Kernel Methods for Dependence and Causality

Kenji Fukumizu

Institute of Statistical Mathematics, Tokyo

Max-Planck Institute for Biological Cybernetics

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

Machine Learning Summer School 2007

August 20-31, Tübingen, Germany

Overview


ISM
統計数理研究所
Image © 2006 Digital Earth Technology
© 2006 Europa Technologies
© 2006 ZENRIN
Google
23 m
ポインタ 35° 39'03.23" N 139° 43'39.34" E 高度 27 m ストリーミング 100% 上空 110 m

Outline of This Lecture

Kernel methodology of inference on probabilities

I. Introduction

II. Dependence with kernels

III. Covariance on RKHS

IV. Representing a probability

V. Statistical test

VI. Conditional independence

VII. Causal inference

I. Introduction

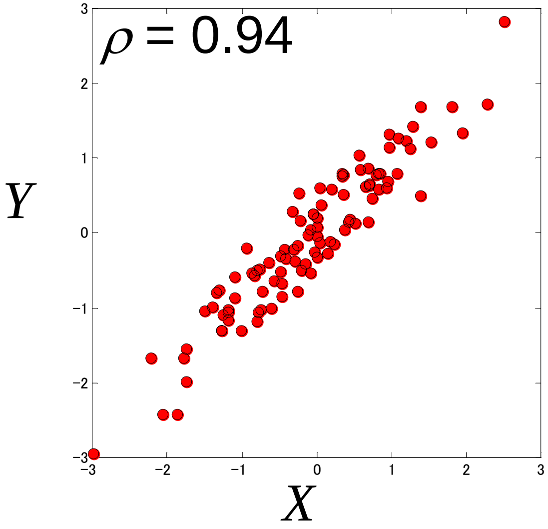
Dependence

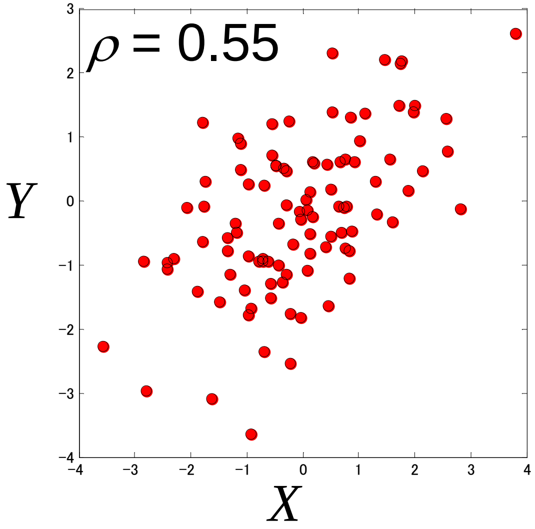
Correlation

– The most elementary and popular indicator to measure the linear relation between two variables.

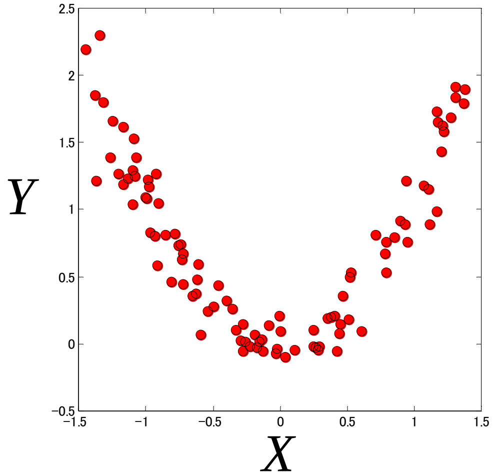
Correlation coefficient (aka Pearson correlation)

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

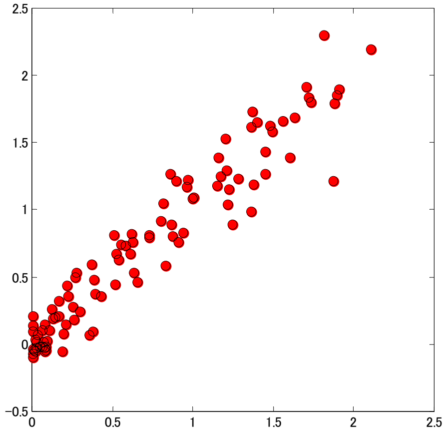




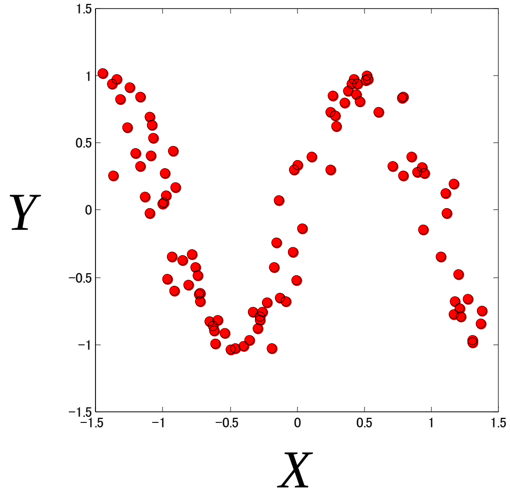
■ Nonlinear dependence



Corr(X, Y)
= 0.17



Corr(X^2, Y)
= 0.96

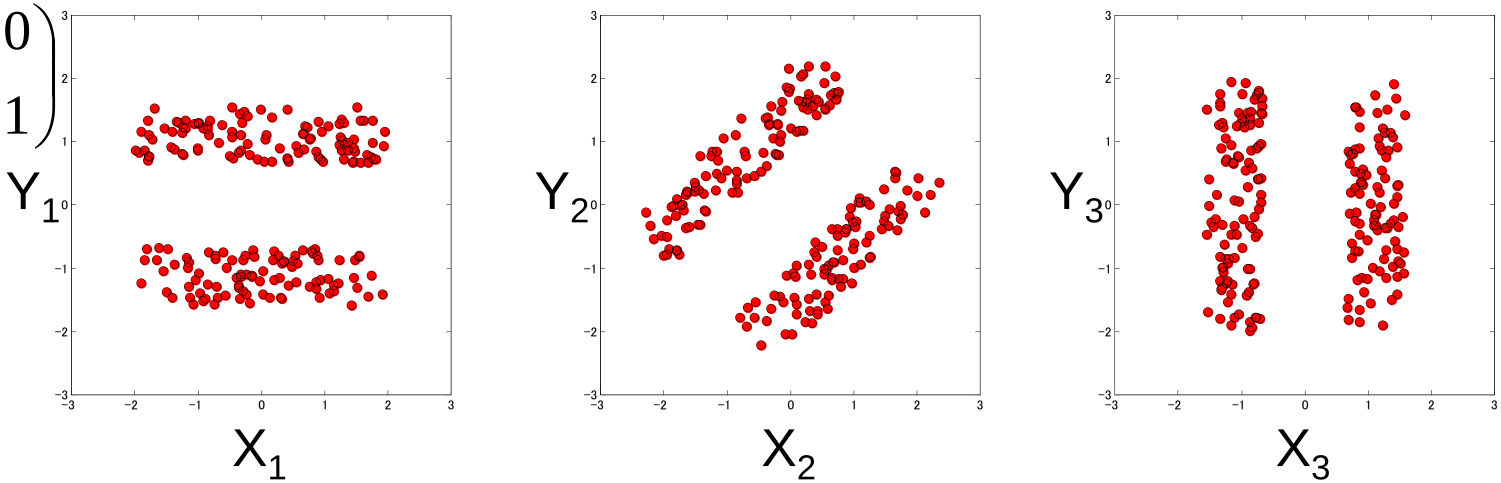


Corr(X, Y) = -0.06
Corr(X^2, Y) = 0.09
Corr(X^3, Y) = -0.38

Corr($\sin(\pi X)$, Y) = 0.93

"Uncorrelated" does not mean "independent"

$$V_{ZZ} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



independent — dependent — independent

They are all uncorrelated!

Note: If $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ and $\tilde{Z} = \begin{pmatrix} \tilde{X} \\ \tilde{Y} \end{pmatrix} = AZ$,

$$V_{\tilde{Z}\tilde{Z}} = E[A(Z - E[Z])(Z - E[Z])^T A^T] = AV_{ZZ}A^T$$

Nonlinear statistics with kernels

– Linear methods can consider only linear relation.

– Nonlinear transform of the original variable may help.

$X \rightarrow (X, X^2, X^3, \ldots)$

But,

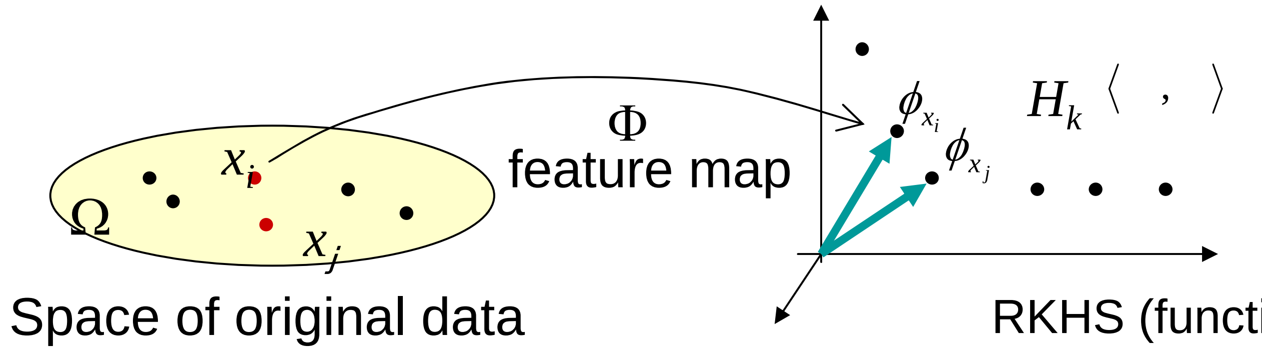
- It is not clear how to make a good transform, in particular, if the data is high-dimensional.
- A transform may cause high-dimensionality.

e.g.) dim $X = 100$ $\rightarrow$ $X_i X_j$ # combinations = 4950

Why not use the kernelization / feature map for the transform?

Kernel methodology for statistical inference

– Transform of the original data by “feature map”.



Let’s do linear statistics in the feature space!

– Is this simply “kernelization”? – Yes, in a big picture.

– But, in this methodology, the methods have clear statistical/probabilistic meaning in the original space, e.g. independence, conditional independence, two-sample test etc.

– From the side of statistics, it is a new approach using p.d. kernels.

Goal: To understand how linear methods in RKHS solve classical inference problems on probabilities.

Remarks on Terminology

- In this lecture, "kernel" means "positive definite kernel".
- In statistics, "kernel" is traditionally used in more general meaning, which does not impose positive definiteness.

  e.g. kernel density estimation (Parzen window approach)

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} k(x, X_i)$$

  $k(x_1, x_2)$ is not necessarily positive definite.

- Statistical jargon
  - "in population": evaluated with probability e.g. $E[X] = \int x dP(x)$
  - "empirical": evaluated with sample e.g. $\frac{1}{N}\sum_{i=1}^{N} X_i$
  - "asymptotic": when the number of data goes to infinity.

    "$\sum_{i=1}^{N} X_i \big/ N$ asymptotically converges to $E[X]$."

II. Dependence with Kernels

Prologue to kernel methodology for inference on probabilities

Independence of Variables

- Definition
  - Random vectors X on $\mathbf{R}^m$ and Y on $\mathbf{R}^n$ are independent ($X \perp\!\!\!\perp Y$)

$$\overset{def.}{\Longleftrightarrow} \quad \Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B)$$

for any $A \in \mathcal{B}_m, B \in \mathcal{B}_n$

- Basic properties
  - If X and Y are independent,

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]$$

  - If further (X,Y) has the joint p.d.f $p_{XY}(x,y)$, and X and Y have the marginal p.d.f. $p_X(x)$ and $p_Y(y)$, resp, then

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad p_{XY}(x,y) = p_X(x)\, p_Y(y)$$

Review: Covariance Matrix

- Covariance matrix

$X = (X_1, \ldots, X_m)^T$, $Y = (Y_1, \ldots, Y_\ell)^T$: *m* and *n* dimensional random vectors

Covariance matrix V_{XY} of X and Y is defined by

$$V_{XY} \equiv E[(X - E[X])(Y - E[Y])^T] = E[XY^T] - E[X]E[Y]^T$$

(*m* x *n* matrix)

In particular, $V_{XX} \equiv E[XX^T] - E[X]E[X]^T$

– $V_{XY} = 0$ if and only if X and Y are uncorrelated.

For a sample $(X^{(1)}, Y^{(1)}), \ldots, (X^{(N)}, Y^{(N)})$

empirical covariance matrix

$$\hat{V}_{XY} = \frac{1}{N}\sum_{i=1}^{N} X^{(i)} Y^{(i)T} - \left(\frac{1}{N}\sum_{i=1}^{N} X^{(i)}\right)\left(\frac{1}{N}\sum_{i=1}^{N} Y^{(i)}\right)^T$$

(*m* x *n* matrix)

Independence of Gaussian variables

■ Multivariate Gaussian (normal) distribution

$X = (X_1, \ldots, X_m) \sim N(\mu, V)$: m-dimensional Gaussian random variable with mean μ and covariance matrix V.

Probability density function (p.d.f.)

$$\phi(x; \mu, V) = \frac{1}{(2\pi)^{m/2} |V|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu) \right)$$

■ Independence of Gaussian variables

– X, Y: Gaussian random vectors of dim p and q (resp.)

"independent" $\Leftrightarrow$ "uncorrelated"

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad V_{XY} = O \quad \Leftrightarrow \quad E[XY^T] = E[X]E[Y]^T$$

$\because$) If $V_{XY} = O$,

$$p_{XY}(y, x) = \frac{1}{(2\pi)^{m/2} |V_{XX}|^{1/2} |V_{YY}|^{1/2}} \exp\left(-\frac{1}{2} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}^T \begin{pmatrix} V_{XX}^{-1} & O \\ O & V_{YY}^{-1} \end{pmatrix} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \right) = p_X(x) p_Y(y)$$

Independence by Nonlinear Covariance

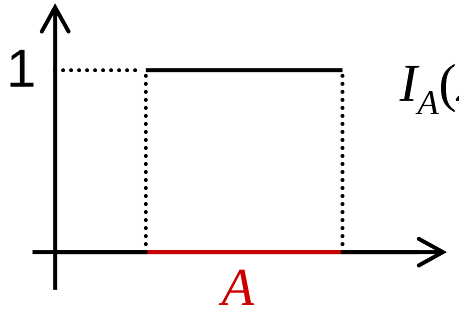
- Independence and nonlinear covariance
  - X and Y are independent

$$\Leftrightarrow \quad Cov[f(X), g(Y)] = 0 \quad \text{for all measurable functions } f \text{ and } g.$$

$\because$) Take $f(x) = I_A(x)$ and $g(y) = I_B(y)$ for measurable sets A and B.

$$E[I_A(X)I_B(Y)] - E[I_A(X)]E[I_B(Y)] = 0$$

$$\Rightarrow \Pr\left(X \in A, Y \in B\right) = \Pr\left(X \in A\right)\Pr\left(Y \in B\right)$$



$I_A(x)$ indicator function of A

Measuring all the nonlinear covariance

$$\sup_{f,g} \left| Cov[f(X), g(Y)] \right|$$

can be used for the dependence measure.

– Questions.
  - How can we calculate the value?

    The space of measurable functions is large, containing noncontinuous and weird functions
  - With finite number of data, how can we estimate the value?

Using Kernels: COCO

- Restrict the functions in RKHS

  X , Y : random variables on Ω_X and Ω_Y , resp.

  Prepare RKHS (H_X, k_X) and (H_X , k_X) defined on Ω_X and Ω_Y, resp

$$\sup_{f \in H_X, g \in H_Y} \frac{\left| Cov[f(X), g(Y)] \right|}{\| f \|_{H_X} \| g \|_{H_Y}}$$

… COnstrained COvariance (COCO, Gretton et al. 05)

- Estimation with data

  $(X_1, Y_1), \ldots, (X_N, Y_N)$: i.i.d. sample

$$\sup_{f \in H_X, g \in H_Y} \frac{\left| Cov_{emp}[f(\hat{X}), g(\hat{Y})] \right|}{\| f \|_{H_X} \| g \|_{H_Y}}$$

$$Cov_{emp}[f(\hat{X}), g(\hat{Y})] = \frac{1}{N} \sum_{i=1}^{N} f(X_i) g(Y_i) - \frac{1}{N} \sum_{i=1}^{N} f(X_i) \frac{1}{N} \sum_{i=1}^{N} g(Y_i)$$

Solution to COCO

– The empirical COCO is reduced to an eigenproblem:

$$\tfrac{1}{N}\max \alpha^T G_X G_Y \beta \qquad \text{subj. to} \qquad \alpha^T G_X \alpha = 1, \quad \beta^T G_Y \beta = 1$$

$$\mathrm{COCO}_{emp} = \sup_{f \in H_X, g \in H_Y} \frac{\left| Cov_{emp}[f(\hat{X}), g(\hat{Y})] \right|}{\| f \|_{H_X} \| g \|_{H_Y}} = \frac{\text{largest singular value of } G_X^{1/2} G_Y^{1/2}}{N}$$

G_X and G_Y are the centered Gram matrices defined by

$$G_X = Q_n K_X Q_n \qquad (N \text{ x } N \text{ matrix})$$

where $K_{X,ij} = k_X(X_i, X_j)$ $\quad Q_n = I_n - \frac{1}{N}\mathbf{1}_N \mathbf{1}_N^T$ (projector on $\mathbf{1}_N^{\perp}$)

$$\mathbf{1}_N = (1, \ldots, 1)^T$$

For a symmetric positive semidefinite matrix A,

$A^{1/2}$ is a symmetric positive semidefinite matrix such that $(A^{1/2})^2 = A$.

