

Kernel Methods for Testing Independence and Goodness of Fit

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Paris, 2019

Testing goodness of fit

Before: comparing two samples

- Given: Samples from unknown distributions P and Q .
- Goal: do P and Q differ?



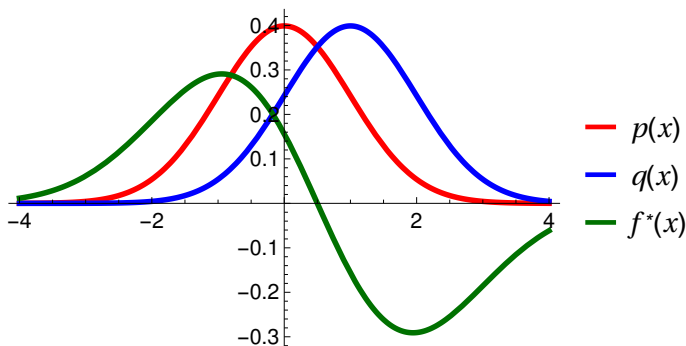
$\sim P$



$\sim Q$

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

$$\begin{aligned} E_p [T_p f] &= \int \left[\frac{1}{p(x)} \frac{d}{dx} (f(x)p(x)) \right] p(x) dx \\ &= \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_p [T_p f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &= \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_p [T_p f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &= \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_p [T_p f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &= \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Stein idea: proof

$$\begin{aligned} E_p [T_p f] &= \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx \\ &= \int \left[\frac{d}{dx} (f(x)p(x)) \right] dx \\ &= [f(x)p(x)]_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

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)

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

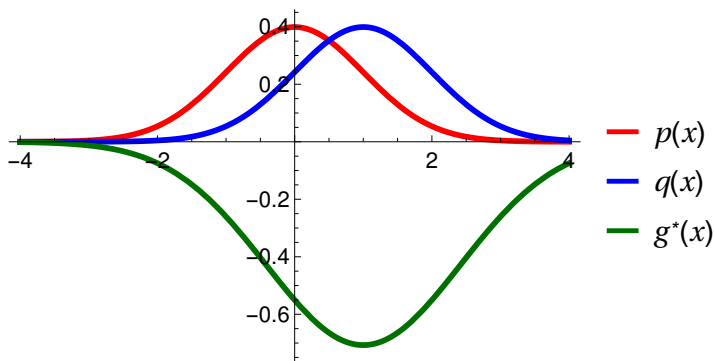
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



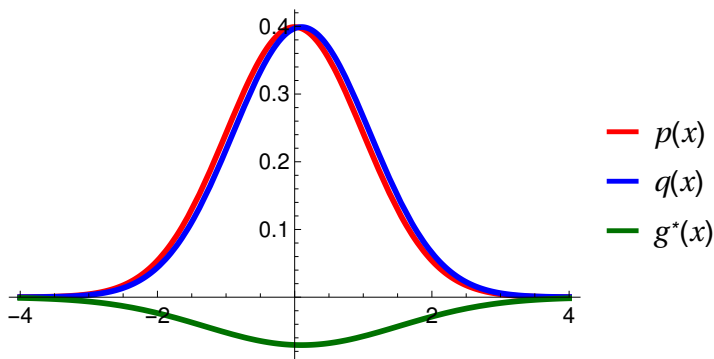
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



Simple expression using kernels

Re-write stein operator as:

$$\begin{aligned} [T_p g](x) &= \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x)) \\ &= \frac{d}{dx} g(x) + g(x) \frac{1}{p(x)} \frac{d}{dx} p(x) \\ &= \frac{d}{dx} g(x) + g(x) \frac{d}{dx} \log p(x) \end{aligned}$$

Can we get a dot product in feature space?

$$\begin{aligned} [T_p g](x) &= \left(\frac{d}{dx} \log p(x) \right) g(x) + \frac{d}{dx} g(x) \\ &=: \langle g, \xi_x \rangle_{\mathcal{F}} \end{aligned}$$

Simple expression using kernels

Re-write stein operator as:

$$\begin{aligned}[T_p g](x) &= \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x)) \\ &= \frac{d}{dx} g(x) + g(x) \frac{1}{p(x)} \frac{d}{dx} p(x) \\ &= \frac{d}{dx} g(x) + g(x) \frac{d}{dx} \log p(x)\end{aligned}$$

Can we get a dot product in feature space?

$$\begin{aligned}[T_p g](x) &= \left(\frac{d}{dx} \log p(x) \right) g(x) + \frac{d}{dx} g(x) \\ &=: \langle g, \xi_x \rangle_{\mathcal{F}}\end{aligned}$$

Simple expression using kernels

Re-write stein operator as:

$$\begin{aligned}[T_p g](x) &= \frac{1}{p(x)} \frac{d}{dx} (g(x)p(x)) \\ &= \frac{d}{dx} g(x) + g(x) \frac{1}{p(x)} \frac{d}{dx} p(x) \\ &= \frac{d}{dx} g(x) + g(x) \frac{d}{dx} \log p(x)\end{aligned}$$

Can we get a dot product in feature space?

$$\begin{aligned}[T_p g](x) &= \left(\frac{d}{dx} \log p(x) \right) g(x) + \frac{d}{dx} g(x) \\ &=: \langle g, \xi_x \rangle_{\mathcal{F}}\end{aligned}$$

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 $z \sim q$,

$$\begin{aligned} [T_p g](z) &= \left(\frac{d}{dx} \log p(z) \right) g(z) + \frac{d}{dx} g(z) \\ &=: \left\langle g, \underbrace{k(z, \cdot) \frac{d}{dz} \log p(z)}_{\xi_z} + \frac{d}{dz} k(z, \cdot) \right\rangle_{\mathcal{F}} \end{aligned}$$

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 $z \sim q$,

$$\begin{aligned} [T_p g](z) &= \left(\frac{d}{dx} \log p(z) \right) g(z) + \frac{d}{dx} g(z) \\ &=: \left\langle g, \underbrace{k(z, \cdot) \frac{d}{dz} \log p(z) + \frac{d}{dz} k(z, \cdot)}_{\xi_z} \right\rangle_{\mathcal{F}} \end{aligned}$$

Kernel stein discrepancy

The kernel Stein discrepancy:

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

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

$$\begin{aligned}h_p(x, y) &:= \partial_x \log p(x) \partial_y \log p(y) k(x, y) \\ &\quad + \partial_y \log p(y) \partial_x k(x, y) + \partial_x \log p(x) \partial_y k(x, y) \\ &\quad + \partial_x \partial_y k(x, y)\end{aligned}$$

Do not need to normalize p , or sample from it.

Kernel stein discrepancy

The kernel Stein discrepancy:

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

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

$$\begin{aligned}h_p(x, y) &:= \partial_x \log p(x) \partial_y \log p(y) k(x, y) \\ &+ \partial_y \log p(y) \partial_x k(x, y) + \partial_x \log p(x) \partial_y k(x, y) \\ &+ \partial_x \partial_y k(x, y)\end{aligned}$$

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:

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

Constructing threshold for a statistical test

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

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

When $q = p$, obtain estimate of null distribution with **wild bootstrap**:

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

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

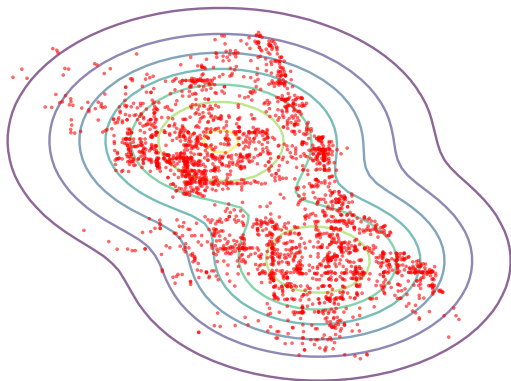
- Consistent estimate of the null distribution when $q = p$
- Consistent test (Type II error goes to zero) under a rich class of alternatives (see Chwialkowski, Strathmann, G., ICML 2016 for details).

