Kernel Methods for Statistical Learning

Kenji Fukumizu

The Institute of Statistical Mathematics /
Graduate University for Advanced Studies

September 6-7, 2012
Machine Learning Summer School 2012, Kyoto

Version 2012.09.04

The latest version of slides is downloadable at
http://www.ism.ac.jp/~fukumizu/MLSS2012/

Lecture Plan

I. Introduction to kernel methods

II. Various kernel methods
kernel PCA, kernel CCA, kernel ridge regression, etc

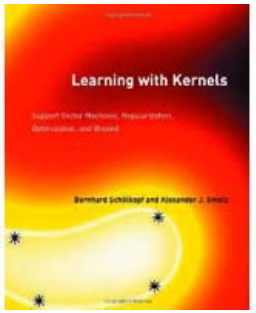
III. Support vector machine
A brief introduction to SVM

IV. Theoretical backgrounds of kernel methods
Mathematical aspects of positive definite kernels

V. Nonparametric inference with positive definite kernels
Recent advances of kernel methods

General references (More detailed lists are given at the end of each section)

- Schölkopf, B. and A. Smola. *Learning with Kernels*. MIT Press. 2002.



- Lecture slides (more detailed than this course)

  This page contains Japanese information, but the slides are written in English.

  Slides: 1, 2, 3, 4, 5, 6, 7, 8

- For Japanese only (Sorry!):
  - 福水「カーネル法入門 ー正定値カーネルによるデータ解析」 朝倉書店（2010）

  - 赤穂「カーネル多変量解析 ー非線形データ解析の新しい展開」 岩波書店（2008）

I. Introduction to Kernel Methods

Kenji Fukumizu

The Institute of Statistical Mathematics /
Graduate University for Advanced Studies

September 6-7
Machine Learning Summer School 2012, Kyoto

Outline

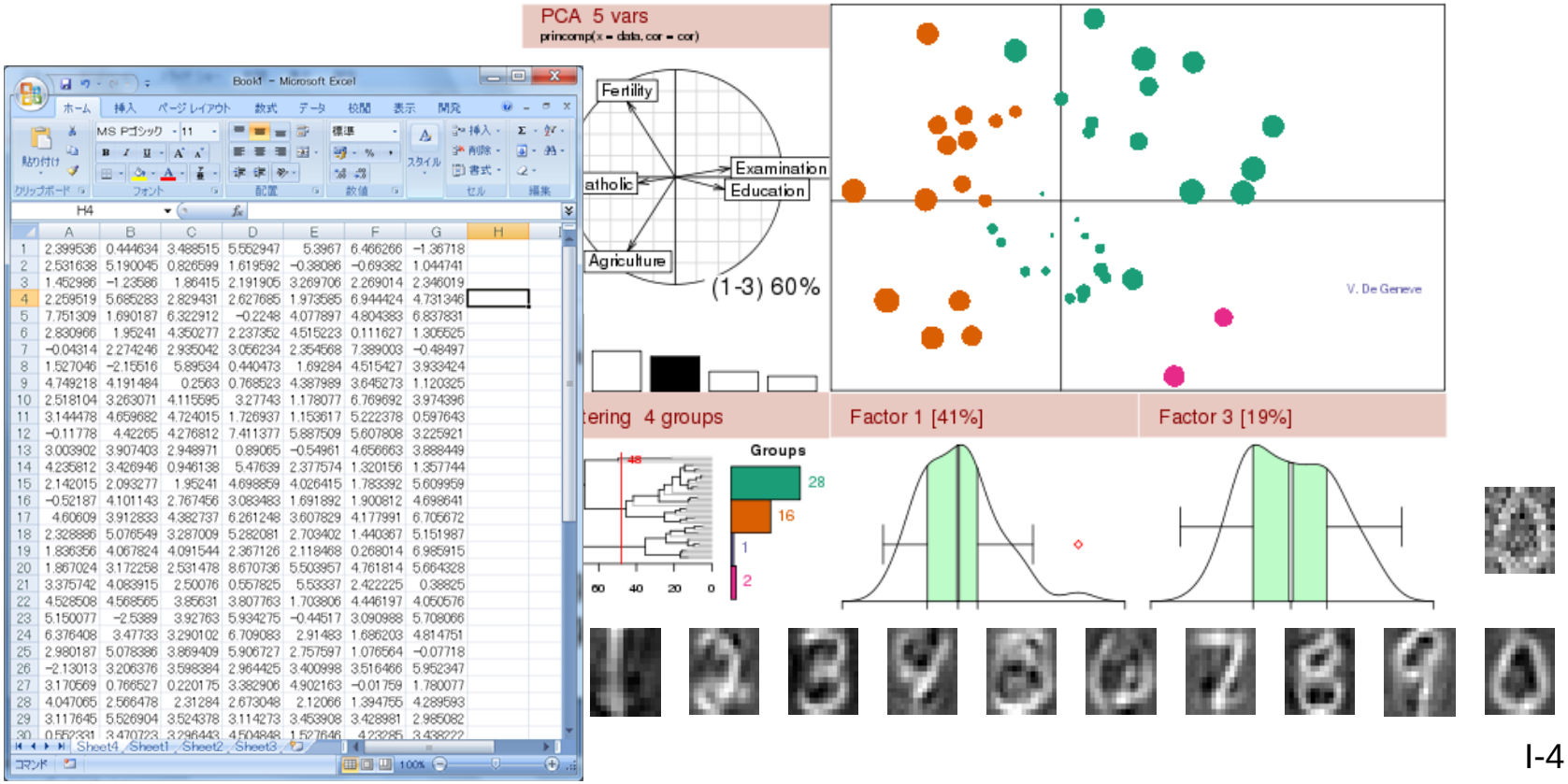
1. Linear and nonlinear data analysis
2. Principles of kernel methods
3. Positive definite kernels and feature spaces

Linear and Nonlinear Data Analysis

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

– 'Table' of numbers → Matrix expression

$$\mathbf{X} = \begin{pmatrix} X_1^{(1)} & \cdots & X_m^{(1)} \\ X_1^{(2)} & \cdots & X_m^{(2)} \\ \vdots & & \vdots \\ X_1^{(N)} & \cdots & X_m^{(N)} \end{pmatrix}$$

m dimensional, N data

– Linear algebra is used for methods of analysis
  - Correlation,
  - Linear regression analysis,
  - Principal component analysis,
  - Canonical correlation analysis, etc.

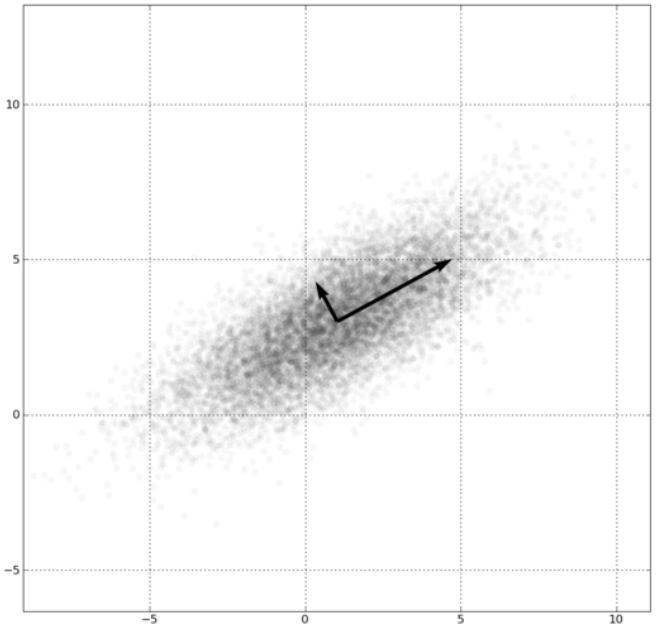
Example 1: Principal component analysis (PCA)

PCA: project data onto the subspace with largest variance.

1st direction = $\operatorname{argmax}_{\|a\|=1} \mathrm{Var}[a^T X]$

$$\mathrm{Var}[a^T X] = \frac{1}{N}\sum_{i=1}^{N}\left\{a^T\left(X^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X^{(j)}\right)\right\}^2$$

$$= a^T V_{XX} a.$$



where

$$V_{XX} = \frac{1}{N}\sum_{i=1}^{N}\left(X^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X^{(j)}\right)\left(X^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X^{(j)}\right)^T$$

(Empirical) covariance matrix of X

– 1st principal direction

$$= \text{argmax}_{\|a\|=1} a^T V_{XX} a$$

$= u_1$ unit eigenvector w.r.t. the largest eigenvalue of V_{XX}

– p-th principal direction =

unit eigenvector w.r.t. the p-th largest eigenvalue of V_{XX}

PCA ⇨ Eigenproblem of covariance matrix V_{XX}

Example 2: Linear classification

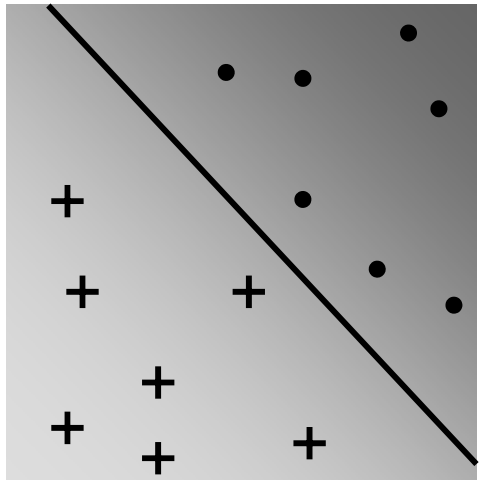
– Binary classification

Input data

$$\mathbf{X} = \begin{pmatrix} X_1^{(1)} & \cdots & X_m^{(1)} \\ X_1^{(2)} & \cdots & X_m^{(2)} \\ \vdots & & \vdots \\ X_1^{(N)} & \cdots & X_m^{(N)} \end{pmatrix}$$

Class label

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(N)} \end{pmatrix} \in \{\pm 1\}^N$$



Find a linear classifier

$$h(x) = \mathrm{sgn}(a^T x + b)$$

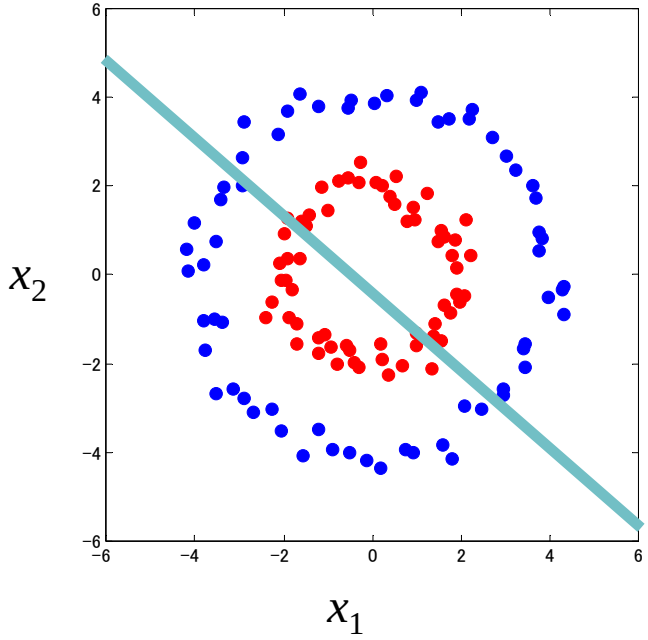
so that

$$h(X^{(i)}) = Y^{(i)} \qquad \text{for all (or most) } i.$$

– Example: Fisher's linear discriminant analyzer, Linear SVM, etc.

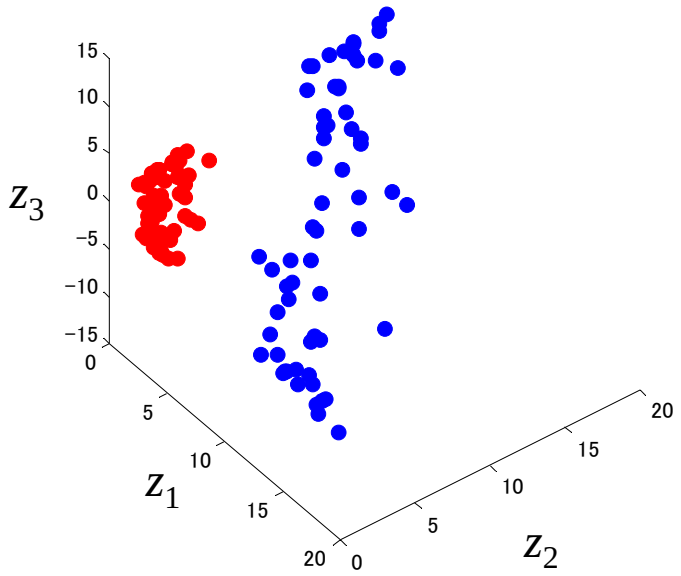
Are linear methods enough?

linearly inseparable



transform

linearly separable



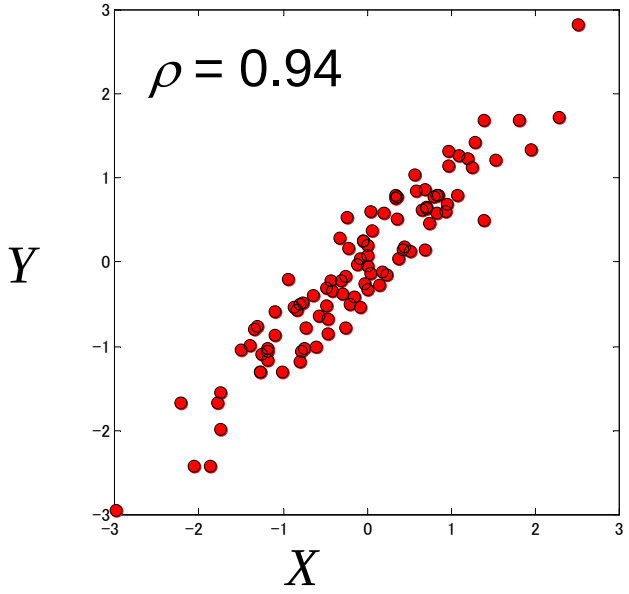
$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

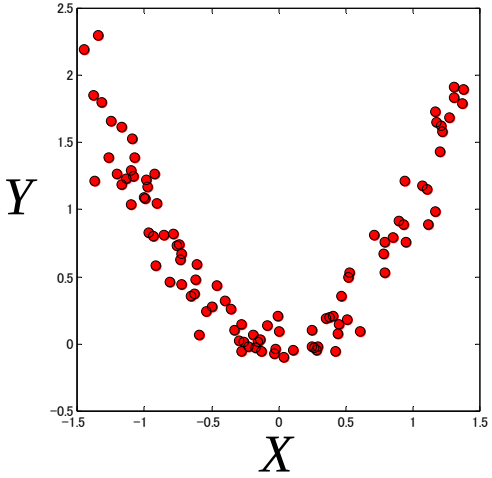
Watch the following movie!

http://jp.youtube.com/watch?v=3liCbRZPrZA

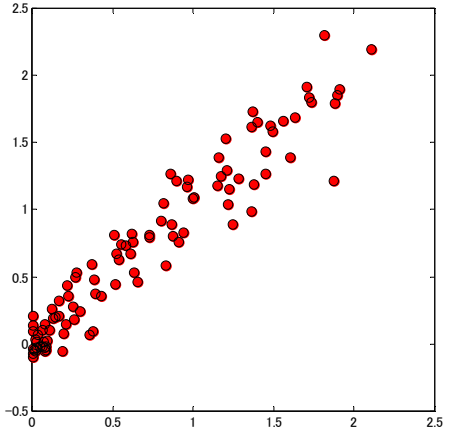
Another example: correlation

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





$\rho(X, Y) = 0.17$



$\rho(X^2, Y) = 0.96$

(X, Y)

transform

(X^2, Y)

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.

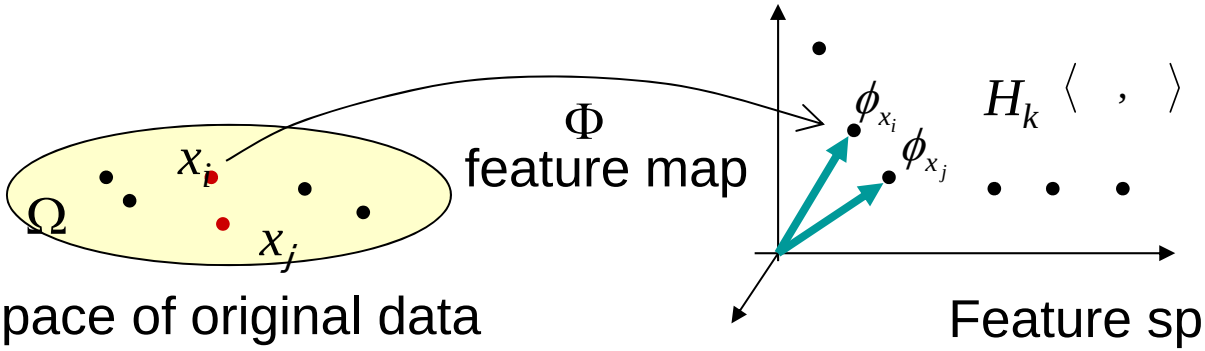
– *Wikipedia.*

Kernel method = a systematic way of transforming data into a high-dimensional feature space to extract nonlinearity or higher-order moments of data.

Principles of kernel methods

Kernel method: Big picture

– Idea of kernel method



Do linear analysis in the feature space!

e.g. SVM

– What kind of space is appropriate as a feature space?
  - Should incorporate various nonlinear information of the original data.
  - The inner product should be computable. It is essential for many linear methods.

Computational issue

– For example, how about using power series expansion?

$$(X, Y, Z) \rightarrow (X, Y, Z, X^2, Y^2, Z^2, XY, YZ, ZX, \ldots)$$

– But, many recent data are high-dimensional.
e.g. microarray, images, etc…

The above expansion is intractable!

e.g. Up to 2nd moments, 10,000 dimension:

Dim of feature space: ${}_{10000}C_1 + {}_{10000}C_2 = 50{,}005{,}000$ (!)

– Need a cleverer way → Kernel method.

Feature space by positive definite kernel

– Feature map: from original space to feature space

$$\Phi: \quad \Omega \quad \to \quad H$$

$$X_1,\ldots,X_n \quad \mapsto \quad \Phi(X_1),\ldots,\Phi(X_n)$$

– With special choice of feature space, we have a function (positive definite kernel) $k(x, y)$ such that

$$\left\langle \Phi(X_i), \Phi(X_j) \right\rangle = k(X_i, X_j)$$ kernel trick.

– Many linear methods use only the inner products of data, and do not need the explicit form of the vector $\Phi(X)$. (*e.g.* PCA)

Positive definite kernel

Definition. Ω: set. $k: \Omega \times \Omega \to \mathbf{R}$ is a positive definite kernel if

1) (symmetry) $k(x, y) = k(y, x)$

2) (positivity) for arbitrary $x_1, \ldots, x_n \in \Omega$

$$\begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$ is positive semidefinite,

(Gram matrix)

i.e., $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ for any $c_i \in \mathbf{R}$

Examples: positive definite kernels on $\mathbf{R}^m$ (proof is give in Section IV)

- Euclidean inner product

$$k(x, y) = x^T y$$

- Gaussian RBF kernel

$$k_G(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right) \qquad (\sigma > 0)$$

- Laplacian kernel

$$k_L(x, y) = \exp\left(-\alpha \sum_{i=1}^{m} |\, x_i - y_i \,|\right) \qquad (\alpha > 0)$$

- Polynomial kernel

$$k_P(x, y) = (c + x^T y)^d \qquad (c \geq 0, d \in \mathbf{N})$$

Proposition 1.1

Let H be a vector space with inner product $\langle \cdot, \cdot \rangle$ and $\Phi: \Omega \to H$ be a map (feature map). If $k: \Omega \times \Omega \to \mathbf{R}$ is defined by

$$\left\langle \Phi(x), \Phi(y) \right\rangle = k(x, y),$$

then $k(x,y)$ is necessarily positive definite.

– Positive definiteness is necessary.

– Proof)

– Positive definite kernel is sufficient.

Theorem 1.2 (Moore-Aronszajn)

For a positive definite kernel k on Ω, there is a Hilbert space H_k (reproducing kernel Hilbert space, RKHS) that consists of functions on Ω such that

1) $k(\cdot, x) \in H_k$ for any x

2) span$\{ k(\cdot, x) \mid x \in \Omega \}$ is dense in H_k

3) (reproducing property)

$$\langle f, k(\cdot, x) \rangle = f(x) \quad \text{for any } f \in H_k,\ x \in \Omega$$

*Hilbert space: vector space with inner product such that the topology is complete.

