Independence and Conditional Independence with Kernels

Statistical Data Analysis with Positive Definite Kernels

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS

Department of Statistical Science, Graduate University for Advanced Studies

October 6-10, 2008, Kyushu University

Outline

1. Covariance operators on RKHS
2. Independence with RKHS
3. Conditional independence with RKHS
4. Summary

Outline

1. Covariance operators on RKHS
2. Independence with RKHS
3. Conditional independence with RKHS
4. Summary

Covariance on RKHS

(X, Y): random variable taking values on $\mathcal{X} \times \mathcal{Y}$. resp.

$(H_{\mathcal{X}}, k_{\mathcal{X}})$, $(H_{\mathcal{Y}}, k_{\mathcal{Y}})$: RKHS with measurable kernels on $\mathcal{X}$ and $\mathcal{Y}$, resp.

Assume $E[k_{\mathcal{X}}(X,X)]E[k_{\mathcal{Y}}(Y,Y)] < \infty$

Cross-covariance operator: $\Sigma_{YX} : H_{\mathrm{X}} \to H_{\mathrm{Y}}$

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

$$= m_{P_{YX}} - m_{P_Y \otimes P_X} \qquad \in H_{\mathcal{Y}} \otimes H_{\mathcal{X}}$$

<u>Proposition</u>

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

for all $f \in H_{\mathcal{X}}, g \in H_{\mathcal{Y}}$

– *c.f.* Euclidean case

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

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

RKHS for product kernel

- RKHS w.r.t. product kernel

k_1, k_2 : positive definite kernel on Ω_1, Ω_2 , resp.

H_1, H_2: corresponding RKHS $\{\phi_i\}_{i=1}^{\infty}, \{\psi_j\}_{j=1}^{\infty}$: CONS of H_1, H_2, resp.

$k_1\, k_2$: product kernel (positive definite)

RKHS corresponding to the product kernel $k_1\, k_2$ is given by $H_1 \otimes H_2$

$H_1 \otimes H_2$ consists of functions

$$f(x, y) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \alpha_{ij}\phi_i(x)\psi_j(y)$$

with $$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |\alpha_{ij}|^2 < \infty.$$

In particular, $$\left\{ \sum_{i=1}^{n} f_i(x) g_i(y) \,\middle|\, f_i \in H_1, g_i \in H_2 \right\} \subset H_1 \otimes H_2.$$

Characterization of independence

- Independence and Cross-covariance operator

<u>Theorem</u>

If the product kernel $k_{\mathcal{X}}k_{\mathcal{Y}}$ is characteristic on $\mathcal{X} \times \mathcal{Y}$, then

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

proof)

$$\Sigma_{XY} = O \quad \Leftrightarrow \quad m_{P_{XY}} = m_{P_X \otimes P_Y}$$

$$\Leftrightarrow \quad P_{XY} = P_X \otimes P_Y$$ (by characteristic assumption)

– *c.f.* for Gaussian variables

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad V_{XY} = O$$ *i.e.* uncorrelated

– *c.f.* Characteristic function

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad E_{XY}[e^{\sqrt{-1}(uX+vY)}] = E_X[e^{\sqrt{-1}uX}]E_Y[e^{\sqrt{-1}vY}]$$

Estimation of cross-cov. operator

$(X_1, Y_1), \ldots, (X_N, Y_N)$: i.i.d. sample on $\mathcal{X} \times \mathcal{Y}$

An estimator of Σ_{YX} is defined by

$$\hat{\Sigma}_{YX}^{(N)} = \frac{1}{N}\sum_{i=1}^{N}\left\{k_{\mathcal{Y}}(\cdot, Y_i) - \hat{m}_Y\right\} \otimes \left\{k_{\mathcal{X}}(\cdot, X_i) - \hat{m}_X\right\}$$

Theorem

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

Corollary to the $\sqrt{N}$ -consistency of the empirical mean, because the norm in $H_{\mathcal{X}} \otimes H_{\mathcal{Y}}$ is equal to the Hilbert-Schmidt norm of the corresponding operator $H_{\mathcal{X}} \to H_{\mathcal{Y}}$.

Hilbert-Schmidt Operator

– Hilbert-Schmidt operator

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

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

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

Hilbert-Schmidt norm: $\|A\|_{HS}^2 = \sum_j \sum_i \langle \psi_j, A\varphi_i \rangle^2$

c.f. Frobenius norm of a matrix

– Fact: $\|A\| \leq \|A\|_{HS}$

Outline

1. Introduction
2. Covariance operators on RKHS
3. Independence with RKHS
4. Conditional independence with RKHS
5. Summary

Measuring Dependence

- Dependence measure

$$M_{YX} = \left\| \Sigma_{YX} \right\|_{HS}^{2}$$

$$M_{YX} = 0 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$ with $k_{\mathcal{X}} k_{\mathcal{Y}}$ characteristic

- Empirical dependence measure

$$\hat{M}_{YX}^{(N)} = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^{2}$$

M_{YX} and $\hat{M}_{YX}^{(N)}$ can be used as measures of dependence.