Statistical model criticism



Chicago crime data

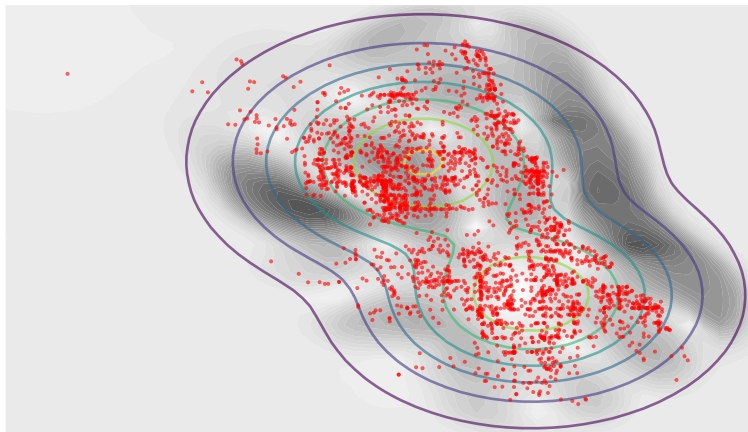
Statistical model criticism



Chicago crime data

Model is Gaussian mixture with **two** components.

Statistical model criticism

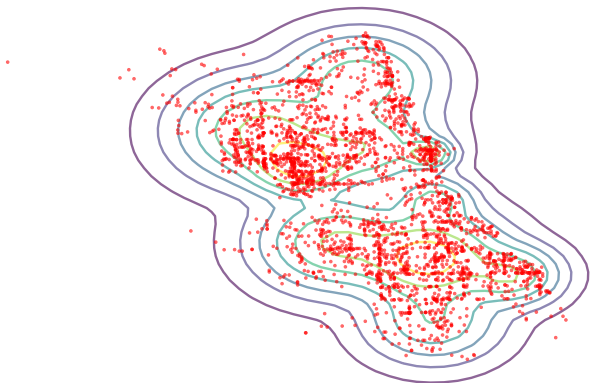


Chicago crime data

Model is Gaussian mixture with **two** components

Stein witness function

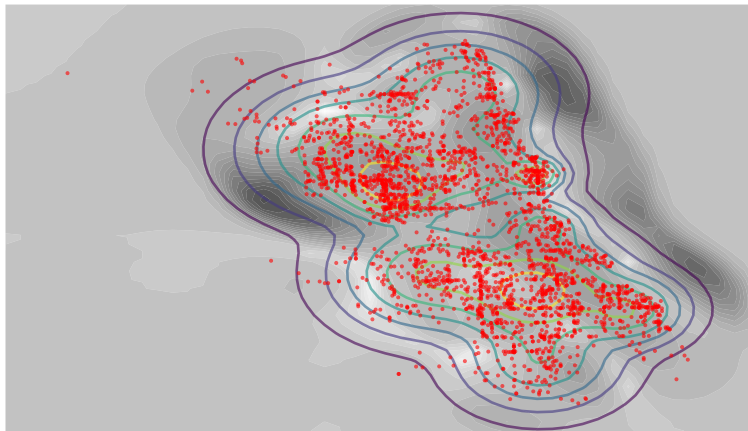
Statistical model criticism



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 $\beta \in (-1, 0)$.

arXiv.org > stat > arXiv:1703.01717

Statistics > Machine Learning

Measuring Sample Quality with Kernels

Jackson Gorham, Lester Mackey




ICML 2017

(Submitted on 6 Mar 2017 (v1), last revised 3 Aug 2017 (this version, v6))

Testing statistical dependence

Dependence testing

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

X	Y
	A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.
	Their noses guide them through life, and they're never happier than when following an interesting scent.
	A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

MMD as a dependence measure?

Could we use MMD?

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

- We don't have samples from $Q := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$
 - **Solution:** simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS \mathcal{H}_κ ?

MMD as a dependence measure?

Could we use MMD?

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

- We don't have samples from $Q := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$
 - **Solution:** simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS \mathcal{H}_κ ?

MMD as a dependence measure?

Could we use MMD?

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

- We don't have samples from $Q := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{XY}$
 - **Solution:** simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS \mathcal{H}_κ ?

MMD as a dependence measure

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

$$k(\text{dog image}, \text{cat image})$$

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

$$l(\text{A large animal who slings slobber, ...}, \text{A responsive, interactive pet ...})$$

MMD as a dependence measure

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

$$k(\text{dog image}, \text{cat image})$$

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

$$l(\text{caption box}, \text{caption box})$$

Kernel κ on **image-text pairs**: **are images and captions similar?**

$$\kappa(\text{dog image}, \text{caption box}, \text{cat image}, \text{caption box})$$

$$= k(\text{dog image}, \text{cat image}) \times l(\text{caption box}, \text{caption box})$$

MMD as a dependence measure

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

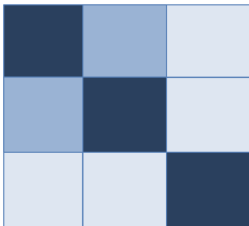
$$MMD^2(\hat{P}_{XY}, \hat{P}_X \hat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2} \text{trace}(KL)$$

(K, L column centered)

MMD as a dependence measure



K

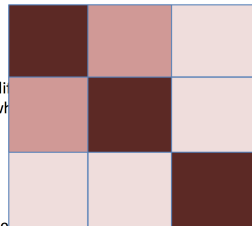


A large animal who slings slobber, exudes a distinctive houndy odor, ...

Their noses guide them through life and they're never happier than when following an interesting scent.

A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

L



Text from dogtime.com and petfinder.com

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?

Illustration: dependence \neq correlation

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

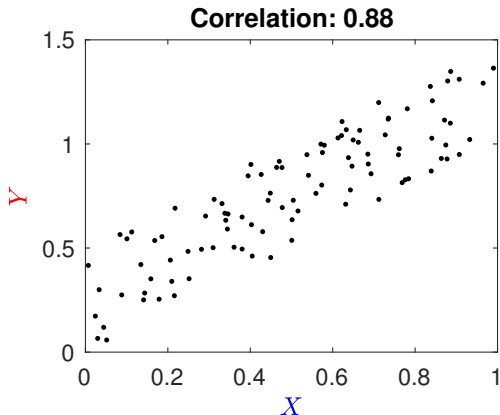


Illustration: dependence \neq correlation

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

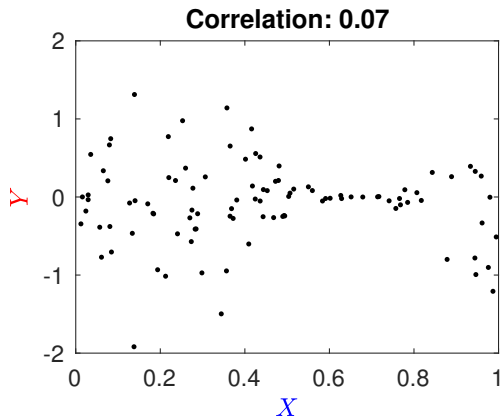
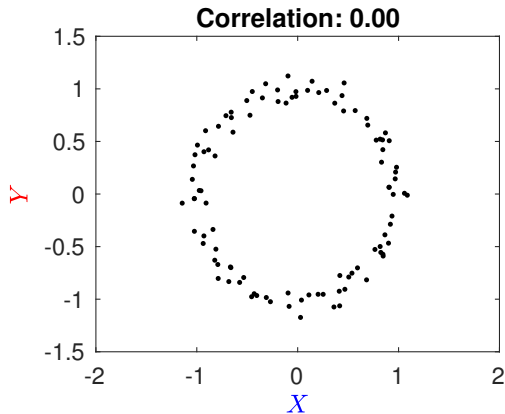