Feature map by positive definite kernel

– Feature space = RKHS.
Feature map:

$$\Phi: \Omega \to H, \qquad x \mapsto k(\cdot, x)$$

$$X_1, \ldots, X_n \mapsto \; k(\cdot, X_1), \ldots, k(\cdot, X_n)$$

– Kernel trick: by reproducing property

$$\langle \Phi(x), \Phi(y) \rangle = \langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y)$$

– Prepare only positive definite kernel:

We do not need an explicit form of feature vector or feature space.

All we need for kernel methods are kernel values $k(X_i, X_j)$.

II. Various Kernel Methods

Kenji Fukumizu

The Institute of Statistical Mathematics /
Graduate University for Advanced Studies

September 6-7
Machine Learning Summer School 2012, Kyoto

Outline

1. Kernel PCA
2. Kernel CCA
3. Kernel ridge regression
4. Some topics on kernel methods

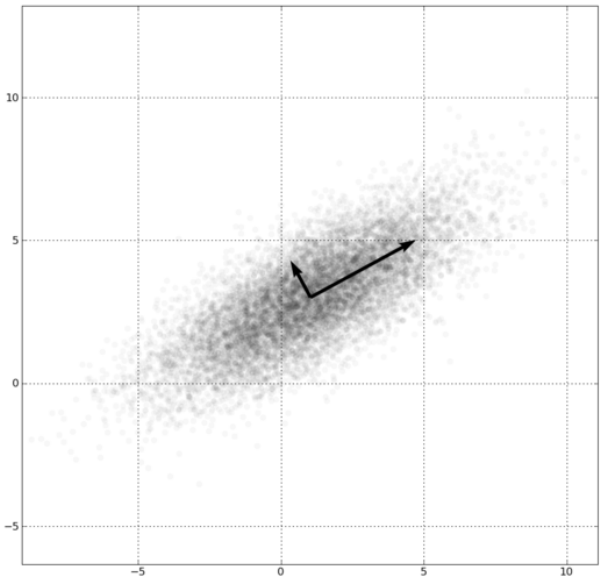
Kernel Principal Component Analysis

Principal Component Analysis

PCA (review)

- Linear method for dimension reduction of data
- Project the data in the directions of large variance.

1st principal axis = $\operatorname{argmax}_{\|a\|=1} \mathrm{Var}[a^T X]$



$$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2$$

$$= a^T V_{XX} a.$$

where

$$V_{XX} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^T$$

From PCA to Kernel PCA

– Kernel PCA: nonlinear dimension reduction of data (Schölkopf et al. 1998).

– Do PCA in feature space

$$\max_{\|a\|=1}: \quad \mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{s=1}^{n} X_s\right)\right\}^2$$

$$\max_{\|f\|_H=1}: \quad \mathrm{Var}[\langle f, \Phi(X)\rangle] = \frac{1}{n}\sum_{i=1}^{n}\left\{\left\langle f, \Phi(X_i) - \frac{1}{n}\sum_{s=1}^{n}\Phi(X_s)\right\rangle\right\}^2$$

It is sufficient to assume

$$f = \sum_{i=1}^{n} c_i \left(\Phi(X_i) - \frac{1}{n} \sum_{s=1}^{n} \Phi(X_s) \right)$$

Orthogonal directions to the data can be neglected, since for $f = \sum_{i=1}^{n} c_i (\Phi(X_i) - \frac{1}{n} \sum_{s=1}^{n} \Phi(X_s)) + f_\perp$, where $f_\perp$ is orthogonal to the $\mathrm{span}\{\Phi(X_i) - \frac{1}{n} \sum_{s=1}^{n} \Phi(X_s)\}_{i=1}^{n}$, the objective function of kernel PCA does not depend on $f_\perp$.

Then,

$$\begin{cases} \mathrm{Var}\left[\langle f, \Phi(X) \rangle\right] = c^T \tilde{K}_X^2 c \\ \| f \|_H^2 = c^T \tilde{K}_X c \end{cases}$$

[Exercise]

where $\tilde{K}_{X,ij} := \left\langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \right\rangle$ (centered Gram matrix)

with $\tilde{\Phi}(X_i) := \Phi(X_i) - \frac{1}{n} \sum_{s=1}^{n} \Phi(X_s)$

(centered feature vector)

Objective function of kernel PCA

$$\max \quad c^T \tilde{K}_X^2 c \qquad \text{subject to} \qquad c^T \tilde{K}_X c = 1$$

The centered Gram matrix $\tilde{K}_X$ is expressed with Gram matrix $K_X = \left(k(X_i, X_j)\right)_{ij}$ as

$$\tilde{K}_X = \left(I_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T\right)^T K \left(I_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T\right)$$

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbf{R}^n$$

$I_n =$ Unit matrix

$$\left(\tilde{K}_X\right)_{ij} = k(X_i, X_j) - \frac{1}{n}\sum_{s=1}^{n} k(X_i, X_s) - \frac{1}{n}\sum_{t=1}^{n} k(X_t, X_j) + \frac{1}{n^2}\sum_{t,s=1}^{N} k(X_t, X_s)$$

[Exercise]

– Kernel PCA can be solved by eigen-decomposition.

– Kernel PCA algorithm

- Compute centered Gram matrix $\tilde{K}_X$
- Eigendecompsoition of $\tilde{K}_X$

$$\tilde{K}_X = \sum_{i=1}^{N} \lambda_i u_i u_i^T$$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ eigenvalues

$u_1, u_2, \ldots, u_N$ unit eigenvectors

- p-th principal component of $X_i = \sqrt{\lambda_p} u_p^T X_i$

Derivation of kernel method in general

- Consider feature vectors with kernels.
- Linear method is applied to the feature vectors. (Kernelization)
- Typically, only the inner products

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

$$\left\langle f, \Phi(X_i) \right\rangle$$

are used to express the objective function of the method.

- The solution is given in the form $f = \sum_{i=1}^{n} c_i \Phi(X_i),$

(representer theorem, see Sec.IV), and thus everything is written by Gram matrices.

These are common in derivation of any kernel methods.

Example of Kernel PCA

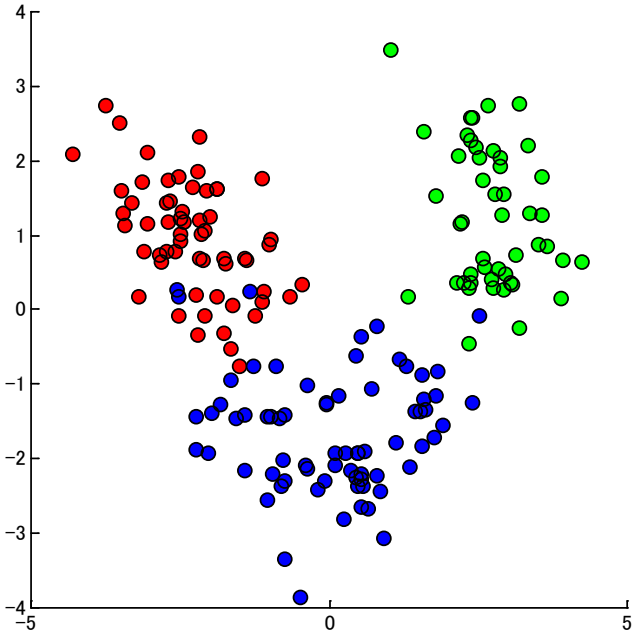
- **Wine data** (from UCI repository)

  13 dim. chemical measurements of for three types of wine. 178 data.

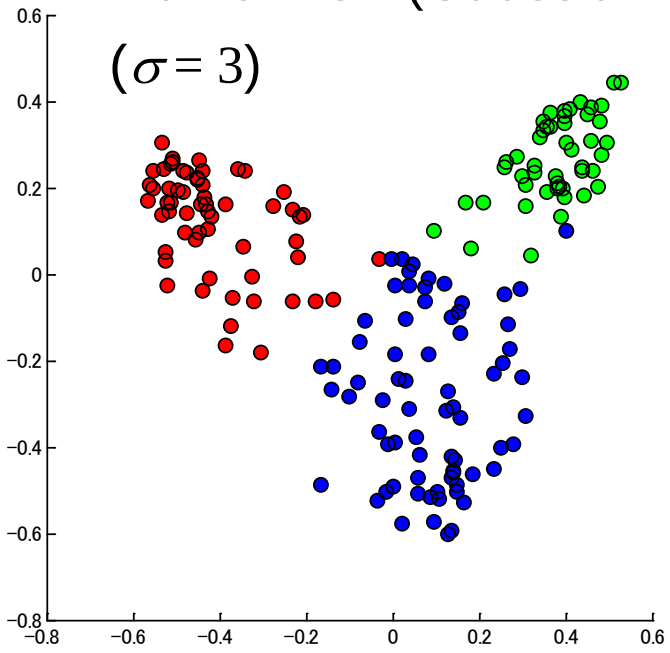
  Class labels are NOT used in PCA, but shown in the figures.

  First two principal components:

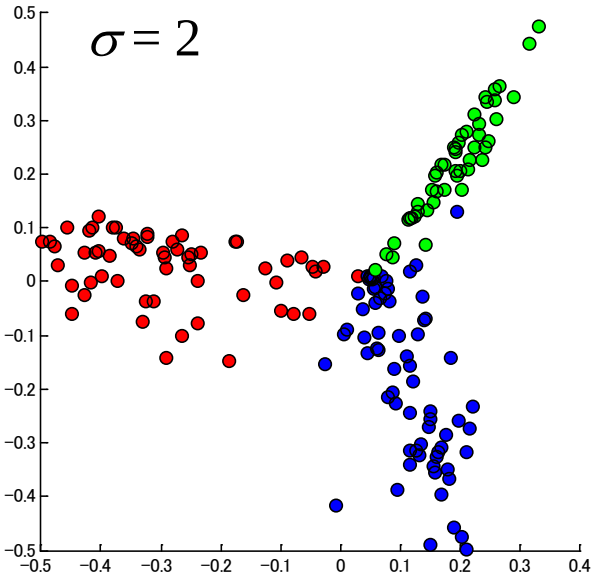
Linear PCA

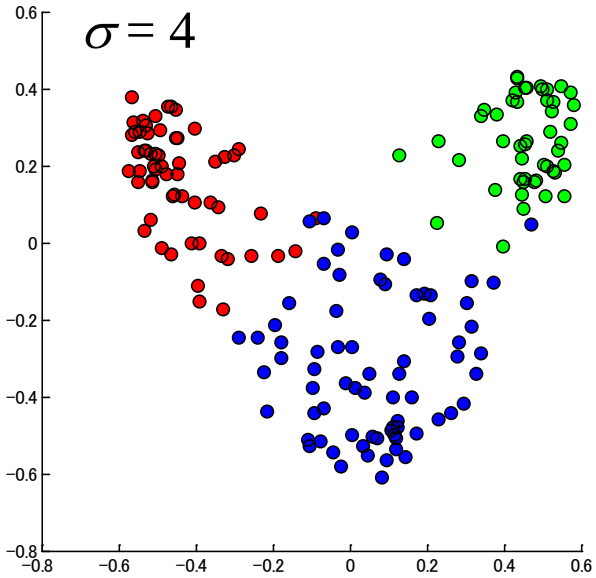


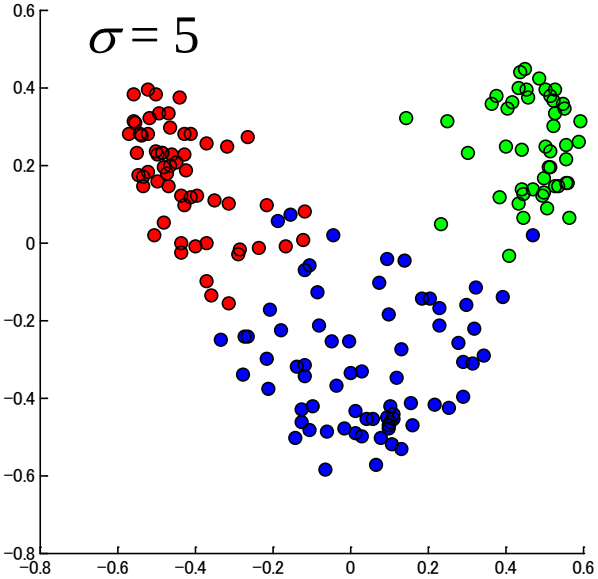
Kernel PCA (Gaussian kernel)



Kernel PCA (Gaussian)







$$k_G(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right)$$

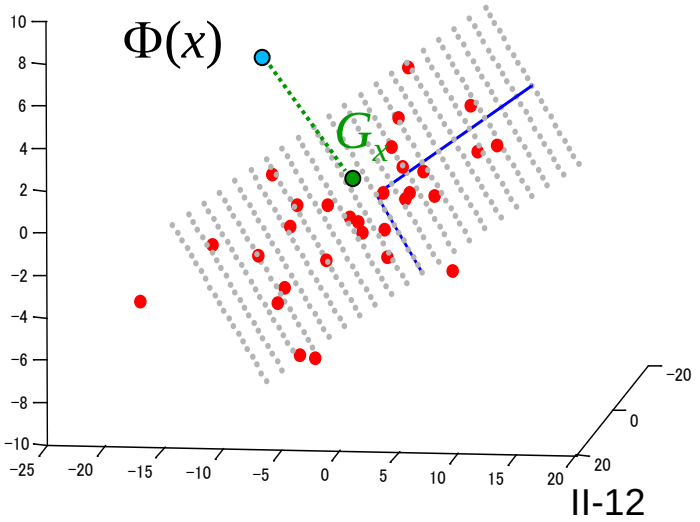
Noise Reduction with Kernel PCA

- PCA can be used for noise reduction
  (principal directions represent signal, other directions noise).
- Apply kernel PCA for noise reduction:
  - Compute d-dim subspace V_d spanned by d-principal directions.
  - For a new data x,

    G_x: Projection of $\Phi(x)$ onto V_d = noise reduced feacure vector.
  - Compute a preimage $\hat{x}$ in data sapce for the noise reduced feature vector G_x.

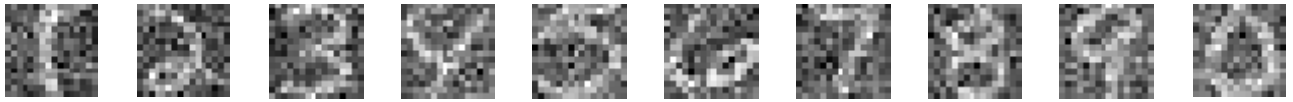
$$\hat{x} = \arg\min_{x'} \left\| \Phi(x') - G_x \right\|^2$$

Note: G_x is not necessariy given as an image of Φ.

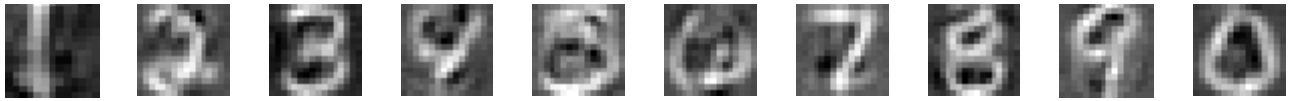


USPS data

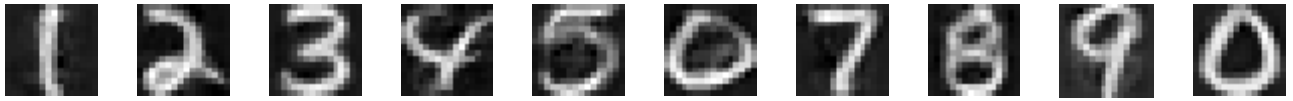
Original data (NOT used for PCA)



Noisy images



Noise reduced images（linear PCA）



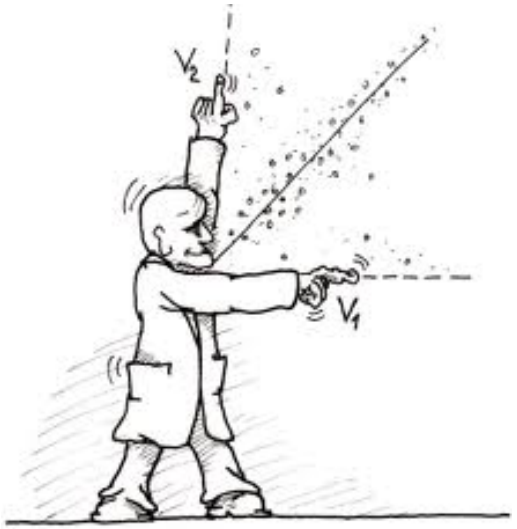
Noise reduced images（kernel PCA, Gaussian kernel）

Generated by Matlab stprtool (by V. Franc)

Properties of Kernel PCA

- Nonlinear features can be extracted.
- Can be used for a preprocessing of other analysis like classification. (dimension reduction / feature extraction)
- The results depend on the choice of kernel and kernel parameters. Interpreting the results may not be straightforward.
- How to choose a kernel and kernel parameter?
  - Cross-validation is not straightforward unlike SVM.
  - If it is a preprocessing, the performance of the final analysis should be maximized.

Kernel Canonical Correlation Analysis



Canonical Correlation Analysis

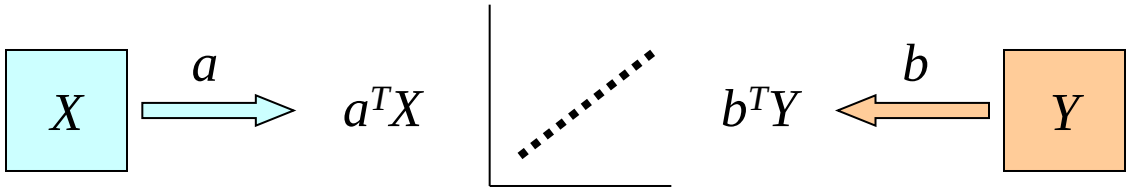
– Canonical correlation analysis (CCA, Hotelling 1936)

Linear dependence of two multivariate random vectors.

- Data $(X_1, Y_1), \ldots, (X_N, Y_N)$
- X_i: m-dimensional, Y_i: ℓ-dimensional

Find the directions a and b so that the correlation of $a^T X$ and $b^T Y$ is maximized.

$$\rho = \max_{a,b} \mathrm{Corr}[a^T X, b^T Y] = \max_{a,b} \frac{\mathrm{Cov}[a^T X, b^T Y]}{\sqrt{\mathrm{Var}[a^T X]\mathrm{Var}[b^T Y]}},$$



Solution of CCA

$$\max_{a,b} a^T \hat{V}_{XY} b \quad \text{subject to} \quad a^T \hat{V}_{XX} a = b^T \hat{V}_{YY} b = 1.$$

– Rewritten as a generalized eigenproblem:

$$\begin{pmatrix} O & \hat{V}_{XY} \\ \hat{V}_{YX} & O \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \rho \begin{pmatrix} \hat{V}_{XX} & O \\ O & \hat{V}_{YY} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

[Exercise: derive this. (Hint. Use Lagrange multiplier method.)]

– Solution:

$$a = V_{XX}^{1/2} u_1, \; b = V_{YY}^{1/2} v_1$$

where u_1 (v_1, resp.) is the left (right, resp.) first eigenvector for the SVD of $\hat{V}_{XX}^{-1/2} \hat{V}_{XY} \hat{V}_{YY}^{-1/2}$.

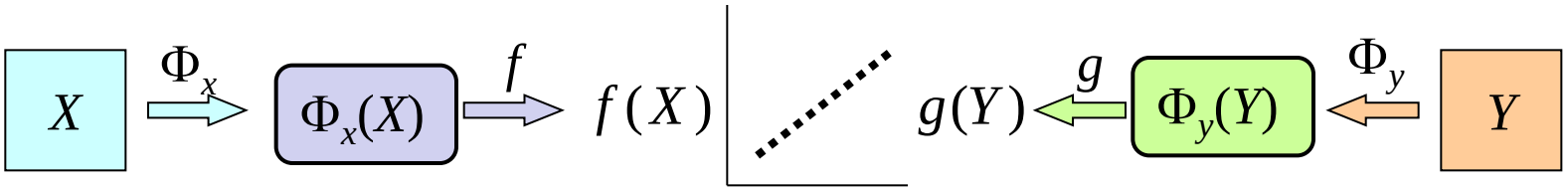
Kernel CCA

– Kernel CCA (Akaho 2000, Melzer et al. 2002, Bach et al 2002)
  - Dependence (not only correlation) of two random variables.
  - Data: $(X_1, Y_1), \ldots, (X_N, Y_N)$ arbitrary variables

  - Consider CCA for the feature vectors with k_X and k_Y:

$$X_1, \ldots, X_N \mapsto \Phi_X(X_1), \ldots, , \Phi_X(X_N) \in H_X,$$
$$Y_1, \ldots, Y_N \mapsto \Phi_Y(Y_1), \ldots, \Phi_Y(Y_N) \in H_Y.$$

$$\max_{f \in H_X, g \in H_Y} \frac{\mathrm{Cov}[f(X), g(Y)]}{\sqrt{\mathrm{Var}[f(X)]\mathrm{Var}[g(Y)]}} = \max_{f \in H_X, g \in H_Y} \frac{\sum_i^N \langle f, \tilde{\Phi}_X(X_i)\rangle\langle \tilde{\Phi}_Y(Y_i), g\rangle}{\sqrt{\sum_i^N \langle f, \tilde{\Phi}_X(X_i)\rangle^2 \sum_i^N \langle g, \tilde{\Phi}_Y(Y_i)\rangle^2}}$$



– We can assume $f = \sum_{i=1}^{N} \alpha_i \widetilde{\Phi}_X(X_i)$ and $g = \sum_{i=1}^{N} \beta_i \widetilde{\Phi}_Y(Y_i)$.

