Kernel Methods for Testing Independence and Goodness of Fit

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Paris, 2019
Testing goodness of fit
Given: Samples from unknown distributions $P$ and $Q$.

Goal: do $P$ and $Q$ differ?
Now: statistical model criticism

\[ MMD(P, Q) = \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_Q f - E_P f] \]

Can we compute MMD with samples from \( Q \) and a model \( P \)?

**Problem:** usually can’t compute \( E_P f \) in closed form.
Stein idea

To get rid of $E_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [E_q f - E_p f]$$

we define the Stein operator

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x))$$

Then

$$E_P T_P f = 0$$

subject to appropriate boundary conditions. (Oates, Girolami, Chopin, 2016)
Stein idea: proof

\[ E_p [ T_p f ] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx \]

\[ \int \left[ \frac{d}{dx} (f(x)p(x)) \right] \, dx \]

\[ = [f(x)p(x)]_{-\infty}^{\infty} \]

\[ = 0 \]
Stein idea: proof

\[
E_p \left[ T_p f \right] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx
\]

\[
= \left[ \frac{d}{dx} (f(x)p(x)) \right] \left. \right|_{-\infty}^{\infty}
\]

\[
= 0
\]
Stein idea: proof

\[ E_p [T_p f] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx \]

\[ = \int \left[ \frac{d}{dx} (f(x)p(x)) \right] \, dx \]

\[ = [f(x)p(x)]_{-\infty}^{\infty} \]

\[ = 0 \]
Stein idea: proof

\[ E_p [T_p f] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx \]
\[ = \int \left[ \frac{d}{dx} (f(x)p(x)) \right] \, dx \]
\[ = [f(x)p(x)]_{-\infty}^{\infty} \]
\[ = 0 \]
Stein idea: proof

\[ E_p [ T_p f ] = \int \left[ \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) \, dx \]

\[ = \int \left[ \frac{d}{dx} (f(x)p(x)) \right] \, dx \]

\[ = [f(x)p(x)]_{-\infty}^{\infty} \]

\[ = 0 \]
Kernel Stein Discrepancy

Stein operator

\[ T_p g = \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x)) \]

Kernel Stein Discrepancy (KSD)

\[ KSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g \]
Kernel Stein Discrepancy

Stein operator

\[ T_p g = \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x)) \]

Kernel Stein Discrepancy (KSD)

\[ \text{KSD}(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g \]
Kernel Stein Discrepancy

Stein operator

\[ T_p g = \frac{1}{p(x)} \frac{d}{dx} \left( g(x) p(x) \right) \]

Kernel Stein Discrepancy (KSD)

\[ KSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g \]

Graphical representation of the functions \( p(x) \), \( q(x) \), and \( g^*(x) \).
Kernel Stein Discrepancy

Stein operator

\[ T_p g = \frac{1}{p(x)} \frac{d}{dx} \left( g(x)p(x) \right) \]

Kernel Stein Discrepancy (KSD)

\[
KSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g
\]
Simple expression using kernels

Re-write stein operator as:

\[
[T_p g](x) = \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x))
\]

\[
= \frac{d}{dx} g(x) + g(x) \frac{1}{p(x)} \frac{d}{dx} p(x)
\]

\[
= \frac{d}{dx} g(x) + g(x) \frac{d}{dx} \log p(x)
\]

Can we get a dot product in feature space?

\[
[T_p g](x) = \left( \frac{d}{dx} \log p(x) \right) g(x) + \frac{d}{dx} g(x)
\]

\[
=: \langle g, \xi_x \rangle_F
\]
Simple expression using kernels

Re-write stein operator as:

\[
[T_p g](x) = \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x))
\]

\[
= \frac{d}{dx} g(x) + g(x) \frac{1}{p(x)} \frac{d}{dx} p(x)
\]

\[
= \frac{d}{dx} g(x) + g(x) \frac{d}{dx} \log p(x)
\]

Can we get a dot product in feature space?

\[
[T_p g](x) = \left( \frac{d}{dx} \log p(x) \right) g(x) + \frac{d}{dx} g(x)
\]

\[
=: \langle g, \xi_x \rangle_F
\]
Simple expression using kernels

Re-write stein operator as:

\[
[T_p g](x) = \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x))
\]

\[
= \frac{d}{dx} g(x) + g(x) \frac{1}{p(x)} \frac{d}{dx} p(x)
\]

\[
= \frac{d}{dx} g(x) + g(x) \frac{d}{dx} \log p(x)
\]

Can we get a dot product in feature space?

\[
[T_p g](x) = \left( \frac{d}{dx} \log p(x) \right) g(x) + \frac{d}{dx} g(x)
\]

\[
=: \langle g, \xi_x \rangle_F
\]
Simple expression using kernels

Reproducing property for derivatives: for differentiable $k(x - x')$,

$$\frac{d}{dx} g(x) = \left\langle g, \frac{d}{dx} k(x, \cdot) \right\rangle_F$$

From previous slide, and denoting $z \sim q$,

$$[T_p g](z) = \left( \frac{d}{dx} \log p(z) \right) g(z) + \frac{d}{dx} g(z)$$

$$=: \left\langle g, \underbrace{k(z, \cdot) \frac{d}{dz} \log p(z) + \frac{d}{dz} k(z, \cdot)}_{\xi_z} \right\rangle_F$$
Simple expression using kernels

Reproducing property for derivatives: for differentiable $k(x - x')$,

$$\frac{d}{dx} g(x) = \left\langle g, \frac{d}{dx} k(x, \cdot) \right\rangle_{\mathcal{F}}$$

From previous slide, and denoting $z \sim q$,

$$[T_p g](z) = \left( \frac{d}{dx} \log p(z) \right) g(z) + \frac{d}{dx} g(z)$$

$$=: \left\langle g, k(z, \cdot) \frac{d}{dz} \log p(z) + \frac{d}{dz} k(z, \cdot) \right\rangle_{\mathcal{F}}$$
Kernel stein discrepancy

The kernel Stein discrepancy:

\[ \text{KSD}(p, q, \mathcal{F}) = \sup \|g\|_{\mathcal{F}} \leq 1 E_{z \sim q} \langle g, \xi_z \rangle_{\mathcal{F}} = \|E_{z \sim q} \xi_z\|_{\mathcal{F}} \]

Closed-form expression for KSD test statistic:

\[ \|E_{z \sim q} \xi_z\|_{\mathcal{F}}^2 = E_{z, z' \sim q} h_p(z, z') \]

where

\[ h_p(x, y) := \partial_x \log p(x) \partial_y \log p(y) k(x, y) + \partial_y \log p(y) \partial_x k(x, y) + \partial_x \log p(x) \partial_y k(x, y) + \partial_x \partial_y k(x, y) \]

Do not need to normalize \( p \), or sample from it.
Kernel stein discrepancy

The kernel Stein discrepancy:

$$\text{KSD}(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_{z \sim q} \langle g, \xi_z \rangle_{\mathcal{F}}$$

$$= \|E_{z \sim q} \xi_z\|_{\mathcal{F}}$$

Closed-form expression for KSD test statistic:

$$\|E_{z \sim q} \xi_z\|_{\mathcal{F}}^2 = E_{z, z' \sim q} h_p(z, z')$$

where

$$h_p(x, y) := \partial_x \log p(x) \partial_y \log p(y) k(x, y) + \partial_y \log p(y) \partial_x k(x, y) + \partial_x \partial_y k(x, y)$$