Derivation

$$Cov_{emp}[f(\hat{X}), g(\hat{Y})] = \frac{1}{N}\sum_{i=1}^{N}\left\{ f(X_i) - \frac{1}{N}\sum_{j=1}^{N} f(X_j) \right\}\left\{ g(Y_i) - \frac{1}{N}\sum_{j=1}^{N} g(Y_j) \right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\langle f, k_X(\cdot, X_i) - \underbrace{\tfrac{1}{N}\textstyle\sum_{j=1}^{N} k_X(\cdot, X_j)}_{\hat{m}_X} \right\rangle \left\langle k_Y(\cdot, Y_i) - \underbrace{\tfrac{1}{N}\textstyle\sum_{j=1}^{N} k_Y(\cdot, Y_j)}_{\hat{m}_Y}, g \right\rangle$$

It is sufficient to consider (representer theorem)

$$f = \sum_{j=1}^{N} \alpha_j \left\{ k_X(\cdot, X_j) - \hat{m}_X \right\}, \qquad g = \sum_{\ell=1}^{N} \beta_\ell \left\{ k_Y(\cdot, Y_\ell) - \hat{m}_Y \right\}$$

$$Cov_{emp}[f(\hat{X}), g(\hat{Y})] = \tfrac{1}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{N}\sum_{j=1}^{N} \alpha_j \beta_\ell \left\langle k_Y(\cdot, Y_\ell) - \hat{m}_Y, k_Y(\cdot, Y_i) - \hat{m}_Y \right\rangle \times \left\langle k_X(\cdot, X_i) - \hat{m}_X, k_X(\cdot, X_j) - \hat{m}_X \right\rangle$$

$$= \tfrac{1}{N} \alpha^T G_X G_Y \beta$$

Maximize it under the constraints

$$\| f \|_{H_X}^2 = \alpha^T G_X \alpha = 1, \quad \| g \|_{H_Y}^2 = \beta^T G_Y \beta = 1$$

By using $u = G_X^{1/2}\alpha, \quad v = G_Y^{1/2}\beta$

$$\tfrac{1}{N} \max_{u,v} u^T G_X^{1/2} G_Y^{1/2} v \quad \text{subj. to} \quad \| u \| = 1, \quad \| v \| = 1$$

Quick Review on RKHS

- Reproducing kernel Hilbert space (RKHS, review)

Ω: set.

$k : \Omega \times \Omega \to \mathbf{R}$ pos. def. kernel

$\Rightarrow$

$\exists 1\ H$: reproducing kernel Hilbert space (RKHS)

such that k is the reproducing kernel of H , *i.e.*

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

2) $\mathrm{Span}\{k(\cdot, x) \mid x \in \Omega\}$ is dense in H.

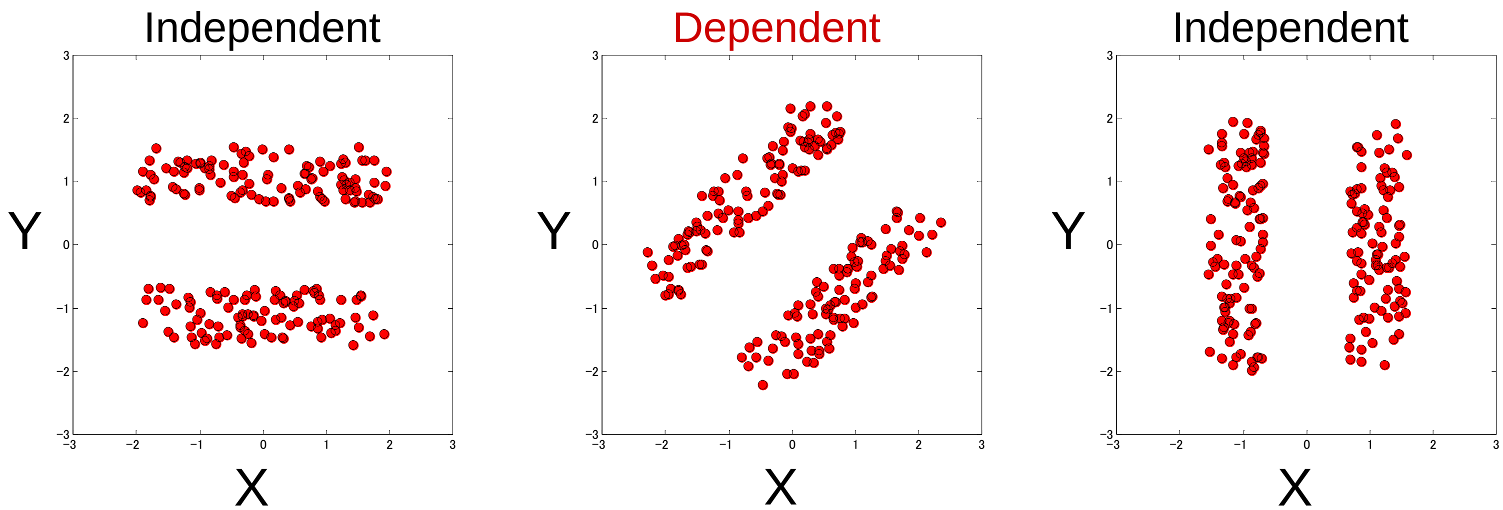
3) $\langle k(\cdot, x), f \rangle_H = f(x)$ (reproducing property)

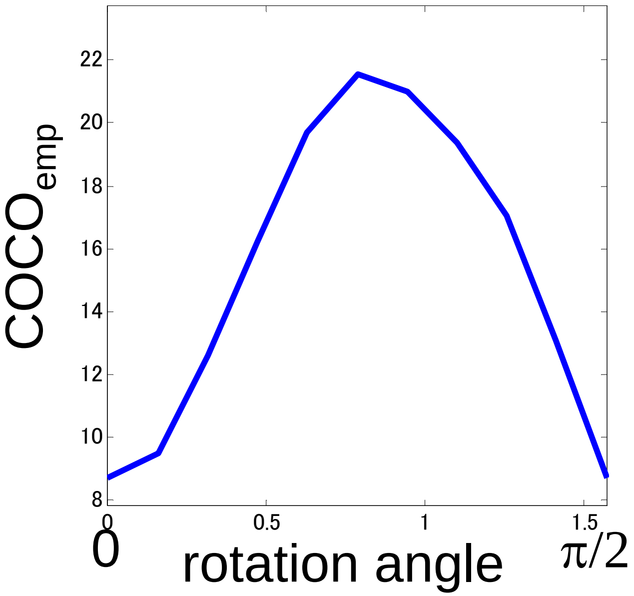
– Feature map

$$\Phi : \Omega \to H, \quad x \mapsto k(\cdot, x) \qquad i.e. \quad \Phi(x) = k(\cdot, x)$$

$$\langle \Phi(x), f \rangle = f(x)$$ (reproducing property)

Example with COCO





Gaussian kernels are used.

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

COCO and Independence

- Characterization of independence

X and Y are independent $\quad\Longleftrightarrow\quad$ $$\sup_{f\in H_X, g\in H_Y} \frac{\left|Cov[f(X), g(Y)]\right|}{\| f \|_{H_X} \| g \|_{H_Y}} = 0$$

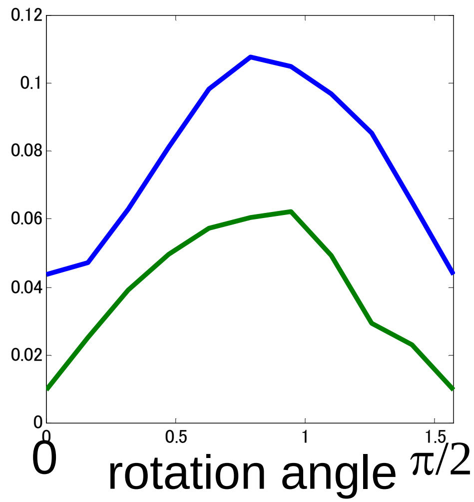
This equivalence holds if the RKHS are “rich enough” to express all the dependence between X and Y. (discussed later in Part IV.)

For the moment, Gaussian kernels are used to guarantee this equivalence.

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

HSIC (Gretton et al. 05)

- How about using other singular values?



1st SV of $G_X^{1/2}G_Y^{1/2}$

2nd SV of $G_X^{1/2}G_Y^{1/2}$

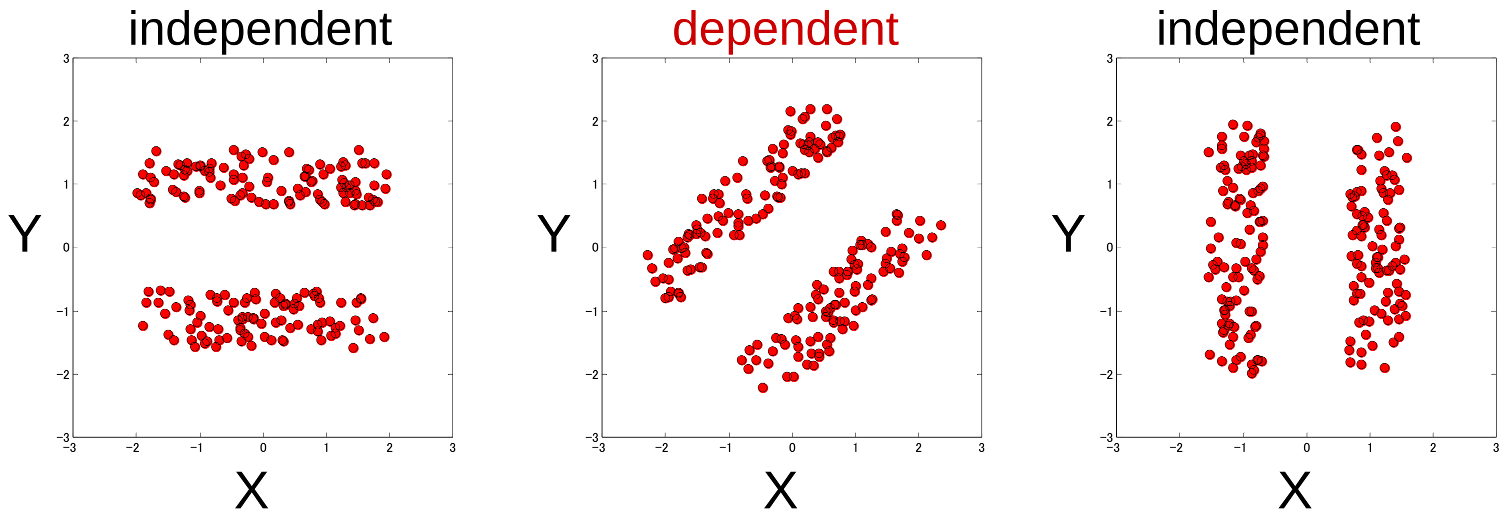
Smaller singular values
also represent dependence.

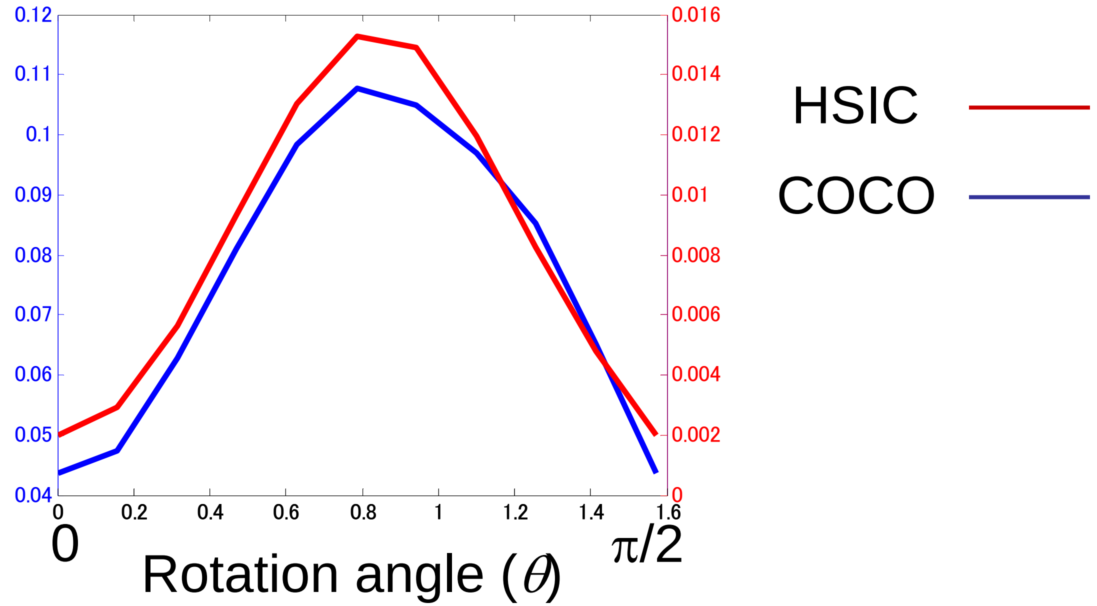
$$\text{HSIC} \equiv \frac{1}{N^2}\sum_{i=1}^{N}\gamma_i^2 = \frac{1}{N^2}\left\| G_X^{1/2}G_Y^{1/2} \right\|_F^2 = \frac{1}{N^2}\mathrm{Tr}\left[G_X G_Y\right]$$

(γ_i: the i-th singular values of $G_X^{1/2}G_Y^{1/2}$)

$\| \; \|_F$: Frobenius norm $\| M \|_F^2 = \sum_{i,j=1}^{N} M_{ij}^2 = \mathrm{Tr}[M^T M]$

Example with HSIC





Summary of Part II

COCO	Empirical	Population
Kernel	1st SV of $G_X^{1/2}G_Y^{1/2}$	$\sup_{\|f\|_{H_X}=\|g\|_{H_Y}=1} Cov[f(X),g(Y)]$
Linear (finite dim.)	1st SV of $\hat{V}_{XY}$	$\max_{\|a\|=\|b\|=1} Cov[a^T X, b^T Y] = \max_{\|a\|=\|b\|=1} a^T V_{XY} b$ = 1st SV of V_{XY}

HSIC	Empirical	Population
Kernel	$\left\|G_X^{1/2}G_Y^{1/2}\right\|_F^2$	What is the population version?
Linear (finite dim.)	$\left\|\hat{V}_{XY}\right\|_F^2$	$\left\|V_{XY}\right\|_F^2$ (Sum of SV^2 of cov. matrix)

III. Covariance on RKHS

Two Views on Kernel Methods

- As a good class of nonlinear functions

Objective functional for a nonlinear method

$$\max_f \Psi(f(X_1),\ldots,f(X_N)) \qquad f\text{: nonlinear function}$$

Find the solution within a RKHS.

– Reproducing property / kernel trick, Representer theorem

c.f. COCO in the previous section.

- Kernelization of linear methods

– Map the data into a RKHS, and apply a linear method

$$X_i \mapsto \Phi(X_i)$$

– Map the random variable into a RKHS, and do linear statistics!

$$X \mapsto \Phi(X) \qquad \text{random variable on RKHS}$$

Covariance on RKHS

– Linear case (Gaussian):

$\mathrm{Cov}[X, Y] = \mathrm{E}[YX^T] - \mathrm{E}[Y]\mathrm{E}[X]^T$: covariance matrix

– On RKHS:

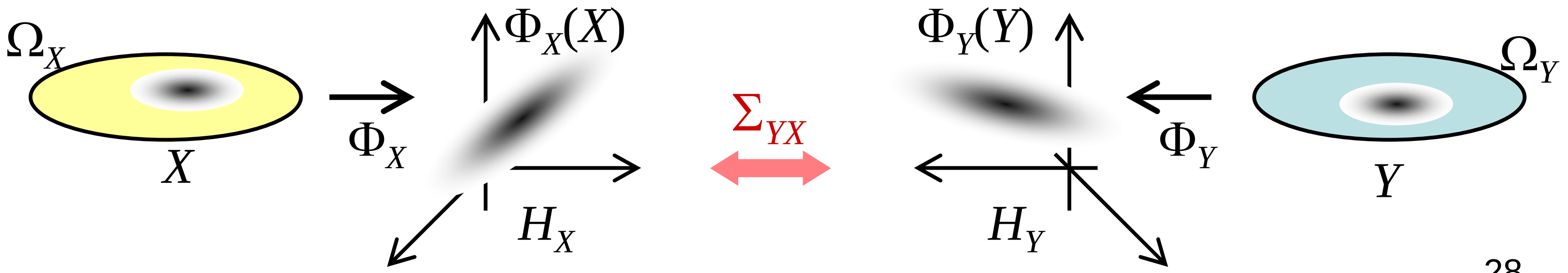
X , Y : random variables on Ω_X and Ω_Y , resp.

Prepare RKHS (H_X, k_X) and (H_Y, k_Y) defined on Ω_X and Ω_Y, resp.

Define random variables on the RKHS H_X and H_Y by

$$\Phi_X(X) = k_X(\cdot, X) \qquad \Phi_Y(Y) = k_Y(\cdot, Y)$$

Define the big (possibly infinite dimensional) covariance matrix Σ_{YX} on the RKHS.



Cross-covariance operator

– Definition

There uniquely exists an operator from H_X to H_Y such that

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

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

Σ_{YX} : Cross-covariance operator

– A bit loose expression

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle \Phi_X(X), \cdot \rangle] - E[\Phi_Y(Y)]E[\langle \Phi_X(X), \cdot \rangle]$$

c.f. Euclidean case

$V_{YX} = \mathrm{E}[YX^T] - \mathrm{E}[Y]\mathrm{E}[X]^T$: covariance matrix

$$(b, V_{YX} a) = Cov[(b, Y), (a, X)]$$

Intuition

Suppose X and Y are $\mathbf{R}$-valued, and $k(x,u)$ admits the expansion

$$k(x,u) = 1 + c_1 xu + c_2 x^2 u^2 + c_3 x^3 u^3 + \cdots \qquad \text{e.g.) } k(x,u) = \exp(xu)$$

With respect to the basis $1, u, u^2, u^3, \ldots$, the random variables on RKHS are expressed by

$$\Phi(X) = k(X,u) \sim (1, c_1 X, c_2 X^2, c_3 X^3, \ldots)^T$$

$$\Phi(Y) = k(Y,u) \sim (1, c_1 Y, c_2 Y^2, c_3 Y^3, \ldots)^T$$

$$\Sigma_{YX} \sim \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & c_1^2 Cov[Y,X] & c_1 c_2 Cov[Y,X^2] & c_1 c_3 Cov[Y^3,X] & \cdots \\ 0 & c_2 c_1 Cov[Y^2,X] & c_2^2 Cov[Y^2,X^2] & c_2 c_3 Cov[Y^2,X^3] & \cdots \\ 0 & c_3 c_1 Cov[Y^3,X] & c_3 c_2 Cov[Y^3,X^2] & c_3^2 Cov[Y^3,X^3] & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The operator Σ_{YX} contains the information on all the higher-order correlation.

Addendum on “operator”

- “Operator” is often used for a linear map defined on a functional space, in particular, of infinite dimension.

- Σ_{YX} is a linear map from H_X to H_Y, as the covariance matrix V_{YX} is a linear map from $\mathbf{R}^m$ to $\mathbf{R}^n$.

- If you are not familiar with the word “operator”, simply replace it with “linear map” or “big matrix”.

- If you are very familiar with the operator terminology, you can easily prove Σ_{YX} is a bounded operator. (Exercise)

Characterization of Independence

- Independence and Cross-covariance operator

If the RKHS's are "rich enough" to express all the moments,

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

$$\Updownarrow$$

$$\mathrm{Cov}[f(X), g(Y)] = 0$$

or

$$E[g(Y)f(X)] = E[g(Y)]E[f(X)]$$

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

($\Rightarrow$ is always true.
$\Leftarrow$ requires the richness assumption. Part IV.)