HS norm of cross-cov. operator I

- Integral expression

$$M_{YX} = \|\Sigma_{YX}\|_{HS}^2 = E[k_{\mathcal{X}}(X,\tilde{X})k_{\mathcal{Y}}(Y,\tilde{Y})] - 2E\big[E[k_{\mathcal{X}}(X,\tilde{X})\,|\,\tilde{X}]E[k_{\mathcal{Y}}(Y,\tilde{Y})\,|\,\tilde{Y}]\big]$$
$$+ E[k_{\mathcal{X}}(X,\tilde{X})]E[k_{\mathcal{Y}}(Y,\tilde{Y})]$$

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

Proof.

$$\begin{aligned}
\|\Sigma_{YX}\|_{HS}^2 &= \|E[k_{\mathcal{X}}(X,\cdot)\otimes k_{\mathcal{Y}}(Y,\cdot)] - m_X\otimes m_Y\|^2 \\
&= \langle E[k_{\mathcal{X}}(X,\cdot)\otimes k_{\mathcal{Y}}(Y,\cdot)], E[k_{\mathcal{X}}(\tilde{X},\cdot)\otimes k_{\mathcal{Y}}(\tilde{Y},\cdot)]\rangle \\
&\quad -2\langle E[k_{\mathcal{X}}(X,\cdot)\otimes k_{\mathcal{Y}}(Y,\cdot)], m_{\tilde{X}}\otimes m_{\tilde{Y}}\rangle + \langle m_X\otimes m_Y, m_{\tilde{X}}\otimes m_{\tilde{Y}}\rangle \\
&= E[k_{\mathcal{X}}(X,\tilde{X})k_{\mathcal{Y}}(Y,\tilde{Y})] - 2E[E[k_{\mathcal{X}}(X,\tilde{X})|\tilde{X}]E[k_{\mathcal{Y}}(Y,\tilde{Y})|\tilde{Y}]] \\
&\quad +E[k_{\mathcal{X}}(X,\tilde{X})]E[k_{\mathcal{Y}}(Y,\tilde{Y})].
\end{aligned}$$

HS norm of cross-cov. operator II

- Empirical estimator

  Gram matrix expression

  HS-norm can be evaluated only in the subspaces

  $\mathrm{Span}\left\{k_{\mathcal{X}}(\cdot, X_i) - \hat{m}_X^{(N)}\right\}_{i=1}^{N}$ and $\mathrm{Span}\left\{k_{\mathcal{Y}}(\cdot, Y_i) - \hat{m}_Y^{(N)}\right\}$.

$$\Rightarrow \quad \hat{M}_{YX}^{(N)} = \frac{1}{N^2}\mathrm{Tr}\left[G_X G_Y\right]$$

where $G_X = Q_N K_X Q_N, \qquad Q_N = I_N - \frac{1}{N}\mathbf{1}_N \mathbf{1}_N{}^T$