Do not need to normalize $p$, or sample from it.
Constructing threshold for a statistical test

Given samples \( \{z_i\}_{i=1}^n \sim q \), empirical KSD (test statistic) is:

\[
\tilde{KSD}(p, q, \mathcal{F}) := \frac{1}{n(n - 1)} \sum_{i=1}^{n} \sum_{j \neq i} h_p(z_i, z_j).
\]

Consistent estimate of the null distribution when \( q = p \) consistent test (Type II error goes to zero) under a rich class of alternatives (see Chwialkowski, Strathmann, G., ICML 2016 for details).
Constructing threshold for a statistical test

Given samples \( \{z_i\}_{i=1}^n \sim q \), empirical KSD (test statistic) is:

\[
\hat{KSD}(p, q, \mathcal{F}) := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} h_p(z_i, z_j).
\]

When \( q = p \), obtain estimate of null distribution with wild bootstrap:

\[
\hat{KSD}(p, q, \mathcal{F}) := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \sigma_i \sigma_j h_p(z_i, z_j).
\]

where \( \{\sigma_i\}_{i=1}^n \) i.i.d, \( E(\sigma_i) = 0 \), and \( E(\sigma_i^2) = 1 \)

- Consistent estimate of the null distribution when \( q = p \)
- Consistent test (Type II error goes to zero) under a rich class of alternatives (see Chwialkowski, Strathmann, G., ICML 2016 for details).
Statistical model criticism

Chicago crime data
Statistical model criticism

Chicago crime data

Model is Gaussian mixture with two components.
Statistical model criticism

Chicago crime data

Model is Gaussian mixture with two components

Stein witness function
Statistical model criticism

Chicago crime data

Model is Gaussian mixture with ten components.
Statistical model criticism

Chicago crime data

Model is Gaussian mixture with ten components

Stein witness function

Code: https://github.com/karlnapf/kernel_goodness_of_fit
Kernel stein discrepancy

Further applications:

- Evaluation of approximate MCMC methods.
  (Chwialkowski, Strathmann, G., ICML 2016; Gorham, Mackey, ICML 2017)

What kernel to use?

- The inverse multiquadric kernel,

\[ k(x, y) = \left( c + ||x - y||_2^2 \right)^{-\beta} \]

for \( \beta \in (-1, 0) \).
Testing statistical dependence
## Dependence testing

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ independent?

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="dog.png" alt="Dog" /></td>
<td>A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.</td>
</tr>
<tr>
<td><img src="beagle.png" alt="Beagle" /></td>
<td>Their noses guide them through life, and they're never happier than when following an interesting scent.</td>
</tr>
<tr>
<td><img src="cat.png" alt="Cat" /></td>
<td>A responsive, interactive pet, one that will blow in your ear and follow you everywhere.</td>
</tr>
</tbody>
</table>

Text from dogtime.com and petfinder.com
MMD as a dependence measure?

Could we use MMD?

\[
\text{MMD}(P_{XY}, P_X P_Y, \mathcal{H}_\kappa)
\]

- We don’t have samples from \( Q := P_X P_Y \), only pairs \( \{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY} \)
  - Solution: simulate \( Q \) with pairs \( (x_i, y_j) \) for \( j \neq i \).

- What kernel \( \kappa \) to use for the RKHS \( \mathcal{H}_\kappa \)?
MMD as a dependence measure?

Could we use MMD?

$$\text{MMD}(P_{XY}, P_X P_Y, \mathcal{H}_\kappa)$$

- We don’t have samples from $Q := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$
  - **Solution**: simulate $Q$ with pairs $(x_i, y_j)$ for $j \neq i$.

- What kernel $\kappa$ to use for the RKHS $\mathcal{H}_\kappa$?
MMD as a dependence measure?

Could we use MMD?

\[ MMD(P_{XY}, P_X P_Y, \mathcal{H}_\kappa) \]

- We don’t have samples from \( Q := P_X P_Y \), only pairs
  \( \{ (x_i, y_i) \}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY} \)
  - **Solution**: simulate \( Q \) with pairs \((x_i, y_j)\) for \( j \neq i \).

- **What kernel \( \kappa \) to use** for the RKHS \( \mathcal{H}_\kappa \)?
MMD as a dependence measure

Kernel $k$ on images with feature space $\mathcal{F}$,

$\mathbf{k}(\cdot, \cdot)$

Kernel $l$ on captions with feature space $\mathcal{G}$,

$\mathbf{l}(\cdot, \cdot)$
MMD as a dependence measure

Kernel $k$ on images with feature space $\mathcal{F}$,

$$k(i, j)$$

Kernel $l$ on captions with feature space $\mathcal{G}$,

$$l(c, d)$$

Kernel $\kappa$ on image-text pairs: are images and captions similar?

$$\kappa(i, j) = k(i, j) \times l(c, d)$$
MMD as a dependence measure

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ independent?

$$MMD^2(\hat{P}_{XY}, \hat{P}_X \hat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2} \text{trace}(KL)$$

($K, L$ column centered)
MMD as a dependence measure

Given:
Samples from a distribution $P_{X \ Y}$

Goal:
Are $X$ and $Y$ independent?

$$\text{MMD}^2(b P_{X \ Y}; b P_X b P_Y; H) := \frac{1}{n^2} \text{trace}(\text{KL})$$
MMD as a dependence measure

Two questions:

- **Why the product kernel?** Many ways to combine kernels - why not eg a sum?
- Is there a more interpretable way of defining this dependence measure?
Given: Samples from a distribution $P_{XY}$

Goal: Are $X$ and $Y$ dependent?
Illustration: dependence $\neq$ correlation

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ dependent?

![Correlation: 0.07](image)
Illustration: dependence $\neq$ correlation

- **Given:** Samples from a distribution $P_{XY}$
- **Goal:** Are $X$ and $Y$ dependent?
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.
Define two spaces, one for each witness

Function in $\mathcal{F}$
$$f(x) = \sum_{j=1}^{\infty} f_j \varphi_j(x)$$

Feature map
$$\varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{bmatrix}$$

Kernel for RKHS $\mathcal{F}$ on $\mathcal{X}$:
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Function in $\mathcal{G}$
$$g(y) = \sum_{j=1}^{\infty} g_j \phi_j(y)$$

Feature map
$$\phi(y) = \begin{bmatrix} \phi_1(y) \\ \phi_2(y) \\ \phi_3(y) \\ \vdots \end{bmatrix}$$

Kernel for RKHS $\mathcal{G}$ on $\mathcal{Y}$:
$$l(x, x') = \langle \phi(y), \phi(y') \rangle_{\mathcal{G}}$$
The constrained covariance

