Independence and Conditional Independence with RKHS
Statistical Inference with Reproducing Kernel Hilbert Space

Kenji Fukumizu
Institute of Statistical Mathematics, ROIS
Department of Statistical Science, Graduate University for Advanced Studies

July 25, 2008 / Statistical Learning Theory II
Outline

1. Introduction

2. Covariance operators on RKHS

3. Independence with RKHS

4. Conditional independence with RKHS

5. Summary
Outline

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

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

\((H_x, k_x), (H_y, k_y)\): RKHS with measurable kernels on \(\mathcal{X}\) and \(\mathcal{Y}\), resp.

Assume \(E[k_x(X, X)]E[k_y(Y, Y)] < \infty\)

Cross-covariance operator: \(\Sigma_{YX}: H_x \to H_y\)

\[\Sigma_{YX} = E[\Phi_Y(Y) \otimes \Phi_X(X)] - m_Y \otimes m_X\]
\[= m_{P_{YX}} - m_{P_Y \otimes P_X} \quad \in H_y \otimes H_x\]

**Proposition**

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

for all \(f \in H_x, g \in H_y\)

– c.f. Euclidean case

\[V_{YX} = E[YX^T] - E[Y]E[X]^T \quad : \text{covariance matrix}\]
\[ (b, V_{YX} a) = \text{Cov}[(b, Y), (a, X)] \]
Characterization of independence

Independence and Cross-covariance operator

**Theorem**
If the product kernel \( k_X k_Y \) is characteristic on \( \mathcal{X} \times \mathcal{Y} \), then

\[ X \text{ and } Y \text{ are independent } \iff \Sigma_{XY} = O \]

**Proof**
\[
\Sigma_{XY} = O \iff m_{P_{XY}} = m_{P_X \otimes P_Y} \\
\iff P_{XY} = P_X \otimes P_Y \quad \text{(by characteristic assumption)}
\]

- *c.f.* for **Gaussian variables**

\[ X \perp Y \iff V_{XY} = O \quad \text{i.e. uncorrelated} \]

- *c.f.* Characteristic function

\[ X \perp Y \iff E_{XY}[e^{-\Im (uX + vY)}] = E_X[e^{-\Im uX}] E_Y[e^{-\Im vY}] \]
Estimation of cross-cov. operator

\((X_1, Y_1), \ldots, (X_N, Y_N)\): i.i.d. sample on \(\mathbf{x} \times \mathbf{y}\).

An estimator of \(\Sigma_{yx}\) is defined by

\[
\hat{\Sigma}_{yx}^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \{k_y(\cdot, Y_i) - \hat{m}_y\} \otimes \{k_x(\cdot, X_i) - \hat{m}_x\}
\]

**Theorem**

\[
\left\| \hat{\Sigma}_{yx}^{(N)} - \Sigma_{yx} \right\|_{HS} = O_p\left(1/\sqrt{N}\right) \quad (N \to \infty)
\]

Corollary to the \(\sqrt{N}\)-consistency of the empirical mean, because the norm in \(H_x \otimes H_y\) is equal to the Hilbert-Schmidt norm of the corresponding operator \(H_x \rightarrow H_y\).
Hilbert-Schmidt Operator

– Hilbert-Schmidt operator
\[ A : H_1 \rightarrow H_2 \] : 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: If \( A : H_1 \rightarrow H_2 \) is regarded as an element \( F_A \in H_1 \otimes H_2, \)

\[ \| A \|_{HS} = \| F_A \| \]

\[ \therefore \] \[ \| A \|_{HS}^2 = \sum_j \sum_i \langle \psi_j, A \varphi_i \rangle_{H_2}^2 = \sum_j \sum_i \langle F_A, \varphi_i \otimes \psi_j \rangle_{H_1 \otimes H_2}^2 = \| F_A \|^2. \]

CONS of \( H_1 \otimes H_2 \)

