RKHS in ML: Testing Statistical Dependence

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

November 20, 2024

Testing statistical dependence

Dependence testing

- Given: Samples from a distribution P_{XY}
- 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.

Text from dogtime.com and petfinder.com

Dependence detection, discrete domain

- How do you detect dependence...
- ... in a discrete domain?

Dependence detection, discrete domain

- How do you detect dependence. . .
- . . . in a discrete domain?

P(A,T)	On time	Late
Alarm	0.27	0.03
No alarm	0.07	0.63

Dependence detection, discrete domain

- How do you detect dependence...
- ... in a discrete domain?

P(A,T)	On time	Late
Alarm	0.10	0.20
No alarm	0.24	0.46

Dependence detection, discrete domain

- How do you detect dependence...
- ... in a discrete domain?

X_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

X_2: No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

...

$\overset{?}{\Longleftrightarrow}$

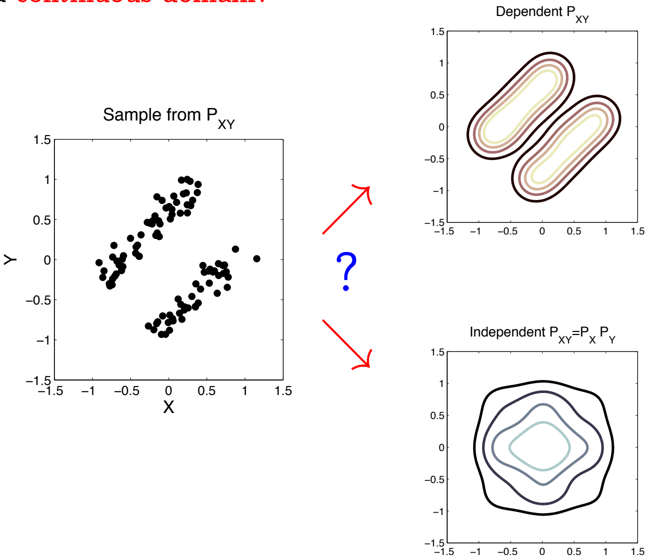
Y_1: Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat et concerne l'aide financiére qu'on a annoncée pour les agriculteurs. La plupart des agriculteurs n'ont encore rien reçu de cet argent.

Y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

...

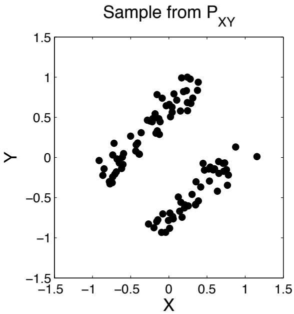
Dependence detection, continuous domain

- How do you detect dependence...
- ... in a continuous domain?

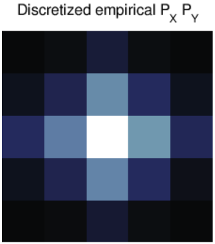


Dependence detection, continuous domain

- How do you detect dependence...
- ... in a continuous domain?

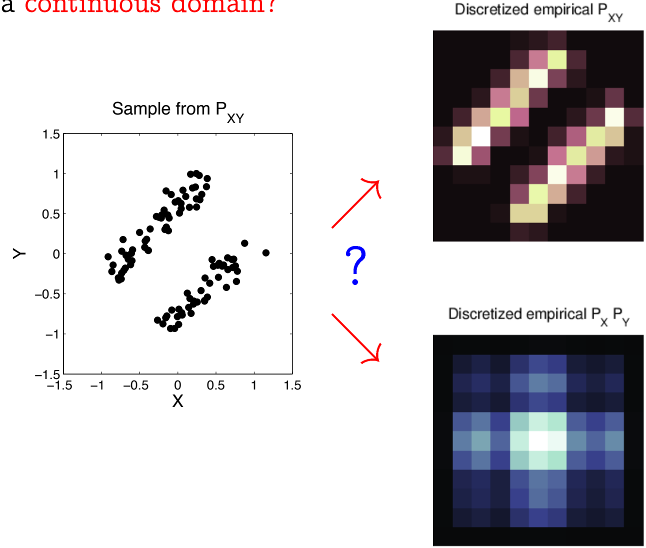


?



