



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

Effective subspace = direction of  $X_1$



## ■ Semiparametric problem

Assume existence of an effective subspace

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

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

### PROBLEM

Find the effective subspace  $B_0$

**without** assuming any models about the conditional probability  $p_{Y|X}$  or the marginal distributions  $p_X$  and  $p_Y$ .

- There is the infinite degree of freedom on unestimated  $p_{Y|X}$ .  
→ Semiparametric problem.
- Once the effective subspace is obtained, any type of regressor can be built on that space.

# Conditional Independence

## ■ Dimensionality reduction and conditional independence

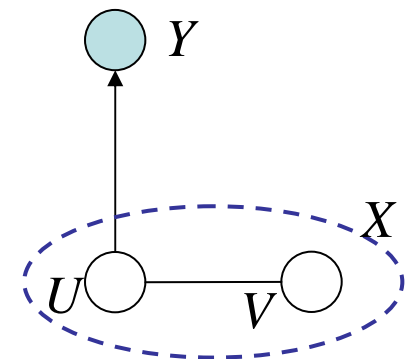
$(U, V) = (B^T X, C^T X)$  for  $(B, C) \in O(m)$  ( $m$ -dim. orthogonal matrix)

$B$  gives the projector onto the effective subspace

$$\Leftrightarrow p_{Y|X}(y|x) = p_{Y|U}(y|B^T x)$$

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

$$\Leftrightarrow \text{Conditional independence} \quad Y \perp V | U$$



## ■ Characterization of conditional independence

➡ Reproducing kernel Hilbert space (RKHS)

# Reproducing Kernel Hilbert Space

## ■ Definition

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

$H$  : reproducing kernel Hilbert space (RKHS)

$\Leftrightarrow_{\text{def}} \exists k : \Omega \times \Omega \rightarrow \mathbf{R}$  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

## ■ Example: Gaussian kernel

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

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



# Reproducing Kernel Hilbert Space

## ■ Properties of RKHS

- Condition of RKHS (Mercer)

If a symmetric function  $k: \Omega \times \Omega \rightarrow \mathbf{R}$  is positive definite, i.e. for any  $x_1, \dots, x_n \in \Omega$ ,

$$\begin{pmatrix} k(x_i, x_j) \end{pmatrix} \geq 0$$

then, there uniquely exists a RKHS with  $k$  its reproducing kernel.

- Advantage of RKHS

- Reproducing property makes computation easy and feasible.

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

$$\langle f, g \rangle_H = \sum_{ij} a_i b_j k(X_i, X_j) \quad \left( k(X_i, X_j) \right)_{ij} : \text{Gram matrix}$$

- RKHS is much smaller than  $L^2(\mathbf{R}^m)$

If  $k$  is continuous, all the functions in  $H$  are continuous, and  $H \subset C(\Omega)$  is a continuous embedding.

# RKHS and Independence

## ■ Independence and characteristic functions

Random variables  $X$  and  $Y$  are independent

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

$e^{\sqrt{-1}\omega^T x}$  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  $\Omega_X$  and  $\Omega_Y$ , respectively.

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

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

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  $\Omega_X$  and  $\Omega_Y$ , respectively.

$H_X$  and  $H_Y$ : RKHS on  $\Omega_X$  and  $\Omega_Y$ , respectively, with bounded kernels.

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

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

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

(Riesz's theorem)

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

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

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

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

– **Monotonicity** of conditional covariance operators

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

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

$\geq$  : 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  $g \in H_Y$ .