– 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**
  \[ M_{YX} = \| \Sigma_{YX} \|_{HS}^2 \]
  \[ M_{YX} = 0 \iff X \perp Y \quad \text{with } k_x k_y \text{ characteristic} \]

- **Empirical dependence measure**
  \[ \hat{M}^{(N)}_{YX} = \| \hat{\Sigma}^{(N)}_{YX} \|_{HS}^2 \]

\( M_{YX} \) and \( \hat{M}^{(N)}_{YX} \) can be used as measures of dependence.
HS norm of cross-cov. operator I

**Integral expression**

\[ M_{yx} = \left\| \Sigma_{yx} \right\|_{HS}^2 = E[k_x(X, \tilde{X})k_y(Y, \tilde{Y})] - 2E[E[k_x(X, \tilde{X})|\tilde{X}]E[k_y(Y, \tilde{Y})|\tilde{Y}]] \\
+ E[k_x(X, \tilde{X})]E[k_y(Y, \tilde{Y})] \]

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

Note: a Hilbert-Schmidt norm always has an integral expression

**Proof.**

\[ \left\| \Sigma_{yx} \right\|_{HS}^2 = \left\| E[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - m_X \otimes m_Y \right\|^2 \]

\[ = \langle E[k_x(X, \cdot) \otimes k_y(Y, \cdot)], E[k_x(\tilde{X}, \cdot) \otimes k_y(\tilde{Y}, \cdot)] \rangle \\
-2\langle E[k_x(X, \cdot) \otimes k_y(Y, \cdot)], m_{\tilde{X}} \otimes m_{\tilde{Y}} \rangle + \langle m_X \otimes m_Y, m_{\tilde{X}} \otimes m_{\tilde{Y}} \rangle \]

\[ = E[k_x(X, \tilde{X})k_y(Y, \tilde{Y})] - 2E[E[k_x(X, \tilde{X})|\tilde{X}]E[k_y(Y, \tilde{Y})|\tilde{Y}]] \\
+ E[k_x(X, \tilde{X})]E[k_y(Y, \tilde{Y})]. \]
Empirical estimator

Gram matrix expression

HS-norm can be evaluated only in the subspaces
\[ \text{Span}\{k_{\hat{x}}(\cdot, X_i) - \hat{m}_X^{(N)}\}_{i=1}^N \] and \[ \text{Span}\{k_{\hat{y}}(\cdot, Y_i) - \hat{m}_Y^{(N)}\} \].

\[ \sum_{i,j=1}^N k_{\hat{x}}(X_i, X_j) k_{\hat{y}}(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^N k_{\hat{x}}(X_i, X_j) k_{\hat{y}}(Y_i, Y_k) \]

\[ + \frac{1}{N^4} \sum_{i,j=1}^N k_{\hat{x}}(X_i, X_j) \sum_{k,l=1}^N k_{\hat{y}}(Y_k, Y_l) \]

\[ = \frac{1}{N^2} \operatorname{Tr}[G_X G_Y] \]

where \[ G_X = Q_N K_X Q_N, \quad Q_N = I_N - \frac{1}{N} 1_N 1_N^T \]

Or equivalently,

\[ \hat{M}_{YX}^{(N)} = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 = \frac{1}{N^2} \sum_{i,j=1}^N k_{\hat{x}}(X_i, X_j) k_{\hat{y}}(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^N k_{\hat{x}}(X_i, X_j) k_{\hat{y}}(Y_i, Y_k) \]

\[ + \frac{1}{N^4} \sum_{i,j=1}^N k_{\hat{x}}(X_i, X_j) \sum_{k,l=1}^N k_{\hat{y}}(Y_k, Y_l) \]
Application: ICA

Independent Component Analysis (ICA)

- Assumption
  - \( m \) independent source signals
  - \( m \) observations of linearly mixed signals

\[
\begin{align*}
  s_1(t) & \quad x_1(t) \\
  s_2(t) & \quad x_2(t) \\
  s_3(t) & \quad x_3(t)
\end{align*}
\]

\[ X(t) = AS(t) \]

- Problem
  - Restore the independent signals \( S \) from observations \( X \).

\[
\hat{S} = BX
\]

\( B: m \times m \) orthogonal matrix
ICA with HSIC

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

Pairwise-independence criterion is applicable.