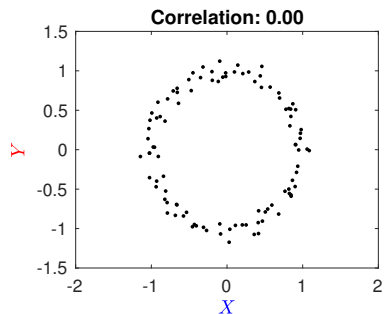
Illustration: dependence \neq correlation

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



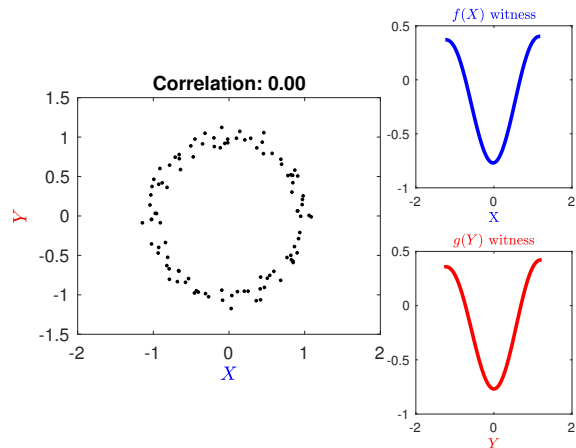
Finding covariance with smooth transformations

Illustration: two variables with no **correlation** but strong **dependence**.



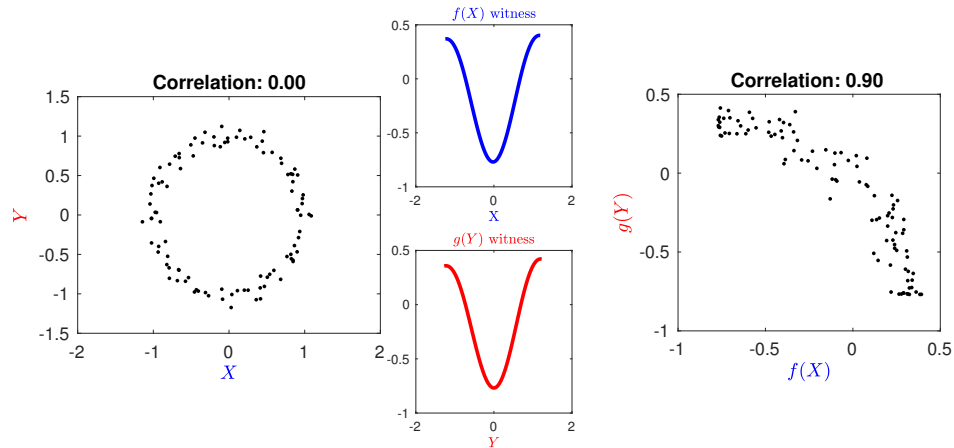
Finding covariance with smooth transformations

Illustration: two variables with no **correlation** but strong **dependence**.



Finding covariance with smooth transformations

Illustration: two variables with no **correlation** but strong **dependence**.

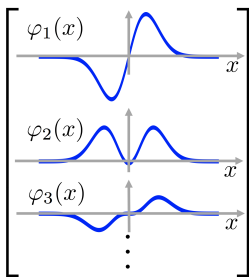


Define two spaces, one for each witness

Function in \mathcal{F}

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

Feature map

$$\varphi(x) = \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \\ \vdots \end{bmatrix}$$
The diagram shows three vertically stacked plots of functions $\varphi_1(x)$, $\varphi_2(x)$, and $\varphi_3(x)$ against x . $\varphi_1(x)$ is a blue curve that is mostly flat with a small negative dip and a positive peak. $\varphi_2(x)$ is a blue curve with two positive peaks. $\varphi_3(x)$ is a blue curve with a negative dip and a positive peak. Vertical arrows point from the labels to their respective curves.

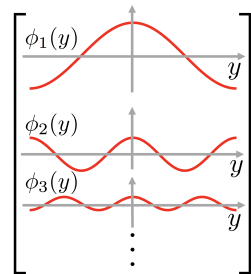
Kernel for RKHS \mathcal{F} on \mathcal{X} :

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Function in \mathcal{G}

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

Feature map

$$\phi(y) = \begin{bmatrix} \phi_1(y) \\ \phi_2(y) \\ \phi_3(y) \\ \vdots \end{bmatrix}$$
The diagram shows three vertically stacked plots of functions $\phi_1(y)$, $\phi_2(y)$, and $\phi_3(y)$ against y . $\phi_1(y)$ is a red curve with a large positive peak. $\phi_2(y)$ is a red curve with two positive peaks. $\phi_3(y)$ is a red curve with a positive peak and a negative dip. Vertical arrows point from the labels to their respective curves.

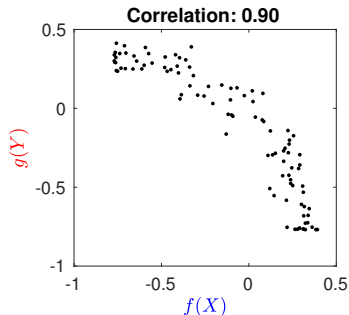
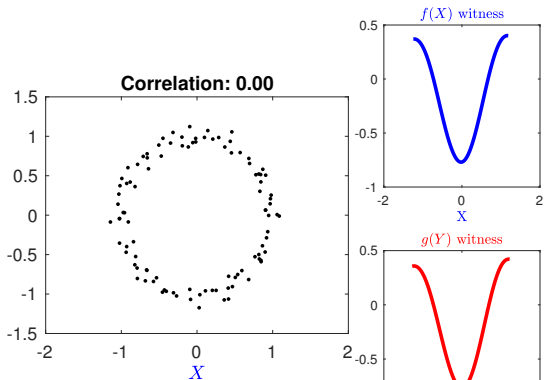
Kernel for RKHS \mathcal{G} on \mathcal{Y} :

$$l(y, y') = \langle \phi(y), \phi(y') \rangle_{\mathcal{G}}$$

The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \text{cov}[f(x)g(y)]$$



The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \text{cov} \left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(x) \right) \left(\sum_{j=1}^{\infty} g_j \phi_j(y) \right) \right]$$

The constrained covariance

The constrained covariance is

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

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

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

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

Rewriting:

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

The constrained covariance

The constrained covariance is

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

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

Rewriting:

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

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 \rightarrow 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 \rightarrow 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 μ_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 $L : \mathcal{G} \rightarrow \mathcal{F}$ 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 $\text{HS}(\mathcal{G}, \mathcal{F})$

- \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 $\text{HS}(\mathcal{G}, \mathcal{F})$

- \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 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 $a b^\top$**)

$$(a \otimes b)g \mapsto \langle g, b \rangle_{\mathcal{G}} a$$

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

$$\begin{aligned}\|a \otimes b\|_{\text{HS}}^2 &= \sum_{j \in J} \|(a \otimes b)g_j\|_{\mathcal{F}}^2 \\ &= \sum_{j \in J} \|a \langle b, g_j \rangle_{\mathcal{G}}\|_{\mathcal{F}}^2 \\ &= \|a\|_{\mathcal{F}}^2 \sum_{j \in J} |\langle b, g_j \rangle_{\mathcal{G}}|^2 \\ &= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2\end{aligned}$$

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 $a b^\top$**)

$$(a \otimes b)g \mapsto \langle g, b \rangle_{\mathcal{G}} a$$

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

$$\begin{aligned}\|a \otimes b\|_{\text{HS}}^2 &= \sum_{j \in J} \|(a \otimes b)g_j\|_{\mathcal{F}}^2 \\ &= \sum_{j \in J} \|a \langle b, g_j \rangle_{\mathcal{G}}\|_{\mathcal{F}}^2 \\ &= \|a\|_{\mathcal{F}}^2 \sum_{j \in J} |\langle b, g_j \rangle_{\mathcal{G}}|^2 \\ &= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2\end{aligned}$$

