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 \(\mathcal{X} \times \mathcal{Y}\). \(\text{resp.}\)

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

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

Cross-covariance operator:

\[\Sigma_{YX} : H_{\mathcal{X}} \rightarrow H_{\mathcal{Y}}\]

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

\[= m_{P_{YX}} - m_{P_{Y} \otimes P_{X}} \quad \in H_{\mathcal{Y}} \otimes H_{\mathcal{X}}\]

**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 \(f \in H_{\mathcal{X}}, g \in H_{\mathcal{Y}}\)

– \(\text{c.f. Euclidean case}\)

\[V_{YX} = E[YY^T] - E[Y]E[X]^T \quad : \text{covariance matrix}\]

\[(b, V_{YX} a) = \text{Cov}[(b, Y), (a, X)]\]
RKHS for product kernel

 RKHS w.r.t. product kernel

\( k_1, k_2: \) positive definite kernel on \( \Omega_1, \Omega_2 \), 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, \( \left\{ \sum_{i=1}^n f_i(x) g_i(y) \bigg| f_i \in H_1, g_i \in H_2 \right\} \subset H_1 \otimes H_2 \).
Characterization of independence

Independence and Cross-covariance operator

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

$$X \perp Y \iff E_{XY}[e^{\sqrt{-1}(uX+vY)}] = E_X[e^{\sqrt{-1}uX}]E_Y[e^{\sqrt{-1}vY}]$$
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 \(\Sigma_{YX}\) is defined by

\[
\hat{\Sigma}_{YX}^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \left\{ k_y(\cdot, Y_i) - \hat{m}_Y \right\} \otimes \left\{ k_x(\cdot, X_i) - \hat{m}_X \right\}
\]

**Theorem**

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

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 \(H_x \to H_y\).
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: \( \| A \|^2_{HS} = \sum_j \sum_i \langle \psi_j, A \varphi_i \rangle^2 \)

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

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

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

**Integral expression**

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

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

**Proof.**

\[
\| \Sigma_{yx} \|_{HS}^2 = \| E[k_x(X, \cdot) \otimes k_y(Y, \cdot)] - m_X \otimes m_Y \|^2 \\
= \langle E[k_x(X, \cdot) \otimes k_y(Y, \cdot)], E[k_x(\tilde{X}, \cdot) \otimes k_y(\tilde{Y}, \cdot)] \rangle \\
- 2\langle E[k_x(X, \cdot) \otimes k_y(Y, \cdot)], m_{\tilde{X}} \otimes m_{\tilde{Y}} \rangle + \langle m_X \otimes m_Y, m_{\tilde{X}} \otimes m_{\tilde{Y}} \rangle \\
= E[k_x(X, \tilde{X})k_y(Y, \tilde{Y})] - 2E[E[k_x(X, \tilde{X})|\tilde{X}]E[k_y(Y, \tilde{Y})|\tilde{Y}]] \\
+ E[k_x(X, \tilde{X})]E[k_y(Y, \tilde{Y})].
\]
Empirical estimator

Gram matrix expression

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

\[ \hat{M}^{(N)}_{YX} = \frac{1}{N^2} \text{Tr}[G_X G_Y] \]

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

Or equivalently,

\[ \hat{M}^{(N)}_{YX} = \left\| \hat{\Sigma}^{(N)}_{YX} \right\|_{HS}^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} k_\mathbf{x}(X_i, X_j)k_\mathbf{y}(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^{N} k_\mathbf{x}(X_i, X_j)k_\mathbf{y}(Y_i, Y_k) 
+ \frac{1}{N^4} \sum_{i,j=1}^{N} k_\mathbf{x}(X_i, X_j) \sum_{k,\ell=1}^{N} k_\mathbf{y}(Y_k, Y_\ell) \]
Application: ICA

Independent Component Analysis (ICA)

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

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

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

- Problem
  - Restore the independent signals $S$ from observations $X$.
  - $\hat{S} = BX$  \hspace{1cm} $B$: $m \times m$ orthogonal matrix

$A$: $m \times m$ invertible matrix
ICA with HS independence measure

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

Pairwise-independence criterion is applicable.