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

## ■ Minimization of conditional covariance operator

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

### – Evaluation

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

# Kernel Dimensionality Reduction

## ■ Estimation of conditional covariance operator

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

Restrict the spaces to the linear hull of  $\left\{ k(\cdot, X^{(i)}) - \frac{1}{n} \sum_{j=1}^n k(\cdot, X^{(j)}) \mid 1 \leq i \leq n \right\}$   
and  $\left\{ k(\cdot, Y^{(i)}) - \frac{1}{n} \sum_{j=1}^n k(\cdot, Y^{(j)}) \mid 1 \leq i \leq n \right\}$

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

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

where

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

$\varepsilon$  : regularization coefficient

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

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

reproducing property and empirical average

# Kernel Dimensionality Reduction

## ■ Kernel dimensionality reduction (KDR)

$$\begin{aligned} & \min_B \quad \hat{\Sigma}_{YY|U} \equiv \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}^{-1} \hat{\Sigma}_{UY} & U = B^T X \\ \Leftrightarrow & \min_B \quad \det \left[ I_n - \hat{\Sigma}_{YY}^{-1/2} \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}^{-1} \hat{\Sigma}_{UY} \hat{\Sigma}_{YY}^{-1/2} \right] \\ \Leftrightarrow & \min_B \quad \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}} & \text{where } \hat{\Sigma}_{[YU][YU]} = \begin{pmatrix} \hat{\Sigma}_{YY} & \hat{\Sigma}_{YU} \\ \hat{\Sigma}_{UY} & \hat{\Sigma}_{UU} \end{pmatrix} \end{aligned}$$

*c.f.* mutual information of Gaussian variables.

### Method of KDR

Kernel Dimensionality Reduction (KDR)

= minimization of  $\frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}$

gradient-based method is used for the minimization.

# Kernel Dimensionality Reduction

## ■ Wide applicability of KDR

- The most general approach for dimensionality reduction:  
no model for  $p(Y|X)$  .
- 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  $n \times n$  matrices is computationally hard.  
→ Incomplete Cholesky decomposition
- Local minimum → annealing is used in gradient method.



# Existing Methods

- Sliced Inverse Regression (SIR, Li 1991)
  - PCA of  $E[X|Y]$   $\rightarrow$  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  $E[Y|X] = g_1(b_1^T X) + \dots + g_d(b_d^T X)$  is used.
- Canonical Correlation Analysis (CCA) / Partial Least Square (PLS)
  - Linear assumption on the regression.
- Nonparametric approach

# Experiments

## ■ Synthesized data

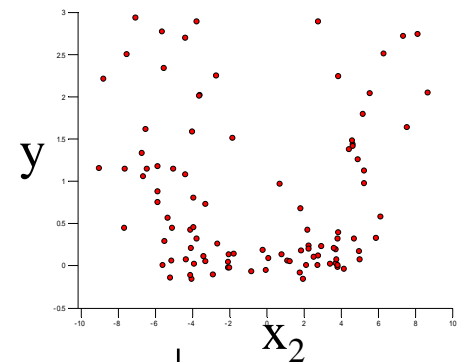
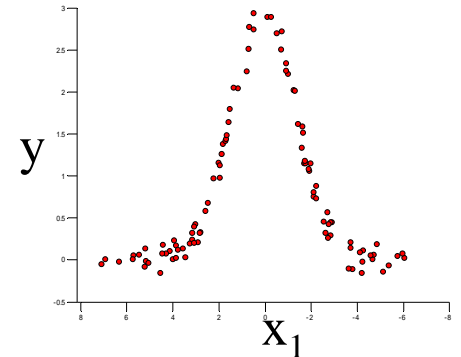
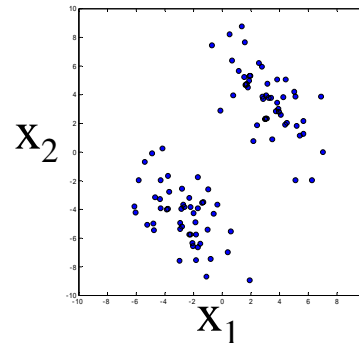
– Data