Dependence detection, continuous domain

- How do you detect dependence...
- ... in a continuous domain?



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 $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$
  - Solution: simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS $\mathcal{H}_\kappa$?

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 $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$
  - Solution: simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS $\mathcal{H}_\kappa$?

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 $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$
  - Solution: simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS $\mathcal{H}_\kappa$?

MMD as a dependence measure

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

$k\Big($  ,  $\Big)$

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

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

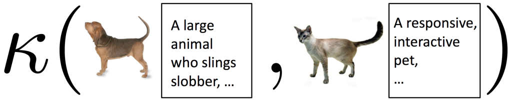
MMD as a dependence measure

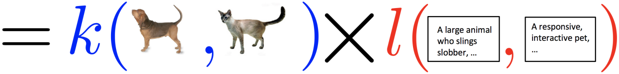
Kernel k on images with feature space $\mathcal{F}$,

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

$l($ A large animal who slings slobber, … $,$ A responsive, interactive pet … $)$

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





MMD as a dependence measure

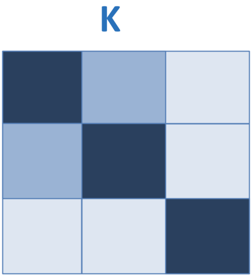
Given: Samples from a distribution P_{XY}

Goal: Are X and Y independent?

$$MMD^2(\widehat{P}_{XY}, \widehat{P}_X \widehat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2}\mathrm{trace}(K H L H)$$

$(H = I_n - \frac{1}{n}1_n 1_n^\top)$

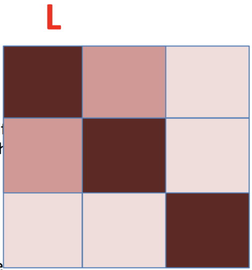
MMD as a dependence measure



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

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

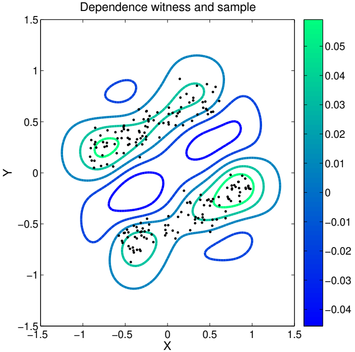
A responsive, interactive pet, one
that will blow in your ear and
follow you everywhere.



Text from dogtime.com and petfinder.com

MMD as a dependence measure

MMD witness, product kernel: $\operatorname*{argmax}_{\|f\|\leq 1} E_{P_{XY}} f - E_{P_X P_Y} f$

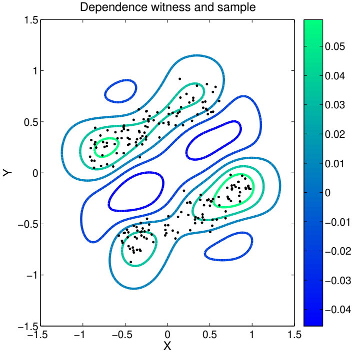


Two questions:

- Why the product kernel? Why not eg a sum?
- Is there a more interpretable definition of the dependence measure?

MMD as a dependence measure

MMD witness, product kernel: $\underset{\|f\|\leq 1}{\operatorname{argmax}} E_{P_{XY}} f - E_{P_X P_Y} f$



Two questions:

- Why the product kernel? Why not eg a sum?
- Is there a more interpretable definition of the dependence measure?