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 $a b^\top$)

$$(a \otimes b)g \mapsto \langle g, b \rangle_{\mathcal{G}} a$$

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

$$\begin{aligned}\|a \otimes b\|_{\text{HS}}^2 &= \sum_{j \in J} \|(a \otimes b)g_j\|_{\mathcal{F}}^2 \\ &= \sum_{j \in J} \left\| a \langle b, g_j \rangle_{\mathcal{G}} \right\|_{\mathcal{F}}^2 \\ &= \|a\|_{\mathcal{F}}^2 \sum_{j \in J} |\langle b, g_j \rangle_{\mathcal{G}}|^2 \\ &= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2\end{aligned}$$

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 $a b^\top$**)

$$(a \otimes b)g \mapsto \langle g, b \rangle_{\mathcal{G}} a$$

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

$$\begin{aligned}\|a \otimes b\|_{\text{HS}}^2 &= \sum_{j \in J} \|(a \otimes b)g_j\|_{\mathcal{F}}^2 \\ &= \sum_{j \in J} \left\| a \langle b, g_j \rangle_{\mathcal{G}} \right\|_{\mathcal{F}}^2 \\ &= \|a\|_{\mathcal{F}}^2 \sum_{j \in J} |\langle b, g_j \rangle_{\mathcal{G}}|^2 \\ &= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2\end{aligned}$$

where we use Parseval's identity. Thus, **the operator is Hilbert-Schmidt**.

Covariance operator in RKHS

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

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

and in particular,

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

Proof: Use Riesz representer theorem. The operator

$$\begin{aligned} C_{\varphi(x)\phi(y)} : \text{HS}(\mathcal{G}, \mathcal{F}) &\rightarrow \mathfrak{R} \\ A &\mapsto E_{xy} \langle \varphi(x) \otimes \phi(y), A \rangle_{\text{HS}} \end{aligned}$$

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

Covariance operator in RKHS

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

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

and in particular,

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

Proof: Use **Riesz** representer theorem. The operator

$$\begin{aligned} C_{\varphi(x)\phi(y)} : \text{HS}(\mathcal{G}, \mathcal{F}) &\rightarrow \mathfrak{R} \\ A &\mapsto E_{xy} \langle \phi(x) \otimes \psi(y), A \rangle_{\text{HS}} \end{aligned}$$

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

Covariance operator in RKHS

Proof (continued): Condition comes from

$$\begin{aligned} |E_{xy} \langle \varphi(x) \otimes \phi(y), A \rangle_{\text{HS}}| &\leq E_{xy} |\langle \varphi(x) \otimes \phi(y), A \rangle_{\text{HS}}| \\ &\leq \|A\|_{\text{HS}} E_{xy} (\|\varphi(x) \otimes \phi(y)\|_{\text{HS}}) \end{aligned}$$

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

Simpler condition:

$$\begin{aligned} E_{xy} (\|\varphi(x) \otimes \phi(y)\|_{\text{HS}}) &= E_{xy} (\|\varphi(x)\|_{\mathcal{F}} \|\phi(y)\|_{\mathcal{G}}) \\ &= E_{xy} \left(\sqrt{k(x, x)l(y, y)} \right) < \infty. \end{aligned}$$

Covariance operator in RKHS

Proof (continued): Condition comes from

$$\begin{aligned} |E_{xy} \langle \varphi(x) \otimes \phi(y), A \rangle_{\text{HS}}| &\leq E_{xy} |\langle \varphi(x) \otimes \phi(y), A \rangle_{\text{HS}}| \\ &\leq \|A\|_{\text{HS}} E_{xy} (\|\varphi(x) \otimes \phi(y)\|_{\text{HS}}) \end{aligned}$$

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

Simpler condition:

$$\begin{aligned} E_{xy} (\|\varphi(x) \otimes \phi(y)\|_{\text{HS}}) &= E_{xy} (\|\varphi(x)\|_{\mathcal{F}} \|\phi(y)\|_{\mathcal{G}}) \\ &= E_{xy} \left(\sqrt{k(x, x)l(y, y)} \right) < \infty. \end{aligned}$$

Covariance operator in RKHS

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:

$$\begin{aligned} \langle f, C_{\varphi(x)\phi(y)} g \rangle_{\mathcal{F}} &= \langle C_{\varphi(x)\phi(y)}, f \otimes g \rangle_{\text{HS}} \\ &\stackrel{(a)}{=} E_{xy} \langle \varphi(x) \otimes \phi(y), f \otimes g \rangle_{\text{HS}} \\ &= E_{xy} [\langle f, \varphi(x) \rangle_{\mathcal{F}} \langle g, \phi(y) \rangle_{\mathcal{F}}] \\ &= E_{xy} [f(x)g(y)] \end{aligned}$$

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:

$$\begin{aligned} \left\langle f, C_{\varphi(x)\phi(y)} g \right\rangle_{\mathcal{F}} &= \left\langle C_{\varphi(x)\phi(y)}, f \otimes g \right\rangle_{\text{HS}} \\ &\stackrel{(a)}{=} E_{xy} \left\langle \varphi(x) \otimes \phi(y), f \otimes g \right\rangle_{\text{HS}} \\ &= E_{xy} [\langle f, \varphi(x) \rangle_{\mathcal{F}} \langle g, \phi(y) \rangle_{\mathcal{F}}] \\ &= E_{xy} [f(x)g(y)] \end{aligned}$$

(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:

$$\begin{aligned} \left\langle f, C_{\varphi(x)\phi(y)} g \right\rangle_{\mathcal{F}} &= \left\langle C_{\varphi(x)\phi(y)}, f \otimes g \right\rangle_{\text{HS}} \\ &\stackrel{(a)}{=} E_{xy} \left\langle \varphi(x) \otimes \phi(y), f \otimes g \right\rangle_{\text{HS}} \\ &= E_{xy} [\langle f, \varphi(x) \rangle_{\mathcal{F}} \langle g, \phi(y) \rangle_{\mathcal{F}}] \\ &= E_{xy} [f(x)g(y)] \end{aligned}$$

(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:

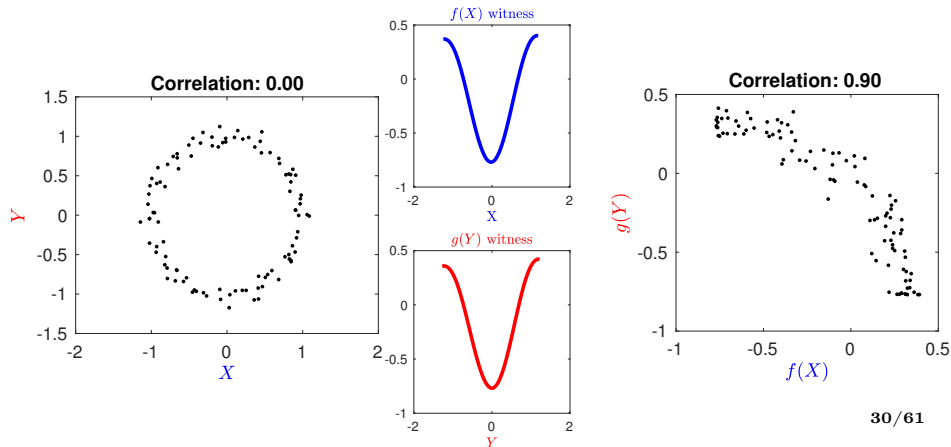
$$\begin{aligned} \left\langle f, C_{\varphi(x)\phi(y)} g \right\rangle_{\mathcal{F}} &= \left\langle C_{\varphi(x)\phi(y)}, f \otimes g \right\rangle_{\text{HS}} \\ &\stackrel{(a)}{=} E_{xy} \left\langle \varphi(x) \otimes \phi(y), f \otimes g \right\rangle_{\text{HS}} \\ &= E_{xy} [\langle f, \varphi(x) \rangle_{\mathcal{F}} \langle g, \phi(y) \rangle_{\mathcal{F}}] \\ &= E_{xy} [f(x)g(y)] \end{aligned}$$