Minimize

\[ L(B) = \sum_{a=1}^{m} \sum_{b>a} HSIC(Y_a, Y_b) \quad Y = BX \]

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

$A$

$B$

Fast KICA

$s_1(t)$

$s_2(t)$

$s_3(t)$

randomly generated
Independence test with kernels 1

- Null hypothesis  $H_0$: $X$ and $Y$ are independent
- Alternative  $H_1$: $X$ and $Y$ are not independent

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

$$\hat{M}^{(N)}_{YX} = \left\| \hat{\Sigma}^{(N)}_{YX} \right\|_{HS}^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} k_x(X_i, X_j)k_y(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^{N} k_x(X_i, X_j)k_y(Y_i, Y_k)$$
$$+ \frac{1}{N^4} \sum_{i,j=1}^{N} k_x(X_i, X_j) \sum_{k,\ell=1}^{N} k_y(Y_k, Y_\ell)$$
Independence test with kernels II

Asymptotic distribution under null-hypothesis

Theorem (Gretton et al. 2008)

If \( X \) and \( Y \) are independent, then

\[
\sum_{i=1}^{\infty} \lambda_i Z_i^2 \quad \text{in law} \quad (N \to \infty)
\]

where

\( Z_i : \text{i.i.d. } \sim N(0,1), \)

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

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

\[
h(U_a,U_b,U_c,U_d) = \frac{1}{4!} \sum_{(a,b,c,d)} k_{a,b}^x k_{a,b}^y - 2k_{a,b}^x k_{a,c}^y + k_{a,b}^x k_{c,d}^y
\]

\[
k_{a,b}^x = k_x(X_a,X_b), \quad U_a = (X_a, Y_a)
\]

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

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

Synthesized data

- Data: two $d$-dimensional samples

$$(X_1^{(1)},...,X_d^{(1)}),..., (X_1^{(N)},...,X_d^{(N)}) \quad (Y_1^{(1)},...,Y_d^{(1)}),..., (Y_1^{(N)},...,Y_d^{(N)})$$

Samp:128, Dim:1

Samp:128, Dim:2

Samp:1024, Dim:4

strength of dependence
**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

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

\( I^0 = \text{MI} \)

\( I^2 = \text{Mean Square Conting.} \)

\( \hat{p}_j \): frequency in \( A_j \)

\( \hat{p}_{rk}^{(k)} \): 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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Match</th>
<th>BOW(N=10)</th>
<th>Spec(N=10)</th>
<th>BOW(N=50)</th>
<th>Spec(N=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HSIC_g</td>
<td>HSIC_p</td>
<td>HSIC_g</td>
<td>HSIC_p</td>
</tr>
<tr>
<td>Agriculture</td>
<td>Random</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>0.99</td>
<td>0.18</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Fishery</td>
<td>Random</td>
<td>1.00</td>
<td>0.96</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>1.00</td>
<td>0.20</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Immigration</td>
<td>Random</td>
<td>1.00</td>
<td>0.96</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>1.00</td>
<td>0.09</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Acceptance rate ($\alpha = 5\%$) (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

\[ \Phi(X) = k(X, X) \]

\[ \Omega \text{ (original space)} \] \[ \Phi \text{ feature map} \] \[ H \text{ (RKHS)} \]

- Basic statistics
  - on Euclidean space
    - Mean
    - Covariance
    - Conditional covariance
  - on RKHS
    - Mean element
    - Cross-covariance operator \( \Sigma_{YX} \)
    - 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

\[
p_{Y|Z}(y \mid z, x) = p_{Y|Z}(y \mid z)
\]  