The constrained covariance is

\[ \text{COCO}(P_{XY}) = \sup \text{cov}[f(x)g(y)] \]

\[ \|f\|_F \leq 1 \]

\[ \|g\|_G \leq 1 \]
The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\|f\|_F \leq 1, \|g\|_G \leq 1} \text{cov} \left[ \left( \sum_{j=1}^{\infty} f_j \varphi_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \phi_j(y) \right) \right]$$
The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup \limits_{\|f\|_\mathcal{F} \leq 1, \|g\|_\mathcal{G} \leq 1} E_{xy} \left[ \left( \sum_{j=1}^{\infty} f_j \tilde{\varphi}_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \tilde{\phi}_j(y) \right) \right]$$

Feature centering: $\tilde{\varphi}(x) = \varphi(x) - E_x \varphi(x)$ and $\tilde{\phi}(y) = \phi(y) - E_y \phi(y)$.
The constrained covariance

The constrained covariance is

\[
\text{COCO}(P_{XY}) = \sup_{\|f\|_{\mathcal{F}} \leq 1, \|g\|_{\mathcal{G}} \leq 1} E_{xy} \left[ \left( \sum_{j=1}^{\infty} f_j \tilde{\varphi}_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \tilde{\phi}_j(y) \right) \right]
\]

Feature centering: \( \tilde{\varphi}(x) = \varphi(x) - E_x \varphi(x) \) and \( \tilde{\phi}(y) = \phi(y) - E_y \phi(y) \).

Rewriting:

\[
E_{xy}[f(x)g(y)] = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top \text{E}_{xy} \left( \begin{bmatrix} \tilde{\varphi}_1(x) \\ \tilde{\varphi}_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \tilde{\phi}_1(y) \\ \tilde{\phi}_2(y) \\ \vdots \end{bmatrix} \right) \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}
\]

\[= C_{\tilde{\varphi}(x)\tilde{\phi}(y)} \]
The constrained covariance

The constrained covariance is

\[
\text{COCO}(P_{XY}) = \sup_{\|f\|_F \leq 1, \|g\|_G \leq 1} E_{xy} \left[ \left( \sum_{j=1}^{\infty} f_j \tilde{\varphi}_j(x) \right) \left( \sum_{j=1}^{\infty} g_j \tilde{\phi}_j(y) \right) \right]
\]

Feature centering: \( \tilde{\varphi}(x) = \varphi(x) - E_x \varphi(x) \) and \( \tilde{\phi}(y) = \phi(y) - E_y \phi(y) \).

Rewriting:

\[
E_{xy}[f(x)g(y)] = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top E_{xy} \begin{bmatrix} \tilde{\varphi}_1(x) \\ \tilde{\varphi}_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \tilde{\phi}_1(y) & \tilde{\phi}_2(y) & \ldots \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}
\]

\[
C_{\tilde{\varphi}(x)\tilde{\phi}(y)}
\]

**COCO:** max singular value of feature covariance \( C_{\tilde{\varphi}(x)\tilde{\phi}(y)} \)
Does feature space covariance exist?

Does an uncentered covariance “matrix” (operator) in feature space exist? I.e. is there some $C_{\varphi(x)\phi(y)} : \mathcal{G} \rightarrow \mathcal{F}$ such that

$$\langle f, C_{\varphi(x)\phi(y)} g \rangle_{\mathcal{F}} = E_{xy}[f(x)g(y)]$$

Does “something” exist → Riesz theorem.
Does feature space covariance exist?

Does an uncentered covariance “matrix” (operator) in feature space exist? I.e. is there some $C_{\varphi(x)\phi(y)} : \mathcal{G} \rightarrow \mathcal{F}$ such that

$$\langle f, C_{\varphi(x)\phi(y)} g \rangle_{\mathcal{F}} = E_{xy}[f(x)g(y)]$$

Does “something” exist → Riesz theorem.

Reminder: Riesz representation theorem

In a Hilbert space $\mathcal{H}$, all bounded linear operators $A$ (meaning $|Ah| \leq \lambda_A \|h\|_{\mathcal{H}}$) can be written

$$Ah = \langle h(\cdot), g_A(\cdot) \rangle_{\mathcal{H}}$$

for some $g_A \in \mathcal{H}$.

We used this theorem to show the mean embedding $\mu_P$ exists.
The Hilbert Space \( \text{HS}(\mathcal{G}, \mathcal{F}) \)

- \( \mathcal{F} \) and \( \mathcal{G} \) separable Hilbert spaces.
- \( (g_j)_{j \in J} \) orthonormal basis for \( \mathcal{G} \).
- Index set \( J \) either finite or countably infinite.

\[
\langle g_i, g_j \rangle_{\mathcal{G}} := \begin{cases} 
1 & i = j, \\
0 & i \neq j
\end{cases}
\]

- Linear operators \( L : \mathcal{G} \to \mathcal{F} \) and \( M : \mathcal{G} \to \mathcal{F} \)
- Hilbert space \( \text{HS}(\mathcal{G}, \mathcal{F}) \)

\[
\langle L, M \rangle_{\text{HS}} = \sum_{j \in J} \langle Lg_j, Mg_j \rangle_{\mathcal{F}}
\]

(independent of orthonormal basis)

- Hilbert-Schmidt norm of the operators \( L \):

\[
\| L \|_{\text{HS}}^2 = \sum_{j \in J} \| Lg_j \|_{\mathcal{F}}^2
\]

\( L \) is Hilbert-Schmidt when this norm is finite.
The Hilbert Space $\text{HS}(\mathcal{G}, \mathcal{F})$

- $\mathcal{F}$ and $\mathcal{G}$ separable Hilbert spaces.
- $(g_j)_{j \in J}$ orthonormal basis for $\mathcal{G}$.
- Index set $J$ either finite or countably infinite.
  
  $$\langle g_i, g_j \rangle_\mathcal{G} := \begin{cases} 
  1 & i = j, \\
  0 & i \neq j 
\end{cases}$$

- Linear operators $L : \mathcal{G} \rightarrow \mathcal{F}$ and $M : \mathcal{G} \rightarrow \mathcal{F}$
- Hilbert space $\text{HS}(\mathcal{G}, \mathcal{F})$
  
  $$\langle L, M \rangle_{\text{HS}} = \sum_{j \in J} \langle Lg_j, Mg_j \rangle_\mathcal{F}$$

  (independent of orthonormal basis)

- Hilbert-Schmidt norm of the operators $L$:
  
  $$\| L \|_{\text{HS}}^2 = \sum_{j \in J} \| Lg_j \|_{\mathcal{F}}^2$$

$L$ is Hilbert-Schmidt when this norm is finite.
The Hilbert Space $\text{HS}(G, \mathcal{F})$

- $\mathcal{F}$ and $G$ separable Hilbert spaces.
- $(g_j)_{j \in J}$ orthonormal basis for $G$.
- Index set $J$ either finite or countably infinite.

\[
\langle g_i, g_j \rangle_g := \begin{cases} 
1 & i = j, \\
0 & i \neq j
\end{cases}
\]

- Linear operators $L : G \to \mathcal{F}$ and $M : G \to \mathcal{F}$
- Hilbert space $\text{HS}(G, \mathcal{F})$

\[
\langle L, M \rangle_{\text{HS}} = \sum_{j \in J} \langle Lg_j, Mg_j \rangle_{\mathcal{F}}
\]

(independent of orthonormal basis)

- Hilbert-Schmidt norm of the operators $L$:

\[
\|L\|_{\text{HS}}^2 = \sum_{j \in J} \|Lg_j\|_{\mathcal{F}}^2
\]