– *c.f.* for Gaussian variables

X and Y are independent $\Leftrightarrow$ $V_{XY} = O$ i.e. uncorrelated

Measures for Dependence

- Kernel measures for dependence/independence

Measure the "norm" of Σ_{YX}.

– Kernel generalized variance (KGV, Bach&Jordan 02, FBJ 04)

$$KGV(X,Y) = \frac{\det \Sigma_{[XY][XY]}}{\det \Sigma_{XX} \det \Sigma_{YY}}$$

– COCO

$$COCO(X,Y) = \left\| \Sigma_{YX} \right\| = \sup_{f \neq 0, g \neq 0} \frac{\left| \langle g, \Sigma_{YX} f \rangle \right|}{\| f \|_{H_X} \| g \|_{H_Y}}$$

– HSIC

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

– HSNIC

$$HSNIC(X,Y) = \left\| \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2} \right\|_{HS}^2$$ (explained later)

Norms of operators

$A : H_1 \to H_2$ operator on a Hilbert space

- Operator norm

$$\|A\| = \sup_{\|f\|=1} \|Af\| = \sup_{\|f\|=1, \|g\|=1} \left| \langle g, Af \rangle \right|$$

c.f. the largest singular value of a matrix

- Hilbert-Schmidt norm

A is called Hilbert-Schmidt if for complete orthonormal systems $\{\varphi_i\}$ of H_1 and $\{\psi_j\}$ of H_2 if

$$\sum_j \sum_i \langle \psi_j, A\varphi_i \rangle^2 < \infty.$$

Hilbert-Schmidt norm is defined by

$$\|A\|_{HS}^2 = \sum_j \sum_i \langle \psi_j, A\varphi_i \rangle^2$$

c.f. Frobenius norm of a matrix

Empirical Estimation

Estimation of covariance operator

i.i.d. sample $(X_1, Y_1), \ldots, (X_N, Y_N)$

An estimator of Σ_{YX} is given by

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

where

$$\hat{m}_X = \frac{1}{N}\sum_{i=1}^{N} k_1(\cdot, X_i), \qquad \hat{m}_Y = \frac{1}{N}\sum_{i=1}^{N} k_2(\cdot, Y_i)$$

– Note
  - This is again an operator.
  - But, it operates essentially on the finite dimensional space spanned by the data $\Phi_X(X_1), \ldots, \Phi_X(X_N)$ and $\Phi_Y(Y_1), \ldots, \Phi_Y(Y_N)$

Empirical cross-covariance operator

Proposition (Empirical mean)

$\hat{m}_X = \frac{1}{N}\sum_{i=1}^{N} k(\cdot, X_i)$ gives the empirical mean:

$$\langle \hat{m}_X, f \rangle = \frac{1}{N}\sum_{i=1}^{N} f(X_i) \equiv \hat{E}[f(X)] \qquad (\forall f \in H_X)$$

Proposition (Empirical covariance)

$\hat{\Sigma}_{YX}^{(N)}$ gives the empirical covariance

$$\langle g, \hat{\Sigma}_{YX}^{(N)} f \rangle = \frac{1}{N}\sum_{i=1}^{N} \{g(Y_i) - \hat{E}[g(Y)]\}\{f(X_i) - \hat{E}[f(X)]\}$$
$$(\forall f \in H_X, \forall g \in H_Y)$$

$\hat{m}_X$: empirical mean element (in RKHS)

$\hat{\Sigma}_{YX}^{(N)}$: empirical cross-covariance operator (on RKHS)

COCO Revisited

- COCO = operator norm

$$COCO(X,Y) = \left\|\Sigma_{YX}\right\| = \sup_{\|f\|=1,\|g\|=1} \left|\langle g, \Sigma_{YX} f \rangle\right|$$

with data $\Rightarrow$

$$COCO_{emp}(\hat{X},\hat{Y}) = \left\|\hat{\Sigma}_{YX}^{(N)}\right\| = \sup_{\|f\|=1,\|g\|=1} \left|\langle g, \hat{\Sigma}_{YX}^{(N)} f \rangle\right|$$

$$= \sup_{\|f\|=\|g\|=1} \left|Cov_{emp}[f(\hat{X}), g(\hat{Y})]\right| \longleftarrow$$ previous definition

$$= \frac{1}{N} \times \text{largest singular value of } G_X^{1/2} G_Y^{1/2}$$

HSIC Revisited

- HSIC = Hilbert-Schmidt Information Criterion

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

with data

$\Rightarrow$

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

$\because$)

$$\left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 = \mathrm{Tr}\left[\hat{\Sigma}_{YX}^{(N)} \hat{\Sigma}_{XY}^{(N)} \right]$$

$$= \mathrm{Tr}\left[\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\{ k_Y(\cdot,Y_i) - \hat{m}_Y \right\} \left\langle k_X(\cdot,X_i) - \hat{m}_X, k_X(\cdot,X_j) - \hat{m}_X \right\rangle \left\langle k_Y(\cdot,Y_j) - \hat{m}_Y, \cdot \right\rangle \right]$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle k_X(\cdot,X_i) - \hat{m}_X, k_X(\cdot,X_j) - \hat{m}_X \right\rangle \left\langle k_Y(\cdot,Y_j) - \hat{m}_Y, k_Y(\cdot,Y_i) - \hat{m}_Y \right\rangle$$

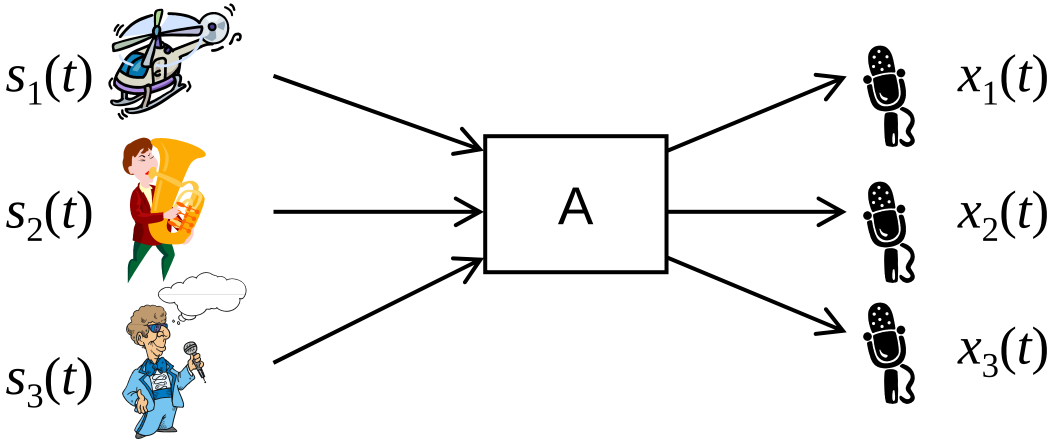
$$= \frac{1}{N^2} \mathrm{Tr}\left[G_X G_Y \right]$$

Application of HSIC to ICA

- Independent Component Analysis (ICA)
  - Assumption

    m independent source signals

    m observations of linearly mixed signals



$$X(t) = AS(t)$$

A: mxm invertible matrix

  - Problem

    Restore the independent signals S from observations X.

$$\hat{S} = BX$$

B: mxm orthogonal matrix

ICA with HSIC

$X^{(1)},\ldots,X^{(N)}$: i.i.d. observation (m-dimensional)

Pairwise-independence criterion is applicable.

Minimize $$L(B)=\sum_{a=1}^{m}\sum_{b>a} HSIC(Y_a,Y_b) \qquad Y=BX$$

Objective function is non-convex. Optimization is not easy.

→ Approximate Newton method has been proposed

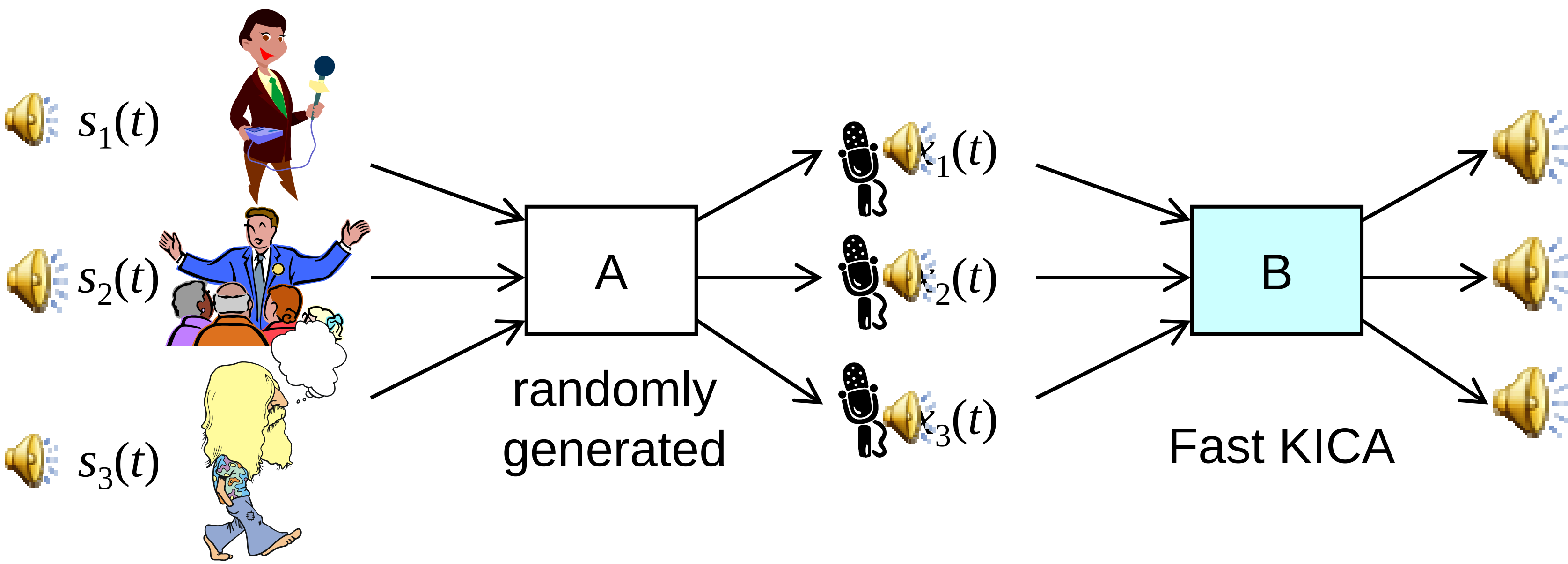
Fast Kernel ICA (FastKICA, Shen et al 07)

(Software downloadable at Arthur Gretton's homepage)

Other methods for ICA

See, for example, Hyvärinen et al. (2001).

■ Experiments (speech signal)



Three speech signals

Normalized Covariance

- **Correlation – normalized variance**

  Covariance is not normalized well: it depends on the variance of X, Y.

  Correlation is better normalized

$$V_{YY}^{-1/2} V_{YX} V_{XX}^{-1/2}$$

- **NOrmalized Cross-Covariance Operator (FBG07)**

NOCCO
$$W_{YX} = \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2}$$

Definition: there is a factorization of the Σ_{YX} such that

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2}$$

– Operator norm is less than or equal to 1, *i.e.* $\| W_{YX} \| \leq 1$

Empirical estimation of NOCCO

$(X_1, Y_1), \ldots, (X_N, Y_N)$: sample

$$\hat{W}_{YX}^{(N)} = \left(\hat{\Sigma}_{YY}^{(N)} + \varepsilon_N I\right)^{-1/2} \hat{\Sigma}_{YX}^{(N)} \left(\hat{\Sigma}_{XX}^{(N)} + \varepsilon_N I\right)^{-1/2}$$

ε_N: regularization coefficient

Note: $\hat{\Sigma}_{XX}^{(N)}$ is of finite rank, thus not invertible

Relation to Kernel CCA

– See Bach & Jordan 02, Fukumizu Bach Gretton 07

Normalized Independence Measure

- HS Normalized Independence Criterion (HSNIC)

Assume $W_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ is Hilbert-Schmidt

$$HSNIC = \| W_{YX} \|_{HS}^2 = \left\| \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2} \right\|_{HS}^2$$

$$HSNIC_{emp} = \left\| \hat{W}_{YX}^{(N)} \right\|_{HS}^2 = \mathrm{Tr}\left[G_X (G_X + N\varepsilon_N I_N)^{-1} G_Y (G_Y + N\varepsilon_N I_N)^{-1} \right]$$

(Confirm this – exercise)

- Characterizing independence

<u>Theorem</u>

Under some “richness” assumptions on kernels (see Part IV).

HSNIC = 0 if and only if X and Y are independent.

Kernel-free Expression

- Integral expression of HSNIC without kernels

<u>Theorem (FGSS07)</u>

Assume that $H_X \otimes H_Y + \mathbf{R}$ is dense in $L^2(P_X \otimes P_Y)$, and the laws P_X and P_Y have p.d.f. w.r.t. the measures μ_1 and μ_2, resp.

$$HSNIC = \| W_{YX} \|_{HS}^2$$

$$= \iint \left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} - 1 \right)^2 p_X(x) p_Y(y) d\mu_1(x) d\mu_2(y)$$

= Mean Square Contingency

- HSNIC is defined by kernels, but it does not depend on the kernels. Free from the choice of kernels!
- $\text{HSNIC}_{\text{emp}}$ gives a kernel estimator for the Mean Square Contingency.

	HSIC	HSNIC
PROS	• Simple to compute • Asymptotic distribution for independence test is known (Part V)	• Does not depend on the kernels in population
CONS	• The value depends on the choice of kernels	• Regularization coefficient is needed. • Matrix inversion is needed. • Asymptotic distribution for independence test is not known.

(Some experimental comparisons are given in Part V.)

Choice of Kernel

- How to choose a kernel?
  - Recall: in supervised learning (e.g. SVM), cross-validation (CV) is reasonable and popular.
  - For unsupervised problems, such as independence measures, there are no theoretically reasonable methods.
  - Some heuristic methods which work:
    - Heuristics for Gaussian kernels

$$\sigma = \text{median}\left\{ \left\| X_i - X_j \right\| \mid i \neq j \right\}$$

    - Make a related supervised problem, if possible, and use CV.
  - More studies are required.

Relation with Other Measures

- Mutual Information

$$MI(X,Y) = \iint p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\,d\mu_X(x)d\mu_Y(y)$$

- MI and HSNIC

$$\boxed{HSNIC(X,Y) \cancel{\leq} MI(X,Y)}$$

>= (correction. June 2014)

∵)

$$HSNIC = \iint p_{XY}(x,y)\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1\right)d\mu_1(x)d\mu_2(y)$$

$$\cancel{\leq} \iint p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\,d\mu_1(x)d\mu_2(y) \quad = MI$$

>= (correction. June 2014)

$$(\log z \leq z - 1)$$

- Mutual Information:
  - Information-theoretic meaning.
  - Estimation is not straightforward for continuous variables.
    Explicit estimation of p.d.f. is difficult for high-dimensional data.
    - Parzen-window is sensitive to the band-width.
    - Partitioning may cause a large number of bins.
  - Some advanced methods: e.g. k-NN approach (Kraskov et al.).
- Kernel method:
  - Explicit estimation of p.d.f. is not required;

    the dimension of data does not appear explicitly, but it is influential in practice.
  - Kernel / kernel parameters must be chosen.
- Experimental comparison

  See Section V (Statistical Tests)

Summary of Part III

- **Cross-Covariance operator**
  - Covariance on RKHS: extension of covariance matrix
  - If the kernel defines a rich RKHS,

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

- **Kernel-based dependence measures**
  - COCO: operator norm of Σ_{XY}
  - HSIC: Hilbert-Schmidt norm of Σ_{XY}
  - HSNIC: Hilbert-Schmidt norm of normalized cross-covariance operator $W_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$

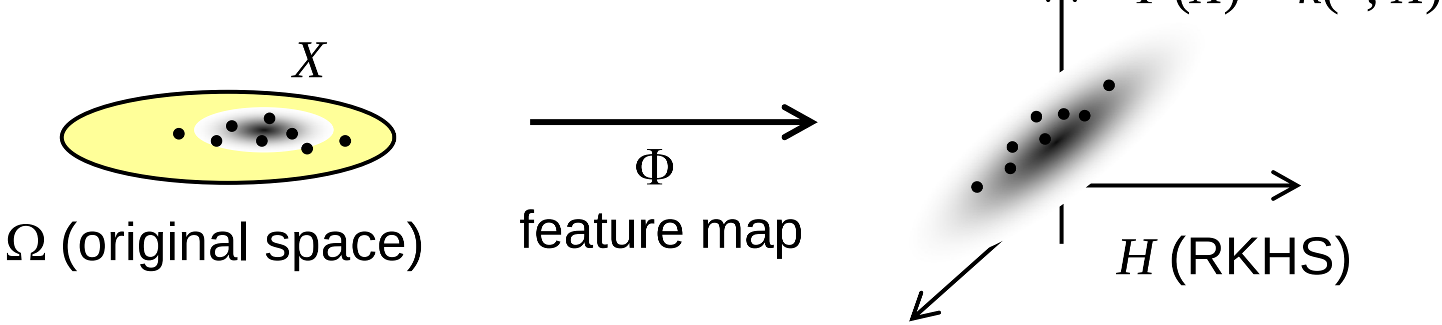
    HSNIC = mean square contingency (in population) kernel free!

  - Application to ICA

IV. Representing a Probability

Statistics on RKHS

- Linear statistics on RKHS



– Basic statistics on Euclidean space		Basic statistics on RKHS
Mean	$\longrightarrow$	Mean element
Covariance	$\longrightarrow$	Cross-covariance operator Σ_{YX}
Conditional covariance	$\longrightarrow$	Conditional-covariance operator (Part VI)

– Plan: define the basic statistics on RKHS and derive nonlinear/ nonparametric statistical methods in the original space.

Mean on RKHS

– Empirical mean on RKHS

$X^{(1)},\dots,X^{(N)}$: i.i.d. sample → $\Phi(X_1),\dots,\Phi(X_N)$: sample on RKHS

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

$$\langle \hat{m}_X, f \rangle = \frac{1}{N}\sum_{i=1}^{N} f(X_i) \equiv \hat{E}[f(X)] \qquad (\forall f \in H_X)$$

– Mean element on RKHS

X : random variable on Ω → $\Phi(X)$: random variable on RKHS.