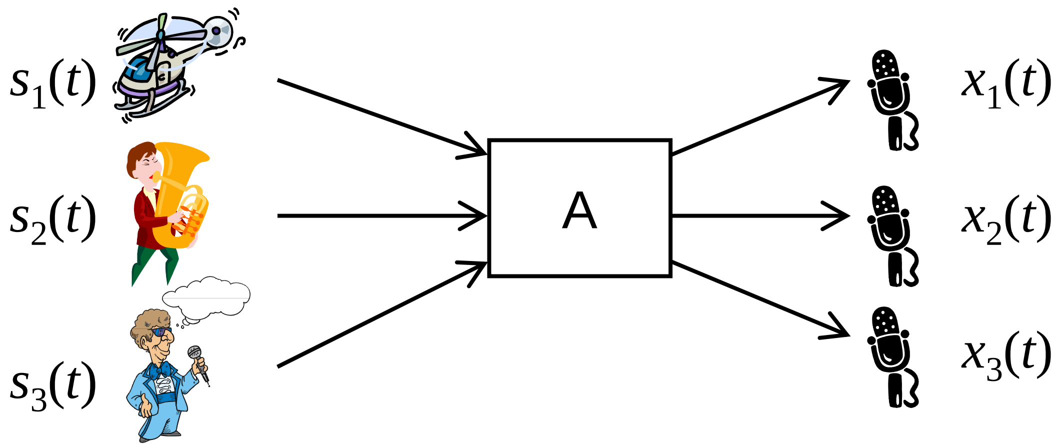
Or equivalently,

$$\hat{M}_{YX}^{(N)} = \left\|\hat{\Sigma}_{YX}^{(N)}\right\|_{HS}^2 = \frac{1}{N^2}\sum_{i,j=1}^{N} k_{\mathcal{X}}(X_i, X_j) k_{\mathcal{Y}}(Y_i, Y_j) - \frac{2}{N^3}\sum_{i,j,k=1}^{N} k_{\mathcal{X}}(X_i, X_j) k_{\mathcal{Y}}(Y_i, Y_k) + \frac{1}{N^4}\sum_{i,j=1}^{N} k_{\mathcal{X}}(X_i, X_j)\sum_{k,\ell=1}^{N} k_{\mathcal{Y}}(Y_k, Y_\ell)$$

Application: ICA

Independent Component Analysis (ICA)

- Assumption
  - m independent source signals
  - m observations of linearly mixed signals



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

A: m x m invertible matrix

- Problem
  - Restore the independent signals S from observations X.

$$\hat{S} = BX$$

B: m x m orthogonal matrix

ICA with HS independence measure

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

Pairwise-independence criterion is applicable.

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

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

→ Approximate Newton method has been proposed

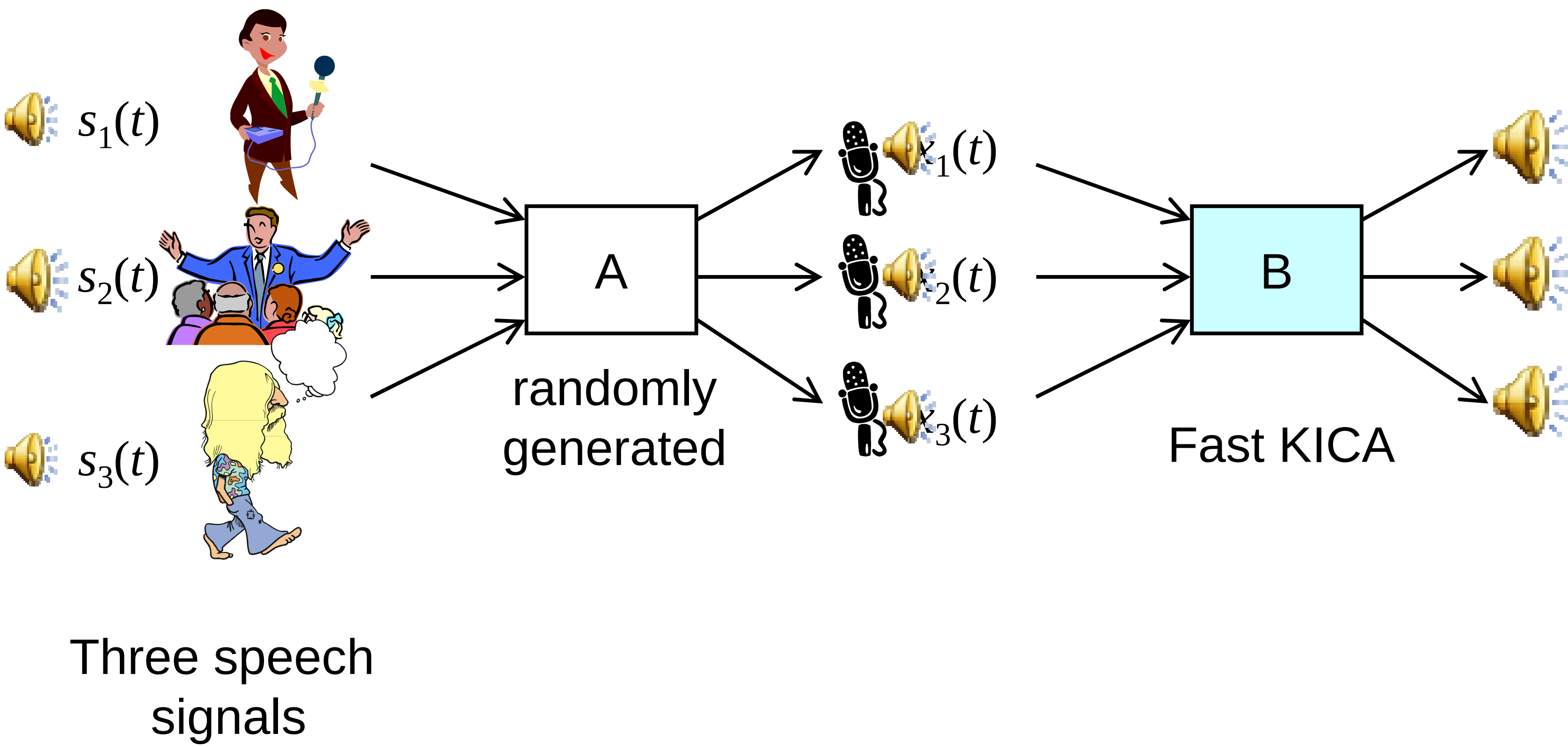
Fast Kernel ICA (FastKICA, Shen et al 07)

(Software downloadable at Arthur Gretton's homepage)

Other methods for ICA

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

■ Experiments (speech signal)



Normalized Covariance Operator

- Normalized Cross-Covariance Operator

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

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

- Characterization of independence

With characteristic kernels,

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

Assume W_{XY} etc. are Hilbert-Schmidt.