$L$ is Hilbert-Schmidt when this norm is finite.
The tensor product $a \otimes b$ is in $\text{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$ (generalize finite case $a b^\top$)

$$(a \otimes b)g \mapsto \langle g, b \rangle_g a$$

Is $a \otimes b \in \text{HS}(\mathcal{G}, \mathcal{F})$?

$$\|a \otimes b\|_{\text{HS}}^2 = \sum_{j \in J} \|(a \otimes b)g_j\|_{\mathcal{F}}^2$$

$$= \sum_{j \in J} \|a \langle b, g_j \rangle_g\|_{\mathcal{F}}^2$$

$$= \|a\|_{\mathcal{F}}^2 \sum_{j \in J} \left|\langle b, g_j \rangle_g\right|^2$$

$$= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2$$

where we use Parseval’s identity. Thus, the operator is Hilbert-Schmidt.
The tensor product $a \otimes b$ is in $\text{HS}(G, F)$

Given $a \in F$ and $b \in G$, the tensor product $a \otimes b$ as a rank-one operator from $G$ to $F$ (generalize finite case $a b^\top$)

$$(a \otimes b)g \mapsto \langle g, b \rangle_g a$$

Is $a \otimes b \in \text{HS}(G, F)$?

$$\|a \otimes b\|_{\text{HS}}^2 = \sum_{j \in J} \|(a \otimes b)g_j\|_F^2$$

$$= \sum_{j \in J} \|a \langle b, g_j \rangle_g\|_F^2$$

$$= \|a\|_F^2 \sum_{j \in J} \left|\langle b, g_j \rangle_g\right|^2$$

$$= \|a\|_F^2 \|b\|_G^2$$

where we use Parseval’s identity. Thus, the operator is Hilbert-Schmidt.
The tensor product $a \otimes b$ is in $\text{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$ (generalize finite case $a b^\top$)

$$(a \otimes b)g \mapsto \langle g, b \rangle_g a$$

Is $a \otimes b \in \text{HS}(\mathcal{G}, \mathcal{F})$?

$$\|a \otimes b\|_{\text{HS}}^2 = \sum_{j \in J} \|(a \otimes b)g_j\|_\mathcal{F}^2$$

$$= \sum_{j \in J} \|a \langle b, g_j \rangle_g\|_\mathcal{F}^2$$

$$= \|a\|_\mathcal{F}^2 \sum_{j \in J} \left|\langle b, g_j \rangle_g \right|^2$$

$$= \|a\|_\mathcal{F}^2 \| b\|_\mathcal{G}^2$$

where we use Parseval’s identity. Thus, the operator is Hilbert-Schmidt.
The tensor product $a \otimes b$ is in $\text{HS}(G, F)$

Given $a \in F$ and $b \in G$, the tensor product $a \otimes b$ as a rank-one operator from $G$ to $F$ (generalize finite case $a b^\top$)

$$(a \otimes b)g \mapsto \langle g, b \rangle_g a$$

Is $a \otimes b \in \text{HS}(G, F)$?

$$
||a \otimes b||^2_{\text{HS}} = \sum_{j \in J} ||(a \otimes b)g_j||^2_F \\
= \sum_{j \in J} ||a \langle b, g_j \rangle_g||^2_F \\
= ||a||^2_F \sum_{j \in J} ||\langle b, g_j \rangle_g||^2 \\
= ||a||^2_F ||b||^2_G
$$

where we use Parseval’s identity. Thus, the operator is Hilbert-Schmidt.
Covariance operator in RKHS

Reminder: does there exist $C_{\varphi(x)\phi(y)} : \mathcal{G} \to \mathcal{F}$ in some Hilbert space $\text{HS}(\mathcal{G}, \mathcal{F})$ such that

$$\left\langle C_{\varphi(x)\phi(y)}, A \right\rangle_{\text{HS}} = E_{xy} \left\langle \varphi(x) \otimes \phi(y), A \right\rangle_{\text{HS}}$$

and in particular,

$$\left\langle C_{\varphi(x)\phi(y)}, f \otimes g \right\rangle_{\text{HS}} = E_{xy} [f(x)g(y)]$$

Proof: Use Riesz representer theorem. The operator

$$C_{\varphi(x)\phi(y)} : \text{HS}(\mathcal{G}, \mathcal{F}) \to \mathbb{R}$$

$$A \mapsto E_{xy} \left\langle \phi(x) \otimes \psi(y), A \right\rangle_{\text{HS}}$$

is bounded when $E_{xy} (||\varphi(x) \otimes \phi(y)||_{\text{HS}}) < \infty$. 

27/61
Covariance operator in RKHS

Reminder: does there exist $\mathcal{C}_{\varphi(x)\phi(y)} : \mathcal{G} \to \mathcal{F}$ in some Hilbert space $\mathcal{HS}(\mathcal{G}, \mathcal{F})$ such that

$$\langle \mathcal{C}_{\varphi(x)\phi(y)}, A \rangle_{\mathcal{HS}} = \mathcal{E}_{xy} \langle \varphi(x) \otimes \phi(y), A \rangle_{\mathcal{HS}}$$

and in particular,

$$\langle \mathcal{C}_{\varphi(x)\phi(y)}, f \otimes g \rangle_{\mathcal{HS}} = \mathcal{E}_{xy} [f(x)g(y)]$$

Proof: Use Riesz representer theorem. The operator

$$\mathcal{C}_{\varphi(x)\phi(y)} : \mathcal{HS}(\mathcal{G}, \mathcal{F}) \to \mathbb{R}$$

$$A \mapsto \mathcal{E}_{xy} \langle \varphi(x) \otimes \phi(y), A \rangle_{\mathcal{HS}}$$

is bounded when $\mathcal{E}_{xy} (\|\varphi(x) \otimes \phi(y)\|_{\mathcal{HS}}) < \infty$. 
Covariance operator in RKHS

Proof (continued): Condition comes from

\[
|E_{xy} \langle \varphi(x) \otimes \phi(y), A \rangle_{HS}| \leq E_{xy} |\langle \varphi(x) \otimes \phi(y), A \rangle_{HS}| \\
\leq ||A||_{HS} E_{xy} (||\varphi(x) \otimes \phi(y)||_{HS})
\]

(first Jensen, then Cauchy-Schwarz). Thus covariance operator exists by Riesz.

Simpler condition:

\[
E_{xy} (||\varphi(x) \otimes \phi(y)||_{HS}) = E_{xy} (||\varphi(x)||_F ||\phi(y)||_G) \\
= E_{xy} \left(\sqrt{k(x, x)l(y, y)}\right) < \infty.
\]
Covariance operator in RKHS

Proof (continued): Condition comes from

\[ |E_{xy} \langle \varphi(x) \otimes \phi(y), A \rangle_{HS} | \leq E_{xy} |\langle \varphi(x) \otimes \phi(y), A \rangle_{HS} | \]
\[ \leq ||A||_{HS} E_{xy} (||\varphi(x) \otimes \phi(y)||_{HS}) \]

(first Jensen, then Cauchy-Schwarz). Thus covariance operator exists by Riesz.