(same as kernel PCA)

$$\max_{\alpha \in \mathrm{R}^N, \beta \in R^N} \frac{\alpha^T \widetilde{K}_X \widetilde{K}_Y \beta}{\sqrt{\alpha^T \widetilde{K}_X^2 \alpha \; \beta^T \widetilde{K}_Y^2 \beta}}$$

$\widetilde{K}_X$ and $\widetilde{K}_Y$: centered Gram matrices.

– Regularization: to avoid trivial solution,

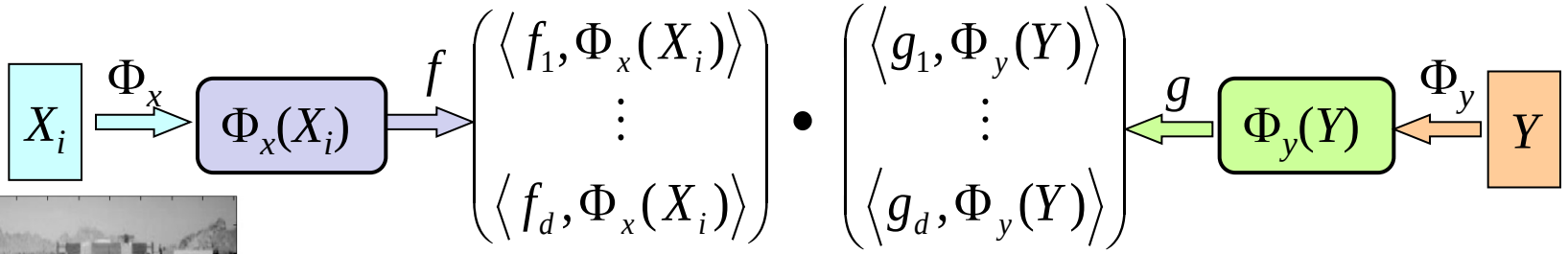
$$\max_{f \in H_X, g \in H_Y} \frac{\sum_{i=1}^{N} \langle f, \widetilde{\Phi}_X(X_i) \rangle \langle \widetilde{\Phi}_Y(Y_i), g \rangle}{\sqrt{\sum_{i=1}^{N} \langle f, \widetilde{\Phi}_X(X_i) \rangle^2 + \varepsilon_N \|f\|_{H_X}^2} \sqrt{\sum_{i=1}^{N} \langle g, \widetilde{\Phi}_Y(Y_i) \rangle^2 + \varepsilon_N \|g\|_{H_Y}^2}}$$

– Solution: generalized eigenproblem

$$\begin{pmatrix} O & \widetilde{K}_X \widetilde{K}_Y \\ \widetilde{K}_Y \widetilde{K}_X & O \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} \widetilde{K}_X^2 + \varepsilon_N \widetilde{K}_X & O \\ O & \widetilde{K}_Y^2 + \varepsilon_N \widetilde{K}_Y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Application of KCCA

– Application to image retrieval (Hardoon, et al. 2004).
  - X_i: image,
    Y_i: corresponding texts (extracted from the same webpages).
  - Idea: use d eigenvectors $f_1,\ldots,f_d$ and $g_1,\ldots,g_d$ as the feature spaces which contain the dependence between X and Y.
  - Given a new word Y_{new}, compute its feature vector, and find the image whose feature has the highest inner product.

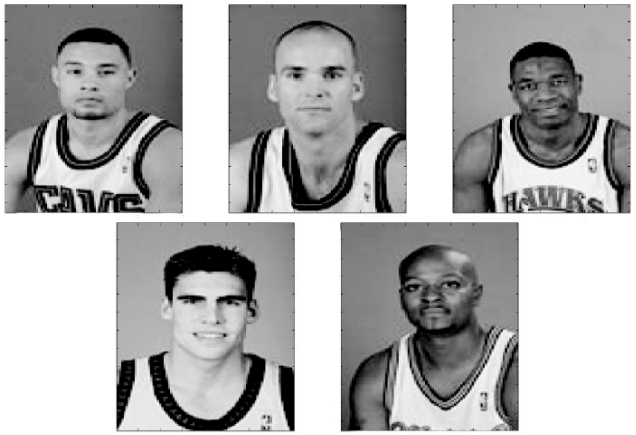


"at phoenix sky harbor on july 6, 1997. 757-2s7, n907wa, ..."

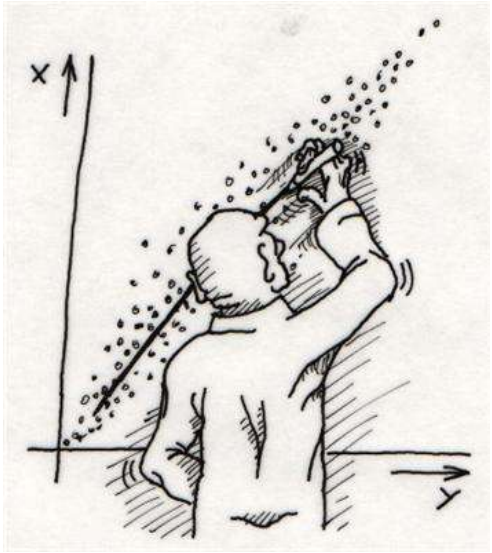
– Example:
  - Gaussian kernel for images.
  - Bag-of-words kernel (frequency of words) for texts.

Text -- “height: 6-11, weight: 235 lbs, position: forward,
born: september 18, 1968, split, croatia college: none”

Extracted images



Kernel Ridge Regression



Ridge regression

$(X_1, Y_1), \ldots, (X_n, Y_n)$: data ($X_i \in \mathbf{R}^m$, $Y_i \in \mathbf{R}$)

Ridge regression: linear regression with L^2 penalty.

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

(The constant term is omitted for simplicity)

– Solution (quadratic function):

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

where

$$V_{XX} = \frac{1}{n} X^T X, \quad X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \in \mathbf{R}^{n \times m}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbf{R}^n$$

– Ridge regression is preferred, when V_{XX} is (close to) singular.

Kernel ridge regression

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: X arbitrary data, $Y \in \mathbf{R}$.
- Kernelization of ridge regression: positive definite kernel k for X

$$\min_{f \in H} \sum_{i=1}^{n} |Y_i - \langle f, \Phi(X_i) \rangle_H|^2 + \lambda \|f\|_H^2$$ Ridge regression on H

equivalently,

$$\min_{f \in H} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \lambda \|f\|_H^2$$ Nonlinear ridge regr.

– Solution is given by the form $f = \sum_{j=1}^{n} c_j \Phi(X_j)$,

Let $f = \sum_{i=1}^{n} c_i \Phi(X_i) + f_\perp = f_\Phi + f_\perp$
($f_\Phi \in Span\{\Phi(X_i)\}_{i=1}^{n}$, $f_\perp$: orthogonal complement)

Objective function $= \sum_{i=1}^{n} | Y_i - \langle f_\Phi + f_\perp, \Phi(X_i) \rangle|^2 + \lambda \|f_\Phi + f_\perp\|^2$

$= \sum_{i=1}^{n} | Y_i - \langle f_\Phi, \Phi(X_i) \rangle|^2 + \lambda (\|f_\Phi\|^2 + \|f_\perp\|^2)$

The 1st term does not depend on $f_\perp$, and 2nd term is minimized in the case $f_\perp = 0$.

– Objective function :

$$\|Y - K_X c\|^2 + \lambda\, c^T K_X c$$

– Solution: $\hat{c} = (K_X + \lambda I_n)^{-1} Y$

Function: $\hat{f}(x) = Y^T (K_X + \lambda I_n)^{-1} \mathbf{k}(x)$ $\quad \mathbf{k}(x) = \begin{pmatrix} k(x, X_1) \\ \vdots \\ k(x, X_n) \end{pmatrix}$

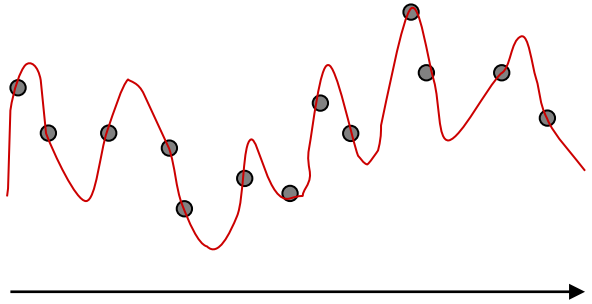
Regularization

– The minimization

$$\min_f |Y_i - f(X_i)|^2$$

may be attained with zero errors.

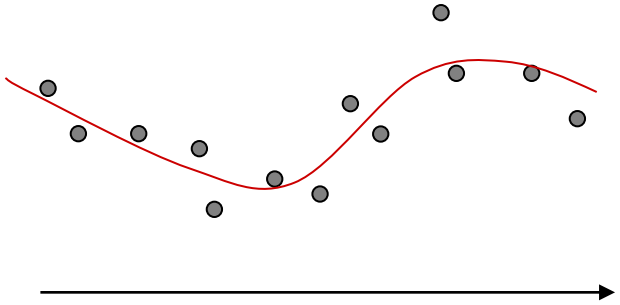
But the function may not be unique.



– Regularization

$$\min_{f \in H} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \lambda \|f\|_H^2$$

- Regularization with smoothness penalty is preferred for uniqueness and smoothness.
- Link with some RKHS norm and smoothness is discussed in Sec. IV.



Comparison

- Kernel ridge regression vs local linear regression

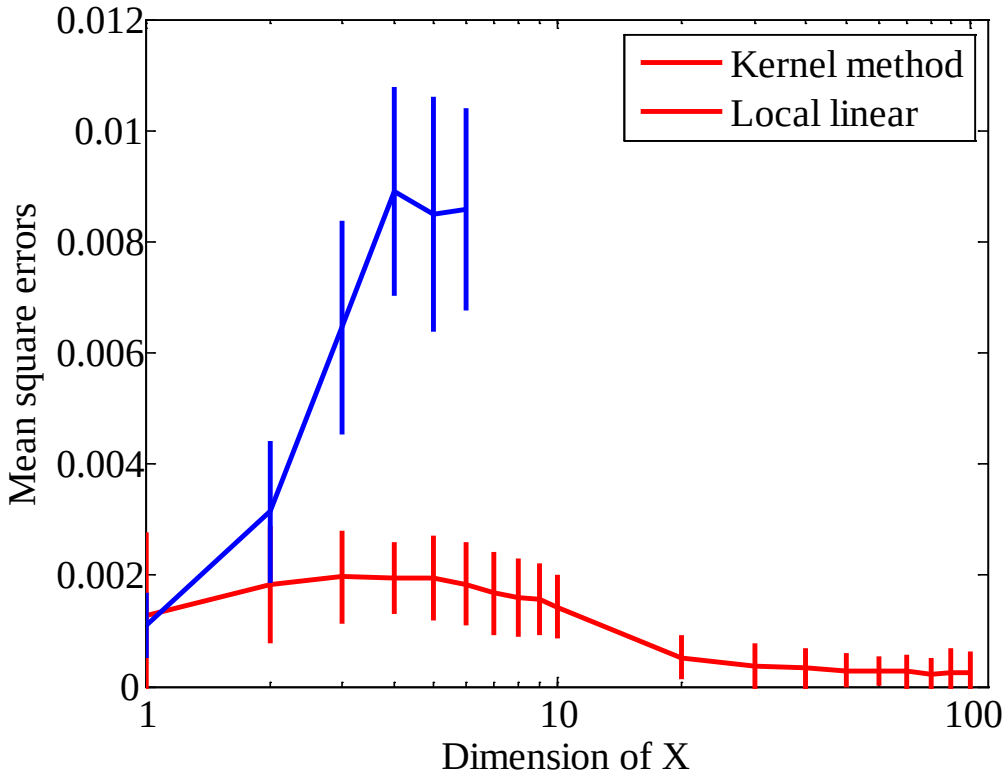
$Y = 1/(1.5 + ||X||^2) + Z, \quad X \sim N(0, I_d),\ Z \sim N(0, 0.1^2)$

n = 100, 500 runs

Kernel ridge regression with Gaussian kernel

Local linear regression with Epanechnikov kernel ('locfit' in R is used)

Bandwidth parameters are chosen by CV.



Local linear regression (e.g., Fan and Gijbels 1996)

- K: smoothing kernel $(K(x) \geq 0,\ \int K(x)dx = 1$, not necessarily positive definite)

- Local linear regression

  $E[Y|X = x_0]$ is estimated by

$$K_h(x) = h^{-d} K\left(\frac{x}{h}\right)$$

$$\min_{a,b} \sum_i |Y_i - a - b^T(X_i - x_0)|^2 K_h(X_i - x_0)$$

  - For each x_0, this minimization can be solved by linear algebra.
  - Statistical property of this estimator is well studied.
  - For one dimensional X, it works nicely with some theoretical optimality.
  - But, weak for high-dimensional data.

Some topics on kernel methods

- Representer theorem
- Structured data
- Kernel choice
- Low rank approximation

Representer theorem

$(X_1, Y_1), \ldots, (X_n, Y_n)$: data

k: positive definite kernel for X, H: corresponding RKHS.

Ψ: monotonically increasing function on R_+.

Theorem 2.1 (representer theorem, Kimeldorf & Wahba 1970)

The solution to the minimization problem:

$$\min_{f \in H} \; F((X_1, Y_1, f(X_1)), \ldots, (X_n, Y_n, f(X_n))) + \Psi(||f||)$$

is attained by

$$f = \sum_{i=1}^{n} c_i \Phi(X_i) \qquad \text{with some } (c_1, \ldots, c_n) \in R^n .$$

The proof is essentially the same as the one for the kernel ridge regression. [Exercise: complete the proof]

Structured Data

– Structured data: non-vectorial data with some structure.

- Sequence data (variable length):

  DNA sequence, Protein (sequence of amino acid)

  Text (sequence of words)

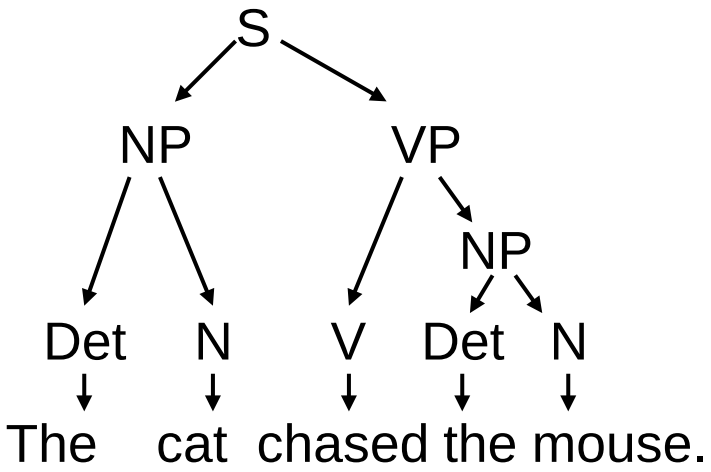
- Graph data (Koji's lecture)

  Chemical compound etc.

- Tree data

  Parse tree

- Histograms / probability measures



– Many kernels uses counts of substructures (Haussler 1999).

Spectrum kernel

– *p*-spectrum kernel (Leslie et al 2002): positive definite kernel for string.

$k_p(s,t) =$ Occurrences of common subsequences of length p.

– Example:

s = “statistics” t = “pastapistan”

3-spectrum

s: sta, tat, ati, tis, ist, sti, tic, ics

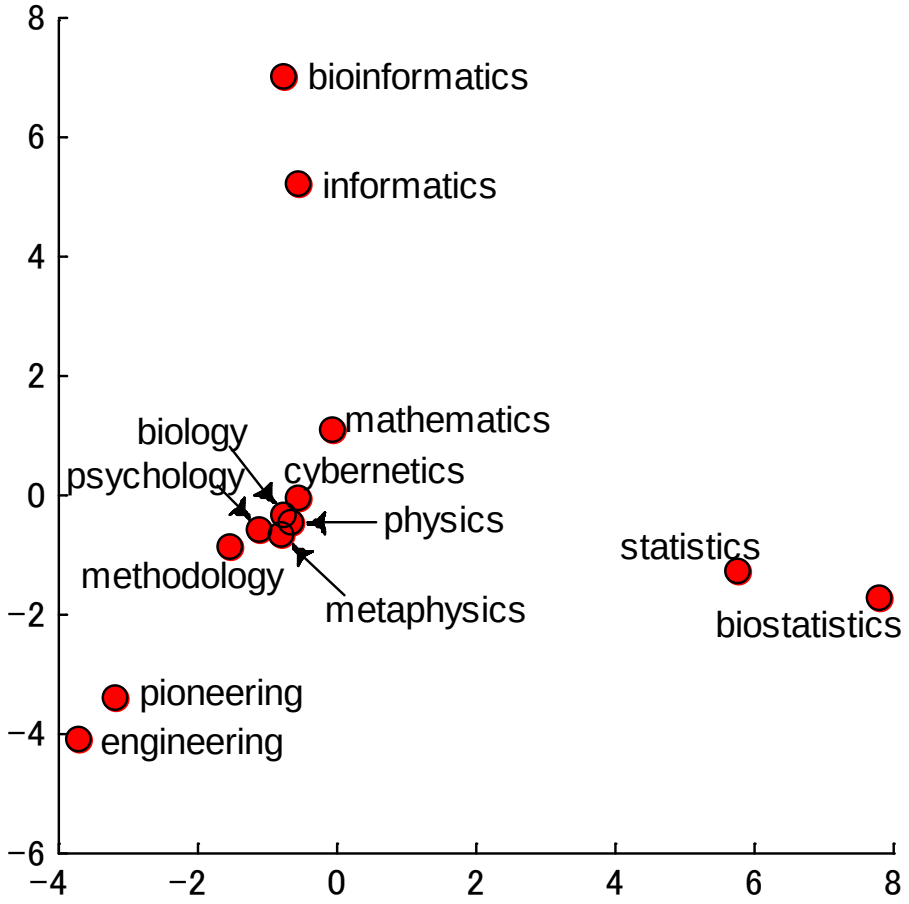
t : pas, ast, sta, tap, api, pis, ist, sta, tan

	sta	tat	ati	tis	ist	sti	tic	ics	pas	ast	tap	api	pis	tan
Φ(s)	1	1	1	1	1	1	1	1	0	0	0	0	0	0
Φ(t)	2	0	0	0	1	0	0	0	1	1	1	1	1	1

$K_3(s,t) = 1 \cdot 2 + 1 \cdot 1 = 3$

– Linear time ($O(p(|S| + |t|)$) algorithm with suffix tree is known (Vishwanathan & Smola 2003).

– Application: kernel PCA of ‘words’ with 3-spectrum kernel



Choice of kernel

Choice of kernel

- Choice of kernel (polyn or Gauss)
- Choice of parameters (bandwidth parameter in Gaussian kernel)

General principles

- Reflect the structure of data (e.g., kernels for structured data)
- For supervised learning (e.g., SVM) → Cross-validation
- For unsupervised learning (e.g. kernel PCA)
  - No general methods exist.
  - Guideline: make or use a relevant supervised problem, and use CV.
- Learning a kernel: Multiple kernel learning (MKL)

$$k(x,y) = \sum_{i=1}^{M} c_i k_i(x,y) \qquad \text{optimize } c_i$$

Low rank approximation

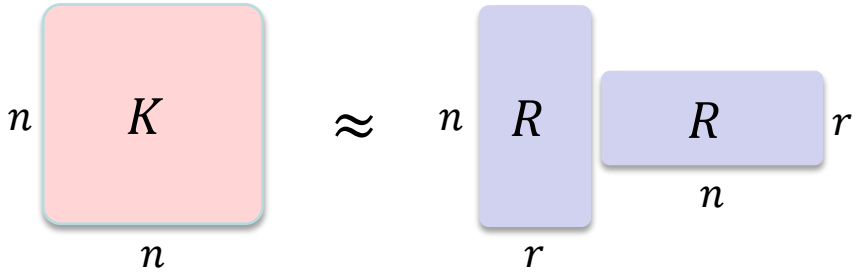
– Gram matrix: $n \times n$ where n is the sample size.

