

# Kernel Method: Data Analysis with Positive Definite Kernels

## 1. Introduction to Kernel Method

Kenji Fukumizu

The Institute of Statistical Mathematics.  
Graduate University of Advanced Studies /  
Tokyo Institute of Technology

Nov. 17-26, 2010

Intensive Course at Tokyo Institute of Technology



# Outline

## Basic idea of kernel methods

- Linear and nonlinear data analysis

- Essence of kernel methods

## Two examples of kernel methods

- Kernel PCA: Nonlinear extension of PCA

- Ridge regression and its kernelization

## Basic idea of kernel methods

Linear and nonlinear data analysis

Essence of kernel methods

## Two examples of kernel methods

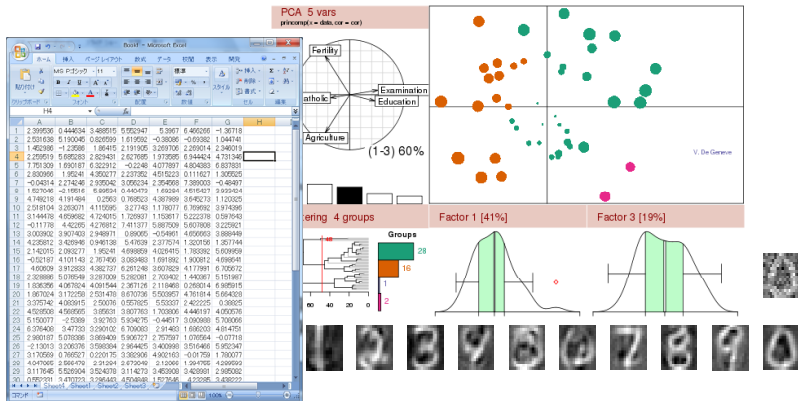
Kernel PCA: Nonlinear extension of PCA

Ridge regression and its kernelization

# What is Data Analysis?

**Analysis of data** is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

– Wikipedia



# Linear Data Analysis

- Typically, data is expressed by a 'table' of numbers →  
Matrix expression:

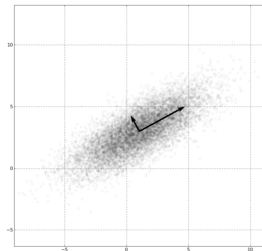
$$X = \begin{pmatrix} X_1^1 & X_1^2 & \cdots & X_1^m \\ X_2^1 & X_2^2 & \cdots & X_2^m \\ & & \vdots & \\ X_N^1 & X_N^2 & \cdots & X_N^m \end{pmatrix} \quad (m \text{ dimensional, } N \text{ data})$$

- Linear operations are used for data analysis. *e.g.*
  - Principal component analysis (PCA)
  - Canonical correlation analysis (CCA)
  - Linear regression analysis
  - Fisher discriminant analysis (FDA)
  - Logistic regression, etc.

- Example 1: Principal Component Analysis (PCA)

$X_1, \dots, X_N$  :  $m$ -dimensional data.

- Find  $d$ -directions to maximize the variance.
- Purpose: represent the structure of the data in a low dimensional space.



- The first principal direction:

$$\begin{aligned}
 u_1 &= \arg \max_{\|u\|=1} \frac{1}{N} \left\{ \sum_{i=1}^N u^T \left( X_i - \frac{1}{N} \sum_{j=1}^N X_j \right) \right\}^2 \\
 &= \arg \max_{\|u\|=1} u^T V u,
 \end{aligned}$$

where  $V$  is the variance-covariance matrix:

$$V = \frac{1}{N} \sum_{i=1}^N \left( X_i - \frac{1}{N} \sum_{j=1}^N X_j \right) \left( X_i - \frac{1}{N} \sum_{j=1}^N X_j \right)^T.$$

- General solution:
  - Eigenvectors  $u_1, \dots, u_m$  of  $V$  (in descending order of eigenvalues).
  - The  $p$ -th principal axis =  $u_p$ .
  - The  $p$ -th principal component of  $X_i$  =  $u_p^T X_i$