$X$ : 2 dim,  $Y$ : 1 dim  
100 data

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

– Problem: find one-dim. effective subspace

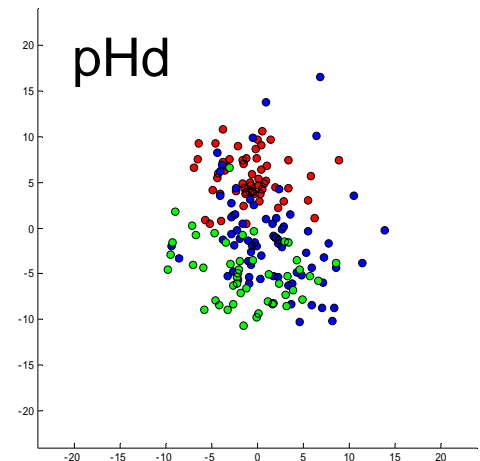
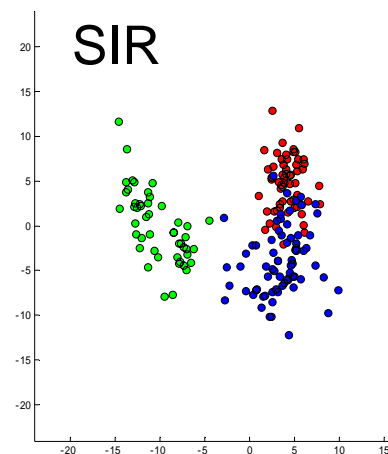
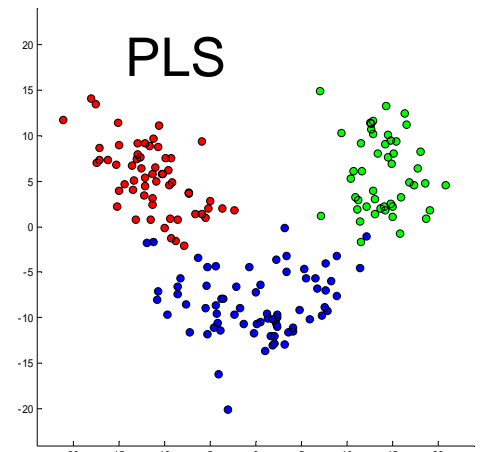
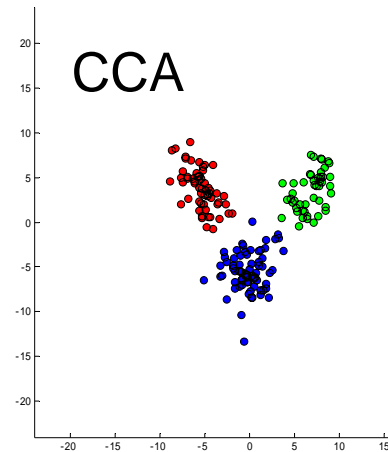
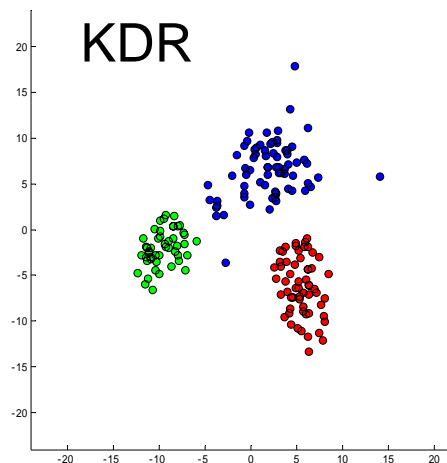
– Results



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

## ■ Wine data (from UCI Machine Learning Repository)

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



## ■ Experiments on 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.

# Experiments

## – Results

Breast-cancer-  
Wisconsin

X: 30 dim.

# training data=200

# test data=369

Heart-disease

X: 13 dim.