– Dependence measure

$$\mathrm{NOCCO} = \left\| W_{YX} \right\|_{HS}^2$$

Kernel-free Integral Expression

Theorem (Fukumizu et al. NIPS 21, 2008)

Assume

P_{XY} have density $p_{XY}(x,y)$

$H_X \otimes H_Y$ are characteristic.

W_{YX} is Hilbert-Schmidt.

Then,

$$\| W_{YX} \|_{HS}^2 = \iint \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1 \right)^2 p_X(x)p_Y(y)dxdy$$

- Kernel-free expression, though the definitions are given by kernels!

– Kernel-free value is desired as a “measure” of dependence.
c.f. If unnormalized operators are used, the measures depend on the choice of kernel.

– Mean square contingency

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

is equal to the mean square contingency, which is a very popular measure of dependence for discrete variables.

Empirical Estimator

– Empirical estimation is straightforward with the empirical cross-covariance operator $\hat{\Sigma}_{YX}^{(N)}$.

– Inversion → regularization: $\Sigma_{XX}^{-1} \to \left(\hat{\Sigma}_{XX}^{(N)} + \varepsilon I\right)^{-1}$

– Replace the covariances in $W_{YX} = \Sigma_{YY}^{-1/2} \Sigma_{YX} \Sigma_{XX}^{-1/2}$ by the empirical ones given by the data $\Phi_X(X_1),\ldots, \Phi_X(X_N)$ and $\Phi_Y(Y_1),\ldots, \Phi_Y(Y_N)$

$$\mathrm{NOCCO}_{emp} = \mathrm{Tr}\left[R_X R_Y\right] \qquad \text{(dependence measure)}$$

where $R_X \equiv G_X \left(G_X + N\varepsilon_N I_N\right)^{-1}$

$$G_X = \left(I_N - \tfrac{1}{N}\mathbf{1}_N \mathbf{1}_N^T\right) K_X \left(I_N - \tfrac{1}{N}\mathbf{1}_N \mathbf{1}_N^T\right) \qquad K_X = \left(k(X_i, X_j)\right)_{i,j=1}^{N}$$

– $\mathrm{NOCCO}_{\mathrm{emp}}$ gives a new kernel estimator for the mean square contingency. Consistency is known.

Independence test with kernels I

- Independence test with positive definite kernels
  - Null hypothesis H0: X and Y are independent
  - Alternative H1: X and Y are not independent

$\hat{M}_{YX}^{(N)}$ and NOCCOemp can be used for test statistics.

$$\hat{M}_{YX}^{(N)} = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^{2} = \frac{1}{N^2} \sum_{i,j=1}^{N} k_{\mathcal{X}}(X_i, X_j) k_{\mathcal{Y}}(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^{N} k_{\mathcal{X}}(X_i, X_j) k_{\mathcal{Y}}(Y_i, Y_k) + \frac{1}{N^4} \sum_{i,j=1}^{N} k_{\mathcal{X}}(X_i, X_j) \sum_{k,\ell=1}^{N} k_{\mathcal{Y}}(Y_k, Y_\ell)$$

Independence test with kernels II

- Asymptotic distribution under null-hypothesis

Theorem (Gretton et al. 2008)

If X and Y are independent, then

$$N\,\hat{M}_{YX}^{(N)} \quad \Rightarrow \quad \sum_{i=1}^{\infty} \lambda_i Z_i^2 \qquad \text{in law} \quad (N \to \infty)$$

where

Z_i : i.i.d. ~ $N(0,1)$,

$\{\lambda_i\}_{i=1}^{\infty}$ is the eigenvalues of the following integral operator

$$\int h(u_a, u_b, u_c, u_d)\varphi_i(u_b)\,dP_{U_b}\,dP_{U_c}\,dP_{U_d} = \lambda_i \varphi_i(u_a)$$

$$h(U_a, U_b, U_c, U_d) = \tfrac{1}{4!}\sum_{(a,b,c,d)} k_{a,b}^{\mathcal{X}} k_{a,b}^{\mathcal{Y}} - 2k_{a,b}^{\mathcal{X}} k_{a,c}^{\mathcal{Y}} + k_{a,b}^{\mathcal{X}} k_{c,d}^{\mathcal{Y}}$$

$$k_{a,b}^{\mathcal{X}} = k_{\mathcal{X}}(X_a, X_b), \quad U_a = (X_a, Y_a)$$

– The proof is easy by the theory of U (or V)-statistics (see e.g. Serfling 1980, Chapter 5).