- Example 2: Linear classification

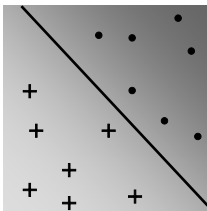
- Binary classification

$$X = \begin{pmatrix} X_1^1 & X_1^2 & \cdots & X_1^m \\ X_2^1 & X_2^2 & \cdots & X_2^m \\ \vdots & \vdots & \ddots & \vdots \\ X_N^1 & X_N^2 & \cdots & X_N^m \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \in \{\pm 1\}.$$

Input Output

- Linear classifier

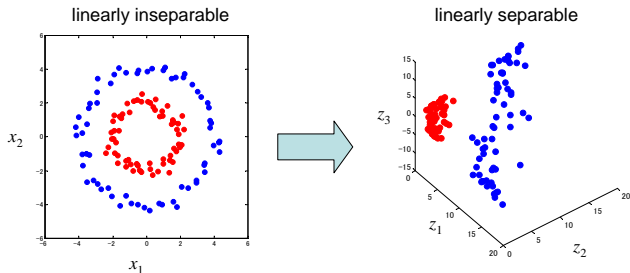
$$h(x) = a^T x + b$$





# Are linear methods enough?

- Example 1: classification



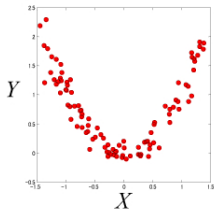
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

(Unclear? Watch <http://jp.youtube.com/watch?v=3liCbRZPrZA>)

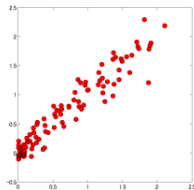
- Example 2: dependence of two data

## Correlation

$$\rho_{XY} = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{E[(X - E[X])^2]E[(Y - E[Y])^2]}}$$



$$\text{Corr}(X, Y) = 0.17$$



$$\text{Corr}(X^2, Y) = 0.96$$

- Transforming data to incorporate high-order moments seems attractive.

# Nonlinear Transform Helps!

- Analysis of data is a process of inspecting, cleaning, **transforming**, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. *– Wikipedia.*
- Kernel method = a systematic way of analyzing data by transforming them into a high-dimensional feature space to extract nonlinearity or higher-order moments of data.

## Basic idea of kernel methods

Linear and nonlinear data analysis

**Essence of kernel methods**

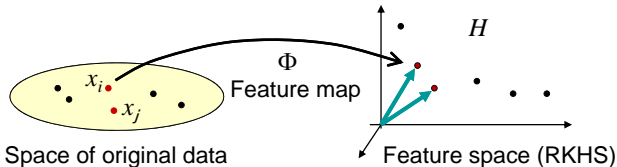
## Two examples of kernel methods

Kernel PCA: Nonlinear extension of PCA

Ridge regression and its kernelization

## Feature Space for Transforming Data

- Kernel methodology = a systematic way of analyzing data by transforming them into a high-dimensional **feature space**.



Apply linear methods on the feature space.

- What space is suitable for a feature space?
  - It should incorporate various nonlinear information of the original data.
  - The inner product of the feature space is essential for data analysis (seen in the next subsection).

# Computational Problem

- For example, how about this?

$$(X, Y, Z) \mapsto (X, Y, Z, X^2, Y^2, Z^2, XY, YZ, ZX, \dots).$$

- But, for high-dimensional data, the above expansion makes the feature space very huge!

*e.g. If  $X$  is 100 dimensional and the moments up to the 3rd order are used, the dimensionality of feature space is*

$${}_{100}C_1 + {}_{100}C_2 + {}_{100}C_3 = 166750.$$

- This causes a serious computational problem in working on the inner product of the feature space.  
We need a cleverer way of computing it.  $\Rightarrow$  [Kernel method](#).

# Inner Product by Positive Definite Kernel

- A positive definite kernel gives efficient computation of the inner product:

With special choice of a feature space  $H$  and feature map  $\Phi$ , we have a function  $k(x, y)$  such that

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j), \quad \text{positive definite kernel}$$

where

$$\Phi : \mathcal{X} \rightarrow \mathcal{H}, \quad x \mapsto \Phi(x) \in \mathcal{H}.$$

- Many linear methods use only the inner product without necessity of the explicit form of the vector  $\Phi(X)$ .