(A)

or

\[
p_{XY|Z}(x, y \mid z) = p_{X|Z}(x \mid z) p_{Y|Z}(y \mid z)
\]  

(B)

With \( Z \) known, the information of \( X \) is unnecessary for the inference on \( Y \)
Review: Conditional Covariance

Conditional covariance of Gaussian variables

- Jointly Gaussian variable
  \[ X = (X_1, \ldots, X_p), \ Y = (Y_1, \ldots, Y_q) \]
  \[ Z = (X, Y) : m \ (= p + q) \ \text{dimensional Gaussian variable} \]
  \[ Z \sim N(\mu, V) \quad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad V = \begin{pmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{pmatrix} \]

- 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 Independence for Gaussian Variables

Two characterizations

\(X, Y, Z\) are Gaussian.

- Conditional covariance

\[X \independent Y \mid Z \iff V_{XY|Z} = O \quad \text{i.e.} \quad V_{YX} - V_{YZ}V_{ZZ}^{-1}V_{ZX} = O\]

- Comparison of conditional variance

\[X \independent Y \mid Z \iff V_{YY[X,Z]} = V_{YY|Z}\]

\[\therefore V_{YY} - V_{Y[X,Z]}V_{[X,Z][X,Z]}V_{[Z,X]Y} = V_{YY} - (V_{YX}, V_{YZ})\begin{pmatrix} V_{XX} & V_{XZ} \\ V_{ZX} & V_{ZZ} \end{pmatrix}^{-1}\begin{pmatrix} V_{XY} \\ V_{YZ} \end{pmatrix} = V_{YY} - (V_{YX}, V_{YZ})\begin{pmatrix} I \\ -V_{XX}[Z] \end{pmatrix}\begin{pmatrix} V_{XZ} & O \\ O & V_{ZZ} \end{pmatrix}\begin{pmatrix} I \\ O \end{pmatrix} - V_{XZ}V_{ZZ}^{-1}V_{XY} = V_{YY|Z} - V_{YX|Z}V_{XX|Z}V_{XY|Z}\]
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:

\[ \min_{A:q \times p \text{ matrix}} E\|\tilde{Y} - A\tilde{X}\|^2 \]

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

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

- For Gaussian variables,

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

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 \( \Omega_X, \Omega_Y, \Omega_Z \) (resp.).
  
  \((H_X, k_X), (H_Y, k_Y), (H_Z, k_Z) : \) RKHS defined on \( \Omega_X, \Omega_Y, \Omega_Z \) (resp.).

  - Conditional cross-covariance operator \( H_X \rightarrow H_Y \)

    \[ \Sigma_{YX|Z} \equiv \Sigma_{YY} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} \]

  - Conditional covariance operator \( H_Y \rightarrow H_Y \)

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

  - \( \Sigma_{ZZ}^{-1} \) may not exist as a bounded operator. But, we can justify the definitions.
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

\[ \text{Range}(W_{YX}) = \text{Range}(\Sigma_{YY}), \quad \text{Ker}(W_{YX}) \perp \text{Range}(\Sigma_{XX}). \]

Rigorous definition of conditional covariance operators

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

\[ \Sigma_{YY|Z} \equiv \Sigma_{YY} - \Sigma_{YY}^{1/2} W_{YZ} W_{ZY} \Sigma_{YY}^{1/2} \]
Two Characterizations of Conditional Independence with Kernels

(1) Conditional covariance operator \((FBJ04, 08)\)

- Conditional variance \((k_Z\text{ is characteristic})\)

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

- Conditional independence (all the kernels are characteristic)

\[
X \perp Y \mid Z \iff \Sigma_{YY[xZ]} = \Sigma_{YY|Z}
\]

\[
X \text{ is not necessary for predicting } g(Y)
\]