(a) by definition of the covariance operator

Back to the constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \text{cov}[f(x)g(y)]$$



Computing COCO from finite data

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

Computing COCO from finite data

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

$\widehat{\text{COCO}}$ is largest eigenvalue γ_{\max} of

$$\begin{bmatrix} 0 & \frac{1}{n} \widetilde{K} \widetilde{L} \\ \frac{1}{n} \widetilde{L} \widetilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

$$\widetilde{K}_{ij} = \langle \varphi(x_i) - \hat{\mu}_x, \varphi(x_j) - \hat{\mu}_x \rangle_{\mathcal{F}} =: \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{F}}$$

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis,
AISTATS'05

Computing COCO from finite data

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

$\widehat{\text{COCO}}$ is largest eigenvalue γ_{\max} of

$$\begin{bmatrix} 0 & \frac{1}{n} \widetilde{K} \widetilde{L} \\ \frac{1}{n} \widetilde{L} \widetilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

$$\widetilde{K}_{ij} = \langle \varphi(x_i) - \hat{\mu}_x, \varphi(x_j) - \hat{\mu}_x \rangle_{\mathcal{F}} =: \langle \check{\varphi}(x_i), \check{\varphi}(x_j) \rangle_{\mathcal{F}}$$

Witness functions:

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

Empirical COCO: proof

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^n \underbrace{\left[\left(f(x_i) - \frac{1}{n} \sum_{j=1}^n f(x_j) \right) \left(g(y_i) - \frac{1}{n} \sum_{j=1}^n g(y_j) \right) \right]}_{\text{covariance}} \\ - \underbrace{\frac{\lambda}{2} \left(\|f\|_{\mathcal{F}}^2 - 1 \right) - \frac{\gamma}{2} \left(\|g\|_{\mathcal{G}}^2 - 1 \right)}_{\text{smoothness constraints}}$$

with Lagrange multipliers $\lambda \geq 0$ and $\gamma \geq 0$.

Empirical COCO: proof

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \frac{1}{n} \sum_{i=1}^n \underbrace{\left[\left(f(x_i) - \frac{1}{n} \sum_{j=1}^n f(x_j) \right) \left(g(y_i) - \frac{1}{n} \sum_{j=1}^n g(y_j) \right) \right]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2} \left(\|f\|_{\mathcal{F}}^2 - 1 \right) - \frac{\gamma}{2} \left(\|g\|_{\mathcal{G}}^2 - 1 \right)}_{\text{smoothness constraints}}$$

with Lagrange multipliers $\lambda \geq 0$ and $\gamma \geq 0$.

Assume:

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

for centered $\tilde{\varphi}(x_i)$, $\tilde{\psi}(y_i)$.

Proof (continued)

First step is **smoothness constraint**:

$$\begin{aligned}\|f\|_{\mathcal{F}}^2 - 1 &= \left\langle \sum_{i=1}^n \alpha_i \tilde{\varphi}(x_i), \sum_{i=1}^n \alpha_i \tilde{\varphi}(x_i) \right\rangle_{\mathcal{F}} - 1 \\ &= \alpha^\top \tilde{K} \alpha - 1\end{aligned}$$

Proof (continued)

First step is **smoothness constraint**:

$$\begin{aligned}\|f\|_{\mathcal{F}}^2 - 1 &= \left\langle \sum_{i=1}^n \alpha_i \tilde{\varphi}(x_i), \sum_{i=1}^n \alpha_i \tilde{\varphi}(x_i) \right\rangle_{\mathcal{F}} - 1 \\ &= \alpha^\top \tilde{K} \alpha - 1\end{aligned}$$

Proof (continued)

Second step is covariance:

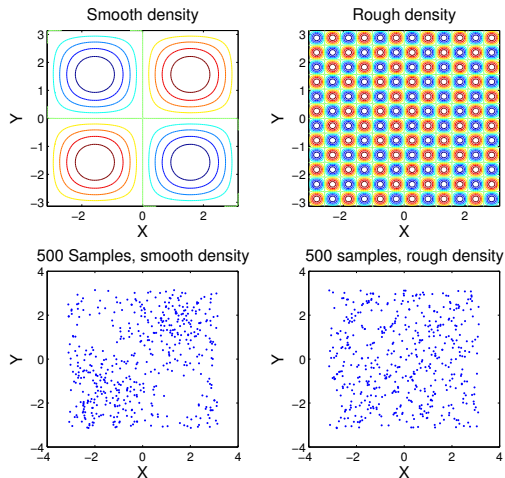
$$\begin{aligned} &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{f}, \tilde{\varphi}(x_i) \rangle_{\mathcal{F}} \langle \mathbf{g}, \tilde{\phi}(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \sum_{i=1}^n \left\langle \underbrace{\sum_{\ell=1}^n \alpha_{\ell} \tilde{\varphi}(x_{\ell})}_{\mathbf{f}}, \tilde{\varphi}(x_i) \right\rangle_{\mathcal{F}} \langle \mathbf{g}, \tilde{\phi}(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \boldsymbol{\alpha}^{\top} \tilde{\mathbf{K}} \tilde{\mathbf{L}} \boldsymbol{\beta} \end{aligned}$$

Proof (continued)

Second step is covariance:

$$\begin{aligned} &= \frac{1}{n} \sum_{i=1}^n \langle \mathbf{f}, \tilde{\varphi}(x_i) \rangle_{\mathcal{F}} \langle \mathbf{g}, \tilde{\phi}(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \sum_{i=1}^n \left\langle \underbrace{\sum_{\ell=1}^n \alpha_{\ell} \tilde{\varphi}(x_{\ell})}_{\mathbf{f}}, \tilde{\varphi}(x_i) \right\rangle_{\mathcal{F}} \langle \mathbf{g}, \tilde{\phi}(y_i) \rangle_{\mathcal{G}} \\ &= \frac{1}{n} \boldsymbol{\alpha}^{\top} \widetilde{\mathbf{K}} \widetilde{\mathbf{L}} \boldsymbol{\beta} \end{aligned}$$

What is a large dependence with COCO?



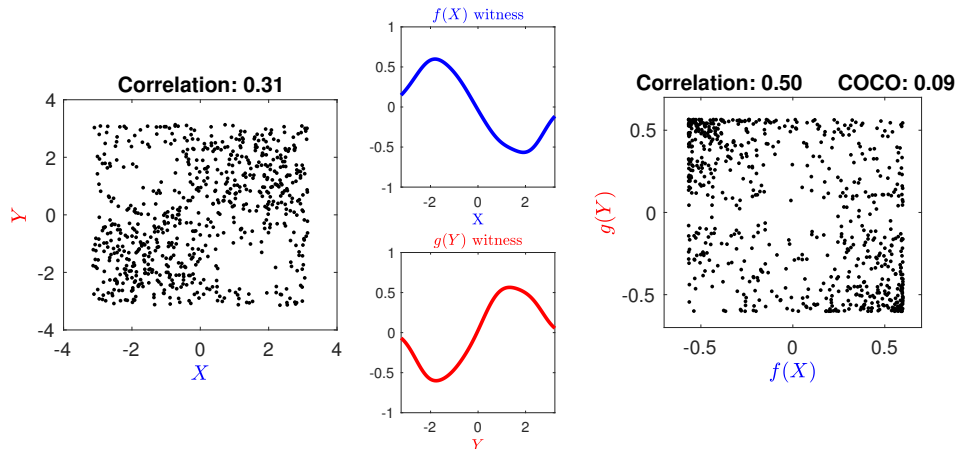
Density takes the form:

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

Which of these is the more "dependent"?

