

Kernel Dimension Reduction for Regression

Kenji Fukumizu

Institute of Statistical Mathematics, Japan
Visiting UC Berkeley

Francis R. Bach & Michael I. Jordan
UC Berkeley

Introduction

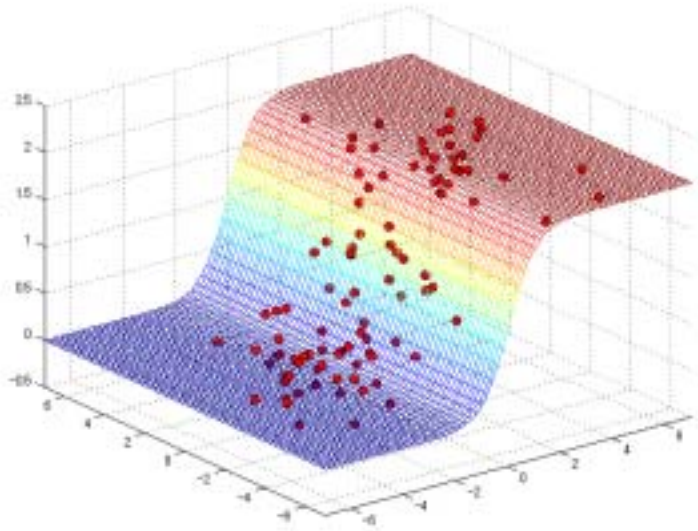
■ PROBLEM: Dimension reduction for regression

- Regression $p(Y | X)$ X : m -dim. explanatory variable / input
 Y : response variable / output,
- Goal: Find the **effective subspace** defined by B .

$$p(Y | X) = \tilde{p}(Y | b_1^T X, \dots, b_d^T X) \quad B = (b_1, \dots, b_d) : \begin{array}{l} m \times d \text{ matrix} \\ d \text{ is fixed.} \end{array}$$

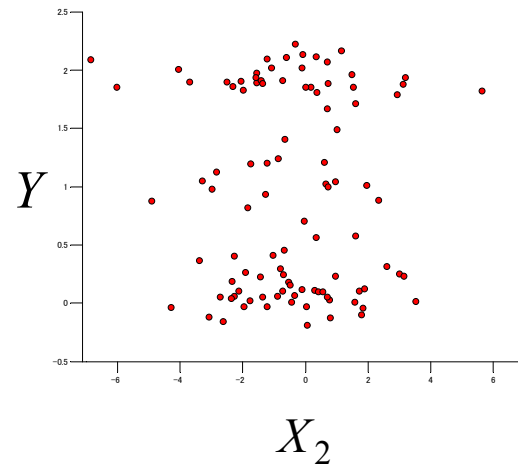
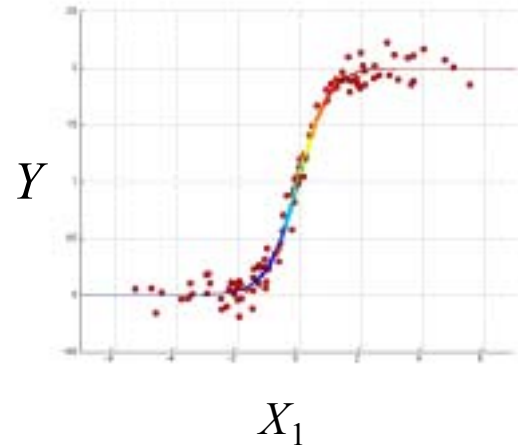
- Effective subspace summarizes all the information of X to explain Y .
Important features. Preprocessing (accuracy and efficiency).
- Semi-parametric problem: **model-free** for regressor
We put almost **no assumptions** on $p(Y|X)$ and $p(X)$.
→ Wide applicability.