Define $m_X = E[\Phi(X)]$

$$\boxed{\langle m_X, f \rangle = E[f(X)] \qquad (\forall f \in H)}$$

Representation of Probability

Moments by a kernel

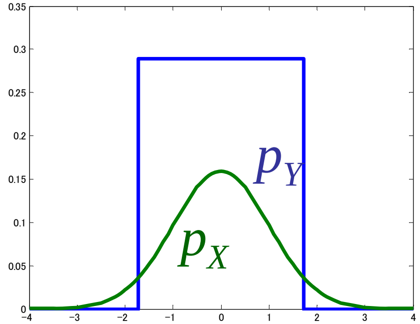
Example of one-variable

$$k(x,u) = \exp(xu) \quad = 1 + c_1 xu + c_2 x^2 u^2 + c_3 x^3 u^3 + \cdots$$

$$\Rightarrow$$

$$m_X(u) = E_X[k(X,u)] = 1 + c_1 \underline{E_X[X]} u + c_2 \underline{E_X[X^2]} u^2 + c_3 \underline{E_X[X^3]} u^3 + \cdots$$

- As a function of u, the mean element m_X contains the information on all the moments – "richness" of RKHS.
- It is natural to expect that m_X "represents" or "characterizes" a probability under "richness" assumption on the kernel.



$$E[X] = 0 \qquad E[Y] = 0$$

$$E[X^2] = 1 \qquad E[Y^2] = 1$$

$$E[X^3] = 0 \qquad E[Y^3] = 0$$

$$E[X^4] = 3 \qquad E[Y^4] = 9/5$$

Characteristic Kernel

- Richness assumption on kernels

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

  H: RKHS on Ω with measurable kernel k.

  m_P: mean element on H for the probability $P \in \mathcal{P}$

  – Definition

    The kernel k is called characteristic if the mapping

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

    is one-to-one.

  – The mean element of a characteristic kernel uniquely determines the probability.

    $$m_X = m_Y \quad \Leftrightarrow \quad P_X = P_Y$$

– "Richness" assumption in the previous sections should be replaced by "kernel is characteristic" or the following denseness assumption.

– Sufficient condition

Theorem

k: kernel on a measurable space $(\Omega, \mathcal{B})$. H: associated RKHS. $q \geq 1$.

If $H + \mathbf{R}$ is dense in $L^q(P)$ for any probability P on $(\Omega, \mathcal{B})$, then k is characteristic

– Examples of characteristic kernel

- Gaussian kernel on the entire $\mathbf{R}^m$

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

- Laplacian kernel on the entire $\mathbf{R}^m$

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

Universal kernel (Steinwart 02)

A continuous kernel k on a compact metric space Ω is called universal if the associated RKHS is dense in $C(\Omega)$, the functional space of the continuous functions on Ω with sup norm.

Example: Gaussian kernel on a compact subset of $\mathbf{R}^m$

Proposition

A universal kernel is characteristic.

- Characteristic kernels are wider class, and suitable for discussing statistical inference of probabilities.
- Universal kernels are defined only on compact sets.

- Gaussian kernels are characteristic either on a compact subset and the entire of Euclidean space.

Two-Sample Problem

Two i.i.d. samples are given;

$$X^{(1)},\ldots,X^{(N_X)} \quad \text{and} \quad Y^{(1)},\ldots,Y^{(N_Y)}.$$

Are they sampled from the same distribution?

- Practically important.

  We often wish to distinguish two things:

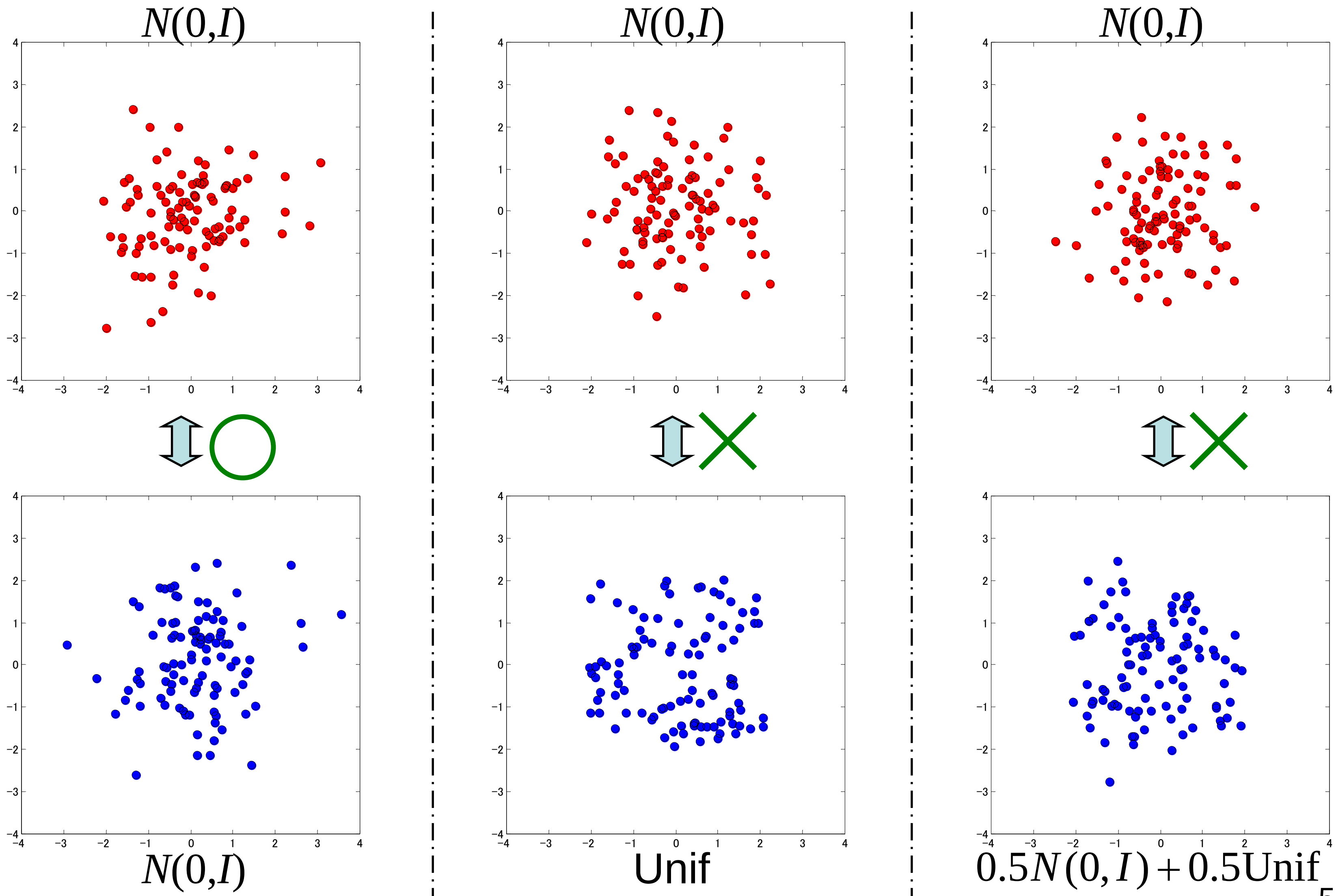
  - Are the experimental results of treatment and control significantly different?
  - Were the plays "*Henry VI*" and "*Henry II*" written by the same author?

- Kernel solution:

  Use the difference $m_X - m_Y$
  with a characteristic kernel such as Gaussian.

– Example: do they have the same distribution?

N = 100



Kernel Method for Two-sample Problem

- Maximum Mean Discrepancy (Gretton etal 07, NIPS19)
  - In population

$$MMD^2 = \left\| m_X - m_Y \right\|_H^2$$

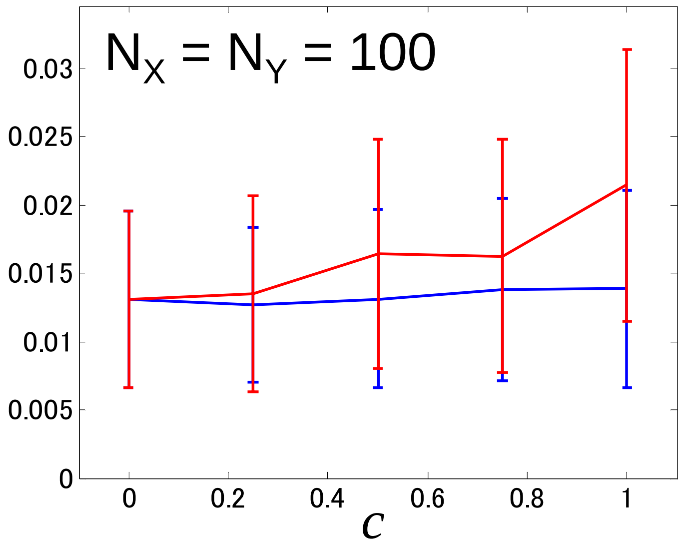
  - Empirically

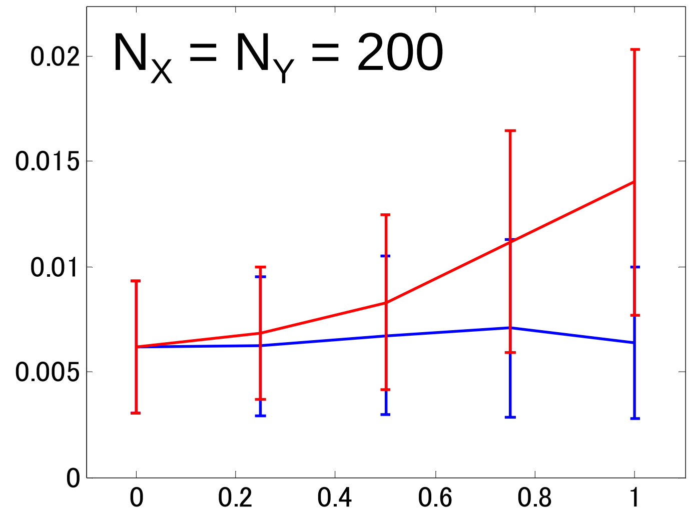
$$MMD_{emp}^2 = \left\| \hat{m}_X - \hat{m}_Y \right\|_H^2$$

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

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

Experiment with MMD

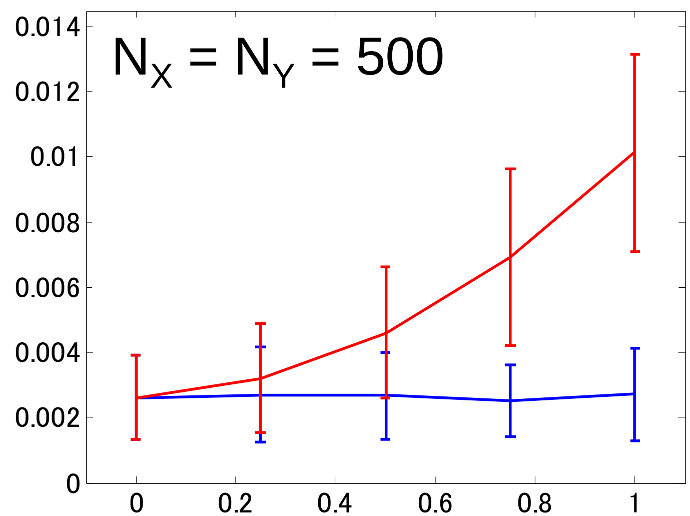




Means of MMD over 100 samples

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

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



Characteristic Function

- Definition

  X: random vector on $\mathbf{R}^m$ with law P_X

  Characteristic function of X is a complex-valued function defined by

$$\xi_X(u) \equiv E\left[e^{\sqrt{-1}u^T X}\right] = \int e^{\sqrt{-1}u^T x} dP_X(x) \qquad (u \in \mathbf{R}^m)$$

  If P_X has p.d.f. $p_X(x)$, the char. function is Fourier transform of $p_X(x)$.

- Moment generating function

$$\frac{1}{\sqrt{-1}^r} \frac{d^r}{du^r} \xi_X(u) = E\left[X^r\right]$$

- Chrac. function is very popular in probability and statistics for characterizing a probability.

Characterizing property

Theorem

X, Y: random vectors on $\mathbf{R}^m$ with prob. law P_X, P_Y (resp.).

$$\xi_X = \xi_Y \quad \Leftrightarrow \quad P_X = P_Y$$

Kernel and Ch. Function

- Fourier kernel is positive definite

$k_F(x, y) = \exp\left(\sqrt{-1}\, x^T y\right)$ is a (complex-valued) pos. def. kernel.

$\xi_X(u) = E[k_F(X, u)]$ = mean element with $k_F(x,y)$!!

  – Characteristic function is a special case of the mean element.

- Generalization of characteristic function approach
  – There are many "characteristic function" methods in the statistical literature (independent test, homogeneity test, etc).
  – The kernel methodology discussed here is generalizing this approach.
    • The data may not be Euclidean, but can be structured.

Re: Representation of Probability

- Various ways of representing a probability
  - Probability density function $p(x)$
  - Cumulative distribution function $F_X(t) = \mathrm{Prob}(\ X < t\ )$
  - All the moments $\mathrm{E}[X],\ \mathrm{E}[X^2],\ \mathrm{E}[X^3],\ \ldots$
  - Characteristic function $\xi_X(u) \equiv E\left[e^{\sqrt{-1}u^T X}\right] = \int e^{\sqrt{-1}u^T x} dP_X(x)$
  - Mean element on RKHS $m_X(u) = \mathrm{E}[k(X, u)]$

Each representation provides methods for statistical inference.

Summary of Part IV

- Statistics on RKHS → Inference on probabilities
  - Mean element → Characterization of probability
    Two-sample problem
  - Covariance operator → Dependence of two variables
    Independence test, Dependence measures
  - Conditional covariance operator → Conditional independence
    (Section VI)

- Characteristic kernel
  - A characteristic kernel gives a “rich” RKHS
  - A characteristic kernel characterizes a probability.
  - Kernel methodology is generalization of characteristic function methods

V. Statistical Test

Statistical Test

- How should we set the threshold?

Example) Based on a dependence measure, we wish to make a decision whether the variables are independent or not.

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

$$I(X,Y) < t \quad \Rightarrow \quad \text{dependent}$$
$$I(X,Y) \geq t \quad \Rightarrow \quad \text{independent}$$

But, the threshold should depend on the property 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 are independent"

- Prepare a test statistic T_N

  e.g. $T_N = HSIC_{emp}$

- Null distribution: Distribution of T_N under the null hypothesis

  This must be computed for $HSIC_{emp}$

- Set significance level α Typically $\alpha = 0.05$ or 0.01

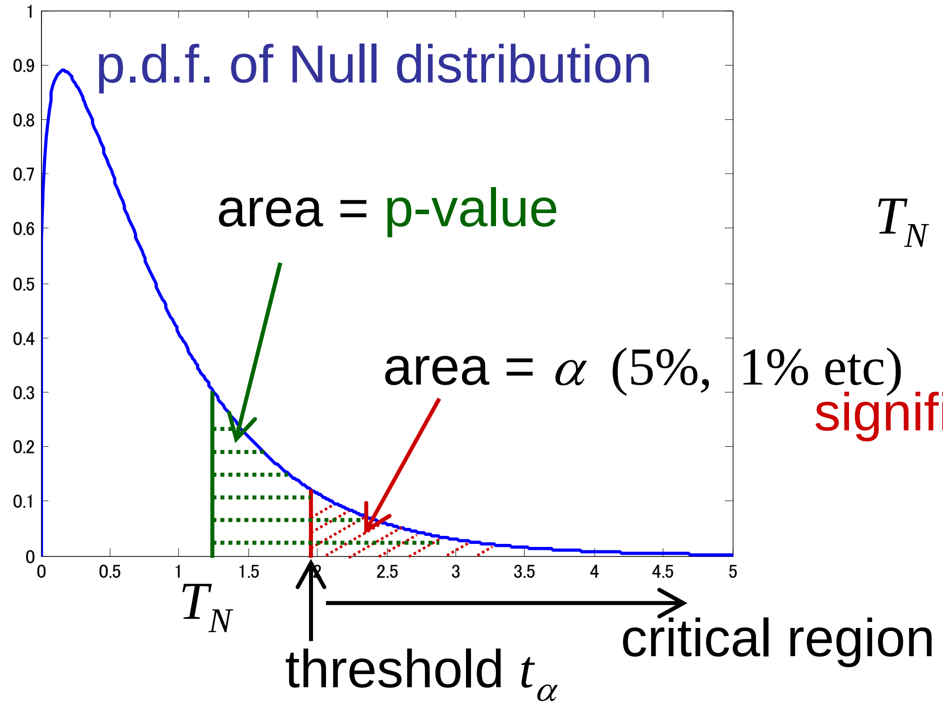
- Compute the critical region: α = Prob. of $T_N > t_\alpha$ under H_0.

- Reject the null hypothesis if $T_N > t_\alpha$,

  The probability that $HSIC_{emp} > t_\alpha$ under independence is very small.

  otherwise, accept the null hypothesis negatively.

One-sided test



$T_N > t_\alpha \iff$ p-value < α

- If the null hypothesis is the truth, the value of T_N should follow the above distribution.
- If the alternative 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. dependence = positive)
- Type II error = false negative

		TRUTH	
TEST RESULT		H_0	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, the type II error should be as small as possible.

Independence Test with HSIC

Independence Test

- Null hypothesis H_0: X and Y are independent

  Alternative H_1: X and Y are not independent (dependent)

- Test statistics

$$T_N = N \times \mathrm{HSIC}_{emp}$$

- Null distribution

  Under H_0 $\quad T_N \Rightarrow \sum_{a=1}^{\infty} \lambda_a Z_a^2$ convergence in distribution

  $(HSIC_{emp} = O_p(1/N))$

  where $Z_a \sim N(0,1)$ i.i.d.