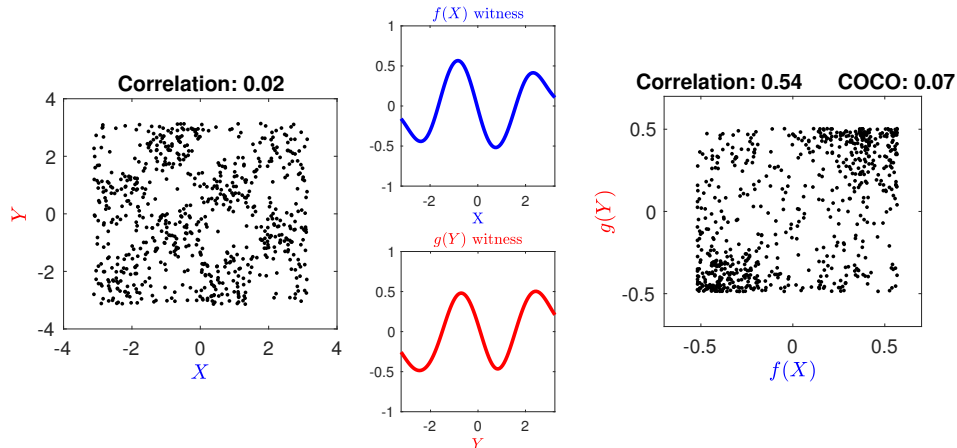
Finding covariance with smooth transformations

Case of $\omega = 1$:



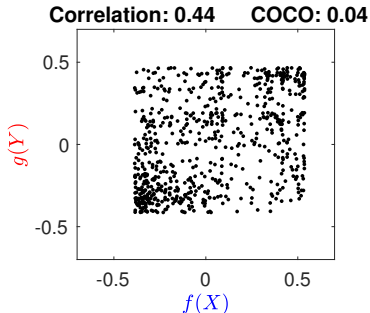
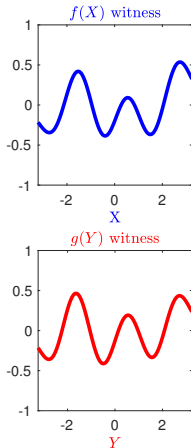
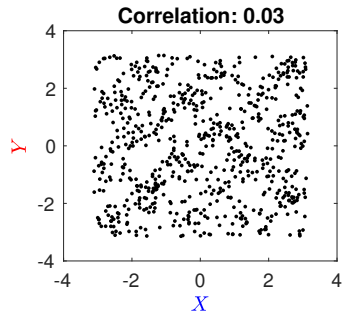
Finding covariance with smooth transformations

Case of $\omega = 2$:



Finding covariance with smooth transformations

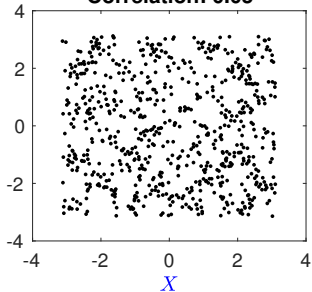
Case of $\omega = 3$:



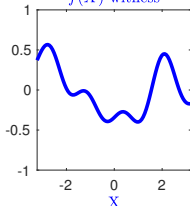
Finding covariance with smooth transformations

Case of $\omega = 4$:

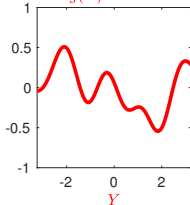
Correlation: 0.05



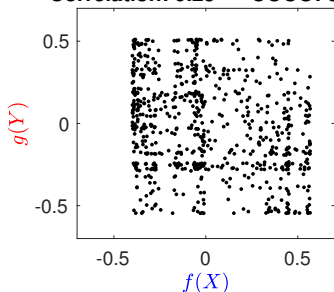
$f(X)$ witness



$g(Y)$ witness



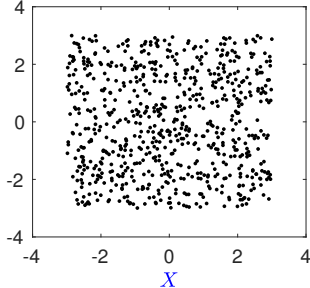
Correlation: 0.25 COCO: 0.02



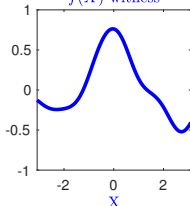
Finding covariance with smooth transformations

Case of $\omega = ??$:

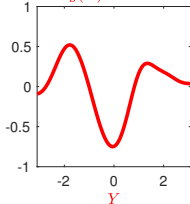
Correlation: 0.01



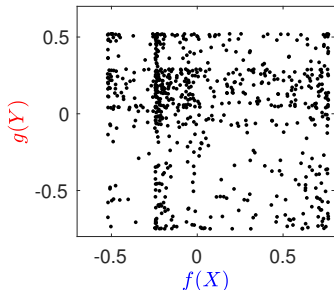
$f(X)$ witness



$g(Y)$ witness

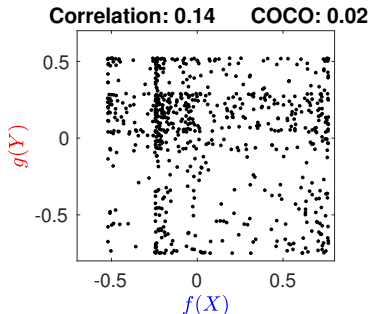
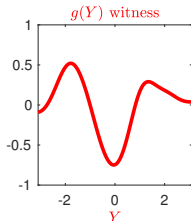
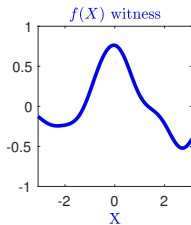
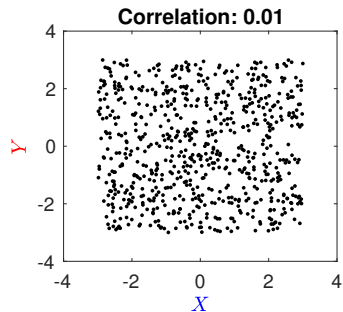


Correlation: 0.14 COCO: 0.02



Finding covariance with smooth transformations

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



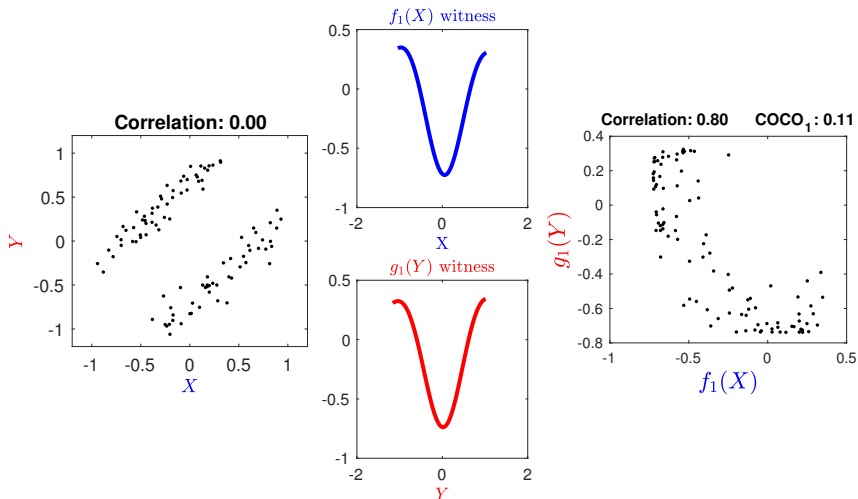
Dependence largest when at “low” frequencies

- 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.