– Example



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

Effective subspace = the direction of x_1



Conditional Independence

■ Dimension reduction and conditional independence

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

B gives effective subspace $\Leftrightarrow p_{Y|U,V}(y|u,v) = p_{Y|U}(y|u)$ for all y, u, v

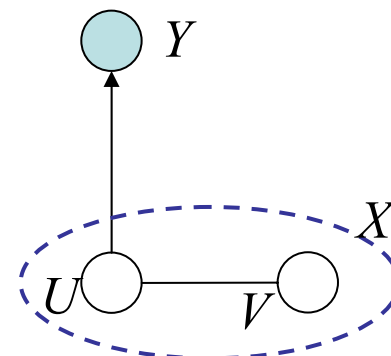
\Leftrightarrow **Conditional independence** $Y \perp V | U$

■ Characterization of cond. indep.

\Rightarrow Reproducing kernel Hilbert space (RKHS)

Gaussian kernel RKHS is used.

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



RKHS and Independence

■ RKHS characterization of Independence

Theorem (B&J2002) \mathcal{H}_X and \mathcal{H}_Y are **Gaussian** RKHS on \mathbf{R}^m and \mathbf{R}^n .

Random vectors $X \in \mathbf{R}^m$ and $Y \in \mathbf{R}^n$ are independent

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

c.f. characteristic functions.

■ Cross-covariance operator

Def. **Cross-covariance operator** $\Sigma_{YX} : \mathcal{H}_X \rightarrow \mathcal{H}_Y$ is defined by

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

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

Theorem For Gaussian RKHS

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

RKHS and Conditional Independence

■ Conditional covariance

X and Y are random vectors. $\mathcal{H}_X, \mathcal{H}_Y$: RKHS with kernel k_X, k_Y , resp.

Assumption: $\exists \Sigma_{XX}^{-1}$, $E_{Y|X}[g(Y)|X] \in \mathcal{H}_X$ for all $g \in \mathcal{H}_Y$.

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

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.

$\mathcal{H}_X, \mathcal{H}_U, \mathcal{H}_Y$: RKHS with **Gaussian kernel** k_X, k_U, k_Y , resp.

$E_{Y|X}[g(Y)|X] \in \mathcal{H}_X$ and $E_{Y|U}[g(Y)|U] \in \mathcal{H}_U$ for all $g \in \mathcal{H}_Y$.

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

■ Minimization of conditional covariance operator

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

– Evaluation

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

Kernel Dimension Reduction

Kernel Dimension Reduction (KDR)

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

Kernel generalized variance (KGV, B&J2002)

Kernel Dimension Reduction (KDR) = minimization of KGV

Minimization – gradient-based method.

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

Restrict the space to the linear hull of $\{k(\cdot, X^{(i)})\}_{i=1}^n$ and $\{k(\cdot, Y^{(i)})\}_{i=1}^n$

Replace $\Sigma_{YY|U}$ by $\hat{\Sigma}_{YY|U} \equiv \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}^{-1} \hat{\Sigma}_{UY}$ ($n \times n$ matrix)

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

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

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

Kernel Dimension Reduction

■ General method of dimension reduction for regression

- KDR needs no assumption on $p(Y|X)$, $p(X)$, $p(Y)$ and dimensionality of Y .
- Applicable for any type of data sets.

c.f. existing methods; sliced inverse regression (SIR), principal Hessian direction (pHd), canonical correlation analysis (CCA), partial least square (PLS), projection pursuit regression (PPR), etc.

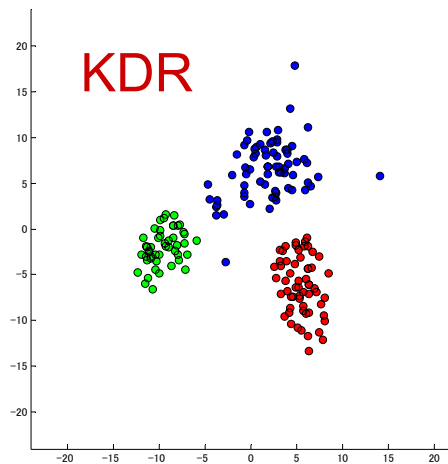
■ Computational cost

- Multiplication of $n \times n$ matrices is computationally hard. (n : # data)
 - Incomplete Cholesky decomposition
- Local minimum → annealing is used in gradient method.

