Kernel covariance operators -
definitions and applications

Arthur Gretton

joint work with Karsten Borgwardt, Kenji Fukumizu, Le Song, Bernhard Schölkopf,
Alex Smola, Choon-Hui Teo

Max Planck Institute for Biological Cybernetics,
Tübingen, Germany
A very short introduction to kernels

- Hilbert space of functions $f \in F$ from $\mathcal{X}$ to $\mathbb{R}$
- **RKHS**: evaluation operator $\delta_x : x \rightarrow \mathbb{R}$ **continuous**
A very short introduction to kernels

- Hilbert space of functions $f \in \mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$
- RKHS: evaluation operator $\delta_x : x \rightarrow \mathbb{R}$ continuous
- Riesz: unique representer of evaluation $k(x, \cdot) \in \mathcal{F}$:

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{F}}$$

- $k(x, \cdot)$ feature map
- $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ kernel function
A very short introduction to kernels

- Hilbert space of functions \( f \in \mathcal{F} \) from \( \mathcal{X} \) to \( \mathbb{R} \)
- RKHS: evaluation operator \( \delta_x : x \rightarrow \mathbb{R} \) continuous
- Riesz: unique representer of evaluation \( k(x, \cdot) \in \mathcal{F} \):
  \[
  f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{F}}
  \]
  - \( k(x, \cdot) \) feature map
  - \( k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \) kernel function
- Inner product between two feature maps:
  \[
  \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{F}} = k(x, x')
  \]
Dependence Detection with Kernels
Kernel dependence measures

- Independence testing
  - Given: $m$ samples $z := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ from $P$
  - Determine: Does $P = P_x P_y$?
Kernel dependence measures

- Independence testing
  - Given: $m$ samples $z := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ from $P$
  - Determine: Does $P = P_x P_y$?

- Kernel dependence measures
  - Zero only at independence
  - Take into account high order moments
  - Make “sensible” assumptions about smoothness
Kernel dependence measures

- **Independence testing**
  - **Given:** $m$ samples $z := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ from $P$
  - **Determine:** Does $P = P_x P_y$?

- **Kernel dependence measures**
  - Zero only at independence
  - Take into account high order moments
  - Make “sensible” assumptions about smoothness

- **Covariance operators** in spaces of features
  - Spectral norm (COCO) [Gretton et al., 2005c,d]
  - Hilbert-Schmidt norm (HSIC) [Gretton et al., 2005b]
Function revealing dependence (1)

- Idea: avoid density estimation when testing $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$ [Rényi, 1959]

\[
\text{COCO}(\mathbf{P}; F, G) := \sup_{f \in F, g \in G} \left( \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)] \right)
\]
Function revealing dependence (1)

- Idea: avoid density estimation when testing $P = P_x P_y$ [Rényi, 1959]

$$\text{COCO}(P; F, G) := \sup_{f \in F, g \in G} \left( E_{x,y} [f(x)g(y)] - E_x[f(x)] E_y[g(y)] \right)$$

- $\text{COCO}(P; F, G) = 0$ iff $x, y$ independent, when $F$ and $G$ are respective unit balls in universal RKHSs $\mathcal{F}$ and $\mathcal{G}$ [via Steinwart, 2001]
  - Examples: Gaussian, Laplace [see also Bach and Jordan, 2002]
Function revealing dependence (1)

- Idea: avoid density estimation when testing $P = P_x P_y$ [Rényi, 1959]

$$
\text{COCO}(P; F, G) := \sup_{f \in F, g \in G} \left( E_{x,y}[f(x)g(y)] - E_x[f(x)]E_y[g(y)] \right)
$$

- $\text{COCO}(P; F, G) = 0$ iff $x, y$ independent, when $F$ and $G$ are respective unit balls in universal RKHSs $\mathcal{F}$ and $\mathcal{G}$ [via Steinwart, 2001]
  - Examples: Gaussian, Laplace [see also Bach and Jordan, 2002]

In geometric terms:

- Covariance operator: $C_{xy} : \mathcal{G} \rightarrow \mathcal{F}$ such that

$$
\langle f, C_{xy}g \rangle_{\mathcal{F}} = E_{x,y}[f(x)g(y)] - E_x[f(x)]E_y[g(y)]
$$
Function revealing dependence (1)

- Idea: avoid density estimation when testing $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$ [Rényi, 1959]

$$\text{COCO}(\mathbf{P}; F, G) := \sup_{f \in F, g \in G} \left( \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)] \right)$$

- $\text{COCO}(\mathbf{P}; F, G) = 0$ iff $x, y$ independent, when $F$ and $G$ are respective unit balls in universal RKHSs $\mathcal{F}$ and $\mathcal{G}$ [via Steinwart, 2001]
  - Examples: Gaussian, Laplace [see also Bach and Jordan, 2002]

In geometric terms:

- Covariance operator: $C_{xy} : \mathcal{G} \rightarrow \mathcal{F}$ such that

$$\langle f, C_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)]$$

- COCO is the spectral norm of $C_{xy}$ [Gretton et al., 2005c,d]:

$$\text{COCO}(\mathbf{P}; F, G) := \|C_{xy}\|_S$$
Function revealing dependence (2)

- Ring-shaped density, correlation approx. zero [example from Fukumizu, Bach, and Gretton, 2005]
Function revealing dependence (2)

- Ring-shaped density, correlation approx. zero [example from Fukumizu, Bach, and Gretton, 2005]
Function revealing dependence (2)

- Ring-shaped density, correlation approx. zero [example from Fukumizu, Bach, and Gretton, 2005]
Function revealing dependence (3)

- **Empirical** $\text{COCO}(z; F, G)$ largest eigenvalue of

\[
\begin{bmatrix}
0 & \frac{1}{m} \tilde{K} \tilde{L} \\
\frac{1}{m} \tilde{L} \tilde{K} & 0
\end{bmatrix}
\begin{bmatrix}
c \\
d
\end{bmatrix}
= \gamma
\begin{bmatrix}
\tilde{K} & 0 \\
0 & \tilde{L}
\end{bmatrix}
\begin{bmatrix}
c \\
d
\end{bmatrix}.
\]

- $\tilde{K}$ and $\tilde{L}$ are matrices of **inner products** between centred observations in respective **feature spaces**:

\[
\tilde{K} = HKH \quad \text{where} \quad H = I - \frac{1}{m} 11^\top
\]

and $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_F$, $l(y_i, y_j) = \langle \psi(y_i), \psi(y_j) \rangle_G$
Function revealing dependence (3)

- **Empirical COCO** \( z; F, G \) largest eigenvalue of
\[
\begin{bmatrix}
0 & \frac{1}{m} \tilde{K}\tilde{L} \\
\frac{1}{m} \tilde{L}\tilde{K} & 0
\end{bmatrix}
\begin{bmatrix}
c \\
d
\end{bmatrix}
= \gamma
\begin{bmatrix}
\tilde{K} & 0 \\
0 & \tilde{L}
\end{bmatrix}
\begin{bmatrix}
c \\
d
\end{bmatrix}.
\]

- \( \tilde{K} \) and \( \tilde{L} \) are matrices of inner products between centred observations in respective feature spaces:
\[
\tilde{K} = HKH \quad \text{where} \quad H = I - \frac{1}{m}11^\top
\]
and \( k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_F \), \( l(y_i, y_j) = \langle \psi(y_i), \psi(y_j) \rangle_G \)

- **Witness function** for \( x \):
\[
f(x) = \sum_{i=1}^{m} c_i \left( k(x_i, x) - \frac{1}{m} \sum_{j=1}^{m} k(x_j, x) \right)
\]
Can we do better?

A second example with zero correlation
Function revealing dependence (4)

- Can we do better?
- A second example with zero correlation
Can we do better?