Simpler condition:

\[ E_{xy} (||\varphi(x) \otimes \phi(y)||_{HS}) = E_{xy} (||\varphi(x)||_F ||\phi(y)||_G) \]
\[ = E_{xy} \left( \sqrt{k(x, x) l(y, y)} \right) < \infty. \]
Covariance operator in RKHS

Does the covariance do what we want? Namely,

$$\left\langle C_\varphi(x)\varphi(y), f \otimes g \right\rangle_{\text{HS}} = E_{xy} [f(x)g(y)]$$

Proof:

$$\left\langle f, C_\varphi(x)\varphi(y) g \right\rangle_{\mathcal{F}} = \left\langle C_\varphi(x)\varphi(y), f \otimes g \right\rangle_{\text{HS}}$$

(a) \quad E_{xy} \left\langle \varphi(x) \otimes \varphi(y), f \otimes g \right\rangle_{\text{HS}}$

$$= E_{xy} [\langle f, \varphi(x) \rangle_{\mathcal{F}} \langle g, \varphi(y) \rangle_{\mathcal{F}}]$$

$$= E_{xy} [f(x)g(y)]$$
Does the covariance do what we want? Namely,

\[ \langle C_\varphi(x)\phi(y), f \otimes g \rangle_{\text{HS}} = E_{xy}[f(x)g(y)] \]

Proof:

\[ \langle f, C_\varphi(x)\phi(y) g \rangle_F = \langle C_\varphi(x)\phi(y), f \otimes g \rangle_{\text{HS}} \]

\[ \overset{(a)}{=} E_{xy} \langle \varphi(x) \otimes \phi(y), f \otimes g \rangle_{\text{HS}} \]

\[ = E_{xy} [\langle f, \varphi(x) \rangle_F \langle g, \phi(y) \rangle_F] \]

\[ = E_{xy} [f(x)g(y)] \]

(a) by definition of the covariance operator
Covariance operator in RKHS

Does the covariance do what we want? Namely,

\[ \langle C_\varphi(x)\phi(y), f \otimes g \rangle_{HS} = E_{xy} [f(x)g(y)] \]

Proof:

\[ \langle f, C_\varphi(x)\phi(y)g \rangle_F = \langle C_\varphi(x)\phi(y), f \otimes g \rangle_{HS} \]

\[ = E_{xy} \langle \varphi(x) \otimes \phi(y), f \otimes g \rangle_{HS} \]

\[ = E_{xy} [\langle f, \varphi(x) \rangle_F \langle g, \phi(y) \rangle_F] \]

\[ = E_{xy} [f(x)g(y)] \]

(a) by definition of the covariance operator
**Covariance operator in RKHS**

Does the covariance do what we want? Namely,

\[
\left\langle C_{\varphi(x)\phi(y)}, f \otimes g \right\rangle_{\text{HS}} = E_{xy} [f(x)g(y)]
\]

Proof:

\[
\left\langle f, C_{\varphi(x)\phi(y)} g \right\rangle_{\mathcal{F}} = \left\langle C_{\varphi(x)\phi(y)}, f \otimes g \right\rangle_{\text{HS}}
\]

\[
\overset{(a)}{=} E_{xy} \left\langle \varphi(x) \otimes \phi(y), f \otimes g \right\rangle_{\text{HS}}
\]

\[
= E_{xy} \left[ \left\langle f, \varphi(x) \right\rangle_{\mathcal{F}} \left\langle g, \phi(y) \right\rangle_{\mathcal{F}} \right]
\]

\[
= E_{xy} [f(x)g(y)]
\]

(a) by definition of the covariance operator
Back to the constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\|f\|_F \leq 1, \|g\|_G \leq 1} \text{cov}[f(x)g(y)]$$
Computing COCO from finite data

Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \widehat{COCO} \) ?
Computing COCO from finite data

Given sample \( \{(x_i, y_i)\}_{i=1}^n \) i.i.d. \( P_{XY} \), what is empirical \( \hat{\text{COCO}} \)?

\( \hat{\text{COCO}} \) is largest eigenvalue \( \gamma_{\text{max}} \) of

\[
\begin{bmatrix}
0 & \frac{1}{n} \bar{K} \bar{L} \\
\frac{1}{n} \bar{L} \bar{K} & 0
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}
= \gamma
\begin{bmatrix}
\bar{K} & 0 \\
0 & \bar{L}
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}.
\]

\( \bar{K}_{ij} = \langle \varphi(x_i) - \hat{\mu}_x, \varphi(x_j) - \hat{\mu}_x \rangle_F =: \langle \bar{\varphi}(x_i), \bar{\varphi}(x_j) \rangle_F \)

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis, AISTATS’05
Computing COCO from finite data

Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \text{ i.i.d. } P_{XY} \), what is empirical \( \widehat{\text{COCO}} \)?

\( \widehat{\text{COCO}} \) is largest eigenvalue \( \gamma_{\text{max}} \) of

\[
\begin{bmatrix}
0 & \frac{1}{n} \widehat{K} \widehat{L} \\
\frac{1}{n} \widehat{L} \widehat{K} & 0 \\
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta \\
\end{bmatrix} = \gamma
\begin{bmatrix}
\widehat{K} & 0 \\
0 & \widehat{L} \\
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta \\
\end{bmatrix}.
\]

\( \widehat{K}_{ij} = \langle \varphi(x_i) - \mu_x, \varphi(x_j) - \mu_x \rangle_{\mathcal{F}} =: \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{F}} \)

Witness functions:

\[
f(x) \propto \sum_{i=1}^{n} \alpha_i \left[ k(x_i, x) - \frac{1}{n} \sum_{j=1}^{n} k(x_j, x) \right]
\]

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis, AISTATS’05
The Lagrangian is

\[ \mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(x_j) \right) \left( g(y_i) - \frac{1}{n} \sum_{j=1}^{n} g(y_j) \right) \right] \]

\[ \quad - \frac{\lambda}{2} \left( \|f\|_F^2 - 1 \right) - \frac{\gamma}{2} \left( \|g\|_G^2 - 1 \right) \]

with Lagrange multipliers \( \lambda \geq 0 \) and \( \gamma \geq 0 \).
Empirical COCO: proof

The Lagrangian is

\[
\mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( f(x_i) - \frac{1}{n} \sum_{j=1}^{n} f(x_j) \right) \left( g(y_i) - \frac{1}{n} \sum_{j=1}^{n} g(y_j) \right) \right]
\]

\[
- \frac{\lambda}{2} \left( \|f\|^2_F - 1 \right) - \frac{\gamma}{2} \left( \|g\|^2_G - 1 \right)
\]

with Lagrange multipliers \( \lambda \geq 0 \) and \( \gamma \geq 0 \).