  Large n causes computational problems.

  e.g. Inversion, eigendecomposition costs $O(n^3)$ in time.

– Low-rank approximation

$$K \approx RR^T, \quad \text{where } R\text{: } n \times r \text{ matrix } (r < n)$$



– The decay of eigenvalues of a Gram matrix is often quite fast

(See Widom 1963, 1964; Bach & Jordan 2002).

– Two major methods

- Incomplete Cholesky factorization (Fine & Sheinberg 2001)

  $O(nr^2)$ in time and $O(nr)$ in space

- Nyström approximation (Williams and Seeger 2001)

  Random sampling + eigendecomposition

– Example: kernel ridge regression

$$Y^T(K_X + \lambda I_n)^{-1}\mathbf{k}(x) \qquad \text{time : } O(n^3)$$

Low rank approximation: $K_X \approx RR^T$. With Woodbury formula*

$$Y^T(K_X + \lambda I_n)^{-1}\mathbf{k}(x) \approx Y^T(RR^T + \lambda I_n)^{-1}\mathbf{k}(x)$$

$$= \frac{1}{\lambda}\{Y^T\mathbf{k}(x) - Y^T R(R^T R + \lambda I_r)^{-1}R^T\mathbf{k}(x)\}$$

$$\text{time : } O(r^2 n + r^3)$$

* Woodbury (Sherman–Morrison–Woodbury) formula:

$$(A + UV)^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}VA^{-1}.$$

Other kernel methods

- Kernel Fisher discriminant analysis（kernel FDA）(Mika et al. 1999)
- Kernel logistic regression（Roth 2001, Zhu&Hastie 2005）
- Kernel partial least square（kernel PLS）（Rosipal&Trejo 2001）
- Kernel K-means clustering（Dhillon et al 2004）

etc, etc, …

- Variants of SVM → Section III.

Summary: Properties of kernel methods

- Various classical linear methods can be kernelized
  → Linear algorithms on RKHS.

- The solution typically has the form

$$f = \sum_{i=1}^{n} c_i \Phi(X_i). \qquad \text{(representer theorem)}$$

- The problem is reduced to manipulation of Gram matrices of size n (sample size).
  - Advantage for high dimensional data.
  - For a large number of data, low-rank approximation is used effectively.

- Structured data:
  - kernel can be defined on any set.
  - kernel methods can be applied to any type of data.

References

Akaho. (2000) Kernel Canonical Correlation Analysis. *Proc. 3rd Workshop on Induction-based Information Sciences (IBIS2000).* (in Japanese)

Bach, F.R. and M.I. Jordan. (2002) Kernel independent component analysis. *Journal of Machine Learning Research,* 3:1–48.

Dhillon, I. S., Y. Guan, and B. Kulis. (2004) Kernel k-means, spectral clustering and normalized cuts. *Proc. 10th ACM SIGKDD Intern. Conf.*

Knowledge Discovery and Data Mining (KDD), 551–556.

Fan, J. and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman Hall/CRC, 1996.

Fine, S. and K. Scheinberg. (2001) Efficient SVM Training Using Low-Rank Kernel Representations. *Journal of Machine Learning Research*, 2:243-264.

Gökhan, B., T. Hofmann, B. Schölkopf, A.J. Smola, B. Taskar, S.V.N. Vishwanathan. (2007) *Predicting Structured Data*. MIT Press.

Hardoon, D.R., S. Szedmak, and J. Shawe-Taylor. (2004) Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664.

Haussler, D. Convolution kernels on discrete structures. *Tech Report UCSC-CRL-99-10, Department of Computer Science, University of California at Santa Cruz,* 1999.

Leslie, C., E. Eskin, and W.S. Noble. (2002) The spectrum kernel: A string kernel for SVM protein classification, in *Proc. Pacific Symposium on Biocomputing*, 564–575.

Melzer, T., M. Reiter, and H. Bischof. (2001) Nonlinear feature extraction using generalized canonical correlation analysis. *Proc. Intern. Conf. .Artificial Neural Networks (ICANN 2001)*, 353–360.

Mika, S., G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. (1999) Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, edits, *Neural Networks for Signal Processing*, volume IX, 41–48. IEEE.

Rosipal, R. and L.J. Trejo. (2001) Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, 2: 97–123.

Roth, V. (2001) Probabilistic discriminative kernel classifiers for multi-class problems. In *Pattern Recognition: Proc. 23rd DAGM Symposium*, 246–253. Springer.

Schölkopf, B., A. Smola, and K-R. Müller. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319.

Schölkopf, B. and A. Smola. *Learning with Kernels*. MIT Press. 2002.

Vishwanathan, S. V. N. and A.J. Smola. (2003) Fast kernels for string and tree matching. *Advances in Neural Information Processing Systems 15*, 569–576. MIT Press.

Williams, C. K. I. and M. Seeger. (2001) Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 13:682–688.

Widom, H. (1963) Asymptotic behavior of the eigenvalues of certain integral equations. *Transactions of the American Mathematical Society*, 109:278{295, 1963.

Widom, H. (1964) Asymptotic behavior of the eigenvalues of certain integral equations II. *Archive for Rational Mechanics and Analysis*, 17:215{229, 1964.

Appendix

Exercise for kernel PCA

- $\| f \|_H^2 = c^T \tilde{K}_X c$

$$\|f\|_H^2 = \left\langle \sum_{i=1}^{n} c_i \tilde{\Phi}(X_i), \sum_{j=1}^{n} c_j \tilde{\Phi}(X_j) \right\rangle = \sum_{i}^{n} c_i c_j \langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle = c^T \tilde{K}_X c.$$

- $\mathrm{Var}\left[\langle f, \Phi(X) \rangle\right] = c^T \tilde{K}_X^2 c$

$$\mathrm{Var}[\langle f, \Phi(X) \rangle] = \frac{1}{n} \sum_{j=1}^{n} \left\langle \sum_{i=1}^{n} c_i \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \right\rangle^2$$

$$= \frac{1}{n} \sum_{j=1}^{n} \left\langle \sum_{i=1}^{n} c_i \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \right\rangle \left\langle \sum_{h=1}^{n} c_h \tilde{\Phi}(X_h), \tilde{\Phi}(X_j) \right\rangle$$

$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{i}^{n} c_i \tilde{K}_{ij} \sum_{hh}^{n} c_h \tilde{K}_{hj} \quad = \frac{1}{n} c^T \tilde{K}_X^2 c$$

■

III. Support Vector Machines
A Brief Introduction

Kenji Fukumizu

The Institute of Statistical Mathematics /
Graduate University for Advanced Studies

September 6-7
Machine Learning Summer School 2012, Kyoto

Large margin classifier

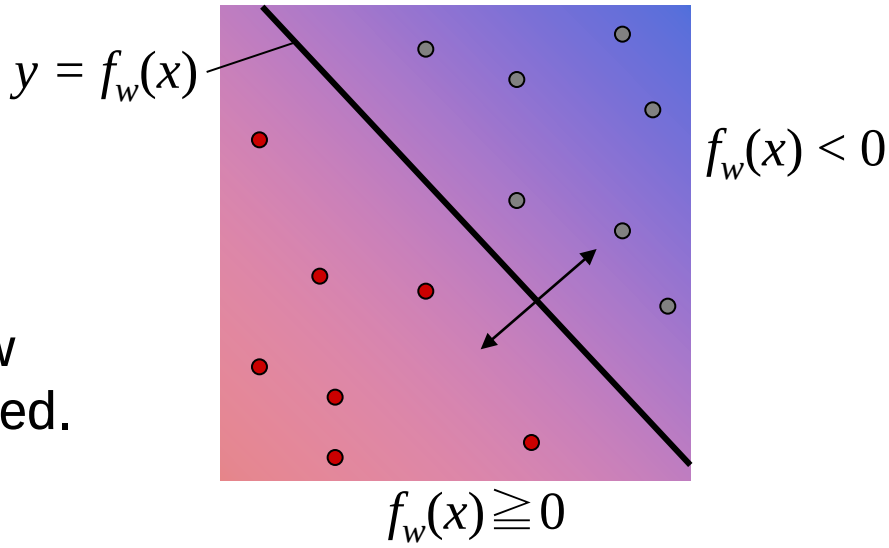
- $(X_1, Y_1), \ldots, (X_n, Y_n)$: training data
  - X_i: input (m-dimensional)
  - $Y_i \in \{\pm 1\}$: binary teaching data,

- Linear classifier

$$f_w(x) = w^T x + b$$

$$h(x) = \mathrm{sgn}(f_w(x))$$

We wish to make $f_w(x)$ with the training data so that a new data x can be correctly classified.



Large margin criterion

Assumption: the data is linearly separable.

Among infinite number of separating hyperplanes, choose the one to give the largest margin.

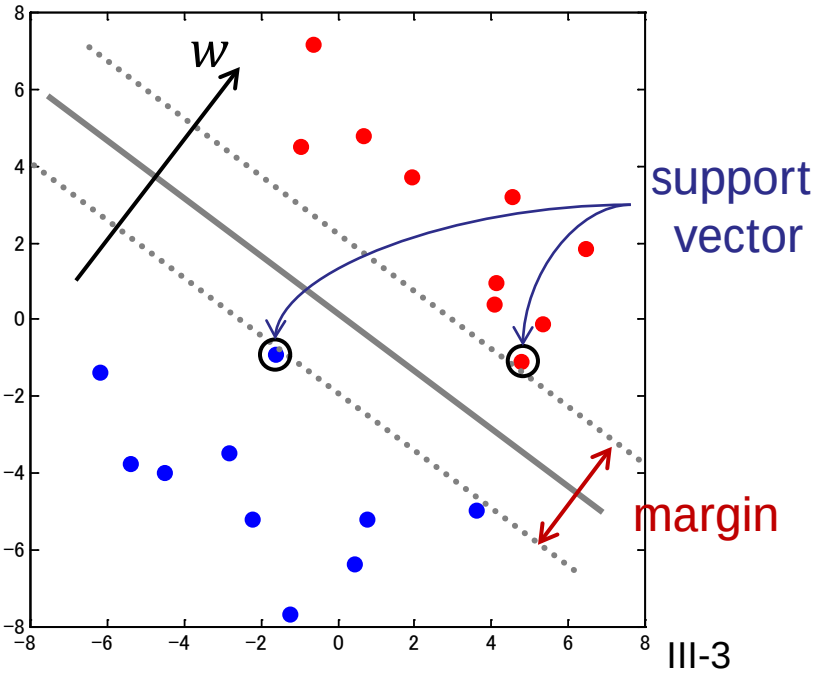
– Margin = distance of two classes measured along the direction of w.

– Support vector machine:

Hyperplane to give the largest margin.

– The classifier is the middle of the margin.

– "Supporting points" on the two boundaries are called support vectors.



– Fix the scale (rescaling of (w, b) does not change the plane)

$$\begin{cases} \min(w^T X_i + b) = 1 & \text{if } Y_i = 1, \\ \max(w^T X_i + b) = -1 & \text{if } Y_i = -1 \end{cases}$$

Then

$$\text{Margin} = \frac{2}{\|w\|}$$

[Exercise] Prove this.

Support vector machine (linear, hard margin)

(Boser, Guyon, Vapnik 1992)

Objective function:

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{subject to} \quad \begin{cases} w^T X_i + b \geq 1 & \text{if } Y_i = 1, \\ w^T X_i + b \geq 1 & \text{if } Y_i = -1. \end{cases}$$

SVM (hard margin)

$$\min_{w,b} \quad \|w\|^2 \qquad \text{subject to} \qquad Y_i(w^T X_i + b) \geq 1 \quad (\forall i)$$

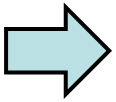
– Quadratic program (QP):
  - Minimization of a quadratic function with linear constraints.
  - Convex, no local optima (Vandenberghe' lecture)
  - Many solvers available (Chih-Jen Lin's lecture)

Soft-margin SVM

– "Linear separability" is too strong. Relax it.

Hard margin

$$Y_i(w^T X_i + b) \geq 1$$



Soft margin

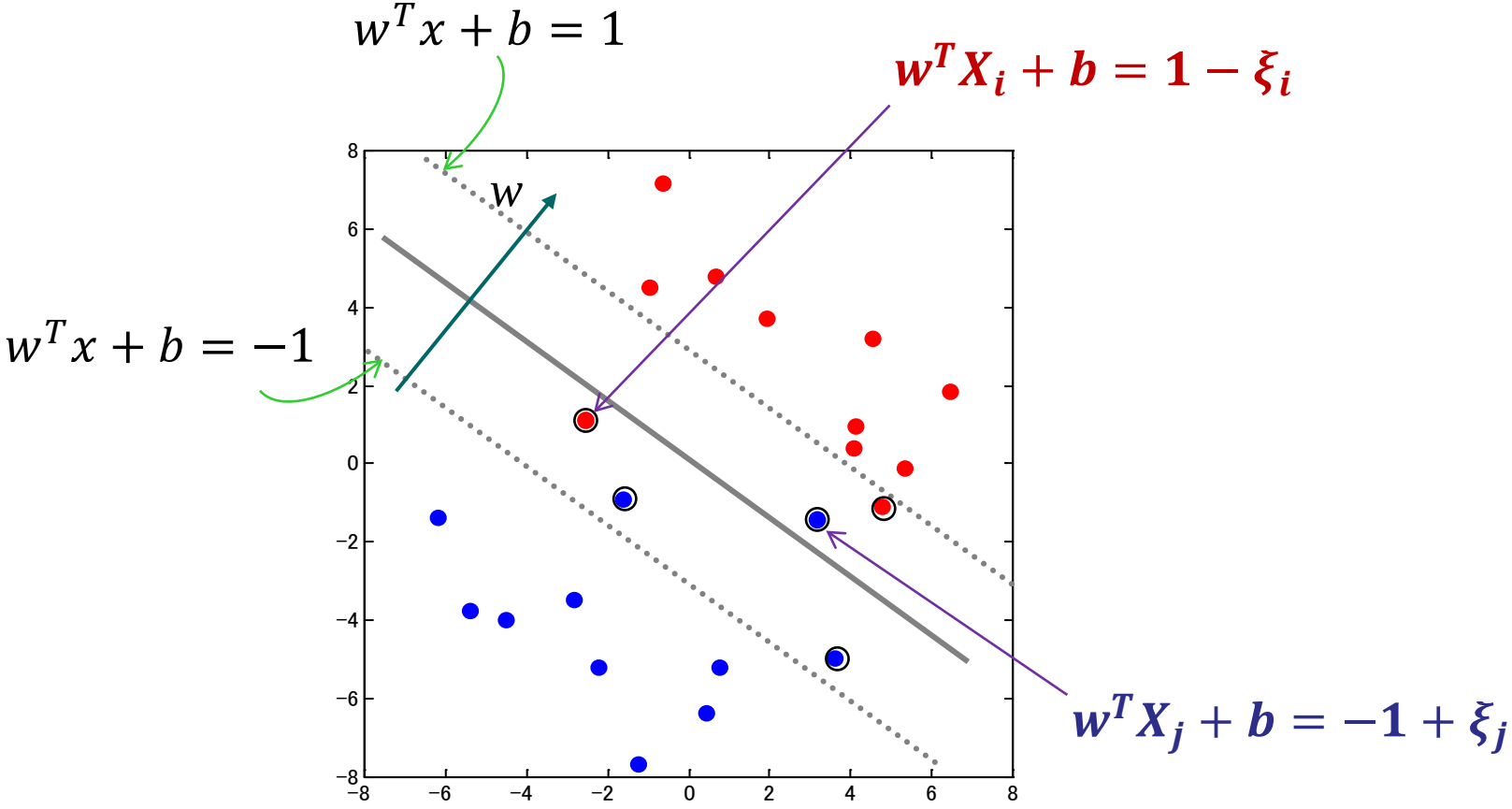
$$Y_i(w^T X_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0$$

SVM (soft margin)

$$\min_{w,b} \quad \|w\|^2 + C\sum_{i=1}^{n} \xi_i \qquad \text{subject to} \qquad \begin{matrix} Y_i(w^T X_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0 \end{matrix} \quad (\forall i)$$

– This is also QP.

– The coefficient C must be given. Cross-validation is often used.


$w^T x + b = 1$
$w^T X_i + b = 1 - \xi_i$
w
$w^T x + b = -1$
$w^T X_j + b = -1 + \xi_j$

SVM and regularization

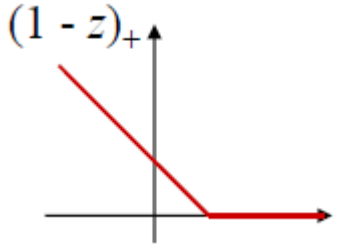
– Soft-margin SVM is equivalent to the regularization problem:

$$\min_{w,b} \underbrace{\sum_{i=1}^{n} \left(1 - Y_i(w^T X_i + b)\right)_+}_{\text{loss}} + \underbrace{\lambda \|w\|^2}_{\text{regularization term}}$$

$$(z)_+ = \max\{0, z\}$$

- loss function: Hinge loss

$$\ell(f(x), y) = \left(1 - f(x)\right)_+$$



- *c.f.* Ridge regression (squared error)

$$\min_{w,b} \sum_{i=1}^{n} (Y_i - (w^T X_i + b))^2 + \lambda \|w\|^2$$

[Exercise] Confirm the above equivalence.

Kernelization: nonlinear SVM

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: training data
  - X_i: input on arbitrary space Ω
  - $Y_i \in \{\pm 1\}$: binary teaching data,

- Kernelization: Positive definite kernel k on Ω (RKHS H),
  Feature vectors $\Phi(X_1), \ldots, \Phi(X_n)$

- Linear classifier on H → nonlinear classifier on Ω

$$h(x) = \operatorname{sgn}(\langle f, \Phi(x) \rangle + b)$$
$$= \operatorname{sgn}(f(x) + b), \qquad f \in H.$$

Nonlinear SVM

$$\min_{f,b} \quad \|f\|^2 + C\sum_{i=1}^{n}\xi_i \qquad \text{subject to} \qquad \begin{matrix} Y_i(\langle f, \Phi(X_i)\rangle + b) \geq 1, \\ \xi_i \geq 0 \end{matrix} \qquad (\forall i)$$

or equivalently,

$$\min_{f,b} \quad \sum_{i=1}^{n}(1 - Y_i(f(X_i) + b))_+ + \lambda\|f\|_H^2$$

By representer theorem, $f = \sum_{j=1}^{n} w_j \Phi(X_j)$.

nonlinear SVM (soft margin)

$$\min_{w,b} \quad w^T K w + C\sum_{i=1}^{n}\xi_i \qquad \text{subject to} \qquad \begin{matrix} Y_i((Kw)_i + b) \geq 1, \\ \xi_i \geq 0 \end{matrix} \qquad (\forall i)$$

- This is again QP.

Dual problem

$$\min_{w,b} \quad w^T K w + C \sum_{i=1}^{n} \xi_i \qquad \text{subject to} \qquad \begin{matrix} Y_i((Kw)_i + b) \geq 1, \\ \xi_i \geq 0 \end{matrix} \qquad (\forall i)$$

– By Lagrangian multiplier method, the dual problem is

SVM (dual problem)

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j K_{ij} \quad \text{subject to} \quad \begin{matrix} 0 \leq \alpha_i \leq C, \\ \sum_{i=1}^{n} Y_i \alpha_i = 0 \end{matrix} \qquad (\forall i)$$

- The dual problem is often preferred.
- The classifier is expressed by

$$f_*(x) + b_* = \sum_{i=1}^{n} \alpha_{i*} Y_i \, K(x, X_i) + b_*$$

– Sparse expression: Only the data with $0 < \alpha_{i*} \leq C$ appear in the summation. → Support vectors.