## Basic idea of kernel methods

Linear and nonlinear data analysis

Essence of kernel methods

## Two examples of kernel methods

Kernel PCA: Nonlinear extension of PCA

Ridge regression and its kernelization



# Review of PCA I

$X_1, \dots, X_N$  :  $m$ -dimensional data.

The first principal direction:

$$\begin{aligned} u_1 &= \arg \max_{\|u\|=1} \text{Var}[u^T X] \\ &= \arg \max_{\|u\|=1} \frac{1}{N} \left\{ \sum_{i=1}^N u^T \left( X_i - \frac{1}{N} \sum_{j=1}^N X_j \right) \right\}^2. \end{aligned}$$

**Observation:** PCA can be done if we can

- compute the inner product between  $u$  and the data,
- solve the optimum  $u$ .

# Kernel PCA I

$X_1, \dots, X_N$  :  $m$ -dimensional data.

Transform the data by a feature map  $\Phi$  into a feature space  $\mathcal{H}$ :

$$X_1, \dots, X_N \mapsto \Phi(X_1), \dots, \Phi(X_N)$$

**Assume** that the feature space has the **inner product**  $\langle \cdot, \cdot \rangle$ .

Apply PCA on the feature space:

- Maximize the variance of the projections onto the direction  $f$ .

$$\max_{\|f\|=1} \text{Var}[\langle f, \Phi(X) \rangle] = \max_{\|f\|=1} \frac{1}{N} \sum_{i=1}^N \left( \left\langle f, \Phi(X_i) - \frac{1}{N} \sum_{j=1}^N \Phi(X_j) \right\rangle \right)^2$$

## Kernel PCA II

- Note: it suffices to use

$$f = \sum_{i=1}^N a_i \tilde{\Phi}(X_i),$$

where

$$\tilde{\Phi}(X_i) = \Phi(X_i) - \frac{1}{N} \sum_{j=1}^N \Phi(X_j).$$

The direction orthogonal to  $\text{Span}\{\tilde{\Phi}(X_1), \dots, \tilde{\Phi}(X_N)\}$  does not contribute.<sup>1</sup>

---

<sup>1</sup>Decompose  $f$  as  $f = f_0 + f_{\perp}$ , where  $f_0 \in \text{Span}\{\tilde{\Phi}(X_i)\}_{i=1}^N$  and  $f_{\perp}$  in its orthogonal complement. The objective function is maximized when  $f_{\perp} = 0$ . [Exercise: confirm the details.]

## Kernel PCA III

- Insert  $f = \sum_{i=1}^N c_i \tilde{\Phi}(X_i)$ .
  - Variance:  $\frac{1}{N} \sum_{i=1}^N \langle f, \tilde{\Phi}(X_i) \rangle^2 = \frac{1}{N} a^T \tilde{K}^2 a$ .
  - constraint:  $\|f\|^2 = a^T \tilde{K} a = 1$ .
- Kernel PCA problem:

$$\max a^T \tilde{K}^2 a \quad \text{subject to} \quad a^T \tilde{K} a = 1,$$

where  $\tilde{K}$  is  $N \times N$  matrix with  $\tilde{K}_{ij} = \langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle$ .

Kernel PCA can be solved by the above **generalized eigenproblem**.

## Kernel PCA IV

- Eigendecomposition:

$$\tilde{K} = \sum_{i=1}^N \lambda_i u^i u^{iT}, \quad (\lambda_1 \geq \dots \lambda_N \geq 0).$$

- Solution of kernel PCA:
  - The first principal direction:

$$f_1 = \sum_{i=1}^N a_i \tilde{\Phi}(X_i), \quad a = \frac{1}{\sqrt{\lambda_1}} u^1,$$

- The first principal component of the data  $X_i$ :

$$\langle \tilde{\Phi}(X_i), f_1 \rangle = \sqrt{\lambda_1} u_i^1,$$

- 2nd, 3rd, ... principal components are similar.