A second example with zero correlation

Correlation: 0

Correlation: −0.37  COCO$^2$: 0.06
Given $\gamma_i := \text{COCO}_i(z; F, G)$, define Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005a]:

$$\text{HSIC}(z; F, G) := \sum_{i=1}^{m} \gamma_i^2$$
Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \text{COCO}_i(z; F, G)$, define Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005a]:

$$\text{HSIC}(z; F, G) := \sum_{i=1}^{m} \gamma_i^2$$

- In limit of infinite samples:

$$\text{HSIC}(\mathbf{P}; F, G) := \|C_{xy}\|_{\text{HS}}^2$$

$$= \langle C_{xy}, C_{xy} \rangle_{\text{HS}}$$

$$= E_{x, x', y, y'}[k(x, x')l(y, y')] + E_{x, x'}[k(x, x')]E_{y, y'}[l(y, y')]$$

$$- 2E_{x, y}[E_{x'}[k(x, x')]E_{y'}[l(y, y')]]$$

- $x'$ an independent copy of $x$, $y'$ a copy of $y$
Link between HSIC and mean difference (1)

- Define the **product space** $\mathcal{F} \times \mathcal{G}$ with kernel

$$\langle \Phi(x, y), \Phi(x', y') \rangle = \mathcal{K}((x, y), (x', y')) = k(x, x')l(y, y')$$
Link between HSIC and mean difference (1)

- Define the product space $\mathcal{F} \times \mathcal{G}$ with kernel

$$\langle \Phi(x, y), \Phi(x', y') \rangle = \mathcal{R}((x, y), (x', y')) = k(x, x')l(y, y')$$

- Define the mean elements

$$\langle \mu_{xy}, \Phi(x, y) \rangle := \mathbf{E}_{x', y'} \langle \Phi(x', y'), \Phi(x, y) \rangle = \mathbf{E}_{x', y'} k(x, x')l(y, y')$$

and

$$\langle \mu_{x \perp y}, \Phi(x, y) \rangle := \mathbf{E}_{x', y''} \langle \Phi(x', y''), \Phi(x, y) \rangle = \mathbf{E}_{x'} k(x, x')\mathbf{E}_{y'} l(y, y')$$
Link between HSIC and mean difference (1)

- Define the product space $\mathcal{F} \times \mathcal{G}$ with kernel

$$\langle \Phi(x, y), \Phi(x', y') \rangle = \mathcal{K}((x, y), (x', y')) = k(x, x')l(y, y')$$

- Define the mean elements

$$\langle \mu_{xy}, \Phi(x, y) \rangle := \mathbf{E}_{x', y'} \langle \Phi(x', y'), \Phi(x, y) \rangle = \mathbf{E}_{x', y'} k(x, x')l(y, y')$$

and

$$\langle \mu_{x \perp y}, \Phi(x, y) \rangle := \mathbf{E}_{x', y''} \langle \Phi(x', y''), \Phi(x, y) \rangle = \mathbf{E}_{x'} k(x, x')\mathbf{E}_{y'} l(y, y')$$

- The distance between these two mean elements is

$$\|\mu_{xy} - \mu_{x \perp y}\|_{\mathcal{F} \times \mathcal{G}}^2 = \langle \mu_{xy} - \mu_{x \perp y}, \mu_{xy} - \mu_{x \perp y} \rangle_{\mathcal{F} \times \mathcal{G}} = \text{HSIC}(\mathcal{P}, \mathcal{F}, \mathcal{G})$$
Link between HSIC and mean difference (2)

- **Witness function:**
  \[
  \sup_{\|f\| \leq 1} \langle f, \mu_{xy} - \mu_{x\perp y} \rangle_{\mathcal{F} \times \mathcal{G}} = \|\mu_{xy} - \mu_{x\perp y}\|_{\mathcal{F} \times \mathcal{G}}
  \]

- **Link to maximum mean discrepancy (MMD)** [Gretton et al., 2007]
Distribution of HSIC at independence

- (Biased) empirical HSIC a v-statistic

\[ HSIC_b = \frac{1}{m^2} \text{trace}(KHLH) \]
Distribution of HSIC at independence

- (Biased) **empirical HSIC** a v-statistic

\[ HSIC_b = \frac{1}{m^2} \text{trace}(K^{HLH}) \]