# training data=149,

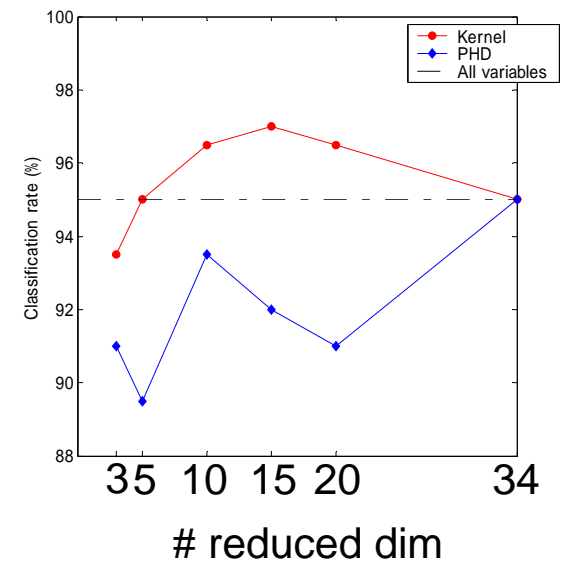
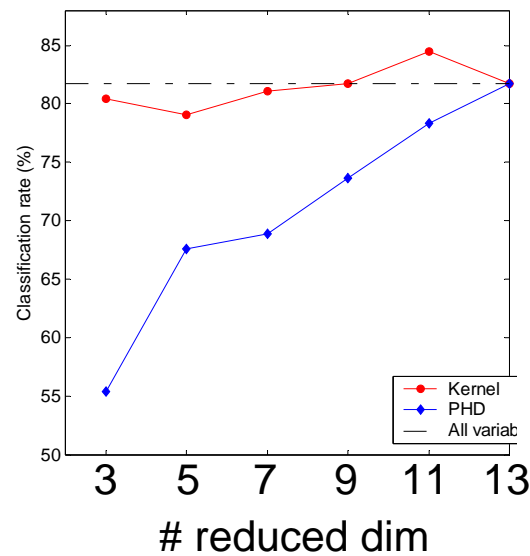
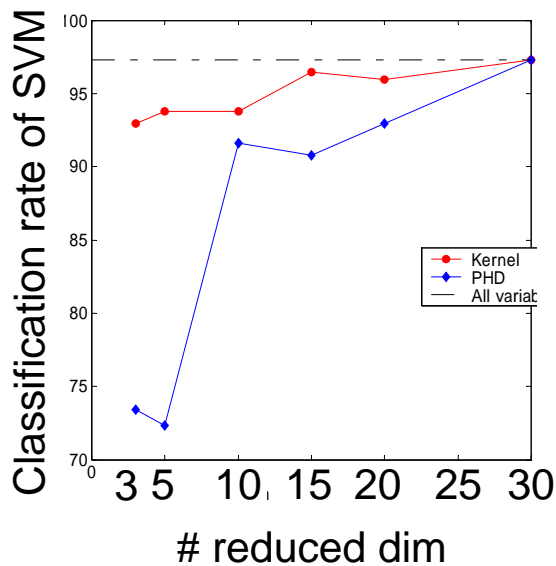
# test data=148

Ionosphere

X: 34 dim.

# training data=151

# test data=200



— Kernel  
— PHD

# Extension to Variable Selection

## ■ Variable selection by KDR

- Select a subset  $(X_{i_1}, \dots, X_{i_d})$  from  $\{X_1, \dots, X_m\}$ .
- Principle

$$Y \perp V | U \iff \Sigma_{YY|U} = \Sigma_{YY|X}$$

- objective function for variable selection.

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

min is taken over all the subsets

$U = (X_{i_1}, \dots, X_{i_d})$  where  $1 \leq i_1 < \dots < i_d \leq m$

- Problem: combinatorial explosion for computation

Evaluation of all the combinations is intractable for large  $m$  and  $d$ .

→ some approximate optimization methods are needed.

# Experiments of Variable Selection

## ■ Small data set

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

ACE: Breiman & Friedman (1985)

	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

# Application to Gene Selection

## ■ AML/ALL classification (Golub et al. *Science* 1999)

- Microarray data: 7129 dim. 38 training data.
- Binary classification: AML / ALL.
- Golub et al. show 50 effective genes using neighborhood analysis.