- c.f. Gaussian variables

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

\[
X \perp Y \mid Z \iff V_{YY[x,Z]} = V_{YY|Z}
\]
(2) Cond. cross-covariance operator (FBJ04)

- Conditional Covariance ($k_z$ is characteristic)
\[
\langle g, \Sigma_{Y|Z} f \rangle = E[\text{Cov}[g(Y), f(X) | Z]]
\]

- Conditional independence
\[
X \perp Y | Z \iff \Sigma_{Y\hat{X}|Z} = O
\]

  \[\iff \Sigma_{\hat{Y}X|Z} = O\]

  where \[\hat{X} = (X,Z), \hat{Y} = (Y,Z)\]

- c.f. Gaussian variables
\[
a^T V_{XY|Z} b = \text{Cov}[a^T X, b^T Y | Z]
\]
\[
X \perp Y | Z \iff V_{XY|Z} = O
\]
– Proof of (1) (partial) : relation between residual error and operator

\[
E[(g(Y) - E[g(Y)]) - (f(Z) - E[f(Z)])]^2
\]

\[
= \langle f, \Sigma_{ZZ} f \rangle - 2 \langle f, \Sigma_{Zy} g \rangle + \langle g, \Sigma_{yy} g \rangle
\]

\[
= \|\Sigma_{ZZ}^{1/2} f\|^2 - 2 \langle f, \Sigma_{ZZ}^{1/2} W_{zy} \Sigma_{yy}^{1/2} g \rangle + \|\Sigma_{yy}^{1/2} g\|^2
\]

\[
= \|\Sigma_{ZZ}^{1/2} f - W_{zy} \Sigma_{YY}^{1/2} g\|^2 + \|\Sigma_{yy}^{1/2} g\|^2 - \|W_{zy} \Sigma_{yy}^{1/2} g\|^2
\]

\[
= \|\Sigma_{ZZ}^{1/2} f - W_{zy} \Sigma_{YY}^{1/2} g\|^2 + \langle g, (\Sigma_{yy} - \Sigma_{yy}^{1/2} W_{zy} W_{zy} \Sigma_{YY}^{1/2}) g \rangle
\]

This part can be arbitrary small by choosing \( f \) because of

\[
\text{Range}(W_{zy}) = \text{Range}(\Sigma_{ZZ}).
\]
Proof of (1): conditional independence

\[
\textbf{Lemma} \quad \text{Var}[Y] = \text{Var}_X [E_{Y|X}[Y | X]] + E_X [\text{Var}_{Y|X}[Y | X]]
\]

From the above lemma

\[
\text{Var}[g(Y) | Z] = \text{E}_{X|Z} [\text{Var}[g(Y) | X, Z] | Z] + \text{Var}_{X|Z} [E[g(Y) | X, Z] | Z]
\]

Take \( E_Z [\cdot] \)

\[
E[\text{Var}[g(Y) | Z]] - E[\text{Var}[g(Y) | X, Z]] = E_Z [\text{Var}_{X|Z} [E[g(Y) | X, Z] | Z]]
\]

L.H.S = 0 from \( \Sigma_{YY|XZ} = \Sigma_{YY|Z} \)

\[ \Rightarrow \quad \text{Var}_{X|Z} [E[g(Y) | X, Z] | Z] = 0 \quad P_z \text{ - almost every } z \]

\[ \Rightarrow \quad E[g(Y) | X, Z] = E[g(Y) | Z] \quad P_{XZ} \text{ - almost every } (x, z) \]

\[ \Rightarrow \quad P_{Y|XZ} = P_{Y|Z} \quad (k_Y \text{ characteristic}) \]
Why is the “extended variable” needed in (2)?