Independence test with kernels III

- Consistency of test

Theorem (Gretton et al. 2008)

If M_{YX} is not zero, then

$$\sqrt{N}\left(\hat{M}_{YX}^{(N)} - M_{YX}\right) \Rightarrow N(0,\sigma^2) \quad \text{in law} \quad (N \to \infty)$$

where

$$\sigma^2 = 16\left(E_a\left[E_{b,c,d}[h(U_a,U_b,U_c,U_d]^2\right] - M_{YX}\right)$$

Choice of Kernel

- How to choose a kernel?
  - No definitive solutions have been proposed yet.
  - For statistical tests, comparison of power or efficiency will be desirable.
  - Other suggestions:
    - Make a relevant supervised problem, and use cross-validation.
    - Some heuristics
      - Heuristics for Gaussian kernels (Gretton et al 2007)

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

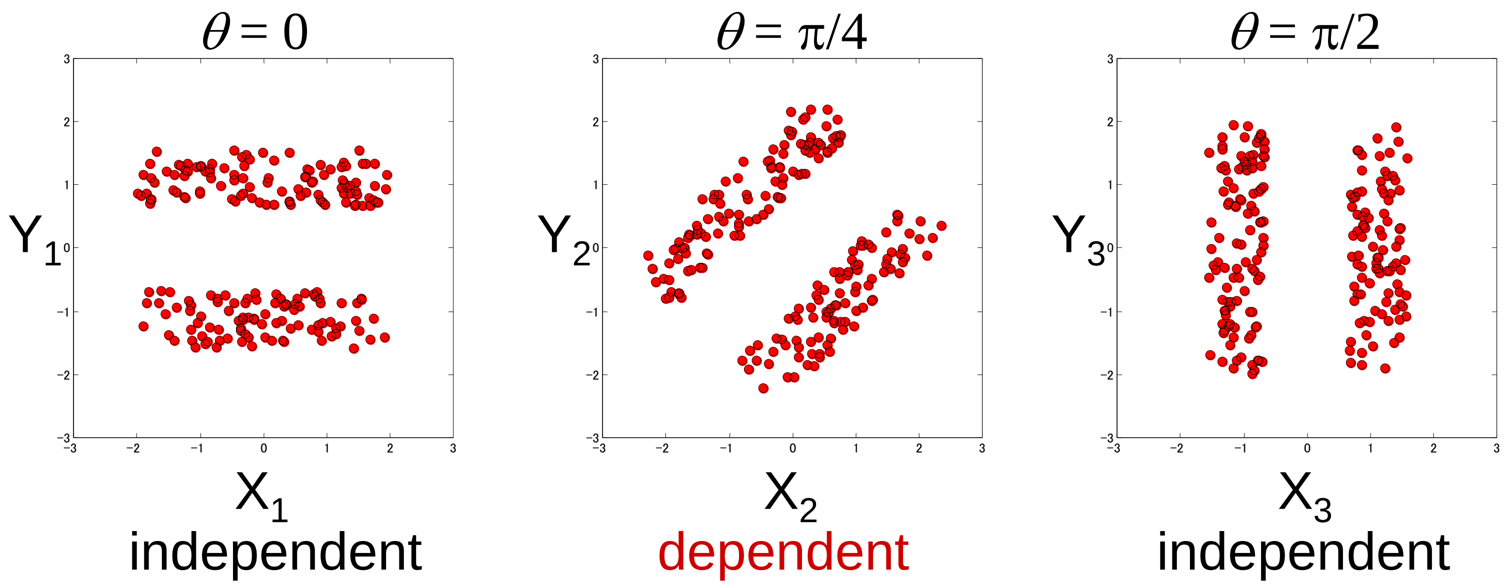
      - Speed of asymptotic convergence (Fukumizu et al. 2008)

$$\lim_{N \to \infty} Var\left[N \times HSIC_{emp}^{(N)} \right] = 2 \left\| \Sigma_{XX} \right\|_{HS}^2 \left\| \Sigma_{YY} \right\|_{HS}^2 \quad \text{under independence}$$

Compare the bootstrapped variance and the theoretical one, and choose the parameter to give the minimum discrepancy.

Application to Independence Test

- Toy example



They are all uncorrelated, but dependent for $0 < \theta < \pi/2$

N = 200.

Permutation test is used for independence test except contingency table.