Exercise: Check the following two relations for

$$f = \sum_{i=1}^N a_i \tilde{\Phi}(X_i).$$

- Variance:  $\frac{1}{N} \sum_{i=1}^N \langle f, \tilde{\Phi}(X_i) \rangle^2 = \frac{1}{N} a^T \tilde{K}^2 a.$
- Squared norm:  $\|f\|^2 = a^T \tilde{K} a.$

## From PCA to Kernel PCA

- The optimum direction is obtained in the form

$$f = \sum_{i=1}^N a_i \tilde{\Phi}(X_i),$$

*i.e.*, in the linear hull of the (centered) data.

- PCA in the feature space is expressed by  $\langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle$  or<sup>2</sup>

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j).$$

---

<sup>2</sup>Exercise: Check the following relation

$$\tilde{K}_{ij} = k(X_i, X_j) - \frac{1}{N} \sum_{b=1}^N k(X_i, X_b) - \frac{1}{N} \sum_{a=1}^N k(X_a, X_j) + \frac{1}{N^2} \sum_{a=1}^N k(X_a, X_b).$$

## Basic idea of kernel methods

Linear and nonlinear data analysis

Essence of kernel methods

## Two examples of kernel methods

Kernel PCA: Nonlinear extension of PCA

Ridge regression and its kernelization

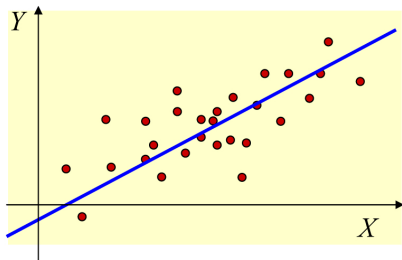


# Review: Linear Regression I

## Linear regression

- Data:  $(X_1, Y_1), \dots, (X_N, Y_N)$ : data
  - $X_i$ : explanatory variable, covariate ( $m$ -dimensional)
  - $Y_i$ : response variable, (1 dimensional)
- Regression model: find the best linear relation

$$Y_i = a^T X_i + \varepsilon_i$$



## Review: Linear Regression II

- Least square method:  $\min_a \sum_{i=1}^N \|Y_i - a^T X_i\|^2$ .
- Matrix expression

$$X = \begin{pmatrix} X_1^1 & X_1^2 & \cdots & X_1^m \\ X_2^1 & X_2^2 & \cdots & X_2^m \\ & & \vdots & \\ X_N^1 & X_N^2 & \cdots & X_N^m \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}.$$

- Solution:

$$\hat{a} = (X^T X)^{-1} X^T Y$$

$$\hat{y} = \hat{a}^T x = Y^T X (X^T X)^{-1} x.$$

**Observation:** Linear regression can be done if we can compute the inner product  $X^T X$ ,  $\hat{a}^T x$  and so on.

# Ridge Regression

Ridge regression:

- Find a linear relation by

$$\min_a \sum_{i=1}^N \|Y_i - a^T X_i\|^2 + \lambda \|a\|^2.$$

$\lambda$ : regularization coefficient.

- Solution

$$\hat{a} = (X^T X + \lambda I_N)^{-1} X^T Y$$

For a general  $x$ ,

$$\hat{y}(x) = \hat{a}^T x = Y^T X (X^T X + \lambda I_N)^{-1} x.$$

- Ridge regression is useful when  $(X^T X)^{-1}$  does not exist, or inversion is numerically unstable.