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

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

randomly generated

B

Fast KICA
**Normalized Covariance Operator**

**Normalized Cross-Covariance Operator**

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

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

**Characterization of independence**

With characteristic kernels,

\[ W_{yx} = 0 \iff X \perp Y \]

Assume \( W_{xy} \) etc. are Hilbert-Schmidt.

- Dependence measure

\[ \text{NOCCO} = \|W_{yx}\|_{HS}^2 \]
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,

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

- 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 $\hat{\Sigma}_{XY}^{(N)}$.

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

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

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

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

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

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

- NOCCO$_{\text{emp}}$ 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 $H_0$: $X$ and $Y$ are independent
  - Alternative $H_1$: $X$ and $Y$ are not independent

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

\[
\hat{M}_{YX}^{(N)} = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} k_x(X_i, X_j)k_y(Y_i, Y_j) - \frac{2}{N^3} \sum_{i,j,k=1}^{N} k_x(X_i, X_j)k_y(Y_i, Y_k) \\
+ \frac{1}{N^4} \sum_{i,j=1}^{N} k_x(X_i, X_j) \sum_{k,\ell=1}^{N} k_y(Y_k, Y_\ell)
\]
Asymptotic distribution under null-hypothesis

**Theorem (Gretton et al. 2008)**

If $X$ and $Y$ are independent, then

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

where

- $Z_i : \text{i.i.d.} \sim N(0,1)$,
- $\{\lambda_i\}_{i=1}^{\infty}$ is the eigenvalues of the following integral operator

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

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

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

Theorem (Gretton et al. 2008)

If $M_{YX}$ is not zero, then

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

where

$$\sigma^2 = 16\left(E_a \left[ E_{b,c,d} \left[ h(U_a, U_b, U_c, U_d) \right]^2 \right] - M_{YX} \right)$$
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)
      \[ \sigma = \text{median}\left\{ \|X_i - X_j\| \mid i \neq j \right\} \]
    - Speed of asymptotic convergence (Fukumizu et al. 2008)
      \[ \lim_{N \to \infty} \text{Var}\left[ N \times HSIC_{emp}^{(N)} \right] = 2 \|\Sigma_{XX}\|_{HS}^2 \|\Sigma_{YY}\|_{HS}^2 \text{ under independence} \]
      Compare the bootstrapped variance and the theoretical one, and choose the parameter to give the minimum discrepancy.
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.

<table>
<thead>
<tr>
<th>Angle</th>
<th>indep.</th>
<th>more dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0</td>
<td>4.5</td>
</tr>
<tr>
<td>HSIC (Median)</td>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td>HSIC (Asymp. Var.)</td>
<td>93</td>
<td>44</td>
</tr>
<tr>
<td>HSNIC ($\varepsilon = 10^4$, Median)</td>
<td>94</td>
<td>23</td>
</tr>
<tr>
<td>HSNIC ($\varepsilon = 10^6$, Median)</td>
<td>92</td>
<td>20</td>
</tr>
<tr>
<td>HSNIC ($\varepsilon = 10^8$, Median)</td>
<td>93</td>
<td>15</td>
</tr>
<tr>
<td>HSNIC (Asymp. Var.)</td>
<td>94</td>
<td>11</td>
</tr>
<tr>
<td>MI (#NN = 1)</td>
<td>93</td>
<td>62</td>
</tr>
<tr>
<td>MI (#NN = 3)</td>
<td>96</td>
<td>43</td>
</tr>
<tr>
<td>MI (#NN = 5)</td>
<td>97</td>
<td>49</td>
</tr>
<tr>
<td>Power Diverg. (#Bins=3)</td>
<td>96</td>
<td>92</td>
</tr>
<tr>
<td>Power Diverg. (#Bins=4)</td>
<td>98</td>
<td>29</td>
</tr>
<tr>
<td>Power Diverg. (#Bins=5)</td>
<td>94</td>
<td>60</td>
</tr>
</tbody>
</table>

# 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

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

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

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

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

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

- Null distribution under independence

\[
T_N \ \Rightarrow \ \chi^2_{q^N - qN + N - 1}
\]
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