Assume:

\[
f = \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i) \quad g = \sum_{i=1}^{n} \beta_i \tilde{\psi}(y_i)
\]

for centered \( \tilde{\phi}(x_i), \tilde{\phi}(y_i) \).
Proof (continued)

First step is **smoothness constraint**:

\[
\|f\|^2_{\mathcal{F}} - 1 = \left\langle \sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i), \sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i) \right\rangle_{\mathcal{F}} - 1
\]

\[
= \alpha^\top \tilde{K} \alpha - 1
\]
First step is smoothness constraint:

\[ \|f\|_F^2 - 1 = \left\langle \sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i), \sum_{i=1}^{n} \alpha_i \tilde{\varphi}(x_i) \right\rangle_F - 1 \]

\[ = \alpha^\top \tilde{K} \alpha - 1 \]
Second step is covariance:

\[
\begin{align*}
&= \frac{1}{n} \sum_{i=1}^{n} \langle f, \tilde{\phi}(x_i) \rangle_{\mathcal{F}} \langle g, \tilde{\phi}(y_i) \rangle_{\mathcal{G}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{\ell=1}^{n} \alpha_\ell \tilde{\phi}(x_\ell), \tilde{\phi}(x_i) \right)_{\mathcal{F}} \langle g, \tilde{\phi}(y_i) \rangle_{\mathcal{G}} \\
&= \frac{1}{n} \alpha^\top \tilde{K} \tilde{L} \beta
\end{align*}
\]
Proof (continued)

Second step is covariance:

\[
\begin{align*}
\quad & = \frac{1}{n} \sum_{i=1}^{n} \langle f, \tilde{\phi}(x_i) \rangle \mathcal{F} \langle g, \tilde{\phi}(y_i) \rangle \\
\quad & = \frac{1}{n} \sum_{i=1}^{n} \langle \sum_{\ell=1}^{n} \alpha_\ell \tilde{\phi}(x_\ell), \tilde{\phi}(x_i) \rangle \mathcal{F} \langle g, \tilde{\phi}(y_i) \rangle \\
\quad & = \frac{1}{n} \alpha^\top \tilde{K} \tilde{L} \beta
\end{align*}
\]
What is a large dependence with COCO?

Density takes the form:

\[ P_{XY} \propto 1 + \sin(\omega x) \sin(\omega y) \]

Which of these is the more “dependent”?
Finding covariance with smooth transformations

Case of $\omega = 1$:

- Correlation: 0.31
- COCO: 0.09
Finding covariance with smooth transformations

Case of $\omega = 2$: 

- Correlation: 0.02
- COCO: 0.07

- Correlation: 0.54

Graphs showing the relationship between $X$ and $Y$ before and after the smooth transformations $f(X)$ and $g(Y)$.
Finding covariance with smooth transformations

Case of $\omega = 3$:
Finding covariance with smooth transformations

Case of $\omega = 4$:

\begin{align*}
\text{Correlation: } & 0.05 \\
\text{Correlation: } & 0.25 \quad \text{COCO: } 0.02
\end{align*}
Finding covariance with smooth transformations

Case of $\omega = ??$:

- Correlation: 0.01
- Correlation: 0.14  COCO: 0.02
Finding covariance with smooth transformations

Case of $\omega = 0$: uniform noise! (shows bias)

Correlation: 0.01

Correlation: 0.14  COCO: 0.02
Dependence largest when at “low” frequencies

- As dependence is encoded at higher frequencies, the smooth mappings $f, g$ achieve lower linear dependence.
- Even for independent variables, COCO will not be zero at finite sample sizes, since some mild linear dependence will be found by $f, g$ (bias)
- This bias will decrease with increasing sample size.
Can we do better than COCO?

A second example with zero correlation.

First singular value of feature covariance $C_{\varphi(x)\phi(y)}$:
Can we do better than COCO?

A second example with zero correlation.

Second singular value of feature covariance $C_{\varphi(x)\varphi(y)}$:
Can we do better than COCO?

A second example with zero correlation.

Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:

Correlation: 0.00

Correlation: 0.37    COCO$^2$: 0.06
The Hilbert-Schmidt Independence Criterion

Writing the $i$th singular value of the feature covariance $C_{\varphi(x)\phi(y)}$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC)

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$
The Hilbert-Schmidt Independence Criterion

Writing the $i$th singular value of the feature covariance $C_\varphi(x)\varphi(y)$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC)

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$ 


HSIC is MMD with product kernel!

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = MMD^2(P_{XY}, P_X P_Y; \mathcal{H}_\kappa)$$

where $\kappa((x, y), (x', y')) = \kappa(x, x') l(y, y').
Asymptotics of HSIC under independence

- Given sample \( \{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \hat{\text{HSIC}} \)?

- **Empirical HSIC** (biased)

  \[
  \hat{\text{HSIC}} = \frac{1}{n^2} \text{trace}(KL)
  \]

  \( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i, y_j) \)  
  
  \((K \text{ and } L \text{ computed with empirically centered features})\)

- **Statistical testing**: given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\hat{\text{HSIC}} > c_\alpha) < \alpha \) for small \( \alpha \)?

- **Asymptotics of \( \hat{\text{HSIC}} \) when \( P_{XY} = P_X P_Y \):**

  \[
  n \hat{\text{HSIC}} \overset{D}{\rightarrow} \sum_{l=1}^{\infty} \lambda_l z_l^2, \quad z_l \sim \mathcal{N}(0, 1)\text{i.i.d.}
  \]

  where \( \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(x_i) dF_{i,q,r} \), \( h_{ijqr} = \frac{1}{4!} \sum_{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv} \)
Asymptotics of HSIC under independence

- Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY} \), what is empirical \( \hat{\text{HSIC}} \)?
- **Empirical HSIC** (biased)

\[
\hat{\text{HSIC}} = \frac{1}{n^2} \text{trace}(KL)
\]

\( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i, y_j) \) (K and L computed with empirically centered features)

- **Statistical testing:** given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\hat{\text{HSIC}} > c_\alpha) < \alpha \) for small \( \alpha \)?
- **Asymptotics of \( \hat{\text{HSIC}} \) when \( P_{XY} = P_X P_Y \):**

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

where \( \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r} \), \( h_{ijqr} = \frac{1}{4!} \sum_{(i,j,q,r)} (k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}) \)
Asymptotics of HSIC under independence

- Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \) i.i.d. \( P_{XY} \), what is empirical \( \hat{HSIC} \)?

- Empirical HSIC (biased)

  \[
  \hat{HSIC} = \frac{1}{n^2} \text{trace}(KL)
  \]

  \( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i y_j) \) (\( K \) and \( L \) computed with empirically centered features)

- Statistical testing: given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\hat{HSIC} > c_\alpha) < \alpha \) for small \( \alpha \)?

- Asymptotics of \( \hat{HSIC} \) when \( P_{XY} = P_X P_Y \):

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

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

- Given sample \( \{(x_i, y_i)\}_{i=1}^{n} \sim^{\text{i.i.d.}} P_{XY} \), what is empirical \( \hat{\text{HSIC}} \)?
- Empirical HSIC (biased)

\[
\hat{\text{HSIC}} = \frac{1}{n^2} \text{trace}(KL)
\]

\( K_{ij} = k(x_i, x_j) \) and \( L_{ij} = l(y_i, y_j) \)

(K and L computed with empirically centered features)

- Statistical testing: given \( P_{XY} = P_X P_Y \), what is the threshold \( c_\alpha \) such that \( P(\hat{\text{HSIC}} > c_\alpha) < \alpha \) for small \( \alpha \)?