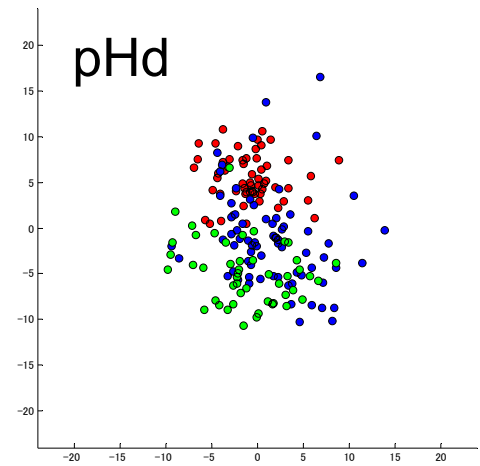
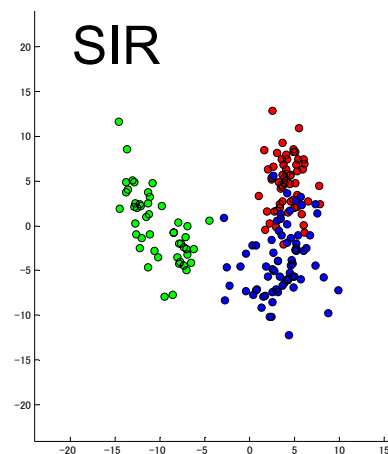
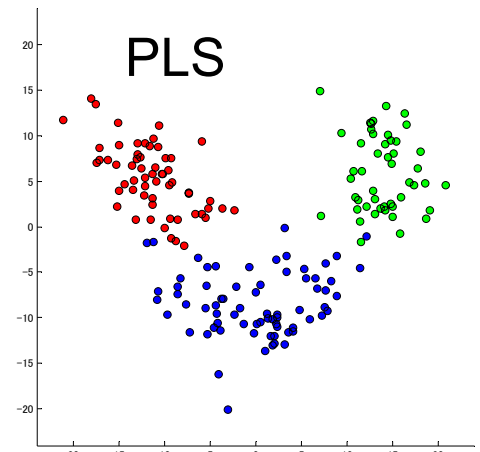
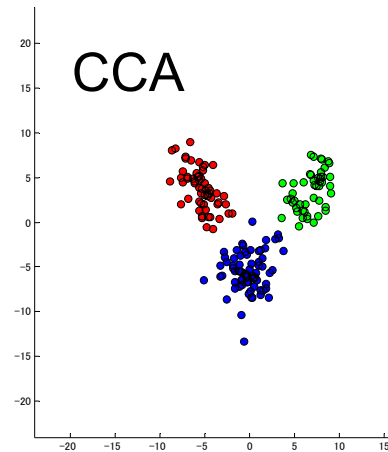
Experiments