<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>Agriculture</td>
<td>Random</td>
<td>0.94</td>
<td>0.95</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>0.18</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Fishery</td>
<td>Random</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>0.20</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Immigration</td>
<td>Random</td>
<td>0.96</td>
<td>0.91</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</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**

  - Basic statistics on Euclidean space
    - Mean
    - Covariance
    - Conditional covariance

  - Basic statistics on RKHS
    - Mean element
    - Cross-covariance operator $\Sigma_{YY}$
    - 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

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

\text{(A)}

or

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

\text{(B)}

\(\text{(A) } X \rightarrow Z \rightarrow Y\)

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

Applications

- Graphical model
- Causal inference, etc.
Conditional Independence for Gaussian Variables

- Two characterizations
  - Conditional covariance
    \[ X \perp Y \mid Z \iff V_{XY|Z} = 0 \quad \text{i.e.} \quad V_{YX} - V_{YZ} V_{ZZ}^{-1} V_{ZX} = 0 \]
  - Comparison of conditional variance
    \[ X \perp Y \mid Z \iff V_{YY[X,Z]} = V_{YY|Z} \]

\[
\begin{align*}
\therefore \quad V_{YY} - V_{Y[X,Z]} V_{[X,Z][X,Z]}^{-1} V_{[Z,X]Y} & = V_{YY} - (V_{YX}, V_{YZ}) \begin{pmatrix} V_{XX} & V_{XZ} \\ V_{ZX} & V_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} V_{XY} \\ V_{ZY} \end{pmatrix} \\
& = V_{YY} - (V_{YX}, V_{YZ}) \begin{pmatrix} I & O \\ -V_{ZZ}^{-1} V_{ZX} & I \end{pmatrix} \begin{pmatrix} V_{XX} & V_{XZ} \\ V_{ZX} & V_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} I & O \\ -V_{ZX} V_{ZZ}^{-1} & I \end{pmatrix} \begin{pmatrix} V_{XY} \\ V_{ZY} \end{pmatrix} \\
& = V_{YY|Z} - V_{YX|Z} V_{XX|Z} V_{XY|Z}
\end{align*}
\]
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,
  \[
  V_{YY[X,Z]} = V_{YY|Z} \quad (\Leftrightarrow \quad X \perp Y \mid Z)
  \]
  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
  \[ X = (X_1, \ldots, X_p), \quad Y = (Y_1, \ldots, Y_q) \]
  \[ Z = (X, Y) : \text{m (}= p + q) \text{ dimensional Gaussian variable} \]
  \[ Z \sim N(\mu, V) \quad \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad V = \begin{pmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{pmatrix} \]

- 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 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.).

  \[ \Sigma_{Y|Z} \equiv \Sigma_{YY} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} : H_X \to H_Y \]

  - Conditional cross-covariance operator

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

  - Conditional covariance operator

  \[ \Sigma_{Z|Z} \text{ may not exist as a bounded operator. But, we can justify the definitions.} \]
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 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

**Proposition (FBJ 2004, 2008+)**

Assume $k_Z$ is characteristic.

$$\left\langle g, \Sigma_{YX|Z} f \right\rangle = E[\text{Cov}[g(Y), f(X) | Z]] \quad (\forall f \in H_X, g \in H_Y)$$

In particular,

$$\left\langle g, \Sigma_{YY|Z} g \right\rangle = E[\text{Var}[g(Y) | Z]] \quad (\forall g \in H_Y)$$

Proof omitted.

Analogy to Gaussian variables:

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

$$b^T (V_{YY} - V_{YZ}V_{ZZ}^{-1}V_{ZY}) b = \text{Var}[b^T Y | Z]$$
Residual error interpretation

Proposition (FBJ 2004, 2008+)

Assume $k_Z$ is characteristic.

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

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

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)

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

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

\[
\text{Range}(W_{ZY}) = \text{Range}(\Sigma_{ZZ}).
\]
Conditional independence with kernels

Theorem (FBJ2004, 2008+)

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

\[
X \perp Y \mid Z \iff \Sigma_{YXZ} = O \quad (\iff \Sigma_{ ZXZ} = O)
\]

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

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

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

– c.f. Gaussian variables

\[
X \perp Y \mid Z \iff V_{XY[Z]} = O
\]

\[
X \perp Y \mid Z \iff V_{YY[XZ]} = V_{YY[Z]}
\]
Why is the “extended variable” needed?