- Associated U-statistic **degenerate** when \( P = P_x P_y \) [Serfling, 1980]:

\[ mHSIC_b \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.} \]

\[ \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2 k_{tu}l_{tv} \]
Distribution of HSIC at independence

- (Biased) **empirical HSIC** a v-statistic

\[
HSIC_b = \frac{1}{m^2} \text{trace}(KHLH)
\]

- Associated U-statistic **degenerate** when \( P = P_x P_y \) [Serfling, 1980]:

\[
mHSIC_b \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}
\]

\[
\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv}
\]

- **First two moments** [Gretton et al., 2008]:

\[
E(HSIC_b) = \frac{1}{m} \text{Tr} C_{xx} \text{Tr} C_{yy}
\]

\[
\text{var}(HSIC_b) = \frac{2(m-4)(m-5)}{(m)_4} \|C_{xx}\|_{\text{HS}}^2 \|C_{yy}\|_{\text{HS}}^2 + O(m^{-3}).
\]
Independence test: verifying ICA and ISA

- **HSICp**: null distribution via **sampling** [Feuerverger, 1993]

- **HSICg**: null distribution via **moment matching** [Kankainen, 1995]:

\[
m_{\text{HSIC}_b}(Z) \sim \frac{x^{\alpha - 1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \quad \alpha = \frac{(\mathbb{E}(\text{HSIC}_b))^2}{\text{var}(\text{HSIC}_b)}, \quad \beta = \frac{\text{var}(\text{HSIC}_b)}{m\mathbb{E}(\text{HSIC}_b)}.
\]
Independence test: verifying ICA and ISA

- **HSICp**: null distribution via **sampling** [Feuerverger, 1993]
- **HSICg**: null distribution via **moment matching** [Kankainen, 1995]
- Compare with contingency table test (PD) [Read and Cressie, 1988]
Independence test: verifying ICA and ISA

- **HSIC_p:** null distribution via sampling
- **HSIC_g:** null distribution via moment matching
- Compare with contingency table test (PD) [Read and Cressie, 1988]
Independence test: verifying ICA and ISA

- **HSICp**: null distribution via sampling
- **HSICg**: null distribution via moment matching
- Compare with contingency table test (PD) [Read and Cressie, 1988]
Independence test: verifying ICA and ISA

- HSICp: null distribution via **sampling** [Feuerverger, 1993]
- HSICg: null distribution via **moment matching** [Kankainen, 1995]
- Compare with contingency table test (PD) [Read and Cressie, 1988]
- Detection of dependence between **text and its translation** [Gretton et al., 2008]
  - String (spectrum) kernel [Leslie et al., 2002] vs bag of words
Applications of covariance operators
HSIC for Microarray feature selection

- Select genes from microarray data for classification
- Different methods choose features optimising different criteria
HSIC for Microarray feature selection

- Select genes from microarray data for classification
- Different methods choose features optimising different criteria
- Several criteria special cases of HSIC: [Song et al., 2007]
  - Pearson’s correlation (normalise by standard deviation) [van’t Veer et al., 2002, Ein-Dor et al., 2006]
  - Mean difference and variants [Bedo et al., 2006, Hastie et al., 2001]
  - Shrunken centroid [Tibshirani et al., 2002, 2003]
  - (Kernel) ridge regression [Li and Yang, 2005]
• Select genes from microarray data for classification

• Different methods choose features optimising different criteria

• Several criteria special cases of HSIC: [Song et al., 2007]
  – Pearson’s correlation (normalise by standard deviation) [van’t Veer et al.,
    2002, Ein-Dor et al., 2006]
  – Mean difference and variants [Bedo et al., 2006, Hastie et al., 2001]
  – Shrunken centroid [Tibshirani et al., 2002, 2003]
  – (Kernel) ridge regression [Li and Yang, 2005]

• When are nonlinear feature maps justified?
Feature selection: BAHSIC (1)

- Backwards elimination of irrelevant features to maximise dependence (HSIC)