  λ_a are the eigenvalues of an integral equation (not shown here)

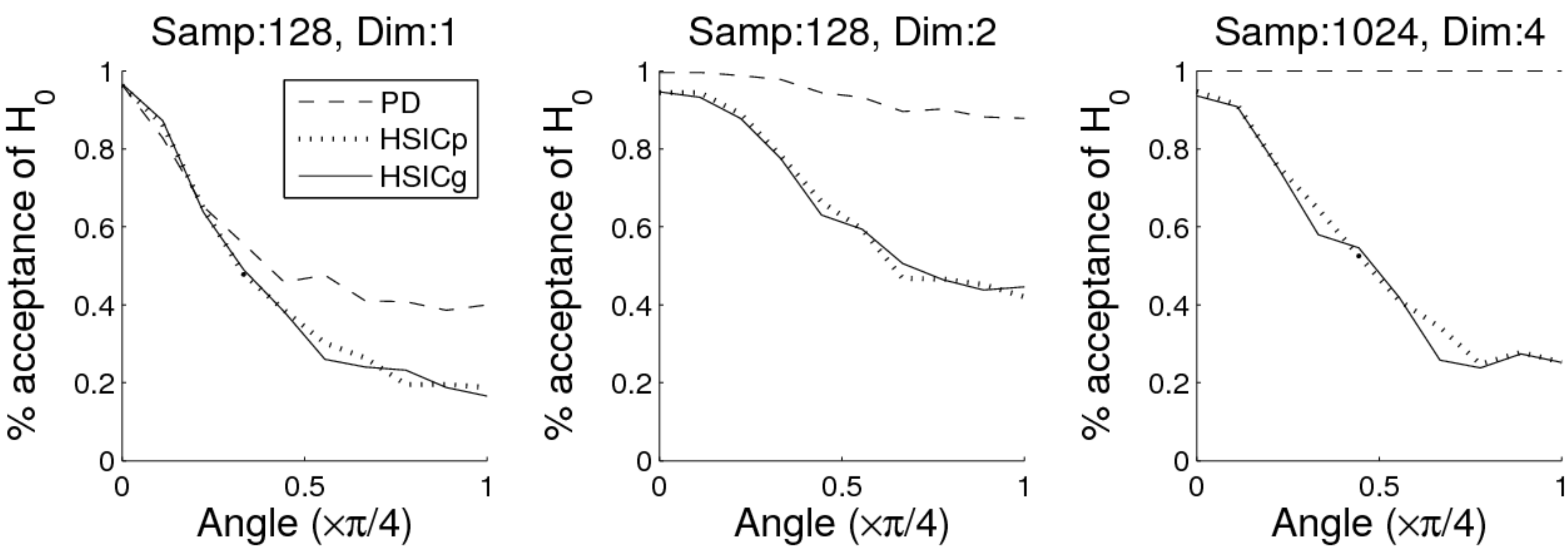
- Under alternative

$$T_N = O_p\left(\sqrt{N}\right) \quad (N \to \infty)$$

Example of Independent Test

- Synthesized data
  - Data: two d-dimensional samples

$$(X_1^{(1)},\dots,X_d^{(1)}),\dots,(X_1^{(N)},\dots,X_d^{(N)}) \qquad (Y_1^{(1)},\dots,Y_d^{(1)}),\dots,(Y_1^{(N)},\dots,Y_d^{(N)})$$



strength of dependence

Traditional Independence Test

- P.d.f.-based
  - Factorization of p.d.f. is used. $p(x_1,\ldots,x_m) = p(x_1)\cdots p(x_m)$
  - Parzen window approach.
  - Estimation accuracy is low for high dimensional data
- Cumulative distribution-based
  - Factorization of c.d.f. is used. $F^X(t_1,\ldots,t_m) = F^{X_1}(t_1)\cdots F^{X_m}(t_m)$
- Characteristic function-based
  - Factorization of characteristic function is used.
- Contingency table-based
  - Domain of each variable is partitioned into a finite number of parts.
  - Contingency table (number of counts) is used.
- And many others

Power Divergence (Ku&Fine05, Read&Cressie)

- Make partition $\{A_j\}_{j\in J}$: Each dimension is divided into q parts so that each bin contains almost the same number of data.
- Power-divergence

$$T_N = 2I^{\lambda}(X, m) = N\frac{2}{\lambda(\lambda+2)}\sum_{j\in J}\hat{p}_j\left\{\left(\hat{p}_j \Big/ \prod_{k=1}^{N}\hat{p}_{j_k}^{(k)}\right)^{\lambda} - 1\right\}$$

I^0 = MI

I^2 = Mean Square Conting.

$\hat{p}_j$: frequency in A_j

$\hat{p}_r^{(k)}$: marginal freq. in r-th interval

- Null distribution under independence

$$T_N \quad \Rightarrow \quad \chi^2_{q^N - qN + N - 1}$$

Limitations

- All the standard tests assume vector (numerical / discrete) data.
- They are often weak for high-dimensional data.

Independent Test on Text

- Data: Official records of Canadian Parliament in English and French.
  - Dependent data: 5 line-long parts from English texts and their French translations.
  - Independent data: 5 line-long parts from English texts and random 5 line-parts from French texts.
- Kernel: Bag-of-words and spectral kernel

Topic	Match	BOW(N=10) $HSIC_g$	BOW(N=10) $HSIC_p$	Spec(N=10) $HSIC_g$	Spec(N=10) $HSIC_p$	BOW(N=50) $HSIC_g$	BOW(N=50) $HSIC_p$	Spec(N=50) $HSIC_g$	Spec(N=50) $HSIC_p$
Agri-culture	Random	1.00	0.94	1.00	0.95	1.00	0.93	1.00	0.95
	Same	0.99	0.18	1.00	0.00	0.00	0.00	0.00	0.00
Fishery	Random	1.00	0.94	1.00	0.94	1.00	0.93	1.00	0.95
	Same	1.00	0.20	1.00	0.00	0.00	0.00	0.00	0.00
Immig-ration	Random	1.00	0.96	1.00	0.91	0.99	0.94	1.00	0.95
	Same	1.00	0.09	1.00	0.00	0.00	0.00	0.00	0.00

Acceptance rate (α = 5%)

(Gretton et al. 07)

Permutation Test

- The theoretical derivation of the null distribution is often difficult even asymptotically.
- The convergence to the asymptotic distribution may be very slow.
- Permutation test – Simulation of the null distribution
  - Make many samples consistent with the null hypothesis by random permutations of the original sample.
  - Compute the values of test statistics for the samples.

Independence test

$$\begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\ | & | & | & | & | & | & | \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 \end{matrix} \quad \Rightarrow \quad \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\ Y_5 & Y_1 & Y_7 & Y_4 & Y_2 & Y_6 & Y_3 \end{matrix}$$

independent

Two-sample test

$$\boxed{X_1\ X_2\ X_3\ X_4\ X_5}\ \boxed{Y_6\ Y_7\ Y_8\ Y_9\ Y_{10}} \Rightarrow \boxed{X_4\ Y_8\ X_2\ Y_9\ Y_6}\ \boxed{X_1\ X_3\ Y_7\ Y_{10}\ X_5}$$

homogeneous

- It can be computationally expensive.

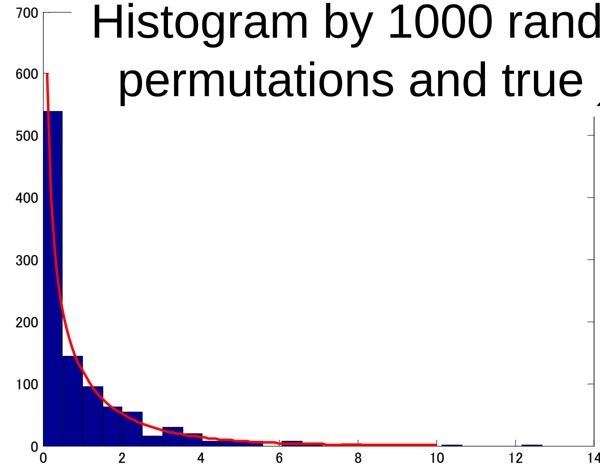
Independence test for 2 x 2 contingency table

– Contingency table

		Y	
		0	1
X	0	175	93
	1	71	161

many random permutations

Histogram by 1000 random permutations and true χ^2.



– Test statistic

$$T_N = N \sum_{i,j=0,1} \frac{(\hat{p}_{ij} - \hat{p}_{X,i}\hat{p}_{Y,j})^2}{\hat{p}_{X,i}\hat{p}_{Y,j}} \quad \Rightarrow \quad \chi^2 \quad (N \to \infty, \quad \text{under } H_0)$$

– Example

		Y	
		0	1
X	0	144	134
	1	102	120

P-value by true χ^2 = 0.193

P-value by permutation = 0.175

Independence is accepted with α = 5%

Independence test with various measures

– Data 1: dependent and uncorrelated by rotation (Part I)

X and Y: one-dimensional, N = 200

indep. ⟶ more dependent

Angle	0.0	4.5	9.0	13.5	18.0	22.5
HSIC (Median)	93	92	63	5	0	0
HSIC (Asymp. Var.)	93	44	1	0	0	0
HSNIC ($\varepsilon = 10^4$, Median)	94	23	0	0	0	0
HSNIC ($\varepsilon = 10^6$, Median)	92	20	1	0	0	0
HSNIC ($\varepsilon = 10^8$, Median)	93	15	0	0	0	0
HSNIC (Asymp. Var.)	94	11	0	0	0	0
MI (#NN = 1)	93	62	11	0	0	0
MI (#NN = 3)	96	43	0	0	0	0
MI (#NN = 5)	97	49	0	0	0	0
Conting. Table (#Bins=3)	100	96	46	9	1	0
Conting. Table (#Bins=4)	98	29	0	0	0	0
Conting. Table (#Bins=5)	98	82	5	0	0	0

acceptance of independence out of 100 tests (α = 5%)

– Data 2: Two coupled chaotic time series (coupled Hénon map)
X and Y: 4-dimensional, $N = 100$

	indep. → more dependent						
Coupling:	0.0	0.1	0.2	0.3	0.4	0.5	0.6
HSIC	75	70	58	52	13	1	0
HSNIC	97	66	21	1	0	1	0
MI (#NN=3)	87	91	83	73	23	6	0
MI (#NN=5)	87	88	75	67	23	5	0
MI (#NN=7)	87	86	75	64	21	5	0

acceptance of independence out of 100 tests (α = 5%)

Two sample test

- **Problem**

  Two i.i.d. samples $X_1,\ldots,X_N \qquad Y_1,\ldots,Y_N$

  Null hypothesis H$_0$: $P_X = P_Y$

  Alternative H$_1$: $P_X \neq P_Y$

- **Homogeneity test with MMD** (Gretton et al NIPS20)

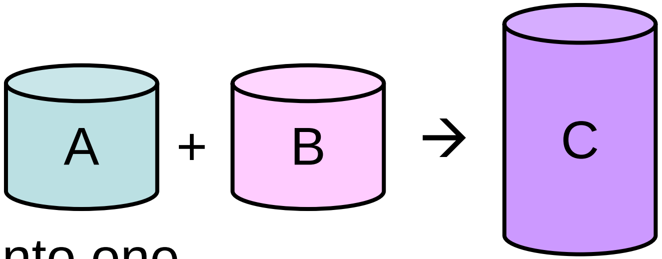
$$T_N = N \times \mathrm{MMD}^2_{\mathrm{emp}}$$

$$= \frac{1}{N}\sum_{i,j=1}^{N}\left\{k(X_i, X_j) - 2k(X_i, Y_j) + k(Y_i, Y_j)\right\}$$

- **Null distribution**
  - Similar to independence test with HSIC (not shown here)

Experiment

– Data integration



We wish to integrate two datasets into one.

The homogeneity should be tested!

% acceptance of homogeneity

Dataset	Attribut.	MMD^2	t-test	FR-WW	FR-KS
Neural I (w/wo spike)	Same	96.5	100.0	97.0	95.0
(N=4000,dim=63)	Diff.	**0.0**	42.0	**0.0**	10.0
Neural II (w/wo spike)	Same	95.2	100.0	95.0	94.5
(N=1000,dim=100)	Diff.	3.4	100.0	**0.8**	31.8
Microarray (health/tumor)	Same	94.4	100.0	94.7	96.1
(N=25,dim=12000)	Diff.	**0.8**	100.0	2.8	44.0
Microarray (subtype)	Same	96.4	100.0	94.6	97.3
(N=25,dim=2118)	Diff.	**0.0**	100.0	**0.0**	28.4

(Gretton et al. *NIPS20*, 2007)

Traditional Nonparametric Tests

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

  One-dimensional variables

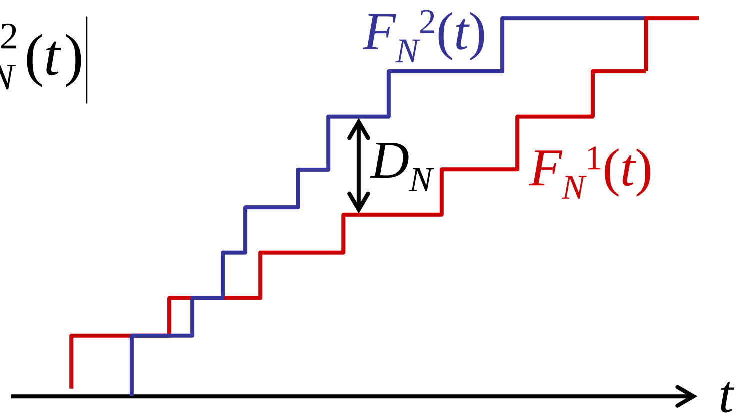
  – Empirical distribution function

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

  – 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

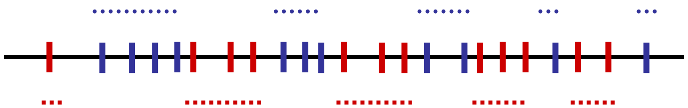
One-dimensional samples

- 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.

R = Number of runs

- Test statistics

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



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)

Summary of Part V

- **Statistical Test**
  - Statistical method of judging significance of a value.
  - It determines a “threshold” with some risk.

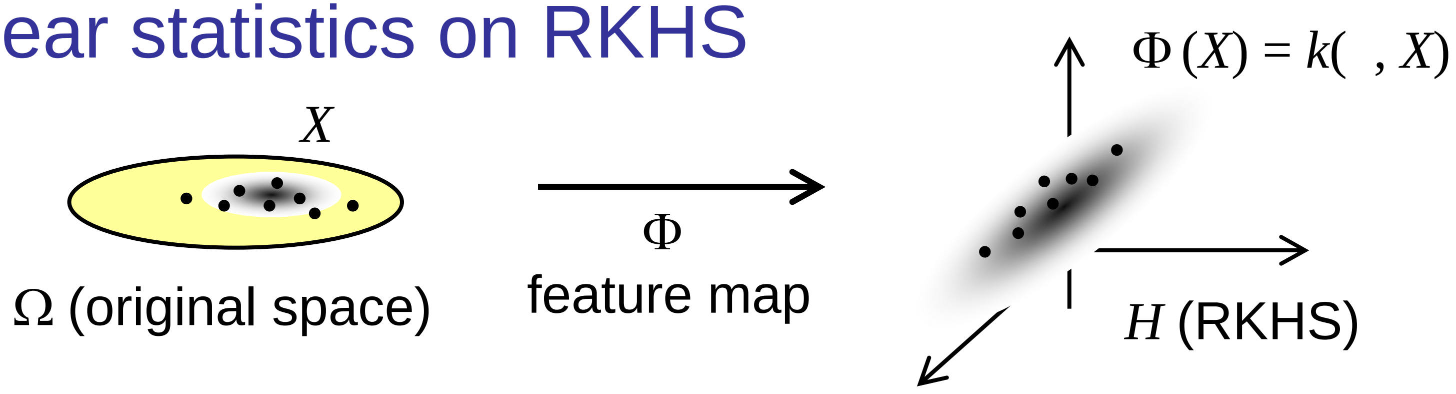
- **Statistical Test with kernels**
  - Independence test with HSIC
  - Two-sample test with MMD^2
  - Competitive with the state-of-art methods of nonparametric tests.
  - Kernel-based statistical tests work for structured data, to which conventional methods cannot be directly applied.

- **Permutation test**
  - It works well, if applicable.
  - Computationally expensive.

VI. Conditional Independence

Re: Statistics on RKHS

- Linear statistics on RKHS



– Basic statistics on Euclidean space $\longrightarrow$ Basic statistics on RKHS

Basic statistics on Euclidean space		Basic statistics on RKHS
Mean	$\longrightarrow$	Mean element
Covariance	$\longrightarrow$	Cross-covariance operator Σ_{YX}
Conditional covariance	$\longrightarrow$	Cond. cross-covariance operator

– Plan: define the basic statistics on RKHS and derive nonlinear/ nonparametric statistical methods in the original space.

Conditional Independence

Definition

X, Y, Z: random variables with joint p.d.f. $p_{XYZ}(x, y, z)$

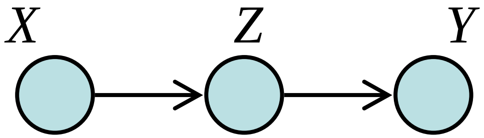
X and Y are conditionally independent given Z, if

$$p_{Y|ZX}(y \mid z, x) = p_{Y|Z}(y \mid z) \qquad \text{(A)}$$

or

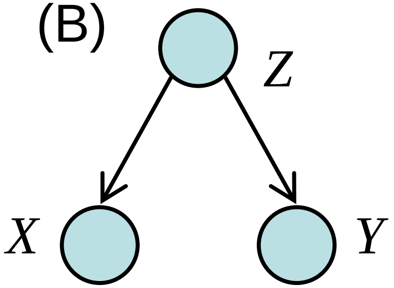
$$p_{XY|Z}(x, y \mid z) = p_{X|Z}(x \mid z)\, p_{Y|Z}(y \mid z) \qquad \text{(B)}$$

(A)



With Z known, the information of X is unnecessary for the inference on Y

(B)



Review: Conditional Covariance

- Conditional covariance of Gaussian variables
  - Jointly Gaussian variable

    $X = (X_1, \ldots, X_p), Y = (Y_1, \ldots, Y_q)$

    $Z = (X, Y)$: $m\ (= p + q)$ dimensional Gaussian variable

    $$Z \sim N(\mu, V) \qquad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \qquad V = \begin{pmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{pmatrix}$$

  - Conditional probability of Y given X is again Gaussian

    $\sim N(\mu_{Y|X}, V_{YY|X})$

    Cond. mean $\quad \mu_{Y|X} \equiv E[Y \mid X = x] = \mu_Y + V_{YX} V_{XX}^{-1} (x - \mu_X)$

    Cond. covariance $\quad V_{YY|X} \equiv Cov[Y \mid X = x] = \underline{V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}}$

    Schur complement of V_{XX} in V

    Note: $V_{YY|X}$ does not depend on x

Conditional Independence for Gaussian Variables

- Two characterizations

X, Y, Z are Gaussian.