KKT condition

Theorem

The solution of the primal and dual problem of SVM is given by the following equations:

(1) $1 - Y_i\,(f^*(X_i) + b^*) - \xi_i^* \leq\ 0\ (\forall\ i)$ [primal constraint]

(2) $\xi_i^* \geq\ 0\ (\forall\ i)$ [primal constraint]

(3) $0 \leq \alpha_i^* \leq C, (\forall\ i)$ [dual constraint]

(4) $\alpha_i^*(1 - Y_i\,(f^*(X_i) + b^*)\ - \xi_i^*)\ =\ 0\ (\forall\ i)$ [complementary]

(6) $\xi_i^*(C - \alpha_i^*)\ =\ 0\ (\forall\ i),$ [complementary]

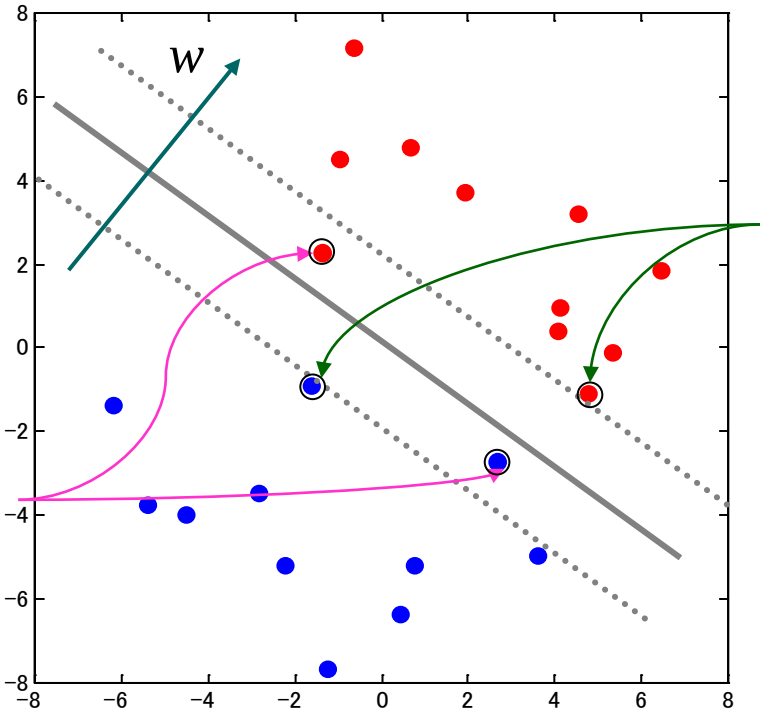
(7) $\sum_{j=1}^{n} K_{ij} w_j^* - \sum_j^n \alpha_j^* Y_j K_{ij} = 0,$

(8) $\sum_{j=1}^{n} \alpha_j^* Y_j = 0,$

Sparse expression

$$\varphi_*(x) = f_*(x) + b_* = \sum_{X_i:\ \text{support vectors}} \alpha_{i*} Y_i K(x, X_i) + b_*$$

– Two types of support vectors.



support vectors
$0 < \alpha_i < C$
$(Y_i \varphi(X_i) = 1)$

support vectors
$\alpha_i = C$
$(Y_i \varphi(X_i) \leq 1)$

Summary of SVM

- One of the kernel methods:
  - kernelization of linear large margin classifier.
  - Computation depends on Gram matrices of size n.
- Quadratic program:
  - No local optimum.
  - Many solvers are available.
  - Further efficient optimization methods are available (e.g. SMO, Plat 1999)
- Sparse representation
  - The solution is written by a small number of support vectors.
- Regularization
  - The objective function can be regarded as regularization with hinge loss function.

- NOT discussed on SVM in this lecture are
  - Many successful applications
  - Multi-class extension
    - Combination of binary classifiers (1-vs-1, 1-vs-rest)
    - Generalization of large margin criterion

      Crammer & Singer (2001), Mangasarian & Musicant (2001), Lee, Lin, & Wahba (2004), etc
  - Other extensions
    - Support vector regression (Vapnik 1995)
    - ν-SVM (Schölkopf et al 2000)
    - Structured-output (Collins & Duffty 2001, Taskar et al 2004, Altun et al 2003, etc)
    - One-class SVM (Schökopf et al 2001)
  - Optimization
    - Solving primal problem
  - Implementation (Chih-Jen Lin's lecture)

References

Altun, Y., I. Tsochantaridis, and T. Hofmann. Hidden Markov support vector machines. In *Proc. 20th Intern. Conf. Machine Learning*, 2003.

Boser, B.E., I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

Crammer, K. and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Collins, M. and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14*, pages 625–632. MIT Press, 2001.

Mangasarian, O. L. and David R. Musicant. Lagrangian support vector machines. *Journal of Machine Learning Research*, 1:161–177, 2001

Lee, Y., Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99: 67–81, 2004.

Schölkopf, B., A. Smola, R.C. Williamson, and P.L. Bartlett. (2000) New support vector algorithms. *Neural Computation, 12(5):1207–1245.*

Schölkopf, B., J.C. Platt, J. Shawe-Taylor, R.C. Williamson, and A.J.Smola. (2001) Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471.

Vapnik, V.N. *The Nature of Statistical Learning Theory*. Springer 1995.

Platt, J. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 185–208. MIT Press, 1999.

Books on Application domains:

- Lee, S.-W., A. Verri (Eds.) *Pattern Recognition with Support Vector Machines: First International Workshop, SVM 2002, Niagara Falls, Canada, August 10, 2002*. Proceedings. Lecture Notes in Computer Science 2388, Springer, 2002.
- Joachims, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Springer, 2002.
- Schölkopf, B., K. Tsuda, J.-P. Vert (Eds.). *Kernel Methods in Computational Biology*. MIT Press, 2004.

IV. Theoretical Backgrounds of Kernel Methods

Kenji Fukumizu

The Institute of Statistical Mathematics /
Graduate University for Advanced Studies

September 6-7
Machine Learning Summer School 2012, Kyoto

$\mathbf{C}$-valued Positive definite kernel

Definition.

Ω: set. $k : \Omega \times \Omega \to \mathbf{C}$ is a positive definite kernel if for arbitrary $x_1, \ldots, x_n \in \Omega$ and $c_1, \ldots, c_n \in \mathbf{C}$,

$$\sum_{i,j=1}^{n} c_i \overline{c_j} k(x_i, x_j) \geq 0.$$

Remark: From the above condition, the Gram matrix $\left(k(x_i, x_j)\right)_{ij}$ is necessarily Hermitian, i.e. $k(y, x) = \overline{k(x, y)}$. [Exercise]

Operations that preserve positive definiteness

Proposition 4.1

If $k_i: X \times X \to \mathbf{C}$ $(i = 1,2, \ldots,)$ are positive definite kernels, then so are the following:

1. (positive combination) $ak_1 + bk_2$ $(a, b \geq 0)$.
2. (product) $k_1 k_2$ $(k_1(x,y)k_2(x,y))$
3. (limit) $\lim_{i\to\infty} k_i(x,y)$, assuming the limit exists.

Proof. 1 and 3 are trivial from the definition. For 2, it suffices to prove that Hadamard product (element-wise) of two positive semidefinite matrices is positive semidefinite.

Remark: The set of positive definite kernels on X is a closed cone, where the topology is defined by the point-wise convergence.

Proposition 4.2

Let A and B be positive semidefinite Hermitian matrices. Then, Hadamard product $K = A * B$ (element-wise product) is positive semidefinite.

Proof)

Eigendecomposition of A: $A = U\Lambda\overline{U}^T = \left(U_p^i\right)\begin{pmatrix}\lambda_1 & 0 \cdots 0 \\ 0 & \lambda_2 \cdots \\ \vdots & \ddots \vdots \\ 0 \cdots 0 & \lambda_n\end{pmatrix}\overline{\left(U_p^j\right)}^T$

i.e.,

$$A_{ij} = \sum_{p=1}^{n} \lambda_p U_p^i \overline{U_p^j} \qquad (\lambda_p \geqq 0 \text{ by the positive semidefiniteness}).$$

Then,

$$\sum_{i,j=1}^{n} c_i \overline{c_j} K_{ij} = \sum_{p=1}^{n} \sum_{i,j=1}^{n} c_i \overline{c_j} \lambda_p U_p^i \overline{U_p^j} B_{ij}$$

$$= \lambda_1 \left(\sum_{i,j=1}^{n} c_i U_1^i \overline{c_j U_1^j} B_{ij} \right) + \cdots + \lambda_n \left(\sum_{i,j=1}^{n} c_i U_n^i \overline{c_j U_n^j} B_{ij} \right) \geq 0.$$

■

Normalization

Proposition 4.3

Let k be a positive definite kernel on Ω, and $f: \Omega \to \mathbf{C}$ be an arbitrary function. Then,

$$\tilde{k}(x,y) := f(x)k(x,y)\overline{f(y)}$$

is positive definite. In particular,

$$f(x)\overline{f(y)}$$

is a positive definite kernel.

- Proof [Exercise]
- Example. Normalization:

$$\tilde{k}(x,y) = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

is positive definite, and $\|\Phi(x)\| = 1$ for any $x \in \Omega$.

Proof of positive definiteness

- Euclidean inner product $x^T y$: easy (Prop. 1.1).
- Polynomial kernel $(x^T y + c)^d$ $(c \geq 0)$:

  $(x^T y + c)^d = (x^T y)^d + a_1 (x^T y)^{d-1} + \cdots + a_d, \quad a_i \geq 0.$

  Product and non-negative combination of p.d. kernels.

- Gaussian RBF kernel $\exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right)$:

  $$\exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right) = e^{-\|x\|^2/\sigma^2} e^{x^T y/\sigma^2} \; e^{-\|y\|^2/\sigma^2}$$

  Note

  $$e^{x^T y/\sigma^2} = 1 + \frac{1}{1!\,\sigma^2}(x^T y) + \frac{1}{2!\,\sigma^4}(x^T y)^2 + \cdots$$

  is positive definite (Prop. 4.1). Proposition 4.3 then completes the proof.

- Laplacian kernel is discussed later.

Shift-invariant kernel

– A positive definite kernel $k(x, y)$ on $\mathbf{R}^m$ is called shift-invariant if the kernel is of the form $k(x, y) = \psi(x - y)$.

– Examples: Gaussian, Laplacian kernel

– Fourier kernel ($\mathbf{C}$-valued positive definite kernel): for each ω,

$$k_F(x, y) = \exp\left(\sqrt{-1}\omega^T(x - y)\right) = \exp\left(\sqrt{-1}\omega^T x\right)\overline{\exp\left(\sqrt{-1}\omega^T y\right)}$$

(Prop. 4.3)

– If $k(x, y) = \psi(x - y)$ is positive definite, the function ψ is called positive definite.

Bochner's theorem

Theorem 4.4 (Bochner)

Let ψ be a continuous function on $\mathbf{R}^m$. Then, ψ is ($\mathbf{C}$-valued) positive definite if and only if there is a finite non-negative Borel measure Λ on $\mathbf{R}^m$ such that

$$\psi(z) = \int \exp\left(\sqrt{-1}\omega^T z\right) d\Lambda(\omega)$$

Bochner's theorem characterizes all the continuous shift-invariant positive definite kernels. $\left\{\exp\left(\sqrt{-1}\omega^T z\right) \middle| \omega \in \mathbf{R}^m\right\}$ is the generator of the cone (see Prop. 4.1).

- Λ is the inverse Fourier (or Fourier-Stieltjes) transform of ψ.
- Roughly speaking, the shift invariant functions are the class that have non-negative Fourier transform.
- Sufficiency is easy: $\sum_{i,j} c_i \overline{c_j}\, \psi\left(z_i - z_j\right) = \int \left|\sum_i c_i e^{\sqrt{-1}\omega^T z_i}\right|^2 d\Lambda(\omega)$. Necessity is difficult.

RKHS in frequency domain

Suppose (shift-invariant) kernel k has a form

$$k(x,y) = \int \exp\left(\sqrt{-1}\omega^T(x-y)\right)\rho(\omega)d\omega\,. \qquad \rho(\omega) > 0.$$

Then, RKHS H_k is given by

$$H_k = \left\{ f \in L^2(\mathbf{R}, dx) \;\middle|\; \int \frac{\left|\hat{f}(\omega)\right|^2}{\rho(\omega)} d\omega < \infty \right\},$$

$$\langle f, g \rangle = \int \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\rho(\omega)} d\omega$$

where $\hat{f}$ is the Fourier transform of f:

$$\hat{f}(\omega) = \frac{1}{(2\pi)^m} \int f(x) \exp\left(-\sqrt{-1}\omega^T x\right) dx\,.$$

Gaussian kernel

$$k_G(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right), \qquad \rho_G(\omega) = \frac{1}{(2\pi)^m}\exp\left(-\frac{\sigma^2\|\omega\|^2}{2}\right)$$

$$H_{k_G} = \left\{ f \in L^2(\mathbf{R}, dx) \;\middle|\; \int \left|\hat{f}(\omega)\right|^2 \exp\left(\frac{\sigma^2\|\omega\|^2}{2}\right) d\omega < \infty \right\}$$

$$\langle f, g \rangle = (2\pi)^m \int \hat{f}(\omega)\overline{\hat{g}(\omega)} \exp\left(\frac{\sigma^2\|\omega\|^2}{2}\right) d\omega$$

Laplacian kernel (on $\mathbf{R}$)

$$k_L(x,y) = \exp(-\beta|x-y|), \qquad \rho_L(\omega) = \frac{1}{2\pi(\omega^2+\beta)}$$

$$H_{k_L} = \left\{ f \in L^2(\mathbf{R}, dx) \;\middle|\; \int \left|\hat{f}(\omega)\right|^2 (\omega^2+\beta) d\omega < \infty \right\}$$

$$\langle f, g \rangle = 2\pi \int \hat{f}(\omega)\overline{\hat{g}(\omega)}(\omega^2+\beta) d\omega$$

– Decay of $f \in H$ for high-frequency is different for Gaussian and Laplacian.

RKHS by polynomial kernel

– Polynomial kernel on $\mathbf{R}$:

$$k_p(x, y) = (x^T y + c)^d, \qquad (c \geq 0, d \in \mathbf{N})$$

$$k_p(x, z_0) = z_0^d x^d + \binom{d}{1} c z_0^{d-1} x^{d-1} + \binom{d}{2} c^2 z_0^{d-2} x^{d-1} + \cdots + c^d.$$

Span of these functions are polynomials of degree d.

Proposition 4.5

If $c \neq 0$, the RKHS is the space of polynomials of degree at most d.

[Proof: exercise. Hint. Find b_i to satisfy $\sum_{i=0}^{d} b_i k(x, z_i) = \sum_{i=0}^{d} a_i x^i$ as a solution to a linear equation.]

Sum and product

$(H_1, k_1), (H_2, k_2)$: two RKHS's and positive definite kernels on Ω.

Sum

RKHS for $k_1 + k_2$:

$$H_1 + H_2 = \{f : \Omega \to R \mid \exists f_1 \in H_1, \exists f_2 \in H_2, f = f_1 + f_2\}$$

$$\|f\|^2 = \left\{\|f_1\|_{H_1}^2 + \|f_2\|_{H_2}^2 \,\middle|\, f = f_1 + f_2, f_1 \in H_1, f_2 \in H_2\right\}$$

Product

RKHS for $k_1 k_2$:

$H_1 \otimes H_2$ =tensor product as a vector space.

$\{f = \sum_{i=1}^n f_i g_i \mid f_i \in H_1, g_i \in H_2\}$ is dense in $H_1 \otimes H_2$.

$$\langle \sum_{i=1}^n f_i^{(1)} g_i^{(1)}, \sum_{j=1}^m f_j^{(2)} g_j^{(2)} \rangle = \sum_{i=1}^n \sum_{j=1}^m \langle f_i^{(1)}, f_j^{(2)} \rangle_{H_1} \langle g_i^{(1)}, g_j^{(2)} \rangle_{H_2}.$$

Summary of Section IV

- Positive definiteness of kernels are preserved by
  - Non-negative combinations,
  - Product
  - Point-wise limit
  - Normalization

- Bochner's theorem: characterization of the continuous shift-invariance kernels on $\mathbf{R}^m$.

- Explicit form of RKHS
  - RKHS with shift-invariance kernels has explicit expression in frequency domain.
  - Polynomial kerns gives RKHS of polynomials.
  - Sum and product can be given.

References

Aronszajn., N. Theory of reproducing kernels. *Trans. American Mathematical Society*, 68(3):337–404, 1950.

Saitoh., S. Integral transforms, reproducing kernels, and their applications. Addison Wesley Longman, 1997.

Solution to Exercises

C-valued positive definiteness

Using the definition for one point, we have $k(x,x)$ is real and non-negative for all x. For any x and y, applying the definition with coefficient $(c,1)$ where $c \in \mathbf{C}$, we have

$$|c|^2 k(x,x) + ck(x,y) + \bar{c}k(y,x) + k(y,y) \geq 0.$$

Since the right hand side is real, its complex conjugate also satisfies

$$|c|^2 k(x,x) + \bar{c}\overline{k(x,y)} + c\overline{k(y,x)} + k(y,y) \geq 0.$$

The difference of the left hand side of the above two inequalities is real, so that

$$\bar{c}\left(k(y,x) - \overline{k(x,y)}\right) - \overline{c\left(k(y,x) - \overline{k(x,y)}\right)}$$

is a real number. On the other hand, since $\alpha - \bar{\alpha}$ must be pure imaginary for any complex number α,

$$\bar{c}\left(k(y,x) - \overline{k(x,y)}\right) = 0$$

holds for any $c \in \mathbf{C}$. This implies $k(y,x) = \overline{k(x,y)}$.

V. Nonparametric Inference with Positive Definite Kernels

Kenji Fukumizu

The Institute of Statistical Mathematics /
Graduate University for Advanced Studies

September 6-7
Machine Learning Summer School 2012, Kyoto

Outline

1. Mean and Variance on RKHS

2. Statistical tests with kernels

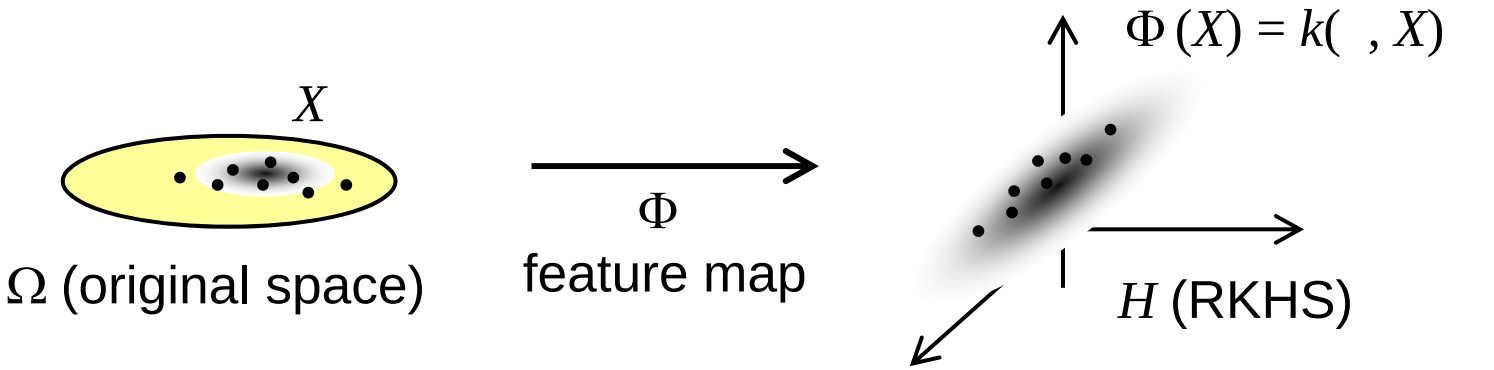
3. Conditional probabilities and beyond.

Introduction

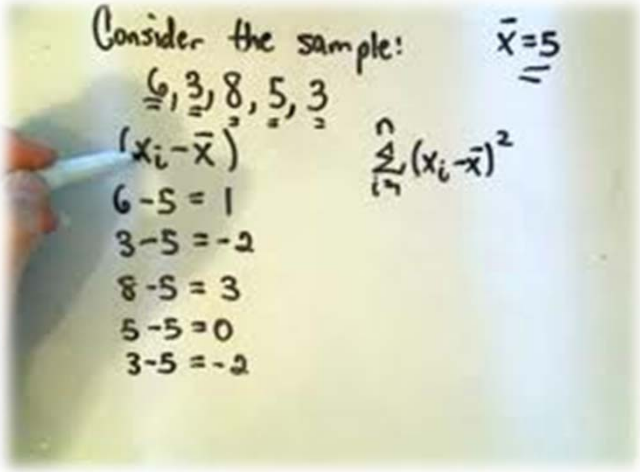
- "Kernel methods" for statistical inference
  - We have seen that kernelization of linear methods provides nonlinear methods, which capture 'nonlinearity' or 'high-order moments' of original data.

    e.g. nonlinear SVM, kernel PCA, kernel CCA, etc.

  - This section discusses more basic statistics on RKHS



Mean and covariance on RKHS



Mean on RKHS

X: random variable taking value on a measurable space Ω_X.

k: measurable positive definite kernel on Ω_X. H: RKHS.

Always assumes "bounded" kernel for simplicity: $\sup_x k(x,x) < \infty$.

$\Phi(X) = k(\cdot, X)$: feature vector = random variable on RKHS.

– Definition. The kernel mean $m_X \in H$ of X on H is defined by

$$m_X = E[\Phi(X)] = \int k(\cdot, x) dP(x).$$

– Reproducing expectation: $\langle m_X, f \rangle = E[f(X)] \quad (\forall f \in H)$

* Notation: m_X depends on k, also. But, we do not show it for simplicity.

Covariance operator

(X, Y): random variable taking values on Ω_X, Ω_Y, resp.

$(H_X, k_X), (H_Y, k_Y)$: RKHS given by kernels on Ω_X and Ω_Y, resp.