# Kernelization of Ridge Regression I

$(X_1, Y_1) \dots, (X_N, Y_N)$  ( $Y_i$ : 1-dimensional)

Transform  $X_i$  by a feature map  $\Phi$  into a feature space  $\mathcal{H}$ :

$$X_1, \dots, X_N \mapsto \Phi(X_1), \dots, \Phi(X_N)$$

**Assume** that the feature space has the **inner product**  $\langle \cdot, \cdot \rangle$ .

Apply ridge regression to the transformed data:

- Find the vector  $f$  such that

$$\min_{f \in \mathcal{H}} \sum_{i=1}^N |Y_i - \langle f, \Phi(X_i) \rangle_{\mathcal{H}}|^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

## Kernelization of Ridge Regression II

- Similarly to kernel PCA, we can assume <sup>3</sup>

$$f = \sum_{j=1}^N c_j \Phi(X_j).$$

- The objective function is

$$\min_c \sum_{i=1}^N \left| Y_i - \left\langle \sum_{j=1}^N c_j \Phi(X_j), \Phi(X_i) \right\rangle_{\mathcal{H}} \right|^2 + \lambda \left\| \sum_{j=1}^N c_j \Phi(X_j) \right\|_{\mathcal{H}}^2.$$

---

<sup>3</sup>[Exercise: confirm this.]

## Kernelization of Ridge Regression III

- Solution:

$$\hat{c} = (K + \lambda I_N)^{-1}Y,$$

where

$$K_{ij} = \langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}} = k(X_i, X_j).$$

For a general  $x$ ,

$$\begin{aligned} \hat{y}(x) = \langle \hat{f}, \Phi(x) \rangle_{\mathcal{H}} &= \langle \sum_j \hat{c}_j \Phi(X_j), \Phi(x) \rangle_{\mathcal{H}} \\ &= Y^T (K + \lambda I_N)^{-1} \mathbf{k}(x), \end{aligned}$$

where

$$\mathbf{k}(x) = \begin{pmatrix} \langle \Phi(X_1), \Phi(x) \rangle \\ \vdots \\ \langle \Phi(X_N), \Phi(x) \rangle \end{pmatrix} = \begin{pmatrix} k(X_1, x) \\ \vdots \\ k(X_N, x) \end{pmatrix}.$$

# Kernelization of Ridge Regression IV

## Outline of Proof.

Matrix expression derives

$$\begin{aligned} & \sum_{i=1}^N \left| Y_i - \left\langle \sum_{j=1}^N c_j \Phi(X_j), \Phi(X_i) \right\rangle_{\mathcal{H}} \right|^2 + \lambda \left\| \sum_{j=1}^N c_j \Phi(X_j) \right\|_{\mathcal{H}}^2 \\ &= (Y - Kc)^T (Y - Kc) + \lambda c^T Kc \\ &= c^T (K^2 + \lambda K) c - 2Y^T Kc + Y^T Y. \end{aligned}$$

Thus, the the objective function is a quadratic form of  $c$ . The solution is given by

$$\hat{c} = (K + \lambda I_N)^{-1} Y.$$

Inserting this to  $\hat{y}(x) = \langle \sum_j \hat{c}_j \Phi(X_j), \Phi(x) \rangle_{\mathcal{H}}$ , we have the claim.  $\square$

# From Ridge Regression to its Kernelization

## Observations:

- The optimum coefficients have the form

$$f = \sum_{i=1}^N c_i \Phi(X_i),$$

*i.e.*, a linear combination of the data.

The orthogonal directions do not contribute to the objective function.

- The objective function of kernel ridge regression can be expressed by the inner products

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j) \quad \text{and} \quad \langle \Phi(X_i), \Phi(x) \rangle = k(X_i, x).$$



# Principles of Kernel Methods

- Observations common in two examples:
  - A feature map transforms data into a feature space  $\mathcal{H}$  with inner product  $\langle \cdot, \cdot \rangle$ .

$$X_1, \dots, X_N \mapsto \Phi(X_1), \dots, \Phi(X_N) \in \mathcal{H}.$$

- Typically, the optimum solution (vector in  $\mathcal{H}$ ) has the form

$$f = \sum_{i=1}^N c_i \Phi(X_i).$$

- The problem is expressed by the inner product  $\langle \Phi(X_i), \Phi(X_i) \rangle$ .
- If the inner product  $\langle \Phi(X_i), \Phi(X_i) \rangle$  is computable, various linear methods can be done on a feature space.
- How can we define such a feature space in general?  
 $\Rightarrow$  **Positive definite kernel!**