– Conditional covariance

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{XY|Z} = O \qquad \text{i.e.} \quad V_{YX} - V_{YZ}V_{ZZ}^{-1}V_{ZX} = O$$

– Comparison of conditional variance

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{YY|[X,Z]} = V_{YY|Z}$$

$$\because) \quad V_{YY} - V_{Y[X,Z]}V_{[X,Z][X,Z]}^{-1}V_{[Z,X]Y} = V_{YY} - (V_{YX}, V_{YZ})\begin{pmatrix} V_{XX} & V_{XZ} \\ V_{ZX} & V_{ZZ} \end{pmatrix}^{-1}\begin{pmatrix} V_{XY} \\ V_{ZY} \end{pmatrix}$$

$$= V_{YY} - (V_{YX}, V_{YZ})\begin{pmatrix} I & O \\ -V_{ZZ}^{-1}V_{ZX} & I \end{pmatrix}\begin{pmatrix} V_{XX|Z}^{-1} & O \\ O & V_{ZZ}^{-1} \end{pmatrix}\begin{pmatrix} I & -V_{XZ}V_{ZZ}^{-1} \\ O & I \end{pmatrix}\begin{pmatrix} V_{XY} \\ V_{ZY} \end{pmatrix}$$

$$= V_{YY|Z} - V_{YX|Z}V_{XX|Z}^{-1}V_{XY|Z}$$

Linear Regression and Conditional Covariance

Review: linear regression

- X, Y: random vector (not necessarily Gaussian) of dim p and q (resp.)

  $$\tilde{X} = X - E[X], \quad \tilde{Y} = Y - E[Y]$$

- Linear regression: predict Y using the linear combination of X. Minimize the mean square error:

  $$\min_{A: q\times p \text{ matrix}} E\left\|\tilde{Y} - A\tilde{X}\right\|^2$$

- The residual error is given by the conditional covariance matrix.

  $$\min_{A: q\times p \text{ matrix}} E\left\|\tilde{Y} - A\tilde{X}\right\|^2 = \mathrm{Tr}\left[V_{YY|X}\right] = \mathrm{Tr}\left[Cov[Y \mid X]\right]$$

– Derivation

$$E\left\|\tilde{Y} - A\tilde{X}\right\|^2 = \mathrm{Tr}\left[E[\tilde{Y}\tilde{Y}^T] - AE[\tilde{X}\tilde{Y}^T] - E[\tilde{Y}\tilde{X}^T]A^T + AE[\tilde{X}\tilde{X}^T]A^T\right]$$

$$= \mathrm{Tr}\left[V_{YY} - AV_{XY} - V_{YX}A^T + AV_{XX}A^T\right]$$

$$= \mathrm{Tr}\left[(A - V_{YX}V_{XX}^{-1})V_{XX}(A - V_{YX}V_{XX}^{-1})^T\right] + \mathrm{Tr}\left[V_{YY} - V_{YX}V_{XX}^{-1}V_{XY}\right]$$

$\Rightarrow$

$$A_{opt} = V_{YX}V_{XX}^{-1}$$

and

$$E\left\|\tilde{Y} - A_{opt}\tilde{X}\right\|^2 = \mathrm{Tr}\left[V_{YY} - V_{YX}V_{XX}^{-1}V_{XY}\right]$$

– For Gaussian variables,

$$V_{YY|[X,Z]} = V_{YY|Z} \qquad (\Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z \quad)$$

can be interpreted as

"If Z is known, X is not necessary for linear prediction of Y."

Conditional Covariance on RKHS

Conditional Cross-covariance operator

X, Y, Z : random variables on Ω_X, Ω_Y, Ω_Z (resp.).

(H_X, k_X), (H_Y, k_Y), (H_Z, k_Z) : RKHS defined on Ω_X, Ω_Y, Ω_Z (resp.).

– Conditional cross-covariance operator $H_X \to H_Y$

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

Note: Σ_{ZZ}^{-1} may not exist. But, we have the decomposition

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2}$$

Rigorously, define

$$\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YY}^{1/2} W_{YZ} W_{ZX} \Sigma_{XX}^{1/2}$$

– Conditional covariance operator

$$\Sigma_{YY|Z} \equiv \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$$

Two Characterizations of Conditional Independence with Kernels

(1) Conditional covariance operator (FBJ04, 06)

Under some “richness” assumptions on RKHS (e.g Gaussian)

– Conditional variance

$$\langle g, \Sigma_{YY|Z} g \rangle = E\left[Var[g(Y) \mid Z]\right] = \inf_{f \in H_Z} E\left|\tilde{g}(Y) - \tilde{f}(Z)\right|^2$$

– Conditional independence

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad \Sigma_{YY|[XZ]} = \Sigma_{YY|Z}$$

X is not necessary for predicting $g(Y)$

– *c.f.* Gaussian variables

$$b^T V_{YY|Z} b = Var[b^T Y \mid Z] = \min_a \left| b^T \tilde{Y} - a^T \tilde{Z} \right|^2$$

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{YY|[X,Z]} = V_{YY|Z}$$

(2) Cond. cross-covariance operator (FBJ04, Sun et al. 07)

Under some "richness" assumptions on RKHS (e.g. Gaussian),

- Conditional Covariance

$$\left\langle g, \Sigma_{YX|Z} f \right\rangle = E\left[Cov[g(Y), f(X) \mid Z]\right]$$

- Conditional independence

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad \Sigma_{Y\ddot{X}|Z} = O \qquad \left(\Leftrightarrow \quad \Sigma_{\ddot{Y}X|Z} = O\right)$$

where $\ddot{X} = (X, Z),\ \ddot{Y} = (Y, Z)$

- *c.f.* Gaussian variables

$$a^T V_{XY|Z} b = Cov[a^T X, b^T Y \mid Z]$$

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

– Why is “extended variable” needed?

$$\left\langle g, \Sigma_{YX|Z} f \right\rangle = E\left[Cov[g(Y), f(X) \mid Z]\right]$$

$$\left\langle g, \Sigma_{YX|Z} f \right\rangle \neq Cov[g(Y), f(X) \mid Z = z]$$

The l.h.s is not a funciton of z.   *c.f*. Gaussian case

$$\Sigma_{YX|Z} = O \quad \Rightarrow \quad p(x, y) = \int p(x \mid z) p(y \mid z) p(z) dz$$

$$\Sigma_{YX|Z} = O \quad \not\Rightarrow \quad p(x, y \mid z) = p(x \mid z) p(y \mid z)$$

However, if X is replaced by $[X, Z]$

$$\Sigma_{Y[X,Z]|Z} = O \quad \Rightarrow \quad p(x, y, z') = \int p(x, z' \mid z) p(y \mid z) p(z) dz$$

where $\quad p(x, z' \mid z) = p(x \mid z) \delta(z' - z)$

$$\Rightarrow \qquad p(x, y, z') = p(x \mid z') p(y \mid z') p(z')$$

i.e. $\quad p(x, y \mid z') = p(x \mid z') p(y \mid z')$

Application to Dimension Reduction for Regression

- Dimension reduction

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

Goal: find an effective subspace spanned by an m x d matrix B s.t.

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

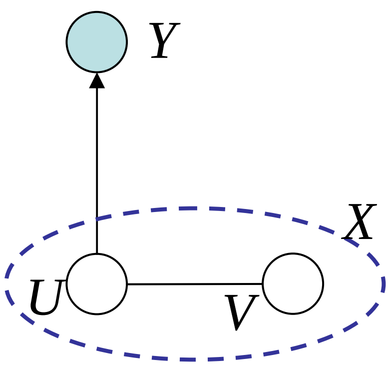
linear feature vector

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

- Conditional independence

B spans effective subspace

$\Longleftrightarrow$ $X \perp\!\!\!\perp Y \mid B^TX$



Kernel Dimension Reduction

(Fukumizu, Bach, Jordan 2004, 2006)

Use d-dimensional Gaussian kernel $k_d(z_1,z_2)$ for $B^T X$, and a characteristic kernel for Y.

$$\Sigma_{YY|B^T X} \geq \Sigma_{YY|X}$$

($\geq$: the partial order of self-adjoint operators)

$$\Sigma_{YY|B^T X} = \Sigma_{YY|X} \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid B^T X$$

$$\boxed{\min_{B:\, B^T B = I_d} \mathrm{Tr}\left[\Sigma_{YY|B^T X}\right]}$$

Very general method for dimension reduction:
No model for regression, no strong assumption on the distributions.
Optimization is not easy.

See FBJ 04, 06 for further details.
(Extension: Nilsson et al. ICML07)

Experiments with KDR

- Wine data

  Data

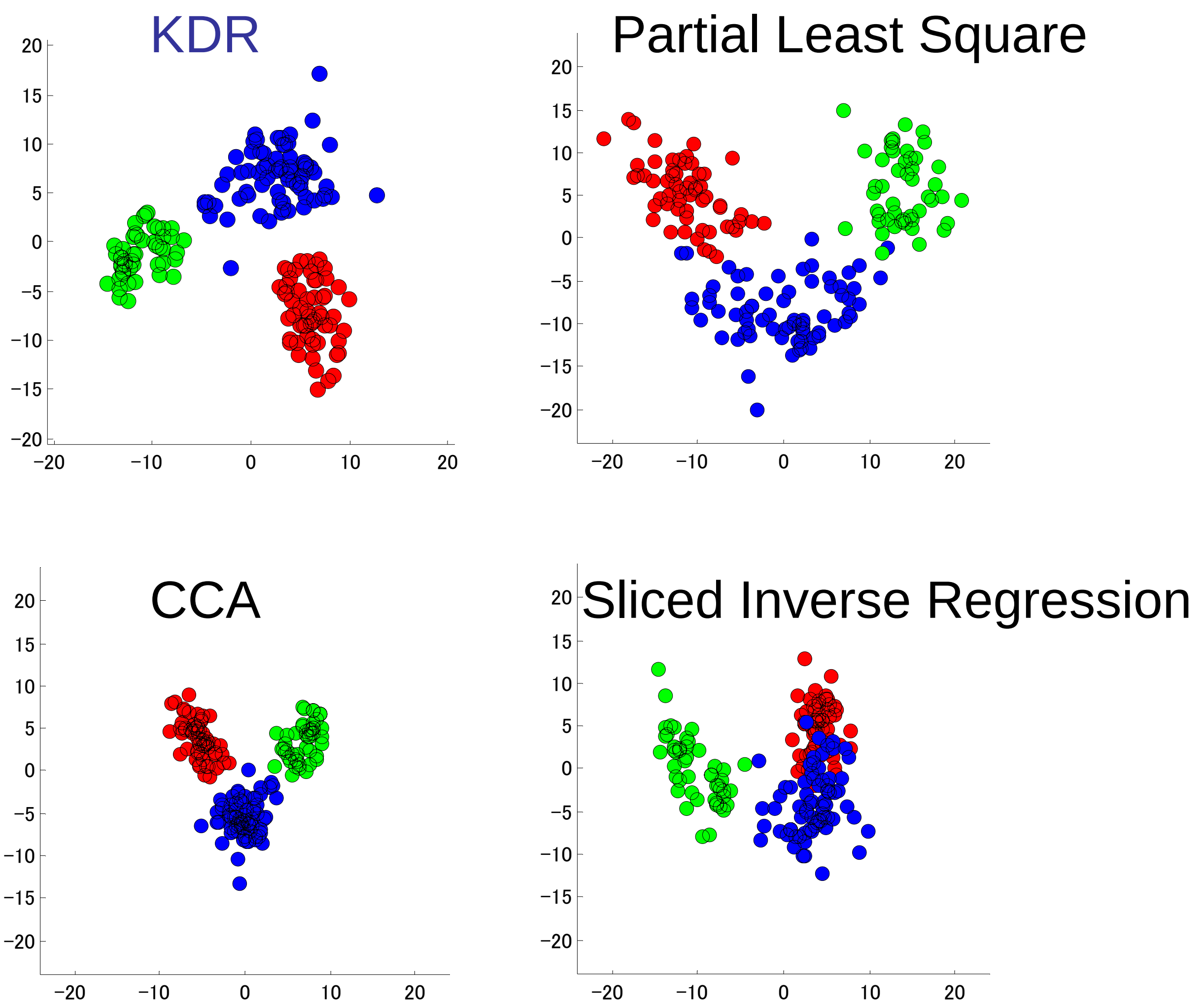
  13 dim. 178 data

  3 classes

  2 dim. projection

$$k(z_1, z_2) = \exp\left(-\|z_1 - z_2\|^2 / \sigma^2\right)$$

$$\sigma = 30$$



Measure of Cond. Independence

- HS norm of cond. cross-covariance operator
  - Measure for conditional dependence

$$HSCIC = \left\| \Sigma_{\ddot{X}\ddot{Y}|Z} \right\|_{HS}^{2} \qquad \ddot{X} = (X, Z), \ddot{Y} = (Y, Z)$$

  - Conditional independence
    Under some "richness" assumptions (e.g. Gaussian),

$$HSCIC = \left\| \Sigma_{\ddot{X}\ddot{Y}|Z} \right\|_{HS}^{2}$$ is zero if and only if $X \perp\!\!\!\perp Y \mid Z$

  - Empirical measure

$$HSCIC_{emp} = \mathrm{Tr}\Big[G_X G_Y - 2 G_X (G_Z + N\varepsilon_N I_N)^{-1} G_Z G_Y + G_Z (G_Z + N\varepsilon_N I_N)^{-1} G_X (G_Z + N\varepsilon_N I_N)^{-1} G_Z G_Y \Big]$$

Normalized Cond. Covariance

- Normalized conditional cross-covariance operator

$$W_{YX|Z} \equiv W_{YX} - W_{YZ}W_{ZX}$$ Recall: $$\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2}$$

$$W_{YX|Z} = \Sigma_{YY}^{-1/2} \Sigma_{YX|Z} \Sigma_{XX}^{-1/2} = \Sigma_{YY}^{-1/2} \left(\Sigma_{YX} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} \right) \Sigma_{XX}^{-1/2}$$

  – Conditional independence

    Under some “richness” assumptions (e.g. Gaussian),

$$W_{Y\ddot{X}|Z} = O \qquad \Leftrightarrow \qquad X \perp\!\!\!\perp Y \mid Z$$

  – HS Normalized Conditional Independence Criteria

$$HSNCIC = \left\| W_{\ddot{X}\ddot{Y}|Z} \right\|_{HS}^{2}$$

$$HSNCIC = 0 \qquad \Leftrightarrow \qquad X \perp\!\!\!\perp Y \mid Z$$

– Kernel-free expression. Under some "richness" assumptions,

$$\| W_{\ddot{Y}\ddot{X}|Z} \|_{HS}^2$$

$$= \iint \left(\frac{p_{XYZ}(x,y,z) - p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z)}{p_{XZ}(x,z)p_{YZ}(y,z)} \right)^2 p_{XZ}(x,z)p_{YZ}(y,z)dxdydz$$

("Conditional" mean square contingency)

– Empirical estimator of HSNCIC

$$HSNCIC_{emp} = \mathrm{Tr}\left[R_{\ddot{X}} R_{\ddot{Y}} - 2R_{\ddot{X}} R_{\ddot{Y}} R_Z + R_{\ddot{X}} R_Z R_{\ddot{Y}} R_Z \right]$$

$R_{\ddot{X}} \equiv G_{\ddot{X}} \left(G_{\ddot{X}} + N\varepsilon_N I_N \right)^{-1}$ etc.

Conditional Independence Test

- Permutation test with the kernel measure

$$T_N = \left\| \hat{\Sigma}^{(N)}_{YX|Z} \right\|^2_{HS} \qquad \text{or} \qquad T_N = \left\| \hat{W}^{(N)}_{YX|Z} \right\|^2_{HS}$$

– If Z takes values in a finite set $\{1, \ldots, L\}$,

set $A_\ell = \{i \mid Z_i = \ell\} \ (\ell = 1,\ldots,L)$,

otherwise, partition the values of Z into L subsets $C_1, \ldots, C_L$, and set

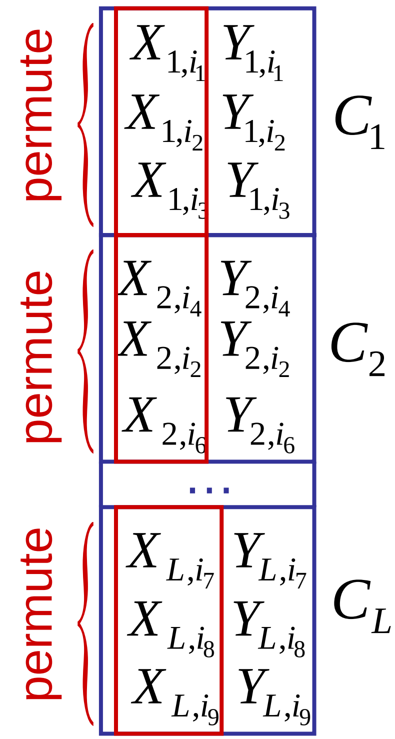
$$A_\ell = \{i \mid Z_i \in C_\ell\} \quad (\ell = 1,\ldots,L).$$

– Repeat the following process B times: ($b = 1, \ldots, B$)

1. Generate pseudo cond. independent data $D^{(b)}$ by permuting X data within each A_ℓ.
2. Compute $T_N^{(b)}$ for the data $D^{(b)}$.

$\longrightarrow$ Approximate null distribution under cond. indep. assumption

– Set the threshold by the $(1-\alpha)$-percentile of the empirical distributions of $T_N^{(b)}$.



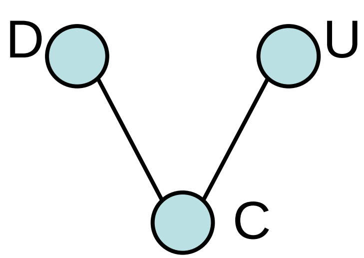
Application to Graphical Modeling

– Three continuous variables of medical measurements. N = 35. (Edwards 2000, Sec.3.1.4)

Creatinine clearance (C), Digoxin clearance (D), Urine flow (U)

Kernel mehod (permut. test)			Linear method		
	HSN(C)IC	P-val.		(partial) cor.	P-val.
$D \perp\!\!\!\perp U \mid C$	1.458	0.924	Parcor(D,U|C)	0.4847	0.0037
$C \perp\!\!\!\perp D$	0.776	<0.001	Cor(C,D)	0.7754	0.0000
$C \perp\!\!\!\perp U$	0.194	0.117	Cor(C,U)	0.3092	0.0707
$D \perp\!\!\!\perp U$	0.343	0.023	Cor(D,U)	0.5309	0.0010

– Suggested undirected graphical model by kernel method



The conditional independence $D \perp\!\!\!\perp U \mid C$ coincides with the medical knowledge.

Statistical Consistency

- Consistency on conditional covariance operator

Theorem (FBJ06, Sun et al. 07)

Assume $\varepsilon_N \to 0$ and $\sqrt{N}\varepsilon_N \to \infty$

$$\left\| \hat{\Sigma}^{(N)}_{YX|Z} - \Sigma_{YX|Z} \right\|_{HS} \to 0 \qquad (N \to \infty)$$

In particular,

$$\left\| \hat{\Sigma}^{(N)}_{YX|Z} \right\|_{HS} \to \left\| \Sigma_{YX|Z} \right\|_{HS} \qquad (N \to \infty)$$

i.e. HSCIC$_{\mathrm{emp}}$ converges to the population value HSCIC.