Illustration: dependence $\neq$correlation

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

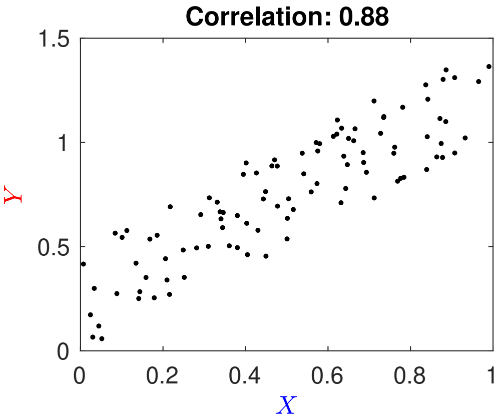


Illustration: dependence $\neq$correlation

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

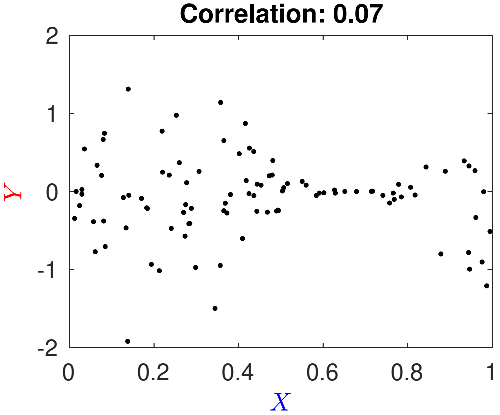
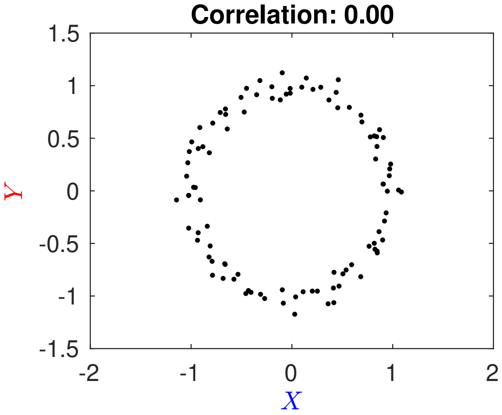


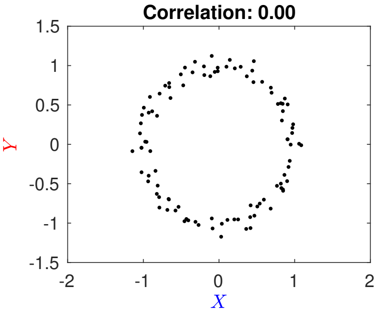
Illustration: dependence $\neq$correlation

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



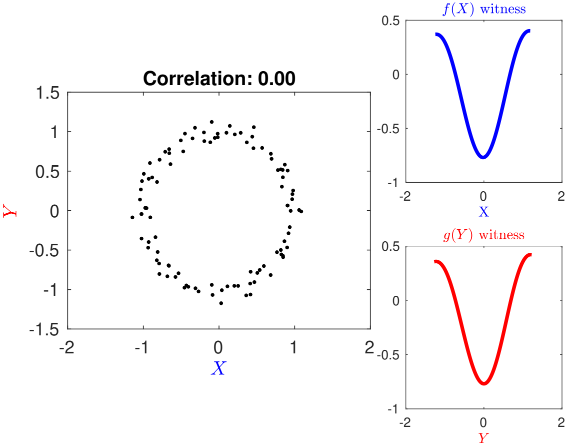
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