Definition. Cross-covariance operator: $\Sigma_{YX} : H_X \to H_Y$

$$\Sigma_{YX} \equiv E[\Phi_Y(Y) \otimes \Phi_X(X)^*] - E[\Phi_Y(Y)] \otimes E[\Phi_X(X)]^*$$

h^* denotes the linear functional $\langle h, \cdot \rangle$: $f \mapsto \langle h, f \rangle$

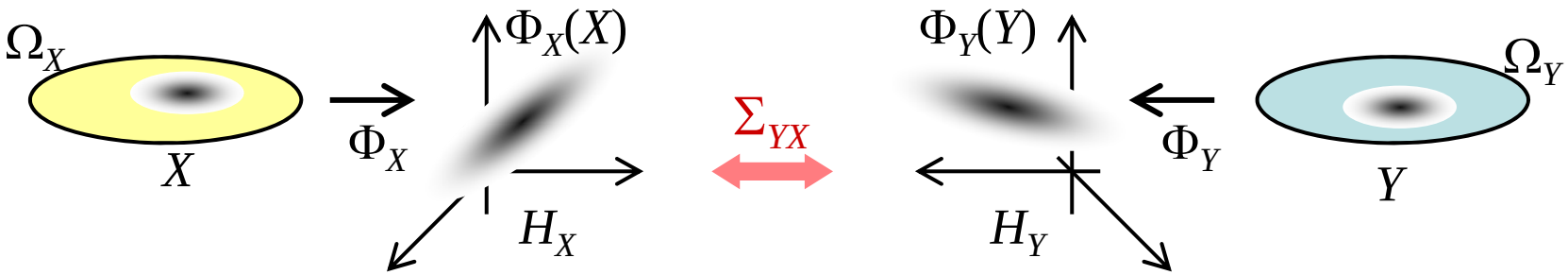
– Simply, covariance of feature vectors.

c.f. Euclidean case $V_{YX} = E[YX^T] - E[Y]E[X]^T$: covariance matrix

– Reproducing covariance:

$$\langle g, \Sigma_{YX} f \rangle = E[g(Y) f(X)] - E[g(Y)]E[f(X)] \ (= \mathrm{Cov}[f(X), g(Y)])$$

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



– Standard identification:

$$H^* \cong H: \quad \langle h, \cdot, \rangle \leftrightarrow \ h.$$

– The operator is regarded as an element in $H_Y \otimes H_X$, i.e.,

$$\Sigma_{YX} \equiv E[\Phi_Y(Y) \otimes \Phi_X(X)^*] - E[\Phi_Y(Y)] \otimes E[\Phi_X(X)]^*$$

can be regarded as

$$\Sigma_{YX} \cong m_{(Y,X)} - m_Y \otimes m_X \quad \in H_X \otimes H_Y$$

Characteristic kernel

(Fukumizu, Bach, Jordan 2004, 2009; Sriperumbudur et al 2010)

– Kernel mean can capture higher-order moments of the variable.

Example

X: $\mathbf{R}$-valued random variable. k: pos.def. kernel on $\mathbf{R}$.

Suppose k admits a Taylor series expansion on $\mathbf{R}$.

$$k(u,x) = c_0 + c_1(xu) + c_2(xu)^2 + \cdots \qquad (c_i > 0)$$

e.g.) $k(x,u) = \exp(xu)$

The kernel mean m_X works as a moment generating function:

$$m_X(u) = E[k(u,X)] = c_0 + c_1 E[X]u + c_2 E[X^2]u^2 + \cdots$$

$$\frac{1}{c_\ell}\frac{d^\ell}{du^\ell} m_X(u)\Big|_{u=0} = E[X^\ell]$$

$\mathcal{P}$: family of all the probabilities on a measurable space $(\Omega, \mathcal{B})$.

H: RKHS on Ω with a bounded measurable kernel k.

m_P: kernel mean on H for a variable with probability $P \in \mathcal{P}$

Definition. A bounded measureable positive definite k is called characteristic (w.r.t. $\mathcal{P}$) if the mapping

$$\mathcal{P} \to H, \qquad P \mapsto m_P$$

is one-to-one.

– The kernel mean with a characteristic kernel uniquely determines a probability.

$$m_P = m_Q \quad \Leftrightarrow \quad P = Q$$

i.e.

$$E_{X \sim P}[f(X)] = E_{X' \sim Q}[f(X')] \quad (\forall f \in H) \quad \Leftrightarrow \quad P = Q.$$

– Analogy to “characteristic function”

With Fourier kernel $F(x, y) = \exp\left(\sqrt{-1}\, x^T y\right)$

$$\text{Ch.f.}_X(u) = E[F(X, u)].$$

- The characteristic function uniquely determines a Borel probability on $\mathbf{R}^m$.
- The kernel mean $m_X(u) = E[k(u, X)]$ with a characteristic kernel uniquely determines a probability on $(\Omega, \mathcal{B})$.

  Note: Ω may not be Euclidean.

– The characteristic RKHS must be large enough!

Examples for $\mathbf{R}^m$ (proved later): Gaussian, Laplacian kernel.

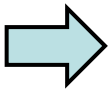
Polynomial kernels are not characteristic.

Statistical inference with kernel means

- Statistical inference: inference on some properties of probabilities.
- With characteristic kernels, they can be cast into the inference on kernel means.

Two sample problem: $P = Q$?

Independence test: $P_{XY} = P_X \otimes P_Q$?



$m_P = m_Q$?

$m_{(XY)} = m_X \otimes m_Y$?

Empirical Estimation

Empirical kernel mean

- An advantage of RKHS approach is easy empirical estimation.
- $X_1,\ldots,X_n$: i.i.d. → $\Phi(X_1),\ldots,\Phi(X_n)$: sample on RKHS

Empirical mean:

$$\hat{m}_X^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\Phi(X_i) = \frac{1}{n}\sum_{i=1}^{n}k(\cdot, X_i)$$

Theorem 5.1 (strong $\sqrt{n}$-consistency)

Assume $E[k(X,X)] < \infty$.

$$\left\|\hat{m}_X^{(n)} - m_X\right\| = O_p\left(1/\sqrt{n}\right) \qquad (n \to \infty).$$

Empirical covariance operator

$(X_1, Y_1), \ldots, (X_n, Y_n)$: i.i.d. sample on $\Omega_X \times \Omega_Y$.

An estimator of Σ_{YX} is defined by

$$\hat{\Sigma}_{YX}^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\{k_Y(\cdot, Y_i) - \hat{m}_Y\} \otimes \{k_X(\cdot, X_i) - \hat{m}_X\}$$

Theorem 5.2

$$\left\| \hat{\Sigma}_{YX}^{(n)} - \Sigma_{YX} \right\|_{HS} = O_p\left(1/\sqrt{n}\right) \qquad (n \to \infty)$$

– Hilbert-Schmidt norm: same as Frobenius norm of a matrix, but often used for infinite dimensional spaces.

$$||A||_{HS}^2 = \sum_{i=1}^{\dim H} \sum_{j=1}^{\dim H} \langle \psi_j, A\varphi_i \rangle^2$$

(sum of squares in matrix expression)

Statistical test with kernels

Two-sample problem

– Two sample homogeneity test

Two i.i.d. samples are given;

$$X_1,\ldots,X_n \sim P \qquad \text{and} \qquad Y_1,\ldots,Y_\ell \sim Q.$$

Q: Are they sampled from the same distribution?

Null hypothesis H_0: $P = Q$.

Alternative H_1: $P \neq Q$.

– Practically important: we often wish to distinguish two things.

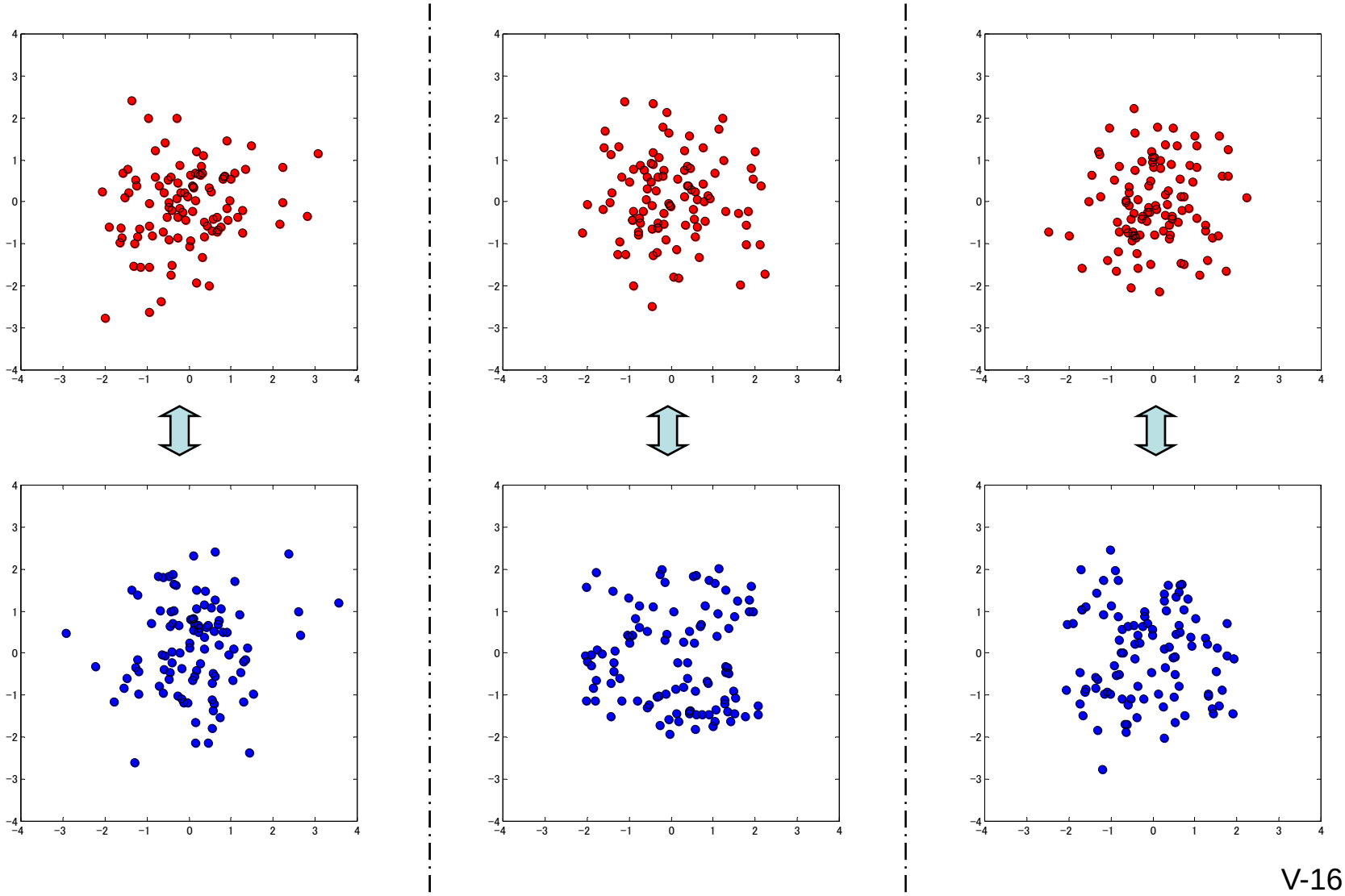
- Are the experimental results of treatment and control significantly different?
- Were the plays "*Henry VI*" and "*Henry II*" written by the same author?

– If then means of X and Y are different, we may use it for test. If they are identical, we need to look at higher order information.

– If mean and variance are the same, it is a very difficult problem.

– Example: do they have the same distribution?

$n = \ell = 100$



Kernel method for two-sample problem

(Gretton et al. NIPS 2007, 2010, JMLR2012).

- Kernel approach
  - Comparison of P_X and P_Y → comparison of m_X and m_Y.
- Maximum Mean Discrepancy
  - In population

$$MMD^2 = \|m_X - m_Y\|_H^2 = E[k(X, \tilde{X})] + E[k(Y, \tilde{Y})] - 2E[k(X, Y)].$$

($\tilde{X}, \tilde{Y}$: independent copy of X, Y)

$$\|m_X - m_Y\|_H = \sup_{f:\|f\|_H=1} |\langle f, m_X - m_Y \rangle| = \sup_{f:\|f\|_H=1} \left| \int f(x) dP(x) - \int f(x) dQ(x) \right|$$

hence, MMD.

  - With characteristic kernel, MMD = 0 if and only if $P_X = P_Y$.

– Test statistics:

Empirical estimator MMD^2_{emp}

$$MMD^2_{emp} = \left\| \hat{m}_X - \hat{m}_Y \right\|^2_H$$

$$= \frac{1}{n^2} \sum_{i,j=1}^{n} k(X_i, X_j) - \frac{2}{n\ell} \sum_{i=1}^{n} \sum_{a=1}^{\ell} k(X_i, Y_a) + \frac{1}{\ell^2} \sum_{a,b=1}^{\ell} k(Y_a, Y_b)$$

– Asymptotic distributions under H_0 and H_1 are known (see Appendix)

- Null distribution: $(n+\ell)MMD^2_{emp}$ → infinite mixture of χ^2
- Alternative: $\sqrt{n+\ell}\left(MMD^2_{emp} - MMD^2\right)$ → normal distribution

– Approximation of null distribution

- Approximation of the mixture coefficients.
- Fitting it with Pearson curve by moment matching.
- Bootstrap (Arcones & Gine 1992)

Experiments

Comparison of two databases.

Data size /	Dim
Neural I: 4000 /	63
Neural II: 1000 /	100
Health: 25 /	12600
Subtype: 25 /	2118

Data set	Attr.	MMD-B	WW	KS
Neural I	Same	96.5	97.0	95.0
	Different	0.0	0.0	10.0
Neural II	Same	94.6	95.0	94.5
	Different	3.3	0.8	31.8
Health	Same	95.5	94.7	96.1
	Different	1.0	2.8	44.0
Subtype	Same	99.1	94.6	97.3
	Different	0.0	0.0	28.4

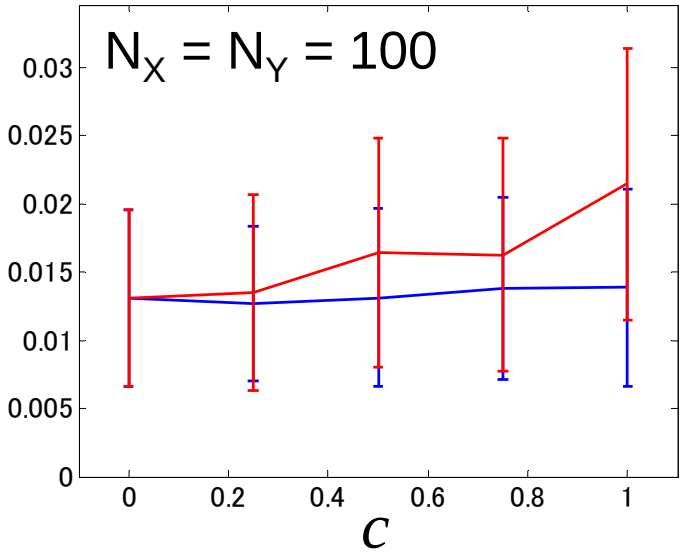
WW: Wald-Walfovitz test
KS: Kolmogorov-Smirnov test

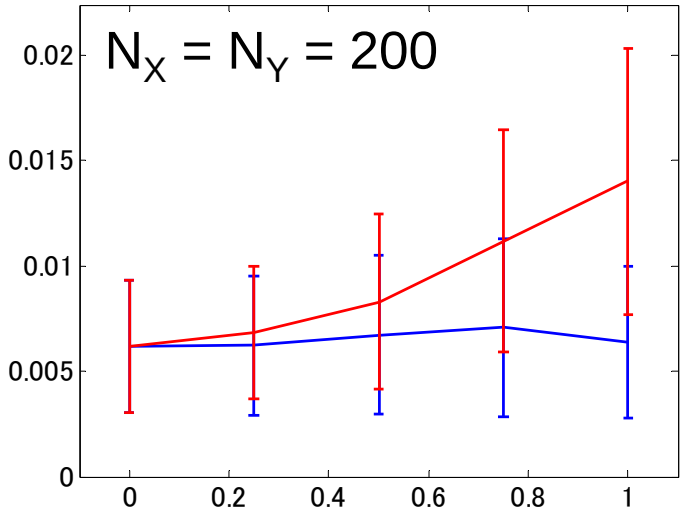
Classical methods (see Appendix)

Percentage of accepting $P = Q$.
Significance level $\alpha = 0.05$.

(Gretton et al. JMLR 2012)

Experiments for mixtures

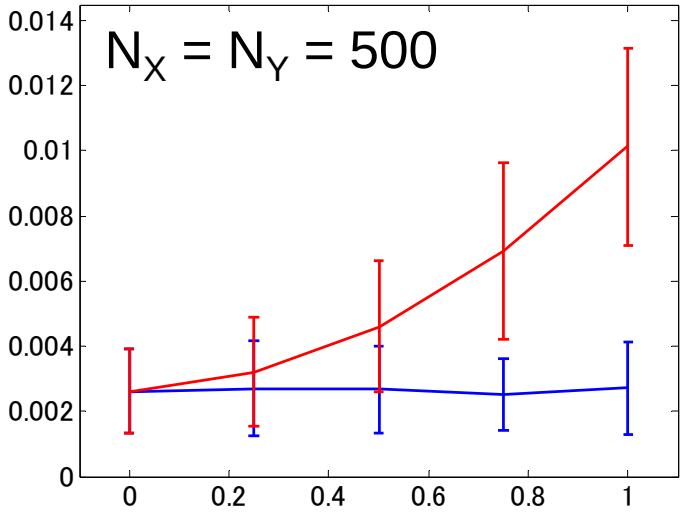




Average values of MMD
over 100 samples

$N(0,1)$ vs
c Unif $+ (1-c)\, N(0,1)$

$N(0,1)$ vs $N(0,1)$



Independence test

Independence

- X and Y are independent if and only if
  - Probability density function: $p_{XY}(x,y) = p_X(x)p_Y(y)$.
  - Cumulative distribution function: $F_{XY}(x,y) = F_X(x)F_Y(y)$.
  - Characteristic function:

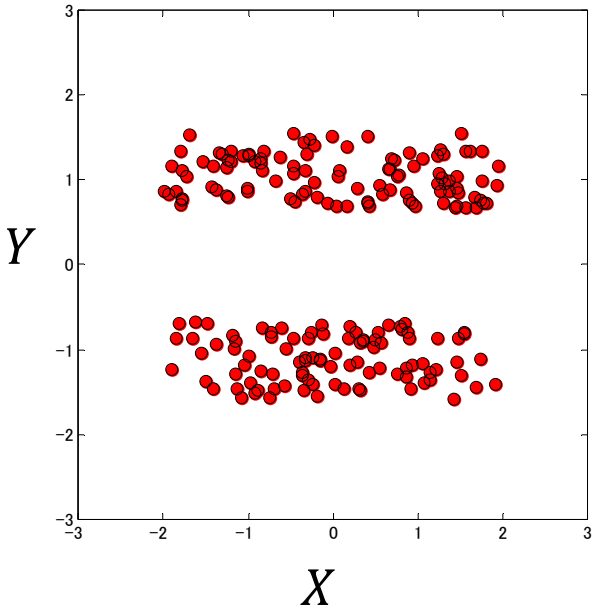
$$E\left[e^{\sqrt{-1}X^T u}e^{\sqrt{-1}Y^T v}\right] = E\left[e^{\sqrt{-1}X^T u}\right]E\left[e^{\sqrt{-1}Y^T v}\right]$$

Independence test

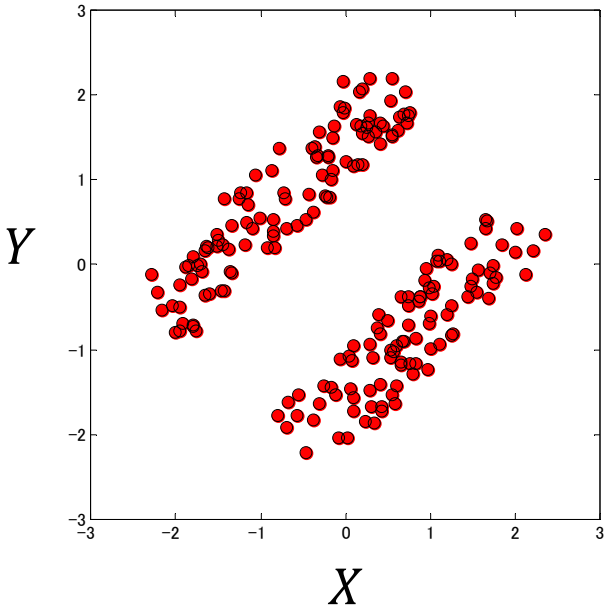
Given i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, are X and Y independent?