Consistency of normalized conditional covariance operator

<u>Theorem (FGSS07)</u>

Assume that $W_{YX|Z}$ is Hilbert-Schmidt, and the regularization coefficient satisfies $\varepsilon_N \to 0$ and $N^{1/3}\varepsilon_N \to \infty$. Then,

$$\left\| \hat{W}_{YX|Z}^{(N)} - W_{YX|Z} \right\|_{HS} \to 0 \qquad (N \to \infty)$$

In particular,

$$\left\| \hat{W}_{YX|Z}^{(N)} \right\|_{HS} \to \left\| W_{YX|Z} \right\|_{HS} \qquad (N \to \infty)$$

i.e. HSNCIC$_{\mathrm{emp}}$ converges to the population value HSNCIC.

– Note: Convergence in HS-norm is stronger than convergence in operator norm.

Summary of Part V

- **Conditional independence by kernels**
  - Conditional independence is characterized in two ways;
    - Conditional covariance operator

$$X \perp\!\!\!\perp Y \mid Z \qquad \Leftrightarrow \qquad \Sigma_{YY|[XZ]} = \Sigma_{YY|Z}$$

    - Conditional cross-covariance operator

$$X \perp\!\!\!\perp Y \mid Z \qquad \Leftrightarrow \qquad \Sigma_{Y\ddot{X}|Z} = O \qquad \text{or} \qquad \Sigma_{\ddot{Y}X|Z} = O$$

- **Kernel Dimensional Reduction**

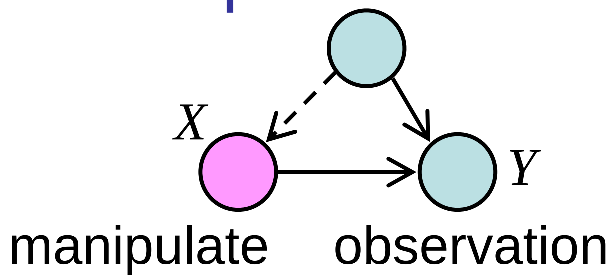
  A very general method for dimension reduction for regression

- **Measures for conditional independence**
  - HS norm of conditional cross-covariance operator
  - HS norm of normalized conditional cross-covariance operator
    Kernel free in population.

VII. Causal Inference

Causal Inference

- With manipulation – intervention



X is a cause of Y?

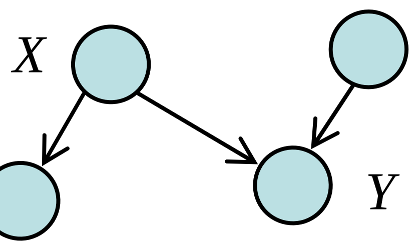
Easier. (*do*-calculus, Pearl 1995)

- No manipulation / with temporal information

$X(t) \quad Y(t)$: observed time series

$X(1), \ldots, X(t)$ are a cause of $Y(t+1)$?

- No manipulation / no temporal information



Causal inference is harder.

Difficulty of causal inference from non-experimental data

- Widely accepted view till 80's

  Causal inference is impossible without manipulating some variables.

  e.g.) *"No causation without manipulation"* (Holland 1986, JASA)

- Temporal information is very helpful, but not decisive.

  e.g.) The barometer falls before it rains, but it does not cause the rain.

- Many philosophical discussions, but not discussed here.

  See Pearl (2000) and the references therein.

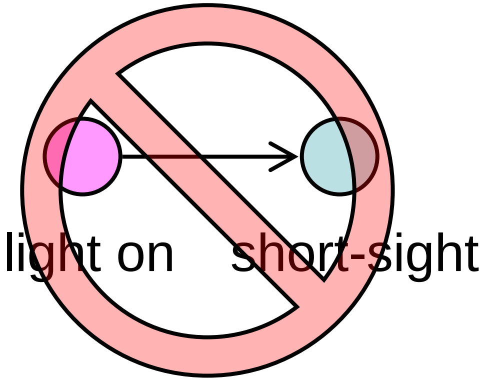
■ Correlation (dependence) and causality

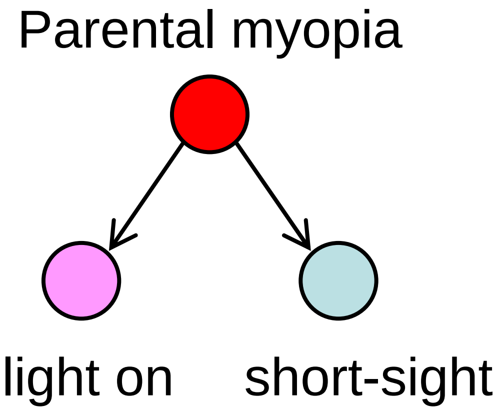
Do not confuse causality with dependence (or correlation)!

Example)

A study shows:

Young children who sleep with the light on are much more likely to develop myopia in later life. (*Nature* 1999)





(*Nature* 2000)

Hidden common cause

Causality of Time Series

- Granger causality (Granger 1969)

  $X(t)$, $Y(t)$: two time series $\quad t = 1, 2, 3, \ldots$

  – Problem:

  Is $\{X(1), \ldots, X(t)\}$ a cause of $Y(t+1)$?

  (No inverse causal relation)

  – Granger causality

  Model: AR

  $$Y(t) = c + \sum_{i=1}^{p} a_i Y(t-i) + \sum_{j=1}^{p} b_j X(t-j) + U_t$$

  Test

  $$H_0: \; b_1 = b_2 = \ldots = b_p = 0$$

  X is called a Granger cause of Y if H_0 is rejected.

- *F*-test
  - Linear estimation

$$Y(t) = c + \sum_{i=1}^{p} a_i Y(t-i) + \sum_{j=1}^{p} b_j X(t-j) + U_t \longrightarrow \hat{c}, \hat{a}_i, \hat{b}_j$$

H_0: $$Y(t) = c + \sum_{i=1}^{p} a_i Y(t-i) + W_t \longrightarrow \hat{\hat{c}}, \hat{\hat{a}}_i$$

$$ERR_1 = \sum_{t=p+1}^{N} \left(\hat{Y}(t) - Y(t)\right) \qquad ERR_0 = \sum_{t=p+1}^{N} \left(\hat{\hat{Y}}(t) - Y(t)\right)^2$$

  - Test statistics

$$T_N \equiv \frac{\left(ERR_0 - ERR_1\right)/p}{ERR_1/(N-2p+1)} \overset{\text{under } H_0}{\Rightarrow} F_{p,N-2p+1} \quad (N \to \infty)$$

p.d.f of $F_{d_1,d_2} = \frac{1}{B(d_1/2, d_2/2)} \left(\frac{d_1 x}{d_1 x + d_2}\right)^{d_1} \left(1 - \frac{d_1 x}{d_1 x + d_2}\right)^{d_2} \frac{1}{x}$

- Software
  - Matlab: Econometrics toolbox (www.spatial-econometrics.com)
  - R: lmtest package

- Granger causality is widely used and influential in econometrics.

  Clive Granger received Nobel Prize in 2003.

- Limitations
  - Linearity: linear AR model is used.
    No nonlinear dependence is considered.
  - Stationarity: stationary time series are assumed.
  - Hidden cause: hidden common causes (other time series) cannot be considered.

  "Granger causality" is not necessarily "causality" in general sense.

- There are many extensions.

- With kernel dependence measures, it is easily extended to incorporate nonlinear dependence.

  Remark: There are few good conditional independence tests
  for continuous variables.

Kernel Method for Causality of Time Series

- Causality by conditional independence
  - Extended notion of Granger causality

    X is NOT a cause of Y if

$$p(Y_t \mid Y_{t-1},\ldots,Y_{t-p},X_{t-1},\ldots,X_{t-p}) = p(Y_t \mid Y_{t-1},\ldots,Y_{t-p})$$

$$\Leftrightarrow$$

$$Y_t \perp\!\!\!\perp X_{t-1},\ldots,X_{t-p} \mid Y_{t-1},\ldots,Y_{t-p}$$

  - Kernel measures for causality

$$HSCIC = \left\| \hat{\Sigma}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$$

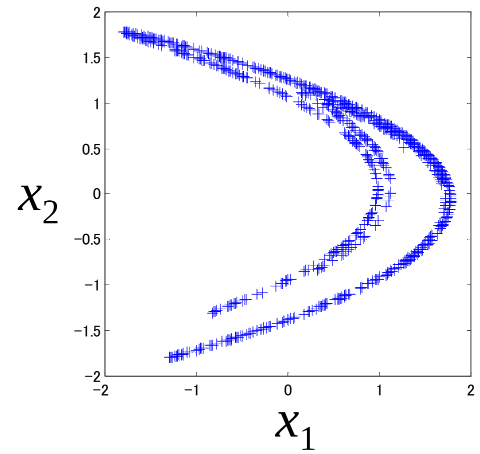
$$HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$$

$$\mathbf{X}_{\mathrm{p}} = \{(X_{t-1,}X_{t-2},\cdots,X_{t-p}) \in \mathbf{R}^p \mid t = p+1,\ldots,N\}$$

$$\mathbf{Y}_{\mathrm{p}} = \{(Y_{t-1,}Y_{t-2},\cdots,Y_{t-p}) \in \mathbf{R}^p \mid t = p+1,\ldots,N\}$$

Example

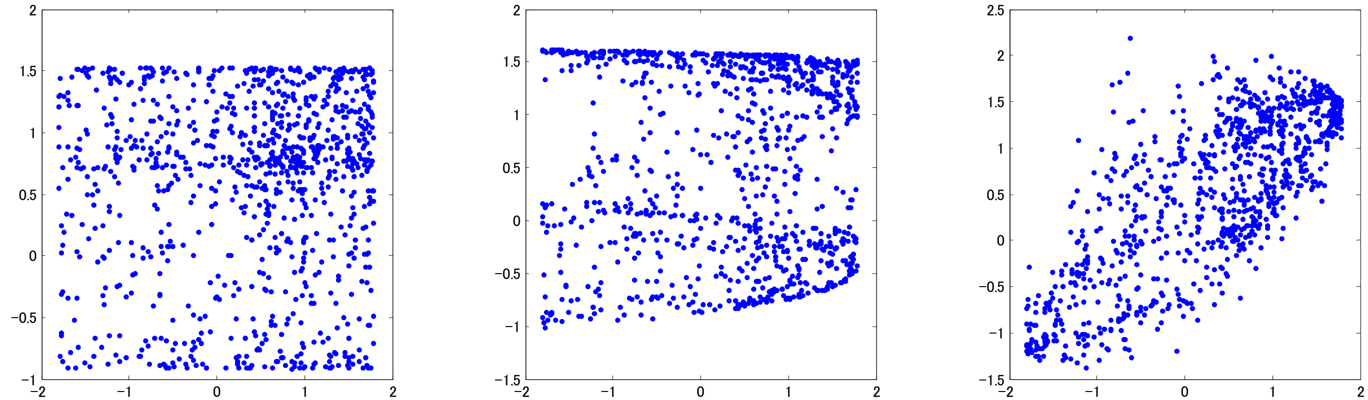
- Coupled Hénon map
  - X, Y:



$$\begin{cases} x_1(t+1) = 1.4 - x_1(t)^2 + 0.3x_2(t) \\ x_2(t+1) = x_1(t) \end{cases}$$

$$\begin{cases} y_1(t+1) = 1.4 - \left\{ \underline{\gamma x_1(t)} y_1(t) + (1-\gamma) y_1(t)^2 \right\} + 0.1 y_2(t) \\ y_2(t+1) = y_1(t) \end{cases}$$

x_1-y_1



$\gamma = 0$ $\quad$ $\gamma = 0.25$ $\quad$ $\gamma = 0.8$

Causality of coupled Hénon map

- X is a cause of Y if $\gamma > 0$. $\quad Y_t \not\perp\!\!\!\perp X_{t-1},\ldots,X_{t-p} \mid Y_{t-1},\ldots,Y_{t-p}$
- Y is not a cause of X for all γ. $\quad X_t \perp\!\!\!\perp Y_{t-1},\ldots,Y_{t-p} \mid X_{t-1},\ldots,X_{t-p}$
- Permutation tests for non-causality with $HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$

N = 100

$x_1 - y_1$	H$_0$: Y_t is not a cause of X_{t+1}							H$_0$: X_t is not a cause of Y_{t+1}						
γ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.0	0.1	0.2	0.3	0.4	0.5	0.6
HSNCIC	94	88	81	63	86	77	62	97	0	0	0	0	0	0
Granger	92	96	95	90	90	94	93	96	92	85	45	13	2	3

1-dimensional independent noise is added to $X(t)$ and $Y(t)$.

HSNCIC	97	96	93	85	81	68	75	96	0	0	0	0	0	0

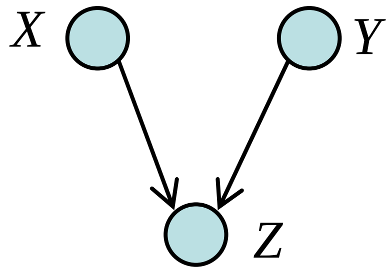
Number of times accepting H$_0$ among 100 datasets (α = 5%)

Causal Inference from Non-experimental Data

Why is it possible?

- DAG of chain X – Z – Y

V-structure



$\Longleftrightarrow$

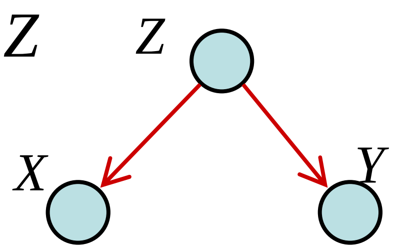
$$X \perp\!\!\!\perp Y$$

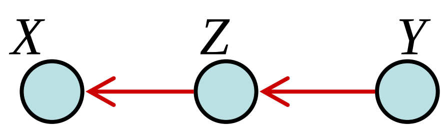
and

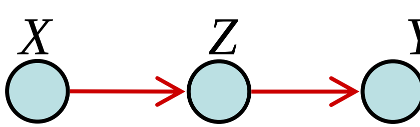
$$X \not\perp\!\!\!\perp Y \mid Z$$

- This is the only detectable directed graph of three variables.
- The following structures cannot be distinguished from the probability.

$$X \perp\!\!\!\perp Y \mid Z$$







p(x,y,z) = p(x|z)p(y|z)p(z) = p(x|z)p(z|y)p(y) = p(x|z)p(z|y)p(x)

Causal Learning Methods

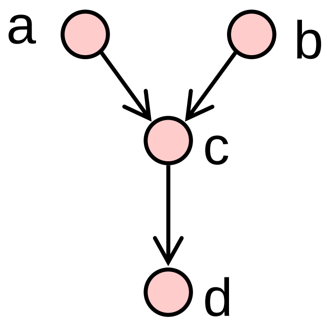
- Constraint-based method (discussed in this lecture)
  - Determine the (cond.) independence of the underlying probability.
  - Relatively efficient for hidden variables.

- Score-based method
  - Structure learning of Bayesian network (Ghahramani's lecture)
  - Able to use informative prior.
  - Optimization in huge search space.
  - Many methods assume discrete variables (discretization) or parametric model.

- Common hidden causes
  - For simplicity, algorithms assuming no hidden variables are explained in this lecture.

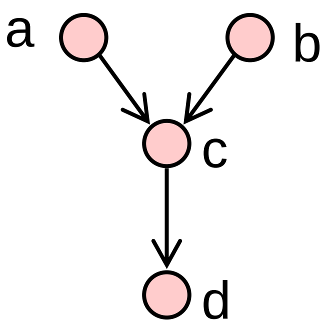
Fundamental Assumptions

- Markov assumption on a DAG
  - Causal relation is expressed by a DAG, and the probability generating data is consistent with the graph.

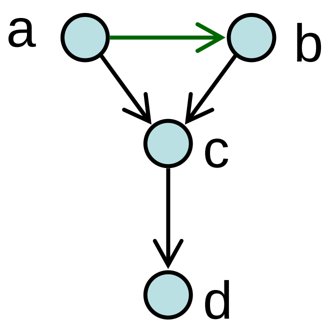


$$p(X) = p(X_a)p(X_b)p(X_c \mid X_a, X_b)p(X_d \mid X_c)$$

- Faithfulness (stability)
  - The inferred DAG (causal structure) must express all the independence relations.



true



unfaithful

This includes the true probability as a special case, but the structure does not express $a \perp\!\!\!\perp b$

Inductive Causation

- IC algorithm (Verma&Pearl 90)

Input – V: set of variables, D: dataset of the variables.

Output – DAG (specifies an equivalence class, directed partially)

1. For each $(a,b) \in V \times V \ (a \neq b)$, search for $S_{ab} \subset V \setminus \{a,b\}$ such that
$$X_a \perp\!\!\!\perp X_b \mid S_{ab}$$
Construct an undirected graph (skeleton) by connecting a and b if and only if no set S_{ab} can be found.

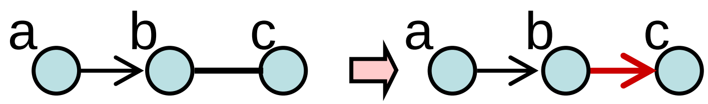
2. For each nonadjacent pair (a,b) with $a - c - b$, direct the edges by $a \to c \leftarrow b$ if $c \notin S_{ab}$

3. Orient as many of undirected edges as possible on condition that neither new v-structures nor directed cycles are created. (See the next slide for the precise implementation)

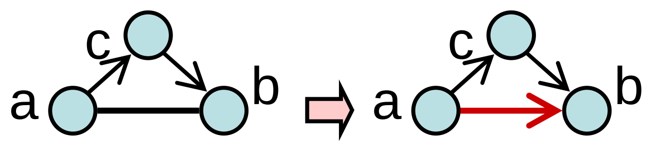
Step 3 of IC algorithm

- The following 4 rules are necessary and sufficient to direct all the possible inferred causal direction (Verma & Pearl 92, Meek 95)

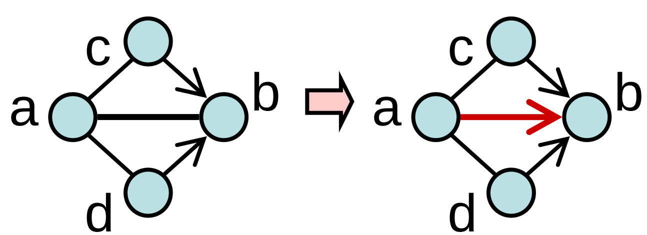
1. If there is a triplet $a \rightarrow b - c$ with a and c nonadjacent, orient $b - c$ into $b \rightarrow c$.