## ■ Greedy optimization algorithm

1. Start from one variable.
2. For already chosen  $t$  variables  $S_t = \{X_{i_1}, \dots, X_{i_t}\}$ , evaluate the objective function values for  $S_t \cup \{X_j, X_k\} - \{X_i\}$  for all combinations of  $X_j, X_k$  not in  $S_t$  and  $X_i$  in  $S_t$ , and select the best one.
3. Repeat this up to  $d$  variables.

## ■ Results

- 50 genes are selected by KDR method



# Application to Gene Selection

KDR	Golub99	Lee03	Szabo02	Li02	Fuj	
1 Leukotriene C4 synthase (LTC4S)	0			0	0	
2 Zyxn	0	0		0	0	
3 FAH Fumarylacetoacetate	0			0	0	
4 LYN V-yes-1 Yamaguchi sarcoma	0	0		0	0	
5 LEPR Leptin receptor	0			0	0	
6 CD33 CD33 antigen (differentiat	0	0		0	0	
7 Liver mRNA for interferon-gamma					0	
8 "PRG1 Proteoglycan 1, secretory	0				0	
9 GB DEF = Homeodomain protein Hox	0					
10 DF D component of complement (ad	0	0	0		0	
11 INTERLEUKIN-8 PRECURSOR	0	0			0	
12 INDUCED MYELOID LEUKEMIA	0				0	
13 "PEPTIDYL-PROLYL CIS-TRANS	0				0	
14 Phosphotyrosine independent liga	0				0	
15 ATP6C Vacuolar H+ ATPase proton	0					
16 CST3 Cystatin C (amyloid angio	0	0	0	0	0	
17 Interleukin 8 (IL8) gene	0	0	0		0	
18 CTSD Cathepsin D (lysosomal aspa	0				0	
19 "ITGAX Integrin, alpha X (antige	0				0	
20 "LGALS3 Lectin, galactoside-bind	0				0	
21 Epb72 gene exon 1	0				0	
22 MAJOR HISTOCOMPATIBILITY	0					
23 LYZ Lysozyme	0				0	
24 Azurocidin gene	0				0	
25 "PFC Properdin P factor, complem	0				0	
26 Lysophospholipase homolog (HU-K5						
27 PPGB Protective protein for beta		0			0	
28 "Catalase (EC 1.11.1.6) 5'flank	0					
29 FTH1 Ferritin heavy chain					0	
30 "CD36 CD36 antigen (collagen typ					0	
31 EUKARYOTIC PEPTIDE CHAIN						
32 GB DEF = CD36 gene exon 15						
33 CSF1 Colony-stimulating factor 1						
34 CA2 Carbonic anhydrase II					0	
35 Hepatocyte growth factor-like pr						
36 MPO Myeloperoxidase		0			0	
37 "CHRNA7 Cholinergic receptor, ni					0	
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			0			
45 "ELA2 Elastatse 2, neutrophil"		0		0	0	
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	25

# Application to Gene Selection

- Classification accuracy
  - Evaluation of classification rate for 34 independent test data.
  - Application of SVM using the selected genes.
  - Results
    - KDR + SVM: 32 correct / 34 samples
    - Golub et al: 29 correct + 5 rejected / 34 samples
    - 50 genes in Golub et al + SVM: 32 correct / 34 samples
    - Furey et al (2000) SVM: 30-32 correct / 34 samples for various number of genes 25 – 1000.
- KDR method provides effective genes.
- KDR method accounts for the combination of genes.
  - Many methods use relation between only  $X_j$  and  $Y$  for selection.

# Summary

- Kernel method is suitable for semiparametric problems
  - Dimensionality reduction for regression = conditional independence.
  - Conditional covariance operators gives the criterion for the conditional independence.
- Kernel dimensionality reduction (KDR)
  - The most general approach for dimensionality reduction.
  - KDR has wide applicability to feature / variable selection.  
*c.f.* other methods have some restrictions.
  - KDR finds effective features / variables in practical problems.
- Future/ongoing studies
  - Theoretical analysis of the estimator: consistency etc.
  - Extension to more general Bayesian networks.

## ■ Tech report etc.

<http://www.ism.ac.jp/~fukumizu/>