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

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

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

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

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

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

\[ p(x, y, z') = p(x \mid z')p(y \mid z')p(z') \]

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

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

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

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

- Empirical conditional covariance operator

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

- Estimator of Hilbert-Schmidt norm

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

\(G_X = Q_N K_X Q_N, \quad Q_N = I_N - \frac{1}{N} 1_N 1_N^T\) 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 $\varepsilon_N \to 0$ and $\sqrt{N} \varepsilon_N \to \infty$

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

In particular,

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

### Permutation test

\[ T_N = \left\| \hat{\Sigma}_{Y|X,Z}^{(N)} \right\|_{HS}^2 \quad \text{or} \quad T_N = \left\| \hat{W}_{Y|X,Z}^{(N)} \right\|_{HS}^2 \]

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

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

- 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 \( t = 1, 2, 3, \ldots \)
  - 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.
\textbf{– F-test}

\textbullet \ Linear estimation

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

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

\[ ERR_1 = \sum_{t=p+1}^{N} (\hat{Y}(t) - Y(t)) \quad ERR_0 = \sum_{t=p+1}^{N} (\hat{Y}(t) - Y(t))^2 \]

\textbullet \ Test statistics

\[ T_N \equiv \frac{(ERR_0 - ERR_1)/p}{ERR_1/(N-2p+1)} \quad \under H_0 \quad F_{p,N-2p+1} \quad (N \rightarrow \infty) \]

\text{p.d.f of } F_{d_1,d_2} = \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}

\textbf{– Software}

\textbullet \ Matlab: Econometrics toolbox (www.spatial-econometrics.com)

\textbullet \ R: lmtest package

45
– 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
    \[
    p(Y_t \mid Y_{t-1}, \ldots, Y_{t-p}, X_{t-1}, \ldots, X_{t-p}) = p(Y_t \mid Y_{t-1}, \ldots, Y_{t-p})
    \]
    \( \iff \)
    \[
    Y_t \perp \!
    \!
    \!
    \!
    \perp\!
    \!
    \!
    \!
    X_{t-1}, \ldots, X_{t-p} \mid Y_{t-1}, \ldots, Y_{t-p}
    \]
  - Kernel measures for causality
    \[
    HSCIC = \left\| \hat{\Sigma}^{(N-p+1)}_{YX_p|Y_p} \right\|_{HS}^2
    \]
    \[
    HSNCIC = \left\| \hat{W}^{(N-p+1)}_{Y|X_p|Y_p} \right\|_{HS}^2
    \]
    \[
    X_p = \{(X_{t-1}, X_{t-2}, \ldots, X_{t-p}) \in \mathbb{R}^p \mid t = p + 1, \ldots, N\}
    \]
    \[
    Y_p = \{(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}) \in \mathbb{R}^p \mid t = p + 1, \ldots, N\}
    \]
Example

**Coupled Hénon map**

- \( X, Y: \)

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

- **Example**
Causality in coupled Hénon map

- $X$ is a cause of $Y$ if $\gamma > 0$. 
  \[ Y_{t+1} \notin X_t \mid Y_t \]

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

- Permutation tests for non-causality with NOC$^3$O

<table>
<thead>
<tr>
<th>$x_1 - y_1$</th>
<th>$H_0$: $Y_t$ is not a cause of $X_{t+1}$</th>
<th>$H_0$: $X_t$ is not a cause of $Y_{t+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>0.0 0.1 0.2 0.3 0.4 0.5 0.6</td>
<td>0.0 0.1 0.2 0.3 0.4 0.5 0.6</td>
</tr>
<tr>
<td>NOC$^3$O</td>
<td>94 88 81 63 86 77 62</td>
<td>97 0 0 0 0 0 0</td>
</tr>
<tr>
<td>Granger</td>
<td>92 96 95 90 90 94 93</td>
<td>96 92 85 45 13 2 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.

- 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).