2. If for $a - b$ there is a chain $a \rightarrow c \rightarrow b$, orient $a - b$ into $a \rightarrow b$.

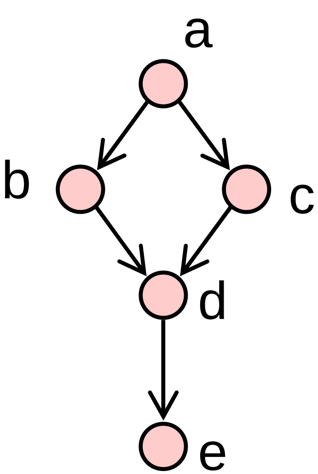


3. If for $a - b$ there are two chains $a - c \rightarrow b$ and $a - d \rightarrow b$ such that c and d are nonadjacent, orient $a - b$ into $a \rightarrow b$.

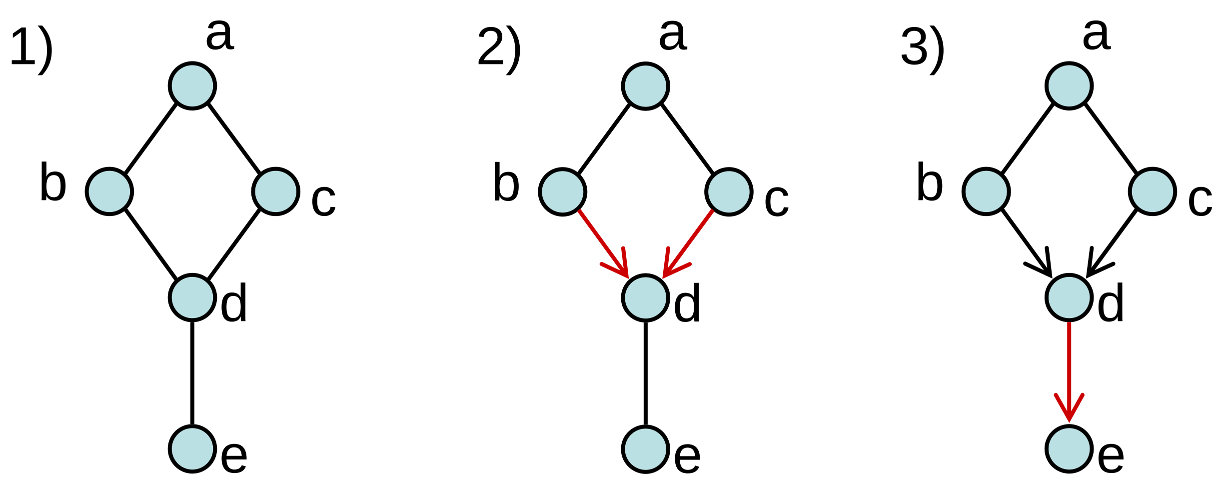


Example

True structure



The output from each step of IC algorithm



$$S_{ad} = \{b, c\}$$
$$S_{ae} = \{d\}$$
$$S_{bc} = \{a\}$$
$$S_{be} = S_{ce} = \{d\}$$

For other pairs,
S does not exist.

For (b,c), $d \notin S_{bc}$

Direction of some edges may be left undetermined.

PC Algorithm

(Peter Sprites & Clark Glymour 91)

- Linear method: partial correlation with χ^2 test is used in Step 1.
- Efficient computation for Step 1.

  Start with complete graph, check $X_a \perp\!\!\!\perp X_b \mid S$ only for $S \subset N_a$, and connect the edge a—b if there is no such S.

  $i = 0.$ $G =$ Complete graph.

  repeat

  for each a in V

  for each b in N_a

  Check $X_a \perp\!\!\!\perp X_b \mid S$ for $S \subset N_a \setminus \{b\}$ with $|S| = i$

  If such S exists,

  set $S_{ab} = S$, and delete the edge a—b from G.

  $i = i + 1$

  until $|N_a| < i$ for all a

- Implemented in TETRAD (http://www.phil.cmu.edu/projects/tetrad/)

Kernel-based Causal Leaning

- Limitations of the previous implementations of IC
  - Linear / discrete assumptions in Step 1.

    Difficulty in testing conditional independence for continuous variables.

    → kernel method!

  - Errors of the skeleton in Step 1 cannot be recovered in the later steps.

    → voting method

■ KCL algorithm (Sun et al. ICML07, Sun et al. 2007)

– Dependence measure: $\hat{\mathbb{H}}_{YX}^{(N)} = HSIC = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2$

– Conditional dependence measure: $\hat{\mathbb{H}}_{YX|Z}^{(N)} \equiv \dfrac{\left\| \hat{\Sigma}_{\ddot{Y}\ddot{X}|Z}^{(N)} \right\|_{HS}^2}{\left\| C_{ZZ} \right\|_{HS}^2}$

where the operator $C_{ZZ} : H_Z \to H_Z$ is defined by

$$\langle f, C_{ZZ} g \rangle = E[f(Z)g(Z)]$$

Motivation: make $\left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2$ and $\left\| \hat{\Sigma}_{\ddot{Y}\ddot{X}|Z}^{(N)} \right\|_{HS}^2$ comparable

<u>Theorem</u>

If $(X, Y) \perp\!\!\!\perp Z$, $\quad \left\| \hat{\Sigma}_{\ddot{Y}\ddot{X}|Z}^{(N)} \right\|_{HS}^2 = \left\| C_{ZZ} \right\|_{HS}^2 \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2$

Outline of the KCL algorithm: IC algorithm is modified as follows:

KCL-1: Skeleton by statistical tests

(1) Permutation tests of conditional independence $X \perp\!\!\!\perp Y \mid S_{XY}$ for all (X, Y, S_{XY}) $(S_{XY} \subset V \setminus \{X, Y\})$ with the measure $\hat{\mathbb{H}}^{(N)}_{YX|Z}$

(2) Connect X and Y if no such S_{XY} exists.

KCL-2: Majority votes for directing edges

For all triplets $X - Z - Y$ (X and Y may be adjacent), give a vote to the direction $X \rightarrow Z$ and $Y \rightarrow Z$ if

$$M_{XY|Z} \equiv \frac{\hat{\mathbb{H}}^{(N)}_{YX|Z}}{\hat{\mathbb{H}}^{(N)}_{YX}} > \lambda$$

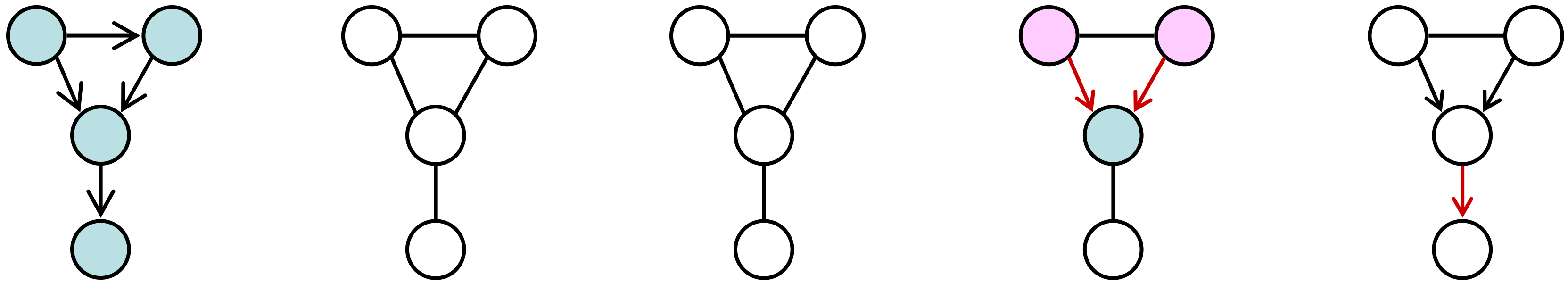
Repeat this for (a) $\lambda >> 1$ (rigorous v-structure)

and (b) $\lambda = \max\{M_{YZ|X}, M_{XZ|Y}\}$ (relative v-structure)

Make an arrow to each edge if a vote is given ("$\leftrightarrow$" is allowed).

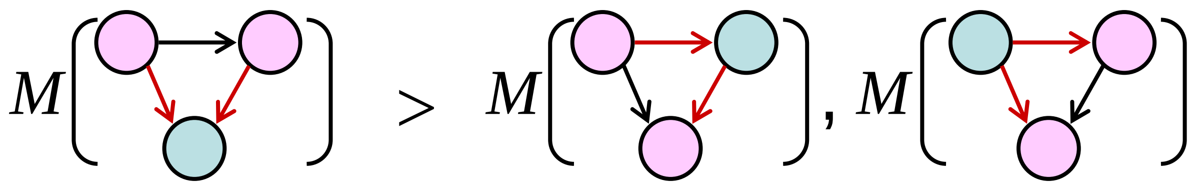
KCL-3: Same as IC-3

Illustration of KCL



true KCL-1 KCL-2 (a) KCL-2 (b) KCL-3

Heuristic assumption:



Conditioning common effect strengthens the dependence between the causes.

Hidden common cause

- FCI (Fast Causal Inference, Spirtes et al. 93) extends PC to allow hidden variables.

- A bi-directional arrow ($\leftrightarrow$) given by KCL may be interpreted as a hidden common cause. Empirically confirmed, but no theoretical justification (Sun et al. 2007).

Experiments with KCL

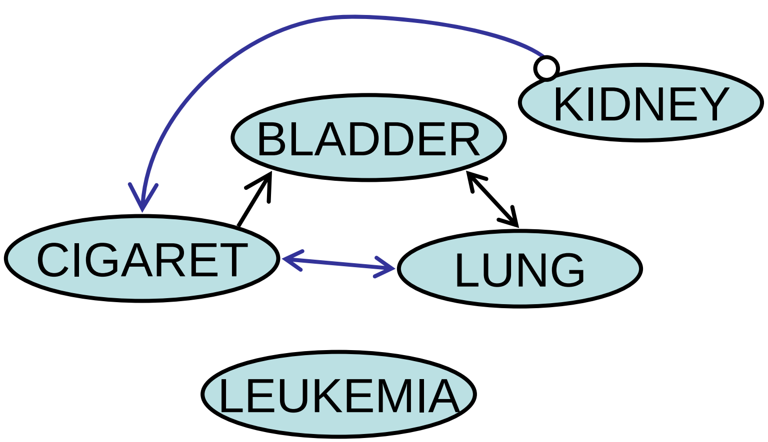
- Smoking and Cancer
  - Data (N = 44)

    CIGARET: Cigarettes sales in 43 states in US and District of Columbia

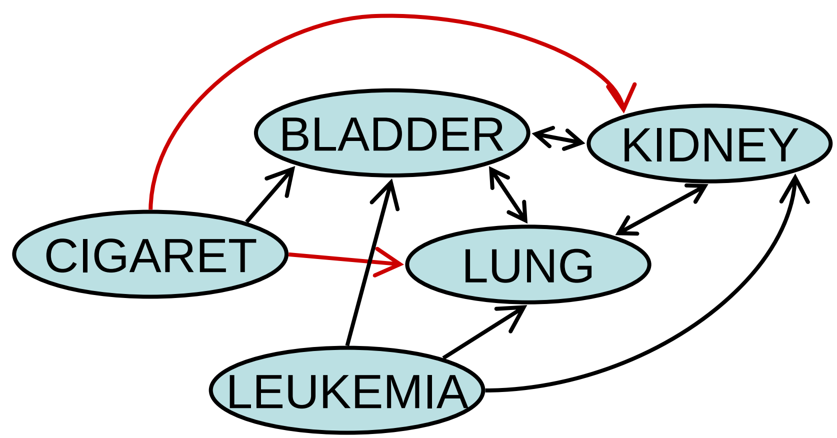
    BLADDER, LUNG, KIDNEY, LEUKEMIA: death rates from various cancers

  - Results

FCI



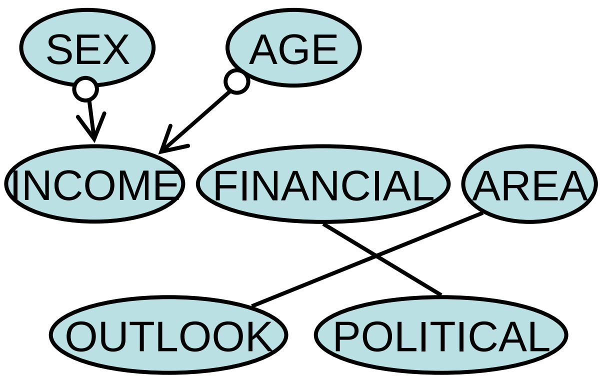
KCL



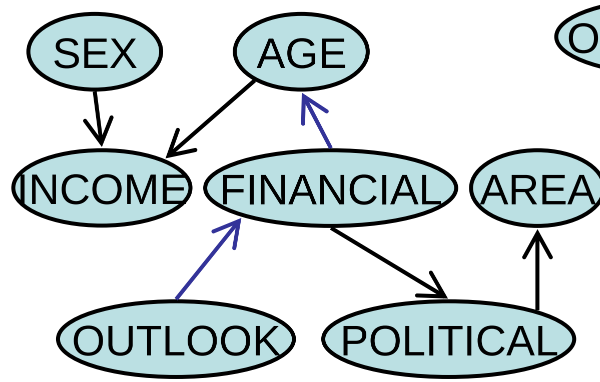
■ Montana Economic Outlook Poll (1992)

- Data: 7 discrete variables, N = 209

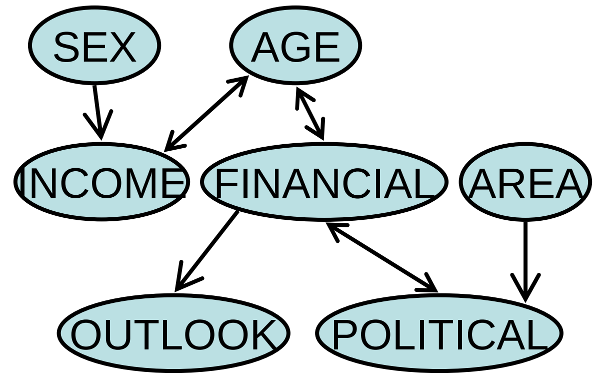
  AGE (3), SEX (2), INCOME (3), POLITICAL (3), AREA (3), FINANCIAL status (3, better/same/worse than a year ago), OUTLOOK (2)



FCI



BN-PC



KCL

BN-PC is a constraint-based method using MI (Chen et al. 2002)

Summary of Part VI

- Causality of time series
  - Kernel-based measures → Nonlinear extension of Granger causality

- Causal inference from non-experimental data
  - Kernel-based Causal Learning (KCL) algorithm
    - Constraint-based method: A variant of Inductive Causation
      - Conditional independence test with kernel measures
      - Voting method for directions
    - More reasonable results are obtained than existing methods. See Sun et al. (2007) for more detailed comparisons.

Bibliography

- Papers

Cheng, J., R. Greiner, J. Kelly, D. A. Bell, and W. Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence Journal*, 137:43–90. (2002).

Friedman, J. and Rafsky, L. Multivariate generalization of the Wald-Wolfovitz and Smirnov two sample tests. *Annals of Stat*. 7:697-717 (1979).

Fukumizu, K., F. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Leaning Research*, 8:361-383 (2007)

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

Fukumizu, K., F. Bach, and M. Jordan. Kernel dimension reduction in regression. Tech Report 715, Dept. Statistics, University of California, Berkeley, 2006.

Fukumizu, K., A. Gretton, X. Sun., and B. Schölkopf. Kernel Measures of Conditional Dependence. *Submitted* (2007)

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424-438 (1969).

Gretton, A., A. J. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf and N. K. Logothetis. Kernel Constrained Covariance for Dependence Measurement. *Proc. 10th Intern. Workshop on Artificial Intelligence and Statistics* (*AISTATS 2005*), pp.112-119 (2005)

Gretton, A., O. Bousquet, A. Smola and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Algorithmic Learning Theory: 16th International Conference, ALT 2005*, pp.63-78 (2005)

Gretton, A., K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems 19*. MIT Press (2007).

Gretton, A., K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A. Smola. A Kernel Statistical Test of Independence. *Submitted* (2007).

Ku, C. and Fine, T. Testing for Stochastic Independence: Application to Blind Source Separation. IEEE Trans. Signal Processing, 53(5):1815-1826 (2005).

Kraskov, A., H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69, 066138-1–16 (2004).

Meek, C. Causal inference and causal explanation with background knowledge. In P.Besnard and S.Hanks (Eds.), *Uncertainty in Artificial Intelligence*, vol. II, pp.403-410. Morgan-Kaufmann.

Nilsson, J. Sha, F. Jordan, M. Regression on manifolds using kernel dimension reduction. *Proc. 24th Intern. Conf. Machine Learning* (*ICML2007*), pp.697-704. (2007)

Pearl, J. Causal diagrams for empirical research. *Biometrika* 82, 669-710 (1995).

Shen, H., S. Jegelka and A. Gretton: Fast Kernel ICA using an Approximate Newton Method. *Proc. 11th Intern. Workshop on Artificial Intelligence and Statistics* (AISTAT2007).

Spirtes, P. and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9:62-72.

Spirtes, P., C. Meek and T. Richardson. Causal inference in the presence of latent variables and selection bias. *Proc. 11th Conf. Uncertainty in Artificial Intelligence*. pp 499-506 (1995).

Steinwart, I. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Leaning Research*, 2, pp.67-93 (2002)

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

Sun, X., D. Janzing, B. Schölkopf, K. Fukumizu, and A. Gretton. Learning Causal Structures via Kernel-based Statistical Dependence Measures. *Submitted* (2007)

Verma, T., J. Pearl. Equivalence and synthesis of causal models. *Proc. 6th Conf. Uncertainty in Artificial Intelligence* (*UAI1990*) pp.220-227 (1990)

Verma, T., J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. *Proc. 8th Conf. Uncertainty in Artificial Intelligence* (*UAI1992*) pp.323-330 (1992)

Holland, P.W. Statistics and causal inference. *J. American Statistical Association* 81: 945-960 (1986).

Graham E.Q., H.S. Chai, and M.G. Maguire, and R.A. Stone. Myopia and ambient lighting at night. *Nature* 399: 113 (May 13, 1999)

Zadnik, K., L.A. Jones, B.C. Irvin, R.N. Kleinstein, R.E. Manny, J.A. Shin, and D.O. Mutti. Myopia and ambient night-time lighting. *Nature* 404: 143-144 (9 March 2000)

Books

Hyvärinen, A. J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience (2001).

Pearl, J. *Causality*. Cambridge University Press (2000)

Edwards, D. *Introduction to graphical modelling*. Springer verlag, New York (2000).

Read, T. and Cressie, N. *Goodness-of-fit Statistics or Discrete Multivariate Analysis*. Wiley, New York (1995)

Spirtes, P., C. Glymour, and R. Scheines. *Causation, prediction, and search.* Springer-Verlag, New York (1993). (2nd ed. 2000)

Many thanks to my collaborators,

Bernhard Schölkopf, Arthur Gretton, Xiaohai Sun, Dominik Janzing,

and to many active students.