- Asymptotics of \( \hat{\text{HSIC}} \) when \( P_{XY} = P_X P_Y \):

\[
n \hat{\text{HSIC}} \overset{D}{\to} \sum_{l=1}^{\infty} \lambda_l z^2_l, \quad z_l \sim \mathcal{N}(0, 1) \text{i.i.d.}
\]

where \( \lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv} \)
A statistical test

Given $P_{XY} = P_X P_Y$, what is the threshold $c_\alpha$ such that $P(\text{HSIC} > c_\alpha) < \alpha$ for small $\alpha$ (prob. of false positive)?

Original time series:

\[
\begin{align*}
X_1 & \quad X_2 & \quad X_3 & \quad X_4 & \quad X_5 & \quad X_6 & \quad X_7 & \quad X_8 & \quad X_9 & \quad X_{10} \\
Y_1 & \quad Y_2 & \quad Y_3 & \quad Y_4 & \quad Y_5 & \quad Y_6 & \quad Y_7 & \quad Y_8 & \quad Y_9 & \quad Y_{10}
\end{align*}
\]

Permutation:

\[
\begin{align*}
X_1 & \quad X_2 & \quad X_3 & \quad X_4 & \quad X_5 & \quad X_6 & \quad X_7 & \quad X_8 & \quad X_9 & \quad X_{10} \\
Y_7 & \quad Y_3 & \quad Y_9 & \quad Y_2 & \quad Y_4 & \quad Y_8 & \quad Y_5 & \quad Y_1 & \quad Y_6 & \quad Y_{10}
\end{align*}
\]

Null distribution via permutation

- Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation $\pi$ of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold $c_\alpha$ is $1 - \alpha$ quantile of empirical CDF
A statistical test

Given $P_{XY} = P_X P_Y$, what is the threshold $c_\alpha$ such that $P(\text{HSIC} > c_\alpha) < \alpha$ for small $\alpha$ (prob. of false positive)?

Original time series:

$X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}$
$Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7 Y_8 Y_9 Y_{10}$

Permutation:

$X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}$
$Y_7 Y_3 Y_9 Y_2 Y_4 Y_8 Y_5 Y_1 Y_6 Y_{10}$

Null distribution via permutation

- Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation $\pi$ of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold $c_\alpha$ is $1 - \alpha$ quantile of empirical CDF
A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold $c_\alpha$ such that $P(\text{HSIC} > c_\alpha) < \alpha$ for small $\alpha$ (prob. of false positive)?

- **Original time series:**

  $X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_6 \ X_7 \ X_8 \ X_9 \ X_{10}$
  $Y_1 \ Y_2 \ Y_3 \ Y_4 \ Y_5 \ Y_6 \ Y_7 \ Y_8 \ Y_9 \ Y_{10}$

- **Permutation:**

  $X_1 \ X_2 \ X_3 \ X_4 \ X_5 \ X_6 \ X_7 \ X_8 \ X_9 \ X_{10}$
  $Y_7 \ Y_3 \ Y_9 \ Y_2 \ Y_4 \ Y_8 \ Y_5 \ Y_1 \ Y_6 \ Y_{10}$

- **Null distribution via permutation**
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation $\pi$ of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold $c_\alpha$ is $1 - \alpha$ quantile of empirical CDF
### Application: dependence detection across languages

**Testing task:** detect dependence between English and French text

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honourable senators, I have a question for the Leader of the Government in the Senate</td>
<td>Honorables sénateurs, ma question s’adresse au leader du gouvernement au Sénat</td>
</tr>
<tr>
<td>No doubt there is great pressure on provincial and municipal governments</td>
<td>Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions</td>
</tr>
<tr>
<td>In fact, we have increased federal investments for early childhood development.</td>
<td>Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes</td>
</tr>
</tbody>
</table>

Text from the aligned hansards of the 36th parliament of Canada, [https://www.isi.edu/natural-language/download/hansard/](https://www.isi.edu/natural-language/download/hansard/)
Application: dependence detection across languages

Testing task: detect dependence between English and French text

\( k \)-spectrum kernel, \( k = 10 \), sample size \( n = 10 \)

\[
\begin{align*}
\overline{\text{HSIC}} &= \frac{1}{n^2} \text{trace}(KL) \\
(K \text{ and } L \text{ column centered})
\end{align*}
\]
Application: Dependence detection across languages

Results (for $\alpha = 0.05$)

- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

Settings: Five line extracts, averaged over 300 repetitions, for “Agriculture” transcripts. Similar results for Fisheries and Immigration transcripts.
Testing higher order interactions
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?

\[ X \perp Y, Y \perp Z, X \perp Z \]

\[ X_1 \text{ vs } Y_1, Y_1 \text{ vs } Z_1, X_1 \text{ vs } Z_1, X_1Y_1 \text{ vs } Z_1 \]

\[ X_1 \text{ vs } Y_1, Y_1 \text{ vs } Z_1 \]

\[ X \sim N(0, 1), X, Y \text{ i.i.d. } Z \mid X, Y \sim \text{sign}(XY) \exp\left( \frac{1}{\sqrt{2}} \right) \]

**Fine print:** Faithfulness violated here!
Assume $X \perp Y$ has been established.

V-structure can then be detected by:

- **Consistent CI test:** $H_0 : X \perp Y | Z$ [Fukumizu et al. 2008, Zhang et al. 2011]
- **Factorisation test:** $H_0 : (X, Y) \perp Z \lor (X, Z) \perp Y \lor (Y, Z) \perp X$
  
  (multiple standard two-variable tests)

How well do these work?
Detecting higher order interaction

Generalise earlier example to $p$ dimensions

$X \perp Y, Y \perp Z, X \perp Z$

$X_1$ vs $Y_1$  $Y_1$ vs $Z_1$

$X_1$ vs $Z_1$  $X_1*Y_1$ vs $Z_1$

$X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$

$Z| X, Y \sim \text{sign}(XY)\ Exp\left(\frac{1}{\sqrt{2}}\right)$

$X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{p-1})$

Fine print: Faithfulness violated here!
V-structure discovery

CI test for $X \perp Y \mid Z$ from Zhang et al. (2011), and a factorisation test, $n = 500$
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[ D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y \]
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that **vanishes** whenever \(P\) can be factorised non-trivially.

\[
D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y
\]

\[
D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\]
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that \textit{vanishes} whenever \(P\) can be factorised non-trivially.

\[
\begin{align*}
D = 2 : \quad & \Delta_L P = P_{XY} - P_X P_Y \\
D = 3 : \quad & \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\end{align*}
\]
Lancaster interaction measure

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that \textbf{vanishes} whenever \(P\) can be factorised non-trivially.

\[
D = 2: \quad \Delta_L P = P_{XY} - P_X P_Y \\
D = 3: \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\]

\[\Delta_L P = 0\]

\[P_{XYZ} \quad \underline{P_X P_{YZ}}\]

\[P_{XZ} P_Y \quad \underline{-P_{XY} P_Z}\]

\[+2P_X P_Y P_Z\]

Case of \(P_X \perp \perp P_{YZ}\)
Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[
\begin{align*}
D = 2 : & \quad \Delta_L P = P_{XY} - P_X P_Y \\
D = 3 : & \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z
\end{align*}
\]

\((X, Y) \perp Z \lor (X, Z) \perp Y \lor (Y, Z) \perp X \Rightarrow \Delta_L P = 0.\)