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

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

The l.h.s is not a function of \( z \). c.f. Gaussian case

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

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

However, if \( X \) is replaced by \([X, Z]\)

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

where \( p(x, z' | z) = p(x | z)\delta(z' - z) \)

\[ \Rightarrow \quad p(x, y, z') = p(x | z')p(y | z')p(z') \]

i.e. \( p(x, y | z') = p(x | z')p(y | z') \)
Empirical Estimator of Cond. Cov. Operator

\[(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)\]

\[\Sigma_{YZ} \rightarrow \hat{\Sigma}_Y \hat{\Sigma}_X \ 	ext{etc.} \quad \text{finite rank operators}\]

\[\Sigma_{ZZ}^{-1} \rightarrow \left(\hat{\Sigma}_Z^{(N)} + \varepsilon_N I\right)^{-1} \quad \text{regularization for inversion}\]

- 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

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

\[G_X = Q_N K X Q_N, \quad Q_N = I_N - \frac{1}{N} 1_N 1_N^T \quad \text{centered Gram matrix}\]

\[S_Z = I_N - (G_Z + N\varepsilon_N I_N)^{-1} G_Z = \left(I_N + \frac{1}{N\varepsilon_N} G_Z\right)^{-1}\]
Statistical Consistency

Consistency on conditional covariance operator

Theorem (FBJ08, Sun et al. 07)
Assume $\epsilon_N \to 0$ and $\sqrt{N} \epsilon_N \to \infty$

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

In particular,

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

**Normalized Cross-Covariance Operator**

\[ W_{yx} = \Sigma^{-1/2}_{yy} \Sigma_{yx} \Sigma^{-1/2}_{xx} \]

Recall:
\[ \Sigma_{yx} = \Sigma^{1/2}_{yy} W_{yx} \Sigma^{1/2}_{xx} \]

**Normalized Conditional cross-covariance operator**

\[ W_{y|x|z} = \Sigma^{-1/2}_{yy} \Sigma_{yx|z} \Sigma^{-1/2}_{xx} = \Sigma^{-1/2}_{yy} \left( \Sigma_{yx} - \Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zx} \right) \Sigma^{-1/2}_{xx} \]
\[ = W_{yx} - W_{yz} W_{zx} \]

**Characterization of conditional independence**

With characteristic kernels,
\[ W_{yx} = 0 \quad \Leftrightarrow \quad X \perp Y \]
\[ W_{y|x|z} = 0 \quad \Leftrightarrow \quad X \perp Y \mid Z \]
Measures for Conditional Independence

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

- Dependence measure
  \[
  \text{NOCCO} = \|W_{YX}\|_{HS}^2
  \]

- Conditional dependence measure
  \[
  \text{NOC}^3\text{O} = \|W_{X\bar{X}|Z}\|_{HS}^2 \quad (X \text{ and } Y \text{ augmented})
  \]

- Independence / conditional independence
  \[
  \text{NOCCO} = 0 \iff X \perp Y \quad \text{NOC}^3\text{O} = 0 \iff X \perp Y \mid Z
  \]
Kernel-free Integral Expression

Theorem

Let

$$E_Z \left[ P_{Y|Z} \otimes P_{X|Z} \right](B \times A) = \int P_{Y|Z}(B | Z = z) P_{X|Z}(A | Z = z) dP_{Z}(z)$$

probability on $\Omega_X \times \Omega_Y$.

Assume

$P_{XY}$ and $E_Z \left[ P_{Y|Z} \otimes P_{X|Z} \right]$ have density $p_{XY}(x, y)$ and $p_{X \perp Y|Z}(x, y)$, resp.

$H_Z$ and $H_X \otimes H_Y$ are characteristic.

$W_{YX}$ and $W_{YZ} W_{ZX}$ are Hilbert-Schmidt.

Then,

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

In the unconditional case

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

- 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.

– In the unconditional case,
  \[ \text{NOCCO} = \| W_{YX} \|_{HS}^2 \]
  is equal to the mean square contingency, which is a very popular measure of dependence for discrete variables.