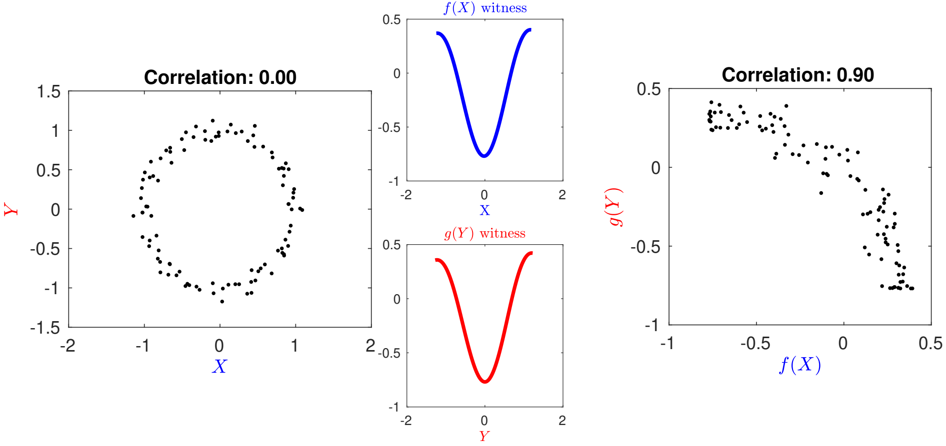
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



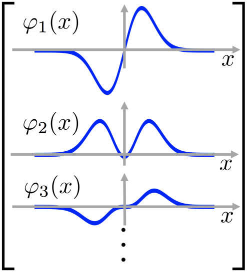
Define two spaces, one for each witness

Function in $\mathcal{F}$

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

Feature map

$\varphi(x) =$



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

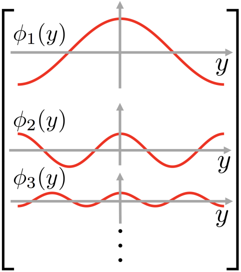
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Function in $\mathcal{G}$

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

Feature map

$\phi(y) =$



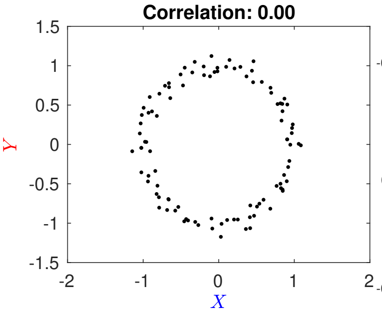
Kernel for RKHS $\mathcal{G}$ on $\mathcal{Y}$:

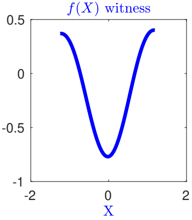
$$l(y, y') = \langle \phi(y), \phi(y') \rangle_{\mathcal{G}}$$

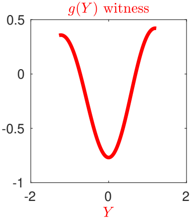
The constrained covariance

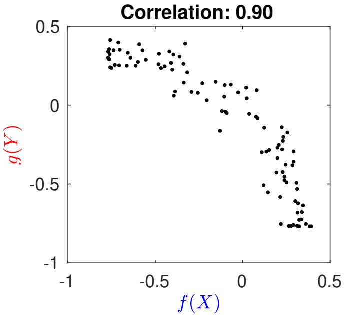
The constrained covariance is

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









The constrained covariance

The constrained covariance is

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

The constrained covariance

The constrained covariance is

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

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

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

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

Rewriting:

$$E_{xy}[f(x)g(y)] - E_x[f(x)]E_y[g(y)]$$

$$= \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top \underbrace{\mathrm{E}_{xy}\left(\begin{bmatrix} \tilde{\varphi}_1(x) \\ \tilde{\varphi}_2(x) \\ \vdots \end{bmatrix}\begin{bmatrix} \tilde{\phi}_1(y) & \tilde{\phi}_2(y) & \ldots \end{bmatrix}\right)}_{C_{\tilde{\varphi}(x)\tilde{\phi}(y)}}\begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}$$

The constrained covariance

The constrained covariance is

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

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

Rewriting:

$$E_{xy}[f(x)g(y)] - E_x[f(x)]E_y[g(y)]$$

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

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

Does feature space covariance exist?

How do we prove existence of feature covariance $C_{\varphi(x)\phi(y)}$?

What is COCO in the finite linear case? Two zero mean random vectors $x \in \Re^d$, $y \in \Re^{d'}$.