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
Power Diverg. (#Bins=3)	96	92	43	9	1	0
Power Diverg. (#Bins=4)	98	29	0	0	0	0
Power Diverg. (#Bins=5)	94	60	2	0	0	0

acceptance of independence out of 100 tests (significance level = 5%)

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

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

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

Results of permutations test with HS measure

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

Acceptance rate (α = 5%)

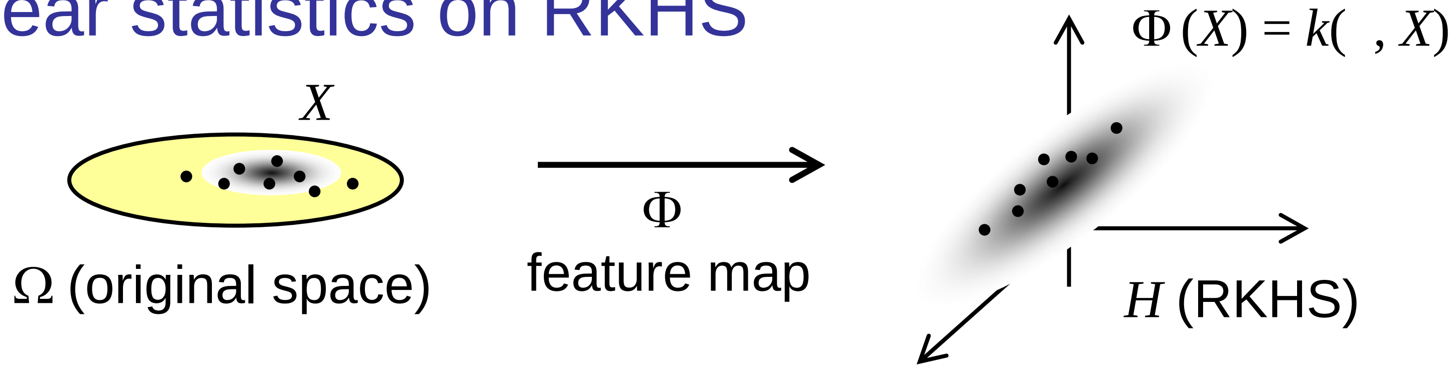
(Gretton et al. 07)

Outline

1. Introduction
2. Covariance operators on RKHS
3. Independence with RKHS
4. Conditional independence with RKHS
5. Summary

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. $p_{XYZ}(x, y, z)$

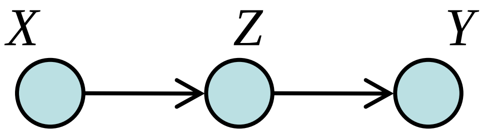
X and Y are conditionally independent given Z, if

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

or

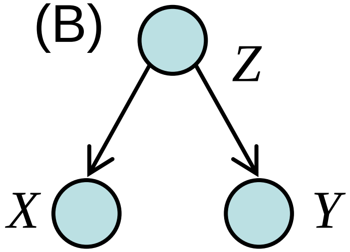
$$p_{XY|Z}(x, y \mid z) = p_{X|Z}(x \mid z)\, p_{Y|Z}(y \mid z) \qquad \text{(B)}$$

(A)



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

(B)



Applications

- Graphical model
- Causal inference, etc.

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.

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

- For Gaussian variables,
  $V_{YY|[X,Z]} = V_{YY|Z} \qquad (\Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z \;)$
  can be interpreted as
  "If Z is known, X is not necessary for linear prediction of Y."

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

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

Schur complement of V_{XX} in V

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

Conditional 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

  $$\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \quad : \quad H_X \to H_Y$$

  – Conditional covariance operator

  $$\Sigma_{YY|Z} \equiv \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY} \quad : \quad H_Y \to H_Y$$

  – Σ_{ZZ}^{-1} may not exist as a bounded operator. But, we can justify the definions.

Decomposition of covariance operator

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

such that W_{YX} is a bounded operator with $\| W_{YX} \| \leq 1$ and

$$\overline{Range(W_{YX})} = \overline{Range(\Sigma_{YY})}, \quad Ker(W_{YX}) \perp \overline{Range(\Sigma_{XX})}.$$

Rigorous definitions

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

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

Conditional Covariance

- Conditional covariance is expressed by operators

<u>Proposition (FBJ 2004, 2008+)</u>

Assume k_Z is characteristic.

$$\langle g, \Sigma_{YX|Z} f \rangle = E\big[\mathrm{Cov}[g(Y), f(X) \mid Z]\big] \qquad (\forall f \in H_X, g \in H_Y)$$

In particular,

$$\langle g, \Sigma_{YY|Z} g \rangle = E\big[\mathrm{Var}[g(Y) \mid Z]\big] \qquad (\forall g \in H_Y)$$

Proof omitted.