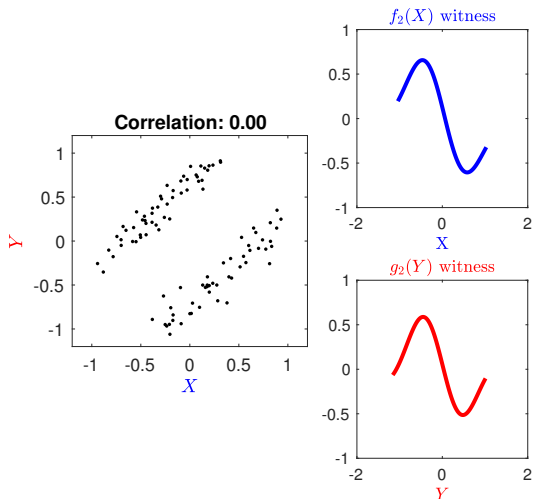
First singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



Can we do better than COCO?

A second example with zero correlation.

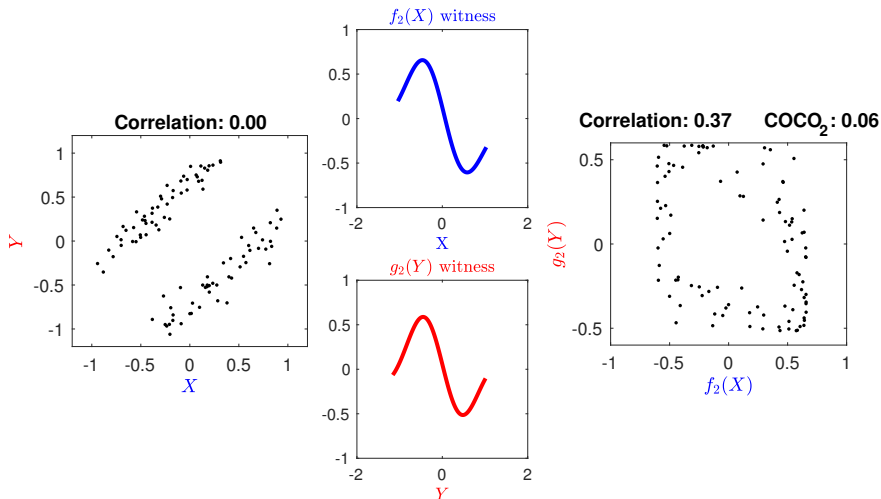
Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



Can we do better than COCO?

A second example with zero correlation.

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



The Hilbert-Schmidt Independence Criterion

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

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

define **Hilbert-Schmidt Independence Criterion (HSIC)**

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

G, Bousquet, Smola, and Schoelkopf, ALT05; G, Fukumizu, Teo, Song, Schoelkopf, and Smola, NIPS 2007,.

The Hilbert-Schmidt Independence Criterion

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

$$\gamma_i := \text{CO}C O_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define **Hilbert-Schmidt Independence Criterion (HSIC)**

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

G, Bousquet, Smola., and Schoelkopf, ALT05; G., Fukumizu, Teo., Song., Schoelkopf., and Smola, NIPS 2007,.

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 $\kappa((x, y), (x', y')) = k(x, x')l(y, y')$.

Asymptotics of HSIC under independence

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

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ 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_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- **Asymptotics** of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

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

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

Asymptotics of HSIC under independence

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

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ 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_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- **Asymptotics** of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

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

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

Asymptotics of HSIC under independence

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

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ 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_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- **Asymptotics** of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

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

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

Asymptotics of HSIC under independence

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

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ 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_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α ?
- **Asymptotics** of \widehat{HSIC} when $P_{XY} = P_X P_Y$:

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

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

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (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 π of indices $\{1, \dots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (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 π of indices $\{1, \dots, n\}$. This gives HSIC for independent variables.
- Repeat for many different permutations, get empirical CDF
- Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (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 π of indices $\{1, \dots, n\}$. This gives HSIC for independent variables.
 - Repeat for many different permutations, get empirical CDF
 - Threshold c_α is $1 - \alpha$ quantile of empirical CDF

Application: dependence detection across languages

Testing task: detect dependence between English and French text

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

Application: dependence detection across languages

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

k -spectrum kernel, $k = 10$, sample size $n = 10$

X

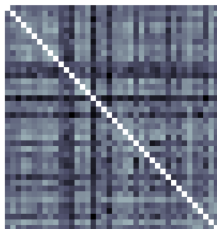
Y

Honourable senators, I
have a question for the
Leader of the Government
in the Senate

No doubt there is great
pressure on provincial and
municipal governments

In fact, we have increased
federal investments for
early childhood
development.

•
•
•



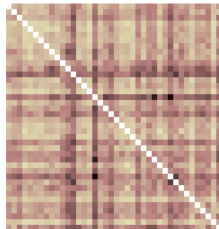
K

Honorables sénateurs, ma
question s'adresse au leader
du gouvernement au Sénat

Les ordres de gouvernements
provinciaux et municipaux
subissent de fortes pressions

Au contraire, nous avons
augmenté
le financement fédéral pour le
développement des jeunes

•
•
•



L

$$\widehat{HSIC} = \frac{1}{n^2} \text{trace}(KL)$$

(K and L column centered)

Application: Dependence detection across languages

Results (for $\alpha = 0.05$)

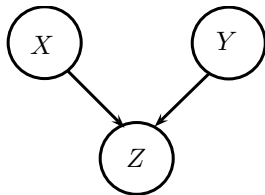
- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

Settings: Five line extracts, averaged over 300 repetitions, for “Agriculture” transcripts. Similar results for Fisheries and Immigration transcripts.

Testing higher order interactions

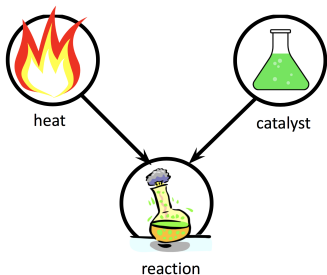
Detecting higher order interaction

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



Detecting higher order interaction

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

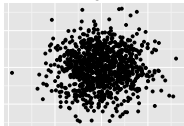


Detecting higher order interaction

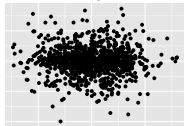
How to detect V-structures with pairwise weak individual dependence?

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

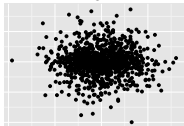
X1 vs Y1



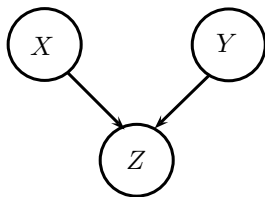
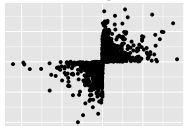
Y1 vs Z1



X1 vs Z1



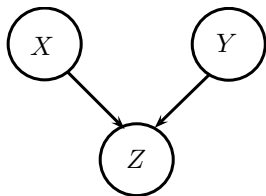
X1*Y1 vs Z1



- $X, Y \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$
- $Z | X, Y \sim \text{sign}(XY) \text{Exp}\left(\frac{1}{\sqrt{2}}\right)$

Fine print: Faithfulness violated here!

V-structure discovery



Assume $X \perp\!\!\!\perp Y$ has been established.

V-structure can then be detected by:

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

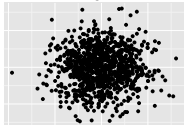
How well do these work?