**Input:** The full set of features $S$

**Output:** An ordered set of features $S^\dagger$

1. $S^\dagger \leftarrow \emptyset$
2. repeat
3. $\sigma_0 \leftarrow \Xi$
4. $I \leftarrow \arg \max_I \sum_{j \in I} \text{HSIC}(\sigma_0, S \setminus \{j\}), \ I \subset S$
5. $S \leftarrow S \setminus I$
6. $S^\dagger \leftarrow S^\dagger \cup I$
7. until $S = \emptyset$

- Application: feature selection in microarrays [Song et al., 2007]
Relation of HSIC to mean difference

- (Biased) empirical HSIC: $\text{HSIC}(X, Y) := \text{Tr}(KHLH)$
Relation of HSIC to mean difference

- **(Biased) empirical HSIC**: \( \text{HSIC}(X, Y) := \text{Tr}(K_{HLH}) \)

- HSIC equivalent to **difference in means**
  - Linear input kernel \( K_x = xx^\top \) (single feature, HSIC is sum of all feature scores)
  - Linear output kernel, \( 1/m_+ \) for one class, \(-1/m_-\) for the other

\[
\text{Tr}(K_xHLH) = \left( \frac{1}{m_+} \sum_{i=1}^{m_+} x_i - \frac{1}{m_-} \sum_{i=1}^{m_-} x_i \right)^2
\]
Relation of HSIC to mean difference

- **(Biased) empirical HSIC**: $\text{HSIC}(X, Y) := \text{Tr}(KHLH)$

- HSIC equivalent to **difference in means**
  - Linear input kernel $K_x = xx^\top$ (single feature, HSIC is sum of all feature scores)
  - Linear output kernel, $1/m_+$ for one class, $-1/m_-$ for the other

\[
\text{Tr}(K_xHLH) = \left( \frac{1}{m_+} \sum_{i=1}^{m_+} x_i - \frac{1}{m_-} \sum_{i=1}^{m_-} x_i \right)^2
\]

- HSIC equivalent to **shrunken centroid**
  - Linear kernels, $Y = \begin{pmatrix} 1_{m_+}/m_+ & -1_{m_+}/m_- \\ -1_{m_-}/m_+ & 1_{m_-}/m_- \end{pmatrix} \in \mathbb{R}^{m\times2}$

\[
\text{Tr}(K_xHLH) = (\bar{x}_+ - \bar{x})^2 + (\bar{x}_- - \bar{x})^2
\]
Relation of HSIC to ridge regression

- Objective: minimize

\[ R = \|y - Vw\|^2 + \lambda\|w\|^2 \]

where

\[ V = \begin{pmatrix} \hat{k}(x_1, \cdot) \\ \hat{k}(x_2, \cdot) \\ \vdots \\ \hat{k}(x_m, \cdot) \end{pmatrix} \quad \text{and} \quad w := \sum_i \alpha_i \hat{k}(x_i, \cdot) \]

- Solution is:

\[ R^* = y^\top y - y^\top (\hat{K} + \lambda I)^{-1} \hat{K} y \]

- Features that minimise \( R^* \) \( \Leftrightarrow \) maximise HSIC with kernel

\[ K = (\hat{K} + \lambda I)^{-1} \hat{K} \]
Linear vs nonlinear kernel: idea

- For microarray data (esp. 2 class), difference in means with linear kernel usually works best.
- Exceptions:
  - Nonlinear dependence between features and labels (e.g. class with multiple subclasses)
  - Features that interact to serve different purposes
Linear vs nonlinear kernel: application

- Three cancer subtypes (diffuse large B-cell lymphoma and leukemia, follicular lymphoma, and chronic lymphocytic leukemia)
Kernel measures of conditional dependence

- We have defined covariance on feature spaces: conditional covariance?
- Gaussian case:
  \[ C_{XY|Z} := C_{XY} - C_{XZ}C_{ZZ}^{-1}C_{ZY} \]
- Does this work in general RKHS?
Kernel measures of conditional dependence

- We have defined covariance on feature spaces: conditional covariance?
- Gaussian case:
  \[ C_{XY|Z} := C_{XY} - C_{XZ}C_{ZZ}^{-1}C_{ZY} \]
- Does this work in general RKHS?
- Problem: \( C_{ZZ}^{-1} \) may not exist! However for all \( C_{XY} \), can define [Baker, 1973]
  \[ C_{XY} = C_{XX}^{1/2}V_{XY}C_{YY}^{1/2} \quad \|V_{XY}\|_2 \leq 1 \]
  - Define \( C_{XY|Z} := C_{XY} - C_{XX}^{1/2}V_{XZ}V_{ZY}C_{YY}^{1/2} \)
We have defined covariance on feature spaces: conditional covariance? 