Analogy to Gaussian variables:

$$b^T (V_{YX} - V_{YZ} V_{ZZ}^{-1} V_{ZX}) a = \mathrm{Cov}[b^T Y, a^T X \mid Z]$$

$$b^T (V_{YY} - V_{YZ} V_{ZZ}^{-1} V_{ZY}) b = \mathrm{Var}[b^T Y \mid Z]$$

Residual error interpretation

<u>Proposition (FBJ 2004, 2008+)</u>

Assume k_Z is characteristic.

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

where $\tilde{f}(X) = f(X) - E[f(X)],\ \tilde{g}(Y) = g(Y) - E[g(Y)].$

c.f. for Gaussian variables

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

– Proof (left = right)

$$E\left|\left(g(Y)-E[g(Y)]\right)-\left(f(Z)-E[f(Z)]\right)\right|^2$$

$$=\langle f, \Sigma_{ZZ} f\rangle - 2\langle f, \Sigma_{ZY} g\rangle + \langle g, \Sigma_{YY} g\rangle$$

$$=\left\|\Sigma_{ZZ}^{1/2} f\right\|^2 - 2\left\langle f, \Sigma_{ZZ}^{1/2} W_{ZY} \Sigma_{YY}^{1/2} g\right\rangle + \left\|\Sigma_{YY}^{1/2} g\right\|^2$$

$$=\left\|\Sigma_{ZZ}^{1/2} f - W_{ZY}\Sigma_{YY}^{1/2} g\right\|^2 + \left\|\Sigma_{YY}^{1/2} g\right\|^2 - \left\|W_{ZY}\Sigma_{YY}^{1/2} g\right\|^2$$

$$=\left\|\Sigma_{ZZ}^{1/2} f - W_{ZY}\Sigma_{YY}^{1/2} g\right\|^2 + \left\langle g, \left(\Sigma_{YY} - \Sigma_{YY}^{1/2} W_{YZ} W_{ZY} \Sigma_{YY}^{1/2}\right) g\right\rangle$$

This part can be arbitrary small by choosing f because of $\overline{Range(W_{ZY})} = \overline{Range(\Sigma_{ZZ})}$.

$\Sigma_{YY|Z}$

Conditional independence with kernels

<u>Theorem (FBJ2004, 2008+)</u>

Assume k_Z and $k_X k_Y k_Z$ are characteristic.

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

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

Assume k_Z , k_Y, $k_X k_Z$ are characteristic.

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

– *c.f.* Gaussian variables

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

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

– Why is the “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 $\quad p(x, z' \mid z) = p(x \mid z) \delta(z' - z)$

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

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

Empirical Estimator of Cond. Cov. Operator

$(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)$

$\Sigma_{YZ} \rightarrow \hat{\Sigma}_{YZ}^{(N)}$ etc. finite rank operators

$\Sigma_{ZZ}^{-1} \rightarrow \left(\hat{\Sigma}_{ZZ}^{(N)} + \varepsilon_N I\right)^{-1}$ regularization for inversion

– Empirical conditional covariance operator

$$\hat{\Sigma}_{YX|Z}^{(N)} := \hat{\Sigma}_{YX}^{(N)} - \hat{\Sigma}_{YZ}^{(N)}\left(\hat{\Sigma}_{ZZ}^{(N)} + \varepsilon_N I\right)^{-1}\hat{\Sigma}_{ZX}^{(N)}$$

– Estimator of Hilbert-Schmidt norm

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

$$G_X = Q_N K_X Q_N, \quad Q_N = I_N - \frac{1}{N}\mathbf{1}_N \mathbf{1}_N^T$$ centered Gram matrix

$$S_Z = I_N - (G_Z + N\varepsilon_N I_N)^{-1} G_Z = \left(I_N + \tfrac{1}{N\varepsilon_N} G_Z\right)^{-1}$$

Statistical Consistency

- Consistency on conditional covariance operator

<u>Theorem (FBJ08, Sun et al. 07)</u>

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

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

In particular,

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

Conditional Independence Test

- Permutation test

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

– 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

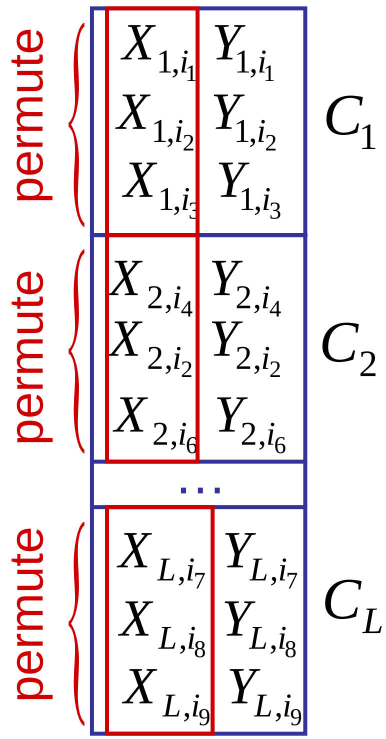
$$A_\ell = \{i \mid Z_i \in C_\ell\} \quad (\ell = 1, \ldots, L).$$

– 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 $T_N^{(b)}$.