Detecting higher order interaction

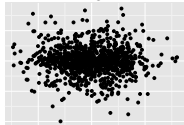
Generalise earlier example to p dimensions

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

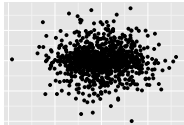
X1 vs Y1



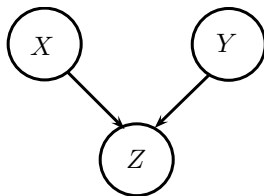
Y1 vs Z1



X1 vs Z1



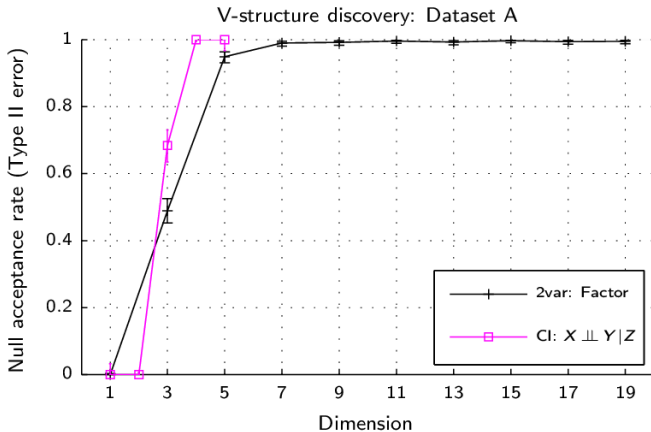
X1*Y1 vs Z1



- $X, Y \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$
- $Z | X, Y \sim \text{sign}(XY) \text{Exp}(\frac{1}{\sqrt{2}})$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$

Fine print: Faithfulness violated here!

V-structure discovery



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

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP 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 $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP 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

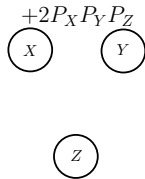
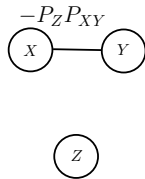
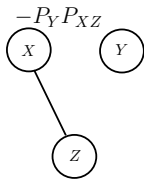
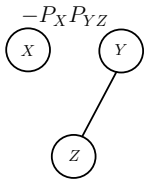
Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that **vanishes** whenever P can be factorised non-trivially.

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

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

$$\Delta_L P =$$

$$P_{XYZ}$$

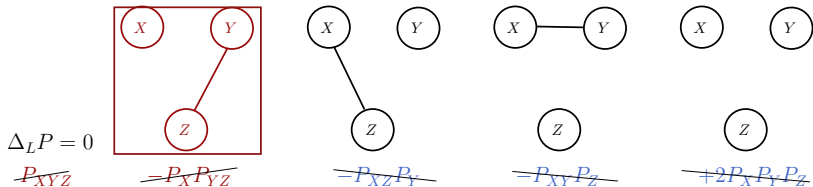


Lancaster interaction measure

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

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

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



Case of $P_X \perp\!\!\!\perp P_{YZ}$

Lancaster interaction measure

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

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

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

$$(X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X \Rightarrow \Delta_L P = 0.$$

...so what might be missed?

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP 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:

$P(0,0,0) = 0.2$	$P(0,0,1) = 0.1$	$P(1,0,0) = 0.1$	$P(1,0,1) = 0.1$
$P(0,1,0) = 0.1$	$P(0,1,1) = 0.1$	$P(1,1,0) = 0.1$	$P(1,1,1) = 0.2$

A kernel test statistic using Lancaster Measure

Construct a test by estimating $\|\mu_\kappa(\Delta_L P)\|_{\mathcal{H}_\kappa}^2$, where $\kappa = k \otimes l \otimes m$:

$$\begin{aligned} & \|\mu_\kappa(P_{XYZ} - P_{XY}P_Z - \dots)\|_{\mathcal{H}_\kappa}^2 = \\ & \langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ} \rangle_{\mathcal{H}_\kappa} - 2 \langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY}P_Z \rangle_{\mathcal{H}_\kappa} \dots \end{aligned}$$

A kernel test statistic using Lancaster Measure

$\nu \setminus \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(K \circ L \circ M)_{++}$	$((K \circ L)M)_{++}$	$((K \circ M)L)_{++}$	$((M \circ L)K)_{++}$	$\text{tr}(K_{++} \circ L_{++} \circ M_{++})$
$P_{XY}P_Z$		$(K \circ L)_{++} M_{++}$	$(MKL)_{++}$	$(KLM)_{++}$	$(KL)_{++} M_{++}$
$P_{XZ}P_Y$			$(K \circ M)_{++} L_{++}$	$(KML)_{++}$	$(KM)_{++} L_{++}$
$P_{YZ}P_X$				$(L \circ M)_{++} K_{++}$	$(LM)_{++} K_{++}$
$P_XP_YP_Z$					$K_{++} L_{++} M_{++}$

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

Lancaster interaction statistic: Sejdinovic, G, Bergsma, NIPS13

$$\|\mu_\kappa(\Delta_L P)\|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} \left[(H K H \circ H L H \circ H M H)_{++} \right].$$

A kernel test statistic using Lancaster Measure

$\nu \setminus \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_X P_Y P_Z$
P_{XYZ}	$(K \circ L \circ M)_{++}$	$((K \circ L)M)_{++}$	$((K \circ M)L)_{++}$	$((M \circ L)K)_{++}$	$\text{tr}(K_{++} \circ L_{++} \circ M_{++})$
$P_{XY}P_Z$		$(K \circ L)_{++} M_{++}$	$(MKL)_{++}$	$(KLM)_{++}$	$(KL)_{++} M_{++}$
$P_{XZ}P_Y$			$(K \circ M)_{++} L_{++}$	$(KML)_{++}$	$(KM)_{++} L_{++}$
$P_{YZ}P_X$				$(L \circ M)_{++} K_{++}$	$(LM)_{++} K_{++}$
$P_X P_Y P_Z$					$K_{++} L_{++} M_{++}$

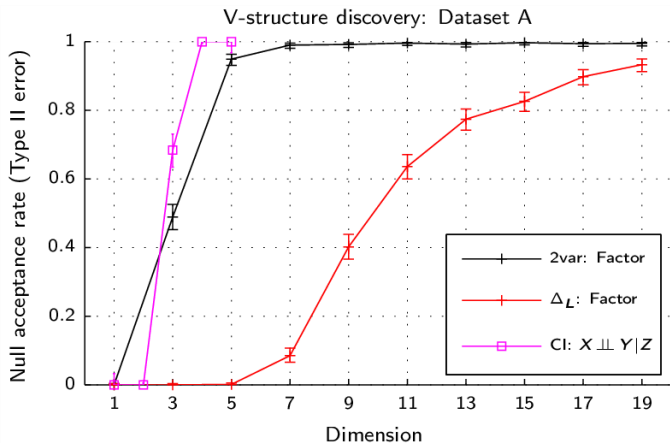
Table: V -statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_X P_Y P_Z$). H is centering matrix $I - n^{-1}$

Lancaster interaction statistic: Sejdinovic, G, Bergsma, NIPS13

$$\|\mu_\kappa(\Delta_L P)\|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} \boxed{(H \mathbf{K} H \circ H \mathbf{L} H \circ H \mathbf{M} H)_{++}}$$

Empirical joint central moment in the feature space

V-structure discovery



Lancaster test, CI test for $X \perp\!\!\!\perp Y|Z$ 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 π , J_{π} associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

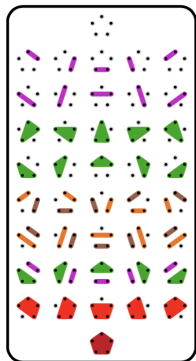
Interaction for $D \geq 4$

- Interaction measure valid for all D :

(Streitberg, 1990)

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

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



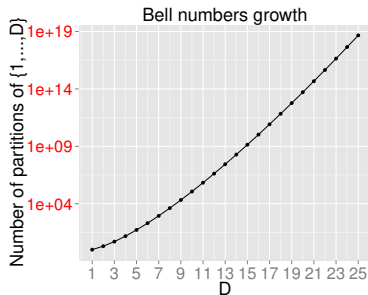
Interaction for $D \geq 4$

■ Interaction measure valid for all D :

(Streitberg, 1990)

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

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



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?