■ Wine data

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



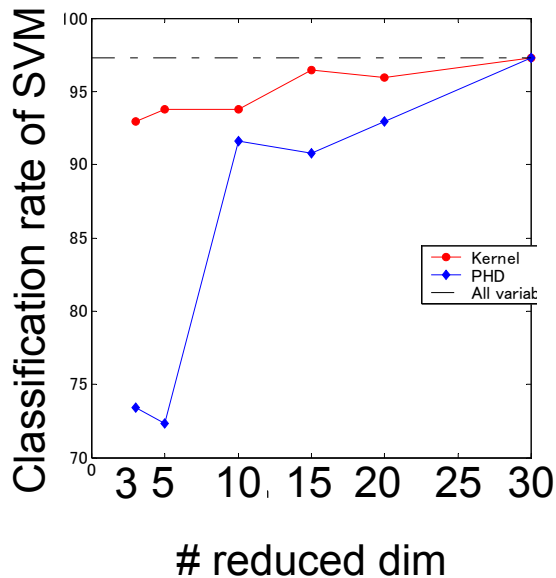
- Class1
- Class2
- Class3



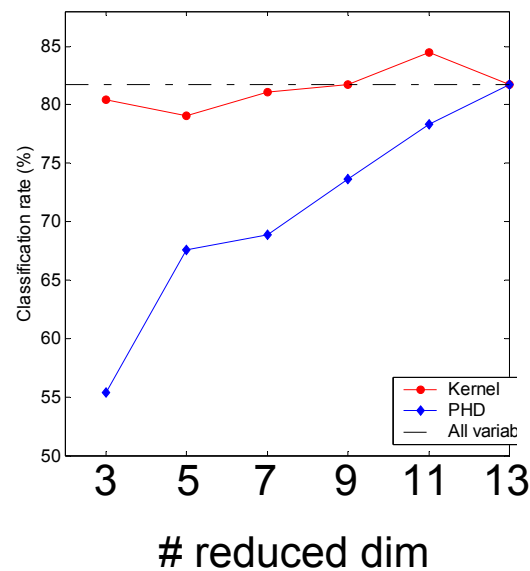
■ Classification after reducing dimensionality

- Purpose: How much information on Y is maintained in the estimated effective subspace?
- SVM is trained and its classification rate is evaluated for the training / test data projected onto the estimated effective subspace.

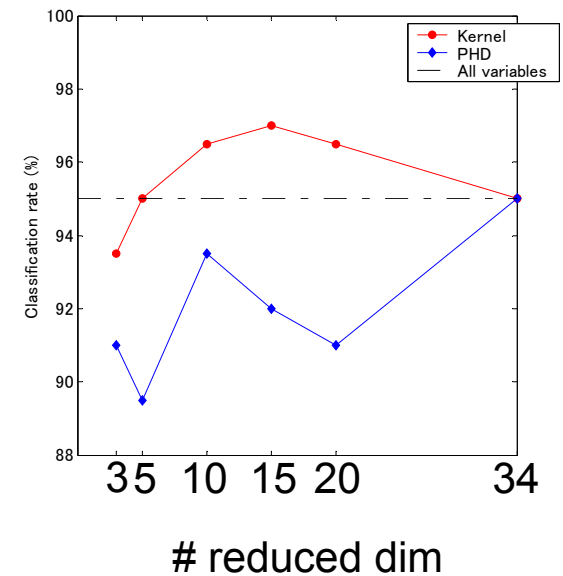
Breast-cancer-Wisconsin



Heart-disease



Ionosphere



Extension : Variable Selection

■ Variable selection by KGV

- Principle $Y \perp V | U \Leftrightarrow \Sigma_{YY|U} = \Sigma_{YY|X}$
- KGV gives an 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 subsets of $\{X_1, \dots, X_m\}$
 $U = (X_{i_1}, \dots, X_{i_d})$ where $1 \leq i_1 < \dots < i_d \leq m$

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

Extension : Variable Selection

■ Small data set

- *Boston Housing*: X :13 dim., Y = house price, 506 data.
- 4 variables are selected. ${}_{13}C_4 = 715$.
- Result: the selected variables are the same as the existing result by Breiman & Friedman (ACE, 1985)

■ Large data set

- Gene Selection : microarray data for ALL/AML classification
6817 dim. 38 data
- Golub et al. (Science 1999) show 50 effective genes using nearest neighbor analysis.
- Optimization: greedy algorithm + Genetic algorithm is used.
- Result: among 50 selected genes, 25 genes are the same as Golub's result.

Summary

■ Kernel Dimension Reduction (KDR)

- Dimension reduction for regression = conditional independence.
- Conditional covariance operators gives the criterion for KDR.
- KDR is the most general framework of dimension reduction for regression. It has wide applicability .
c.f. other methods have some restrictions.

■ Extension

- Variable selection

■ Future/ongoing studies

- Statistical analysis of the estimator: consistency etc.
- How to choose the number of dimensions.