– In the conditional case, if the augmented variables are used,
  \[
  \| W_{Y\bar{X}|Z} \|_{HS}^2 = \int\int\int \left( \frac{p_{XYZ}(x, y, z) - p_{X|Z}(x | z)p_{Y|Z}(y | z)p_{Z}(z)}{p_{XZ}(x, z)p_{YZ}(y, z)} \right)^2 p_{XZ}(x, z)p_{YZ}(y, z)dx\,dy\,dz
  \]
  (conditional mean square contingency)
Empirical Estimators

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

- Inversion $\rightarrow$ regularization: $\Sigma_{XX}^{-1} \rightarrow \left(\hat{\Sigma}_{XX}^{(N)} + \epsilon I\right)^{-1}$

- 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 $\Phi_Y(Y_1), \ldots, \Phi_Y(Y_N)$

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

  $\text{NOC}^3\text{O}_{emp} = \text{Tr}\left[R_X \hat{\varphi}_Y - 2R_X \hat{\varphi} R_Z + R_Z R_{\hat{\varphi}} R_Z\right]$ (conditional dependence measure)

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

  $G_X = \left(I_N - \frac{1}{N} 1_N 1_N^T\right)K_X \left(I_N - \frac{1}{N} 1_N 1_N^T\right)$

  $K_X = \left(k(X_i, X_j)\right)_{i,j=1}^N$

- $\text{NOCCO}_{emp}$ and $\text{NOC}^3\text{O}_{emp}$ give kernel estimates for the mean square contingency and conditional mean square contingency, resp.
Consistency

Theorem (Fukumizu et al. 2008)

Assume that $W_{YX|Z}$ is Hilbert-Schmidt, and the regularization coefficient satisfies

$$\varepsilon_N \to 0 \quad N^{1/3} \varepsilon_N \to \infty.$$ 

Then,

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

In particular,

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

i.e. $\text{NOC}^3\text{O}_{\text{emp}}$ ($\text{NOCCO}_{\text{emp}}$) converges to the population value $\text{NOC}^3\text{O}$ ($\text{NOCCO}$, resp).
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)
      \[
      \sigma = \text{median}\left\{\|X_i - X_j\| \mid i \neq j\right\}
      \]
    – Speed of asymptotic convergence (Fukumizu et al. 2008)
      \[
      \lim_{N \to \infty} \text{Var}\left[ N \times HSIC^{(N)}_{\text{emp}} \right] = 2 \|\Sigma_{XX}\|_{HS}^2 \|\Sigma_{YY}\|_{HS}^2 \text{ under independence}
      \]
      Compare the bootstrapped variance and the theoretical one, and choose the parameter to give the minimum discrepancy.
Conditional Independence Test

- **Permutation test**

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

- If \( Z \) takes values in a finite set \{1, ..., L\}, set \( A_\ell = \{ i \mid Z_i = \ell \} \) \( (\ell = 1, ..., L) \), otherwise, partition the values of \( Z \) into \( L \) subsets \( C_1, ..., C_L \), and set \( A_\ell = \{ i \mid Z_i \in C_\ell \} \) \( (\ell = 1, ..., L) \).

- Repeat the following process \( B \) times: \( (b = 1, ..., B) \)
  1. Generate pseudo cond. independent data \( D^{(b)} \) by permuting \( X \) data within each \( A_\ell \).
  2. Compute \( T_N^{(b)} \) for the data \( D^{(b)} \).

\[ \rightarrow \text{Approximate null distribution} \]

\[ \rightarrow \text{under cond. indep. assumption} \]

- 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 \( t = 1, 2, 3, \ldots \)

  - 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

    \[
    H_0: \quad b_1 = b_2 = \ldots = b_p = 0
    \]

    \( X \) is called a Granger cause of \( Y \) if \( H_0 \) is rejected.
- **F-test**