Gaussian case:

\[ C_{XY|Z} := C_{XY} - C_{XZ}C_{ZZ}^{-1}C_{ZY} \]

Does this work in general RKHS? 

Problem: \( C_{ZZ}^{-1} \) may not exist! However for all \( C_{XY} \), can define [Baker, 1973]

\[ C_{XY} = C_{XX}^{1/2}V_{XY}C_{YY}^{1/2} \quad \|V_{XY}\|_2 \leq 1 \]

- Define \( C_{XY|Z} := C_{XY} - C_{XX}^{1/2}V_{XZ}V_{ZY}C_{YY}^{1/2} \)

Applications:

- Subspace projection [Fukumizu, Bach, and Jordan, 2004, 2006]
- Testing to fit graphical models [Sun et al., 2007]
Kernel measures of conditional dependence

- We have defined covariance on feature spaces: conditional covariance?
- Gaussian case:
  \[ C_{XY|Z} := C_{XY} - C_{XZ}C_{ZZ}^{-1}C_{ZY} \]
- Does this work in general RKHS?
- Problem: \( C_{ZZ}^{-1} \) may not exist! However for all \( C_{XY} \), can define \([Baker, 1973]\)
  \[ C_{XY} = C_{XX}^{1/2}V_{XY}C_{YY}^{1/2} \quad \|V_{XY}\|_2 \leq 1 \]
  - Define \( C_{XY|Z} := C_{XY} - C_{XX}^{1/2}V_{XZ}V_{ZY}C_{YY}^{1/2} \)
- Applications:
  - Subspace projection \([Fukumizu, Bach, and Jordan, 2004, 2006]\)
  - Testing to fit graphical models \([Sun et al., 2007]\)
- Modified requirement: \( \mathcal{F} + \mathbb{R} \) dense in \( L^q(\mathbb{P}) \) \( \forall \mathbb{P}, q \geq 1 \)
Properties of kernel conditional covariance (1)

• For **Gaussian** cond. covariance: predict $b^\top Y$ from $a^\top X$, where covariance is $C_{XY}$ and $EX = EY = 0$:

$$\min_a (b^\top Y - a^\top X)^2 = b^\top C_{YY|X} b$$

• For **Kernel** cond. covariance [Fukumizu et al., 2006]:

$$\inf_{f \in \mathcal{F}} E[(g(Y) - E_Y g(Y)) - (f(X) - E_X f(X))]^2 = \langle g, C_{YY|Z} g \rangle_{\mathcal{G}} \quad \forall g \in \mathcal{G}$$
Properties of kernel conditional covariance (1)

- For **Gaussian** cond. covariance: predict \( b^\top Y \) from \( a^\top X \), where covariance is \( C_{XY} \) and \( \mathbf{E}X = \mathbf{E}Y = 0 \):

  \[
  \min_a (b^\top Y - a^\top X)^2 = b^\top C_{YY|X} b
  \]

- For **Kernel** cond. covariance [Fukumizu et al., 2006]:

  \[
  \inf_{f \in \mathcal{F}} \mathbb{E}[(g(Y) - \mathbb{E}_Y g(Y)) - (f(X) - \mathbb{E}_X f(X))]^2 = \langle g, C_{YY|Z} g \rangle_{\mathcal{G}} \quad \forall g \in \mathcal{G}
  \]