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

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

H$_0$: $$Y(t) = c + \sum_{i=1}^{p} a_i Y(t-i) + W_t \quad \longrightarrow \quad \hat{\hat{c}}, \hat{\hat{a}}_i$$

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

- Test statistics

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

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

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

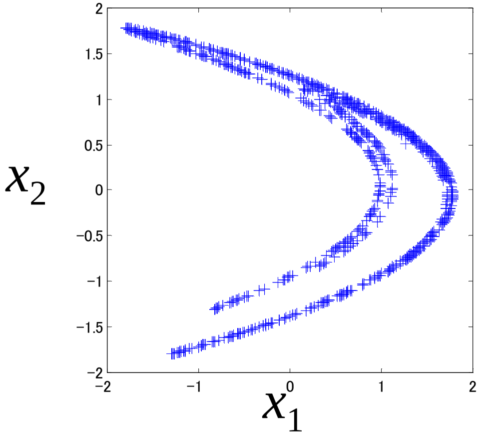
$$HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$$

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

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

Example

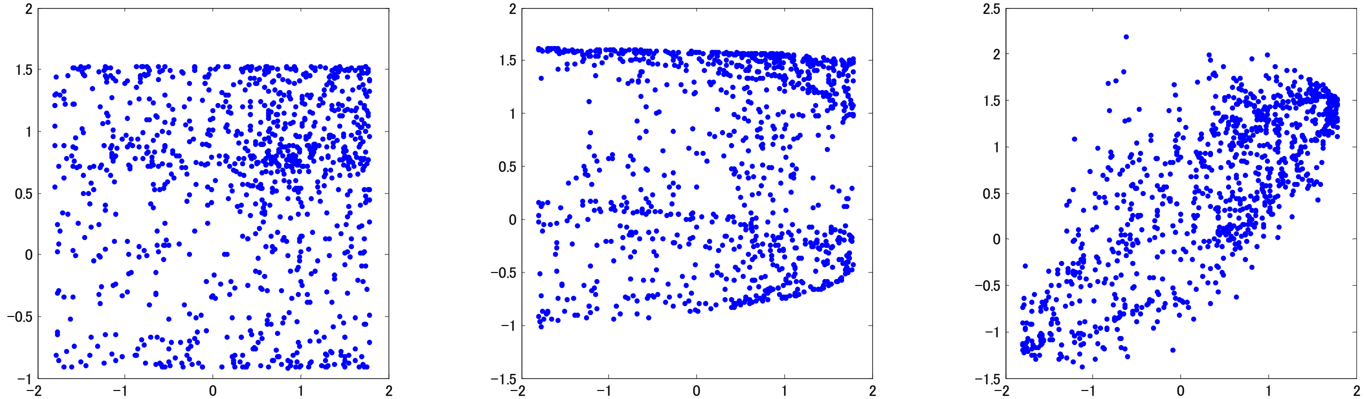
- Coupled Hénon map
  - X, Y:



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

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

x_1-y_1



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

Causality in coupled Hénon map

- X is a cause of Y if $\gamma > 0$. $Y_{t+1} \not\perp\!\!\!\perp X_t \mid Y_t$
- Y is not a cause of X for all γ. $X_{t+1} \perp\!\!\!\perp Y_t \mid X_t$
- Permutation tests for non-causality with NOC³O

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
NOC³O	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

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

Summary

- Dependence analysis with RKHS
  - Covariance and conditional covariance on RKHS can capture the (in)dependence and conditional (in)dependence of random variables.
  - Easy estimators can be obtained for the Hilbert-Schmidt norm of the operators.
  - If the normalized covariance is used, the Hilbert-Schmidt norm is independent of kernel, assuming it is characteristic.
  - Statistical tests of independence and conditional independence are possible with kernel measures.
    - Applications: dimension reduction for regression (FBJ04, FBJ08), causal inference (Sun et al. 2007).

References

Fukumizu, K. Francis R. Bach and M. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*. To appear, 2008.

Fukumizu, K., A. Gretton, X. Sun, and B. Schoelkopf: Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems 21*, 489-496, MIT Press (2008).

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

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

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

Gretton, A., O. Bousquet, A. Smola and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. Proc. Algorithmic Learning Thoery (ALT2005), 63-78. 2005.

Shen, H., S. Jegelka and A. Gretton: Fast Kernel ICA using an Approximate Newton Method. AISTATS 2007.

Serfling, R. J. *Approximation Theorems of Mathematical Statistics*. Wiley-Interscience 1980.

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