  - **Linear estimation**
    \[
    Y(t) = c + \sum_{i=1}^{p} a_i Y(t - i) + \sum_{j=1}^{p} b_j X(t - j) + U_t \quad \rightarrow \quad \hat{c}, \hat{a}_i, \hat{b}_j
    \]
    \[
    H_0: \quad Y(t) = c + \sum_{i=1}^{p} a_i Y(t - i) + W_t \quad \rightarrow \quad \hat{c}, \hat{a}_i
    \]
    \[
    ERR_1 = \sum_{t=p+1}^{N} (\hat{Y}(t) - Y(t)) \quad ERR_0 = \sum_{t=p+1}^{N} (\hat{Y}(t) - Y(t))^2
    \]

  - **Test statistics**
    \[
    T_N \equiv \frac{(ERR_0 - ERR_1)/p}{ERR_1/(N - 2p + 1)} \quad \text{under } H_0 \quad F_{p,N-2p+1} \quad (N \to \infty)
    \]
    The p.d.f of \( F_{d_1,d_2} \) is
    \[
    \frac{1}{B(d_1/2,d_2/2)} \left( \frac{d_1x}{d_1x + d_2} \right)^{d_1} \left( 1 - \frac{d_1x}{d_1x + d_2} \right)^{d_2} \frac{1}{x}
    \]

- **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
  
  \[ X \text{ is NOT a cause of } Y \text{ if} \]
  
  \[ p(Y_t \mid Y_{t-1}, \ldots, Y_{t-p}, X_{t-1}, \ldots, X_{t-p}) = p(Y_t \mid Y_{t-1}, \ldots, Y_{t-p}) \]

  \[ \iff \]

  \[ Y_t \perp X_{t-1}, \ldots, X_{t-p} \mid Y_{t-1}, \ldots, Y_{t-p} \]

- Kernel measures for causality

  \[ HSCIC = \left\| \hat{\Sigma}^{(N-p+1)}_{\hat{y}X_p | Y_p} \right\|_{HS}^2 \]

  \[ HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\hat{y}X_p | Y_p} \right\|_{HS}^2 \]

  \[ X_p = \{(X_{t-1}, X_{t-2}, \ldots, X_{t-p}) \in \mathbb{R}^p \mid t = p + 1, \ldots, N\} \]

  \[ Y_p = \{(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}) \in \mathbb{R}^p \mid t = p + 1, \ldots, N\} \]
Example

- Coupled Hénon map

\[
\begin{align*}
    x_1(t+1) &= 1.4 - x_1(t)^2 + 0.3x_2(t) \\
    x_2(t+1) &= x_1(t) \\
    y_1(t+1) &= 1.4 - \left\{ \gamma x_1(t)y_1(t) + (1 - \gamma)y_1(t)^2 \right\} + 0.1y_2(t) \\
    y_2(t+1) &= y_1(t)
\end{align*}
\]

\(\gamma = 0\) 
\(\gamma = 0.25\) 
\(\gamma = 0.8\)
Causality in coupled Hénon map

- \( X \) is a cause of \( Y \) if \( \gamma > 0 \).
  \[ Y_{t+1} \not\rightarrow X_t \mid Y_t \]

- \( Y \) is not a cause of \( X \) for all \( \gamma \).
  \[ X_{t+1} \not\leftarrow Y_t \mid X_t \]

- Permutation tests for non-causality with NOC\(^3\)O

\[
\begin{array}{l|cccccccc}
\text{NOC\(^3\)O} & 0.0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 \\
\hline
0.0 & 94 & 88 & 81 & 63 & 86 & 77 & 62 \\
0.1 & 97 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 96 & 92 & 85 & 45 & 13 & 2 & 3 \\
\end{array}
\]

<table>
<thead>
<tr>
<th>\text{Granger}</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>96</td>
<td>92</td>
<td>85</td>
<td>45</td>
<td>13</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Number of times accepting \( H_0 \) among 100 datasets (\( \alpha = 5\% \))
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.
- 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).

- Further studies are required for kernel choice.