- Null hypothesis H_0: $P_{XY} = P_X \otimes P_Y$ (independent)
- Alternative H_1: $P_{XY} \neq P_X \otimes P_Y$ (not independent)

Independent



Dependent



Independence with kernel

Theorem 5.3. (Fukumizu, Bach, Jordan, JMLR 2004)

If the product kernel $k_X k_Y$ is characteristic, then

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad \Sigma_{YX} = O$$

Recall $\Sigma_{YX} \cong m_{YX} - m_Y \otimes m_X \in H_X \otimes H_Y$.

Comparison between m_{YX} (kernel mean of P_{YX}) and $m_Y \otimes m_X$ (kernel mean of $P_Y P_X$).

– Dependence measure: Hilbert-Schmidt independence criterion

$$HSIC(X,Y) := \|\Sigma_{YX}\|_{HS}^2$$

$$= \|m_{YX} - m_Y \otimes m_X\|_{H_Y \otimes H_X}^2 \qquad \text{[Exercise]}$$

$$= E\left[k_X(X,\tilde{X})k_Y(Y,\tilde{Y})\right] + E\left[k_X(X,\tilde{X})\right]E\left[k_Y(Y,\tilde{Y})\right]$$
$$-2E\left[E\left[k_X(X,\tilde{X})|X\right]E\left[k_Y(Y,\tilde{Y})|Y\right]\right]$$

$(\tilde{X},\tilde{Y})$: independent copy of (X,Y).

Independence test with kernels

(Gretton, Fukumizu, Teo, Song, Schölkopf, Smola. NIPS 2008)

– Test statistic:

$$\mathrm{HSIC}_{emp}(X,Y) := \left\|\hat{\Sigma}_{YX}^{(n)}\right\|_{HS}^2 = \frac{1}{n}\mathrm{Tr}[G_X G_Y]$$

G_X, G_Y: centered Gram matrices

Or equivalently,

$$\mathrm{HSIC}_{emp}(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} k_X(X_i,X_j)k_Y(Y_i,Y_j) - \frac{2}{n^3}\sum_{i,j,k=1}^{n} k_X(X_i,X_j)k_Y(Y_i,Y_k) + \frac{1}{n^4}\sum_{i,j=1}^{n} k_X(X_i,X_j)\sum_{k,\ell=1}^{n} k_Y(Y_k,Y_\ell)$$

– This is a special case of MMD comparing P_{XY} and $P_X \otimes P_Y$ with product kernel $k_X k_Y$.

– Asymptotic distributions are given similarly to MMD (Gretton et al 2009).

Comparison with existing method

- Distance covariance (Székely et al., 2007; Székely & Rizzo, 2009)
  - Distance covariance / distance correlation has gained popularity in statistical community as a dependence measure beyond Pearson correlation.

Definition. X, Y: m and ℓ-dimensional random vectors .

Distance covariance:

$$V^2(X,Y) := E[\|X - X'\| \, \|Y - Y'\|] + E\|X - X'\| \, E\|Y - Y'\| - 2E[E[\|X - X'\| \, |X] E[\|Y - Y'\| \, |Y]]$$

where (X', Y') and (X, Y) are i.i.d. $\sim P_{XY}$.

Distance correlation:

$$R(X,Y) := \frac{V(X,Y)}{\sqrt{V(X,X)V(Y,Y)}}$$

Relation between distance covariance and MMD

Proposition 5.4 (Sejdinovic, Gretton, Sriperumbudur, F. ICML2012)

A kernel on Euclidean spaces

$$k(x, y) = \|x\| + \|y\| - \|x - y\|.$$

is positive definite, and

$$\mathrm{HSIC}_k^2(X, Y) = V^2(X, Y).$$

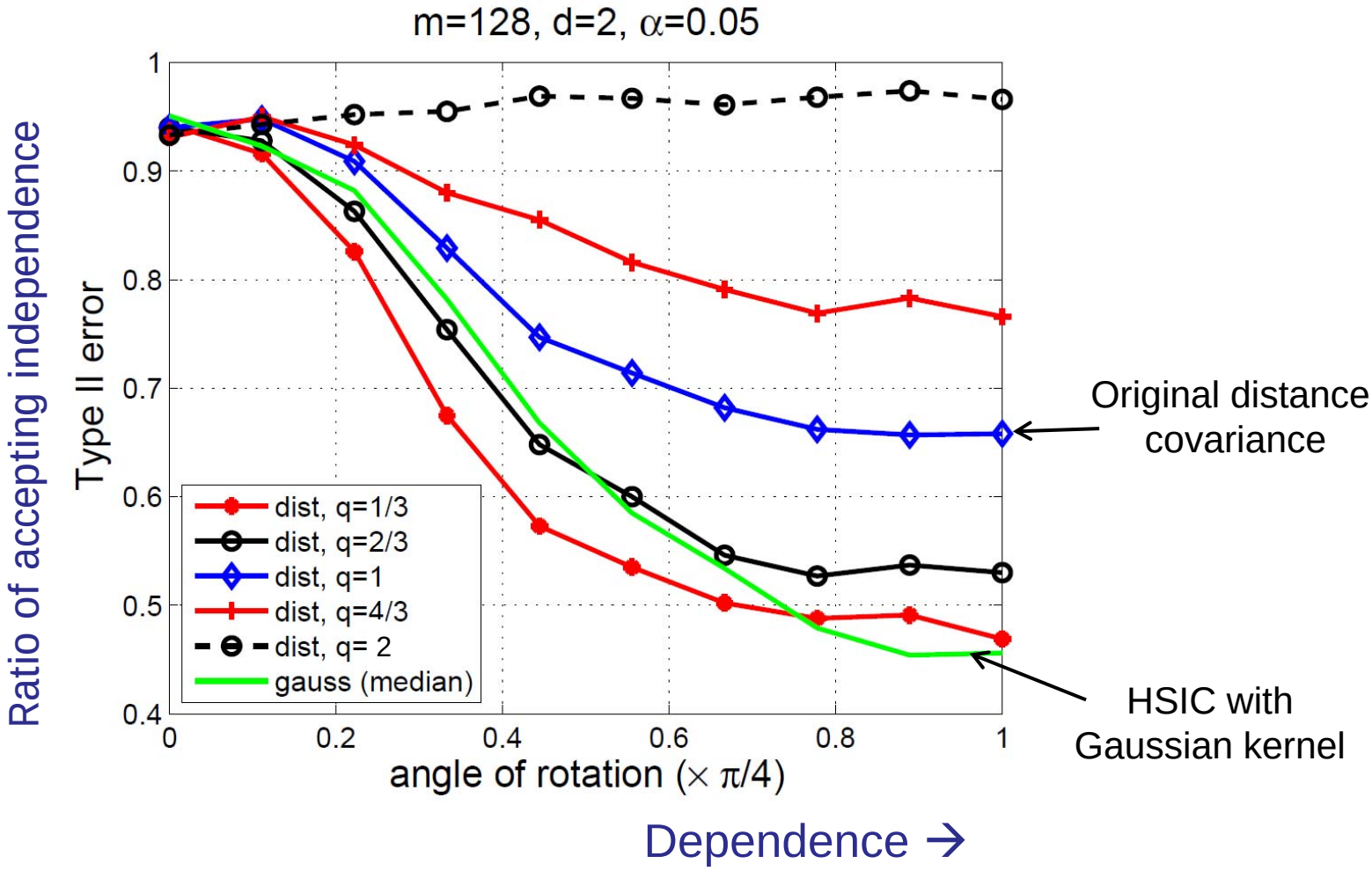
- Distance covariance is a specific instance of MMD.
- Positive definite kernel approach is more general in choosing kernels, and thus may perform better.
- Extension

$$k_q(x, y) = \|x\|_q + \|y\|_q - \|x - y\|_q \qquad \left(\|z\|_q = \left(\sum_{j=1}^m |z_j|^q\right)^{\frac{1}{q}}\right)$$

k_q is positive definite for $0 < q \leq 2$. We can define

$$V_q^2(X, Y) \coloneqq \mathrm{HSIC}_{k_q}^2(X, Y)$$

Experiments



Choice of kernel for MMD

- Heuristics for Gaussian kernel:

$$\sigma = \text{median}\left\{ \left\| X_i - X_j \right\| \mid i, j = 1, \ldots, n \right\}$$

- Using performance of statistical test:

  Type II error of the test statistics (Gretton et al NIPS 2012).

Challenging open questions.

Conditional probabilities and beyond

Conditional probability

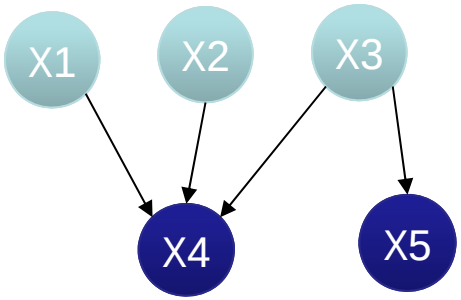
– Conditional probabilities appear in many machine learning problems

- Regression / classification: direct inference of $E[Y|X]$ or $p(y|x)$.
  → already seen in Section II.

- Bayesian inference

$$q(y|x) = \frac{p(y|x)\pi(x)}{\int p(y|x')\pi(x')dx'}$$

- Conditional independence / dependence
  – Graphical modeling
    Conditional independent relations among variables.
  – Causality
  – Dimension reduction / feature extraction



Conditional kernel mean

Conditional kernel mean

Definition.

$$E[\Phi_Y(Y)|X = x] = \int k_Y(\cdot, Y)p(y|x)dy$$

- Simply, kernel mean of $p(y|x)$.
- It determines the conditional probability with a characteristic kernel.
- Again, inference problems on conditional probabilities can be solved as inference on conditional kernel means.

- But, how can we estimate it?

Covariance operator revisited

(Uncentered) cross-covariance operator

X, Y: random variables on Ω_X, Ω_Y,

k_X, k_Y: positive definite kernels on Ω_X, Ω_Y, $E[k_X(X,X)k_Y(Y,Y)] < \infty$.

Definition. Uncentered cross-covariance operator

$$C_{YX} \equiv E[\Phi_Y(Y) \otimes \Phi_X(X)^*] \quad : \quad H_X \to H_Y$$

(Or $C_{YX} \in H_Y \otimes H_X^* \cong H_Y \otimes H_X$)

– Reproducing property:

$$\langle g, C_{YX} f \rangle_{H_Y} = E[g(Y)f(X)] \qquad (\forall g \in H_Y, f \in H_X)$$

$$\langle f_2, C_{XX} f_1 \rangle_{H_Y} = E[f_2(X)f_1(X)] \qquad (\forall f_1, f_2 \in H_X)$$

Conditional probability by regression

– Recall for zero-mean Gaussian random variable (X, Y),

$$E[Y|X = x] = V_{YX}V_{XX}^{-1}x.$$

- Given by the solution to the least mean square

$$\int \|Y - AX\|^2 dP(X, Y)$$

– For the feature vector

$$E[\Phi_Y(Y)|X = x] = C_{YX}C_{XX}^{-1}\Phi_X(x).$$

- Given by the solution to the least mean square

$$\int \|\Phi_Y(Y) - A\Phi_X(X)\|_{H_Y}^2 dP(X, Y)$$

C_{XX}^{-1} is not well defined in infinite dimensional cases, but regularized estimator can be justified.

Estimator for conditional kernel mean

– Empirical estimation: given $(X_1, Y_1), \ldots, (X_n, Y_n)$,

$$\hat{C}_{YX}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1} \Phi_X(x)$$

In Gram matrix expression,

$$\mathbf{k}_Y(*)^T (G_X + n\varepsilon_n I_n)^{-1} \mathbf{k}_X(x)$$

$$\mathbf{k}_X(x) = \begin{pmatrix} k_X(X_1, x) \\ \vdots \\ k_X(X_n, x) \end{pmatrix}, \quad \mathbf{k}_Y(*) = \begin{pmatrix} k_Y(*, Y_1) \\ \vdots \\ k_Y(*, Y_n) \end{pmatrix}.$$

Proposition 5.5 (Consistency)

If k_X is characteristic, $\frac{p(y,x)}{p(y)p(x)} \in H_Y \otimes H_X$, and $\varepsilon_n \to 0, \varepsilon_n \sqrt{n} \to \infty$ as $n \to \infty$, then for every x

$$\hat{C}_{YX}\left(\hat{C}_{XX} + \varepsilon_n I\right)^{-1} k_X(\cdot, x) \to E[\Phi_Y(Y)|X = x] \text{ in } H_Y$$

in probability.

Conditional covariance

Review: Gaussian variables

Conditional covariance matrix: $V_{YX|Z} := V_{YX} - V_{YZ}V_{ZZ}^{-1}V_{ZX}$

Fact: $V_{YX|Z} = \mathrm{Cov}[Y, X|Z = z]$ for any z

Conditional cross-covariance operator

Definition: X, Y, Z: random variables on $\Omega_X, \Omega_Y, \Omega_Z$.

k_X, k_Y, k_Z: positive definite kernel on $\Omega_X, \Omega_Y, \Omega_Z$.

Conditional cross-covariance operator

$$C_{YX|Z} \equiv C_{YX} - C_{YZ}C_{ZZ}^{-1}C_{ZX}$$

– Reproducing averaged conditional covariance

Proposition 5.6 If k_Z is characteristic, then for $\forall f \in H_X$, $g \in H_Y$,

$$\begin{aligned}\langle g, C_{YX|Z} f \rangle_{H_Y} &= E\big[\mathrm{Cov}[f(X), g(Y)|Z]\big] \\ &= E[f(X)g(Y)] - E\big[E[f(X)|Z]E[g(Y)|Z]\big]\end{aligned}$$

– An interpretation: Compare the conditional kernel means for $p(y,x|z)$ and $p(y|z)p(x|z)$.

$$E[\Phi_Y(Y) \otimes \Phi_X(X)|Z=z] - E[\Phi_Y(Y)|Z=z] \otimes E[\Phi_X(X)|Z=z]$$

Dependence on z is not easy to handle → Average it out.

$$E[\Phi_Y(Y) \otimes \Phi_X(X)] - E[\, E[\Phi_Y(Y)|Z] \otimes E[\, \Phi_X(X) \mid Z \,]]$$

$$C_{YX} \quad - \quad C_{YZ}C_{ZZ}^{-1} \cdot C_{ZZ} \cdot C_{ZZ}^{-1}C_{ZX}$$

– Empirical estimator:

$$\hat{C}_{YX|Z} \equiv \hat{C}_{YX} - \hat{C}_{YZ}\left(\hat{C}_{ZZ} + \varepsilon_n I\right)^{-1}\hat{C}_{ZX}$$

Conditional independence

– Recall: for Gaussian random variable

$$V_{YX|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z.$$

– By average over Z, $C_{YX|Z} = O$ does not imply conditional independence, which requires $p(y, x|z) = p(y|z)p(x|z)$ for each z.

– Trick: consider

$$C_{\ddot{Y}X|Z} \coloneqq C_{\ddot{Y}X} - C_{\ddot{Y}Z}\, C_{ZZ}^{-1} C_{ZX},$$

where $\ddot{Y} = (Y, Z)$ and the product kernel $k_Y k_Z$ is used for $\ddot{Y}$.

Theorem 5.7 (Fukumizu et al. JMLR 2004)

Assume k_X, k_Y, k_Z are characteristic, then

$$C_{\ddot{Y}X|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z.$$

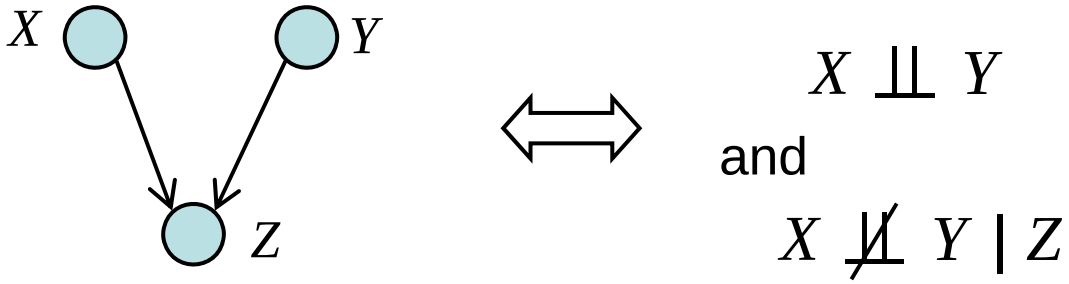
– $C_{Y\ddot{X}|Z}$, $C_{\ddot{Y}\ddot{X}|Z}$ can be similarly used.

Applications of conditional dependence measures

- Conditional independence test (Fukumizu et al. NIPS2007)
  - Squared HS-norm $\left\|\hat{C}_{\ddot{Y}\ddot{X}|Z}\right\|_{HS}^2$ can be used for conditional independence test.
  - Unlike the independence test, the asymptotic null distribution is not available. Permutation test is needed.
  - Background: The conditional independent test with continuous non-Gaussian variables is not easy, and a challenging open problem.

Causal inference

- Directional acyclic graph (DAG) is used for representing the causal structure among variables.
- The structure can be learned by conditional independence tests. The above test can be applied (Sun, Janzing, Schölkopf, F. ICML2007).



Feature extraction / dimension reduction for supervised learning → see next.

Dimension reduction and conditional independence

Dimension reduction for supervised learning

Input: $X = (X_1, \ldots, X_m)$, Output: Y (either continuous or discrete)

Goal: find an effective dimension reduction space (EDR space) spanned by an m x d matrix B s.t.

$$p_{Y|X}(Y \mid X) = p_{Y|B^T X}(Y \mid B^T X) \qquad \text{where } B^T X = (b_1{}^T X, \ldots, b_d{}^T X)$$

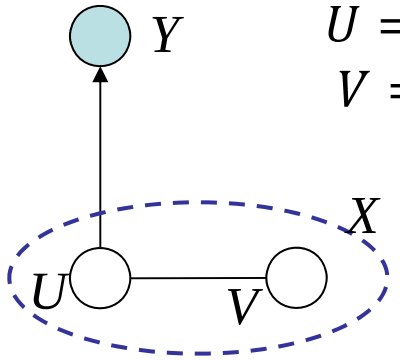
linear feature vector

No further assumptions on cond. p.d.f. p.

Conditional independence

B spans effective subspace

$$\Longleftrightarrow \qquad X \perp\!\!\!\perp Y \mid B^T X$$



$$U = B^T X,$$
$$V = B_\perp^T X$$

Gradient-based method

(Samarov 1993; Hristache et al 2001)

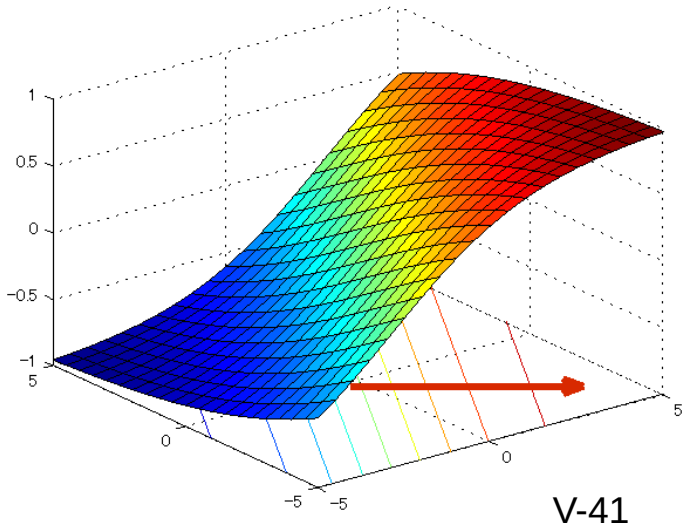
Average Derivative Estimation (ADE)

- Assumptions:
  - Y is one dimensional
  - EDR space $\quad p(Y \mid X) = \tilde{p}(Y \mid B^T X)$

$$\frac{\partial}{\partial x} E[Y \mid X = x] = \frac{\partial}{\partial x} \int y \tilde{p}(y \mid B^T x) dy = B \int y \frac{\partial}{\partial z} \tilde{p}(y \mid z)\Big|_{z=B^T x} dy$$

- Gradient of the regression function lies in the EDR space at each x

→ $\hat{B}$ = average or PCA of the gradients at many x.



Kernel Helps!