- Given three RKHSs \((k, \mathcal{F})\) on \( \mathcal{X} \), \((l, \mathcal{G})\) on \( \mathcal{Y} \), \((q, \mathcal{H})\) on \( \mathcal{Z} \), then [Fukumizu et al., 2006, Sun et al., 2007]:

  \[
  \langle g, C_{Y|Z} f \rangle_{\mathcal{G}} = \mathbb{E}_{\mathcal{Z}} \text{cov}[f(X), g(Y)|Z] \quad \forall f \in \mathcal{F}, g \in \mathcal{G}
  \]
Properties of kernel conditional covariance (2)

- From previous result: $C_{XY|Z} = 0 \iff P_{XY} = E_Z[P_{X|Z} \otimes P_{Y|Z}]$
  
  - where $E_Z[P_{X|Z} \otimes P_{Y|Z}](A \times B) = E_Z[P_{X|A|Z}P_{y|B|Z}]$

- Weaker than: $P_{XY} = P_{X|Z} \otimes P_{Y|Z}$
Properties of kernel conditional covariance (2)

- From previous result: $C_{XY|Z} = 0 \iff P_{XY} = E_Z[P_{X|Z} \otimes P_{Y|Z}]
  \quad$ where $E_Z[P_{X|Z} \otimes P_{Y|Z}](A \times B) = E_Z[P_{X \in A|Z}P_{y \in B|Z}]

- Weaker than: $P_{XY} = P_{X|Z} \otimes P_{Y|Z}

- How to fix: Define $\tilde{X} := (X, Z)$ with kernel $k(\tilde{x}, \cdot) := k(x, \cdot)q(z, \cdot):
  \quad C_{\tilde{X}Y|Z} = 0 \iff X \perp\!
\!
\!
\!
\perp Y|Z

- Independence criterion:
  \quad HSCIC = \|C_{\tilde{X}\tilde{Y}|Z}\|^2_{HS}$
Statistical test of conditional independence

- Partition \( Z \) data into \( L \) intervals \( C_\ell \), index sets \( S_\ell : z_i \in C_\ell \iff i \in S_\ell \)
- Repeat \( B \) times:
  - Within each \( S_\ell \), generate simulated conditionally independent data \((X_{\pi(i)}, Y_i)\)
  - Compute the HSCIC for permuted data
- Construct approximate null distribution from \( B \) values of HSCIC
- Threshold is \((1 - \alpha)th\) quantile

More details in [Sun et al., 2007, Fukumizu et al., 2008]
Conclusions (dependence measures)

- **HSIC** generalises many linear and nonlinear criteria for feature selection

- HSIC with non-linear kernel can account for:
  - Nonlinear dependence between features and labels
  - Features that work together to separate classes

- **Conditional independence**
  - Measured by combining covariance operators
  - Used in a conditional independence test
Questions?
References


Hard-to-detect dependence (1)

- COCO can be $\approx 0$ for dependent RVs with highly non-smooth densities
Hard-to-detect dependence (1)

- COCO can be $\approx 0$ for dependent RVs with highly non-smooth densities.

- Reason: norms in the denominator.

$$\text{COCO}(\mathbf{P}; F, G) := \sup_{f \in F, g \in G} \frac{\text{cov} (f(x), g(y))}{\|f\|_{F} \|g\|_{G}}$$

- RESULT: not detectable with finite sample size.

- More formally: see Ingster [1989].
Hard-to-detect dependence (2)

Density takes the form:

\[ P_{x,y} \propto 1 + \sin(\omega x) \sin(\omega y) \]
Hard-to-detect dependence (3)

- Example: sinusoids of increasing frequency

\[ \omega = 1, 2, 3, 4, 5, 6 \]

Graph showing the COCO (empirical average, 1500 samples) as a function of the frequency of the non-constant density component.
Choosing kernel size (1)

- The RKHS norm of $f$ is $\|f\|_{\mathcal{H}_X}^2 := \sum_{i=1}^{\infty} \tilde{f}_i^2 \left( \tilde{k}_i \right)^{-1}$.
- If kernel decays quickly, its spectrum decays slowly:
  - then non-smooth functions have smaller RKHS norm
- Example: spectrum of two Gaussian kernels
Choosing kernel size (2)

- Could we just decrease kernel size?
- Yes, but only up to a point