...so what might be missed?
**Lancaster interaction measure**

Lancaster interaction measure of \((X_1, \ldots, X_D) \sim P\) is a signed measure \(\Delta P\) that vanishes whenever \(P\) can be factorised non-trivially.

\[
D = 2 : \quad \Delta_L P = P_{XY} - P_X P_Y \\
D = 3 : \quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z
\]

\(\Delta_L P = 0 \not\Rightarrow (X, Y) \perp\!\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\!\perp X\)

**Example:**

<table>
<thead>
<tr>
<th>(P(0, 0, 0) = 0.2)</th>
<th>(P(0, 0, 1) = 0.1)</th>
<th>(P(1, 0, 0) = 0.1)</th>
<th>(P(1, 0, 1) = 0.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P(0, 1, 0) = 0.1)</td>
<td>(P(0, 1, 1) = 0.1)</td>
<td>(P(1, 1, 0) = 0.1)</td>
<td>(P(1, 1, 1) = 0.2)</td>
</tr>
</tbody>
</table>
Construct a test by estimating $\| \mu_\kappa (\Delta_L P) \|^2_{\mathcal{H}_\kappa}$, where $\kappa = k \otimes l \otimes m$:

$$
\| \mu_\kappa (P_{XYZ} - P_{XY}P_Z - \cdots) \|^2_{\mathcal{H}_\kappa} = \\
\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ} \rangle_{\mathcal{H}_\kappa} - 2 \langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY}P_Z \rangle_{\mathcal{H}_\kappa} \cdots
$$
A kernel test statistic using Lancaster Measure

<table>
<thead>
<tr>
<th>$\nu \setminus \nu'$</th>
<th>$P_{XYZ}$</th>
<th>$P_{XY}P_{YZ}$</th>
<th>$P_{XZ}P_{Y}$</th>
<th>$P_{YZ}P_{X}$</th>
<th>$P_{X}P_{Y}P_{Z}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{XYZ}$</td>
<td>$(K \circ L \circ M)_{++}$</td>
<td>$((K \circ L) M)_{++}$</td>
<td>$((K \circ M) L)_{++}$</td>
<td>$((M \circ L) K)_{++}$</td>
<td>$\text{tr}(K_+ \circ L_+ \circ M_+)$</td>
</tr>
<tr>
<td>$P_{XY}P_{Z}$</td>
<td>$(K \circ L)<em>{++} M</em>{++}$</td>
<td>$(MKL)_{++}$</td>
<td>$(KLM)_{++}$</td>
<td>$(KL)<em>{++} M</em>{++}$</td>
<td>$(KL)<em>{++} M</em>{++}$</td>
</tr>
<tr>
<td>$P_{XZ}P_{Y}$</td>
<td>$(K \circ M)<em>{++} L</em>{++}$</td>
<td>$(KML)_{++}$</td>
<td>$(KML)_{++}$</td>
<td>$(KM)<em>{++} L</em>{++}$</td>
<td>$(KM)<em>{++} L</em>{++}$</td>
</tr>
<tr>
<td>$P_{YZ}P_{X}$</td>
<td>$(L \circ M)<em>{++} K</em>{++}$</td>
<td>$(LM)<em>{++} K</em>{++}$</td>
<td>$(LM)<em>{++} K</em>{++}$</td>
<td>$K_{++} L_{++} M_{++}$</td>
<td>$K_{++} L_{++} M_{++}$</td>
</tr>
<tr>
<td>$P_{X}P_{Y}P_{Z}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: $V$-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{H_\kappa}$ (without terms $P_{X}P_{Y}P_{Z}$). $H$ is centering matrix $I - n^{-1}$

**Lancaster interaction statistic:** Sejdinovic, G, Bergsma, NIPS13

$$\| \mu_\kappa (\Delta_L P) \|^2_{H_\kappa} = \frac{1}{n^2} \left( H K H \circ H L H \circ H M H \right)_{++}.$$
## A kernel test statistic using Lancaster Measure

<table>
<thead>
<tr>
<th>$\nu \backslash \nu'$</th>
<th>$P_{XYZ}$</th>
<th>$P_{XY}P_Z$</th>
<th>$P_{XZ}P_Y$</th>
<th>$P_{YZ}P_X$</th>
<th>$P_XPYP_Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{XYZ}$</td>
<td>$(K \circ L \circ M)_{++}$</td>
<td>$((K \circ L)M)_{++}$</td>
<td>$((K \circ M)L)_{++}$</td>
<td>$((M \circ L)K)_{++}$</td>
<td>$\text{tr}(K_+ \circ L_+ \circ M_+)$</td>
</tr>
<tr>
<td>$P_{XY}P_Z$</td>
<td>$(K \circ L)<em>{++}M</em>{++}$</td>
<td>$(MKL)_{++}$</td>
<td>$(KLM)_{++}$</td>
<td>$(KL)<em>{++}M</em>{++}$</td>
<td></td>
</tr>
<tr>
<td>$P_{XZ}P_Y$</td>
<td>$(K \circ M)<em>{++}L</em>{++}$</td>
<td>$(KML)_{++}$</td>
<td>$(KM)<em>{++}L</em>{++}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{YZ}P_X$</td>
<td>$(L \circ M)<em>{++}K</em>{++}$</td>
<td>$(LM)<em>{++}K</em>{++}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_XP_YP_Z$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$K_{++}L_{++}M_{++}$</td>
</tr>
</tbody>
</table>

Table: $V$-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_XPYP_Z$). $H$ is centering matrix $I - n^{-1}$

### Lancaster interaction statistic:

Sejdinovic, G, Bergsma, NIPS13

$$
\| \mu_\kappa (\Delta_L P) \|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} (HKH \circ HLLH \circ HMMH)_{++}.
$$

Empirical joint central moment in the feature space
Lancaster test, CI test for $X \perp Y \mid Z$ from Zhang et al. (2011), and a factorisation test, $n = 500$
Interaction for $D \geq 4$

- Interaction measure valid for all $D$:

  \[(Streitberg, 1990)\]

  \[\Delta_S P = \sum_{\pi}(-1)^{|\pi|-1}(|\pi| - 1)!J_\pi P\]

- For a partition $\pi$, $J_\pi$ associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$. 
Interaction for $D \geq 4$

- Interaction measure valid for all $D$:
  \[(\text{Streitberg, 1990})\]
  \[
  \Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_{\pi} P
  \]

- For a partition $\pi$, $J_{\pi}$ associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1} P_{X_3} P_{X_2} P_{X_4}$. 

59/61
Interaction for $D \geq 4$

- **Interaction measure valid for all $D$:**
  
  $\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi|-1)! J_\pi P$

- For a partition $\pi$, $J_\pi$ associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1} P_{X_3} P_{X_2} P_{X_4}$.
Co-authors

From Gatsby:

- Kacper Chwialkowski
- Wittawat Jitkrittum
- Heiko Strathmann

External collaborators:

- Wicher Bergsma
- Olivier Bousquet
- Kenji Fukumizu
- Bernhard Schoelkopf
- Dino Sejdinovic
- Bharath Sriperumbudur
- Alex Smola
- Le Song
- Zoltan Szabo
Questions?