– Weakness of ADE:
  - Difficulty of estimating gradients in high dimensional space.
    ADE uses local polynomial regression.
    » Sensitive to bandwidth
  - May find only a subspace of the effective subspace.
    e.g. $Y \sim f(X_1) + Z, \qquad Z \sim N(0, \sigma(X_2)^2).$

– Kernel method
  - Can handle conditional probability in regression form $E[\Phi(Y)|X = x]$
  - Characterizes conditional independence $X \perp\!\!\!\perp Y \mid B^T X$

Derivative with kernel

– Reproducing the derivative (e.g. Steinwart & Christmann, Chap. 4):

Assume $k_X(x, \tilde{x})$ is differentiable and $\frac{\partial k_X(\cdot, x)}{\partial x} \in H_X$ then

$$\left\langle f, \frac{\partial k_X(\cdot, x)}{\partial x} \right\rangle = \frac{\partial f(x)}{\partial x} \qquad \text{for any } f \in H_X$$

– Combining with the estimation of conditional kernel mean,

$$\hat{M}_{ij}(x) = \left\langle \frac{\partial \hat{E}[\Phi_Y(Y) \mid X = x]}{\partial x^i}, \frac{\partial \hat{E}[\Phi_Y(Y) \mid X = x]}{\partial x^j} \right\rangle$$

$$= \left\langle \hat{C}_{YX}(\hat{C}_{XX} + \varepsilon_n I)^{-1} \frac{\partial k_X(\cdot, x)}{\partial x_i}, \hat{C}_{YX}(\hat{C}_{XX} + \varepsilon_n I)^{-1} \frac{\partial k_X(\cdot, x)}{\partial x_j} \right\rangle$$

The top d eigenvectors of $\widehat{M}(x)$ estimates the EDR space

Gradient-based kernel dimension reduction (gKDR)

(Fukumizu & Leng, NIPS 2012)

– Method

- Compute

$$\tilde{M}_n = \frac{1}{n}\sum_{i=1}^{n} \nabla \mathbf{k}_X(X_i)^T (G_X + n\varepsilon_n I)^{-1} G_Y (G_X + n\varepsilon_n I)^{-1} \nabla \mathbf{k}_X(X_i).$$

$$\nabla \mathbf{k}_X(X_i) = \left(\frac{\partial k_X(X_1, x)}{\partial x}, \ldots, \frac{\partial k_X(X_n, x)}{\partial x} \right)^T_{x=X_i} \qquad G_X = \left(k_X(X_i, X_j) \right)$$

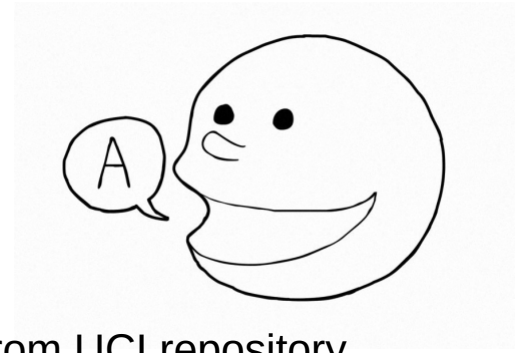
- Compute top d eigenvectors of $\tilde{M}_n$. → Estimator $\hat{B}$.

– gKDR estimate the subspace to realize the conditional independence

– Choice of kernel:

Cross-validation with some regressor/classifier, e.g. kNN method.

Experiment: ISOLET



- Speech signals of 26 alphabets
- 617 dim. (continuous)
- 6238 training data / 1559 test data

Data from UCI repository.

- Results

Dim	10	15	20	25	30	35	40	45	50
gKDR	**14.43**	**7.50**	**5.00**	**4.75**	-	-	-	-	-
gKDR-v	**16.87**	**7.57**	**4.75**	**4.30**	**3.85**	**3.85**	**3.59**	**3.53**	**3.08**
CCA	**13.09**	**8.66**	**6.54**	**6.09**	-	-	-	-	-

Classification errors for test data by SVM (%)

c.f. C4.5 + ECOC: 6.61%
Neural Networks (best): 3.27%

Experiment: Amazon Commerce Reviews

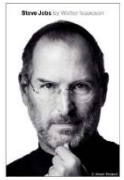
- Author identification for Amazon commerce reviews.
- Dim = 10000 (linguistic style: e.g. usage of digit, punctuation, words and sentences' length, and frequency of words, etc)
- n = #authors x 30

(Data from UCI repository.)

★☆☆☆☆ **Worst book I've ever read**, October 31, 2011

By [redacted] - See all my reviews

REAL NAME



amazon.com

Amazon Verified Purchase (What's this?)

This review is from: Steve Jobs (Hardcover)

I want to save you the time so here's a one line piece of advice: grab it from your library instead of buying it and you'll thank yourself later.

Like many people, I pre-ordered this book, read it as soon as it came and was very disappointed. The book is literally a description of events in chronological order. If you are the type that reads the news, especially tech news, you'll learn very little here. There is no attempt at analysis and the author never bothered to ask Steve (or his colleagues, friends or even his wife and kids) "Why?". If anything, it proves that the author Isaacson is a terrible interviewer. It takes great skill to draw out from a human being his/her motivations and reasoning. I expect a biographer to possess that skill. Unfortunately, Isaacson has done nothing more than compile a large number of facts and put them together in order of Apple product releases. You'll find yourself saying "Is that all Steve was?" because he comes across as an extremely shallow person.

Perhaps the issue lies in Isaacson's view of Steve Jobs' place in history. Elsewhere, he's speaks of Einstein and Jobs as being comparable. I wish I'd read that before.

This is really just a terrible piece of work.

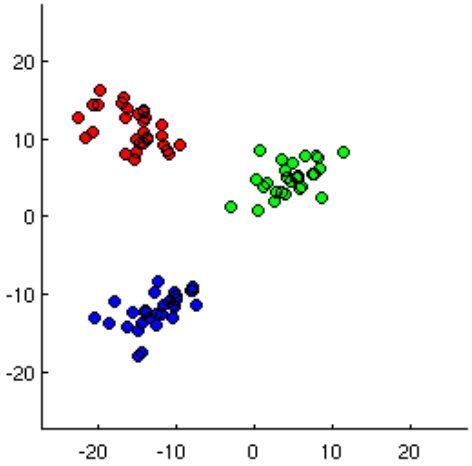
Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report abuse | Permalink

Comments (9)

– Example of 2-dim plots
  for 3 authors



– gKDR (dim = #authors)
  vs correlation-based variable selection (dim=500/2000)

#Authors	gKDR + 5NN	gKDR + SVM	Corr (500) + SVM	Corr (2000) + SVM
10	9.3	12.0	15.7	8.3
20	16.2	16.2	30.2	18.0
30	20.1	18.0	29.2	24.0
40	22.8	21.8	35.4	25.0
50	22.7	19.5	41.1	29.0

10-fold cross-validation errors (%)

Bayesian inference with kernels

(Fukumizu, Song, Gretton NIPS 2011)

Bayes' rule

$$q(x|y) = \frac{p(y|x)\pi(x)}{q(y)}, \qquad q(y) = \int p(y|x)\pi(x)dx.$$

Kernel Bayes' rule

- π → m_Π: kernel mean of prior, $\hat{m}_\Pi = \sum_{j=1}^{\ell} \gamma_j \; k_X(\cdot, U_j)$
- $p(y|x)$ → C_{YX}: kernel representation of relation between X and Y, $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$, i.i.d.
- Goal: compute kernel mean of posterior $q(x|y)$.

$$\hat{m}_{q_{X|Y=y}} = \sum_{i=1}^{n} \; w_i(y) k_X(\cdot, X_i)$$

$$w(y) = R_{X|Y}\mathbf{k}_Y(y)$$
$$R_{X|Y} = \Lambda G_Y \left((\Lambda G_Y)^2 + \delta_n I_n\right)^{-1}\Lambda$$
$$\Lambda = \mathrm{Diag}\left((G_X + n\varepsilon_n I_n)^{-1} G_{XU}\gamma\right)$$

Bayesian inference using kernel Bayes' rule

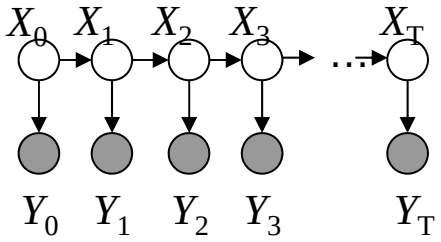
- NO PARAMETRIC MODELS, BUT SAMPLES!

- When is it useful?
  - Explicit form of cond. p.d.f. $p(y|x)$ or prior $\pi(x)$ is unavailable, but sampling is easy.
    - Approximate Bayesian Computation (ABC), Process prior.
  - Cond. p.d.f. $p(y|x)$ is unknown, but sample from $p(x, y)$ is given in training phase.
    - Example: nonparametric HMM (shown later).
  - If both of $p(y|x)$ and $\pi(x)$ are known, there are many good sampling methods, such as MCMC, SMC, etc. But, they may take long time. KBR uses matrix operations.

Application: nonparametric HMM

Model: $p(X,Y) = \pi(X_0) \prod_{t=0}^{T} p(Y_t|X_t) \prod_{t=0}^{T-1} q(X_{t+1}|X_t)$

– Assume:
  $p(y_t|x_t)$ and/or $q(x_t|x_{t-1})$ is not known.
  But, data $(X_t, Y_t)_{t=0}^{T}$ is available
  in training phase.



  Examples:
  - Measurement of hidden states is expensive,
  - Hidden states are measured with time delay.

– Testing phase (e.g., filtering, e.g.):
  given $\tilde{y}_0, \ldots, \tilde{y}_t,$ estimate hidden state x_t.

– Sequential filtering/prediction uses Bayes' rule ➔ KBR applied.

Camera angles

- Hidden X_t: angles of a video camera located at a corner of a room.
- Observed Y_t: movie frame of a room + additive Gaussian noise.
- X_t: 3600 downsampled frames of 20 x 20 RGB pixels (1200 dim.).
- The first 1800 frames for training, and the second half for testing

noise	KBR (Trace)	Kalman filter(Q)
$\sigma^2 = 10^{-4}$	$0.15 \pm < 0.01$	0.56 ± 0.02
$\sigma^2 = 10^{-3}$	0.21 ± 0.01	0.54 ± 0.02

Average MSE for camera angles (10 runs)

To represent SO(3) model, $\mathrm{Tr}[AB^{-1}]$ for KBR, and quaternion expression for Kalman filter are used .

Summary of Part V

- Kernel mean embedding of probabilities
  - Kernel mean gives a representation of probability distribution.
  - Inference on probabilities can be cast into inference on kernel means. e.g. two sample test, independent test

- Conditional probabilities
  - Conditional probabilities can be handled with kernel means and covariances
    - Conditional independence test
    - Graphical modeling
    - Causal inference
    - Dimension reduction
    - Bayesian inference

References

Fukumizu, K., Bach, F.R., and Jordan, M.I. (2004) Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*. 5:73-99,

Fukumizu, K., F.R. Bach and M. Jordan. (2009) Kernel dimension reduction in regression. *Annals of Statistics*. 37(4), pp.1871-1905

Fukumizu, K., L. Song, A. Gretton (2011) Kernel Bayes' Rule. *Advances in Neural Information Processing Systems 24* (NIPS2011) 1737-1745.

Gretton, A., K.M. Borgwardt, M.Rasch, B. Schölkopf, A.J. Smola (2007) A Kernel Method for the Two-Sample-Problem. *Advances in Neural Information Processing Systems 19*, 513-520.

Gretton, A., Z. Harchaoui, K. Fukumizu, B. Sriperumbudur (2010) A Fast, Consistent Kernel Two-Sample Test. *Advances in Neural Information Processing Systems 22*, 673-681.

Gretton, A., K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, A. Smola. (2008) A Kernel Statistical Test of Independence. *Advances in Neural Information Processing Systems 20*, 585-592.

Gretton, A., K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, A. Smola. (2008) A Kernel Statistical Test of Independence. *Advances in Neural Information Processing Systems 20*, 585-592.

Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny. (2001) Structure Adaptive Approach for Dimension Reduction. Annals of Statsitics, 29(6):1537-1566.

Samarov, A. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of American Statistical Association*. 88, 836-847.

Sejdinovic, D., A. Gretton, B. Sriperumbudur, K. Fukumizu. (2012) Hypothesis testing using pairwise distances and associated kernels. *Proc. 29th International Conference on Machine Learning* (ICML2012).

Sriperumbudur, B.K., A. Gretton, K. Fukumizu, B. Schölkopf, G.R.G. Lanckriet. (2010) Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*. 11:1517-1561.

Sun, X., D. Janzing, B. Schölkopf and K. Fukumizu. (2007) A kernel-based causal learning algorithm. *Proc. 24th Annual International Conference on Machine Learning (ICML2007)*, 855-862.

Székely, G.J., M.L. Rizzo, and N.K. Bakirov. (2007) Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6): 2769-2794.

Székely, G.J. and M.L. Rizzo. (2009) Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236-1265.

Appendices

Statistical Test: quick introduction

How should we set the threshold?

Example) Based on MMD, we wish to make a decision whether the two variables have the same distribution.

Simple-minded idea: Set a small value like $t = 0.001$

$$MMD(X,Y) < t \Longrightarrow \text{Perhaps, same}$$

$$MMD(X,Y) \geq t \Longrightarrow \text{Different}$$

But, the threshold should depend on the properties of X and Y.

Statistical hypothesis test

- A statistical way of deciding whether a hypothesis is true or not.
- The decision is based on sample → We cannot be 100% certain.

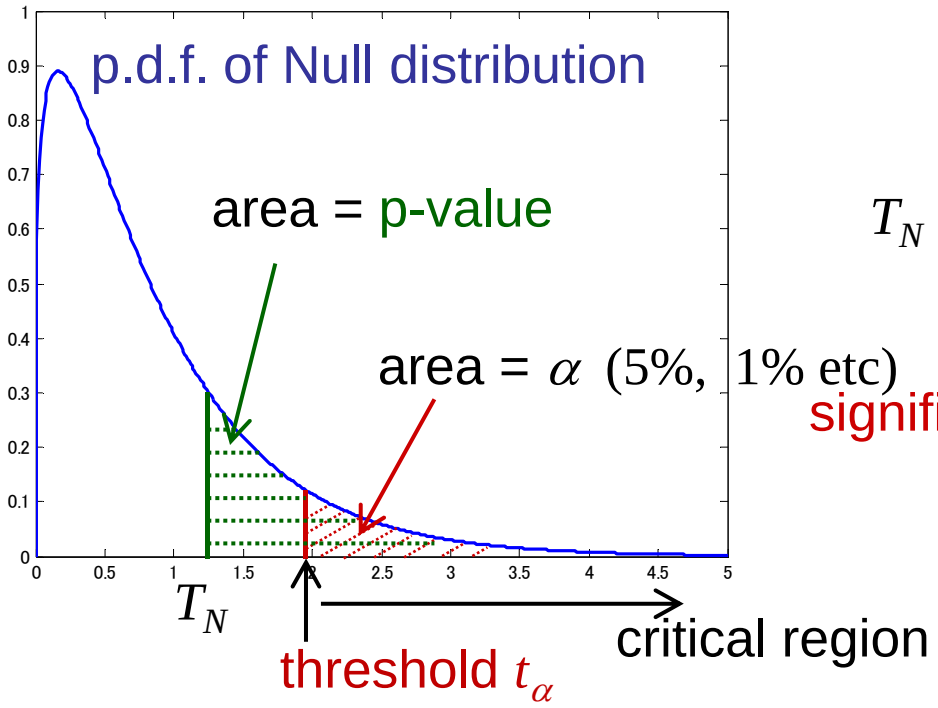
Procedure of hypothesis test

- Null hypothesis H_0 = hypothesis assumed to be true
  "X and Y have the same distribution"
- Prepare a test statistic T_N
  e.g. $T_N = MMD_{emp}^2$
- Null distribution: Distribution of T_N under H_0
- Set significance level α Typically $\alpha = 0.05$ or 0.01
- Compute the critical region: α = Pr($T_N > t_\alpha$ under H_0)
- Reject the null hypothesis if $T_N > t_\alpha$

  The probability that $MMD_{emp}^2 > t_\alpha$ under H0 is very small.

otherwise, accept H_0 negatively.

One-sided test



$T_N > t_\alpha \Leftrightarrow$ p-value < α

- If H_0 is the truth, the value of T_N should follow the null distribution.
- If H_1 is the truth, the value of T_N should be very large.
- Set the threshold with risk α.
- The threshold depends on the distribution of the data.

Type I and Type II error

- Type I error = false positive (e.g. "$P \neq Q$" = positive)
- Type II error = false negative

TEST RESULT	TRUTH: H_0	TRUTH: Alternative
Accept H_0	True negative	Type II error False negative
Reject H_0	Type I error False positive	True positive

- Significance level controls the type I error.
- Under a fixed type I error, a good test statistics should give small type II error.

MMD: Asymptotic distribution

Under H_0

$X_1, \ldots, X_n \sim P$, $Y_1, \ldots, Y_\ell \sim Q$: i.i.d. Let $N = n + \ell$.

Assume $\frac{n}{N} \to \gamma$, $\frac{\ell}{N} \to (1-\gamma)$ $(0 < \gamma < 1)$ as $N \to \infty$.

Under the null hypothesis of $P = Q$,

$$N\,\mathrm{MMD}^2_{emp} \underset{law}{\Longrightarrow} \sum_i^{\infty} \lambda_i \left(Z_i^2 - \frac{1}{\gamma(1-\gamma)} \right) \quad (N \to \infty),$$

where $Z_1, Z_2, \ldots$ are i.i.d. with law $N(0; 1/\gamma(1-\gamma))$, and $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of the integral operator on $L^2(P)$

$$Tf = \int \tilde{k}(x,y) f(y) dP(y)$$

with $\tilde{k}$ the centered kernel

$$\tilde{k}(x,y) = k(x,y) - E[k(x,X)] - E[k(X,y)] + E\left[k\left(X,\tilde{X}\right)\right].$$

$\tilde{X}$: independent copy of X.

Under H_1

Under the alternative $P \neq Q$,

$$\sqrt{N}\left(\mathrm{MMD}^2_{\mathrm{emp}} - MMD^2\right) \underset{law}{\Longrightarrow} N(0;\sigma^2) \qquad (N \to \infty),$$

where

$$\sigma^2 = 4\left(\frac{\mathrm{Var}\left[E[k(X,\tilde{X}) - k(X,Y)|X]\right]}{\gamma} + \frac{\mathrm{Var}\left[E[k(Y,\tilde{Y}) - k(X,Y)|Y]\right]}{1-\gamma}\right).$$

- The asymptotic distributions are derived by the general theory of U-statistics (e.g. see van der Vaart 1998, Chapter 12).
- Estimation of the null distribution:
  - Estimation of λ_i
  - Approximation by Pearson curve with moment matching.
  - Bootstrap MMD (Arcones & Gine 1992)

Conventional methods for two sample problem

- Kolmogorov-Smirnov (K-S) test for two samples

  One-dimensional variables

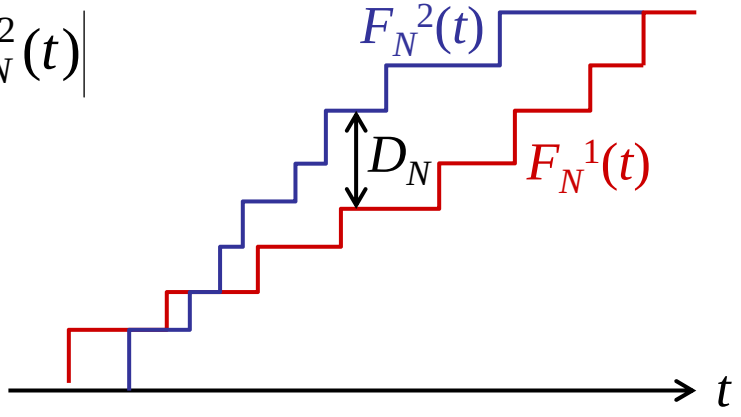
  – Empirical distribution function

$$F_N(t) = \frac{1}{N}\sum_{i=1}^{N} I(X_i \le t)$$

  – KS test statistics

$$D_N = \sup_{t \in \mathbf{R}} \left| F_N^1(t) - F_N^2(t) \right|$$

  – Asymptotic null distribution is known (not shown here).

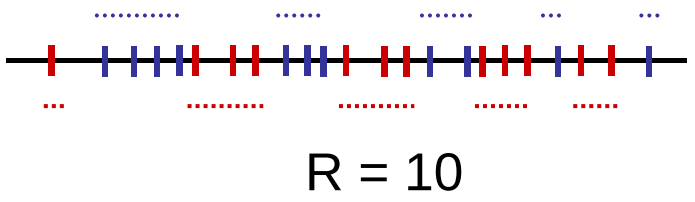


Wald-Wolfowitz run test

One-dimensional samples

- Combine the samples and plot the points in ascending order.
- Label the points based on the original two groups.
- Count the number of “runs”, i.e. consecutive sequences of the same label.
- Test statistics R = Number of runs

$$T_N = \frac{R - E[R]}{\sqrt{Var[R]}} \quad \Rightarrow \quad N(0,1)$$



R = 10

- In one-dimensional case, less powerful than KS test

Multidimensional extension of KS and WW test

- Minimum spanning tree is used (Friedman Rafsky 1979)