Compute their covariance matrix:

$$C_{xy} = E_{xy}\left(xy^\top\right)$$

...which is a $d \times d'$ matrix! How to get a single "summary" number?

Solve for vectors $f \in \Re^d$, $g \in \Re^{d'}$

$$\begin{aligned}\operatorname*{argmax}_{\|f\|=1,\|g\|=1} f^\top C_{xy} g &= \operatorname*{argmax}_{\|f\|=1,\|g\|=1} E_{xy}\left[\left(f^\top x\right)\left(y^\top g\right)\right] \\ &= \operatorname*{argmax}_{\|f\|=1,\|g\|=1} E_{xy}[f(x)g(y)]\end{aligned}$$

Maximum singular value of C_{xy}.

Does feature space covariance exist?

How do we prove existence of feature covariance $C_{\varphi(x)\phi(y)}$?

What is COCO in the finite linear case? Two zero mean random vectors $x \in \Re^d$, $y \in \Re^{d'}$.

Compute their covariance matrix:

$$C_{xy} = E_{xy}\left(xy^\top\right)$$

...which is a $d \times d'$ matrix! How to get a single "summary" number?

Solve for vectors $f \in \Re^d$, $g \in \Re^{d'}$

$$\begin{aligned}\operatorname*{argmax}_{\|f\|=1,\|g\|=1} f^\top C_{xy} g &= \operatorname*{argmax}_{\|f\|=1,\|g\|=1} E_{xy}\left[\left(f^\top x\right)\left(y^\top g\right)\right] \\ &= \operatorname*{argmax}_{\|f\|=1,\|g\|=1} E_{xy}[f(x)g(y)]\end{aligned}$$

Maximum singular value of C_{xy}.

Does feature space covariance exist?

How do we prove existence of feature covariance $C_{\varphi(x)\phi(y)}$?

What is COCO in the finite linear case? Two zero mean random vectors $x \in \Re^d$, $y \in \Re^{d'}$.

Compute their covariance matrix:

$$C_{xy} = E_{xy}\left(xy^\top\right)$$

...which is a $d \times d'$ matrix! How to get a single "summary" number?

Solve for vectors $f \in \Re^d$, $g \in \Re^{d'}$

$$\operatorname*{argmax}_{\|f\|=1,\|g\|=1} f^\top C_{xy} g = \operatorname*{argmax}_{\|f\|=1,\|g\|=1} E_{xy}\left[\left(f^\top x\right)\left(y^\top g\right)\right]$$
$$= \operatorname*{argmax}_{\|f\|=1,\|g\|=1} E_{xy}[f(x)g(y)]$$

Maximum singular value of C_{xy}.

Does feature space covariance exist?

How do we prove existence of feature covariance $C_{\varphi(x)\phi(y)}$?

What is COCO in the finite linear case? Two zero mean random vectors $x \in \Re^d$, $y \in \Re^{d'}$.

Compute their covariance matrix:

$$C_{xy} = E_{xy}\left(xy^\top\right)$$

...which is a $d \times d'$ matrix! How to get a single "summary" number?

Solve for vectors $f \in \Re^d$, $g \in \Re^{d'}$

$$\begin{aligned}\underset{\|f\|=1,\|g\|=1}{\operatorname{argmax}} f^\top C_{xy} g &= \underset{\|f\|=1,\|g\|=1}{\operatorname{argmax}} E_{xy}\left[\left(f^\top x\right)\left(y^\top g\right)\right] \\ &= \underset{\|f\|=1,\|g\|=1}{\operatorname{argmax}} E_{xy}[f(x)g(y)]\end{aligned}$$

Maximum singular value of C_{xy}.

Does feature space covariance exist?

Given features $\varphi(x) \in \mathcal{F}$, $\phi(y) \in \mathcal{G}$

Challenge 1: Can we define a feature space analog to $x\, y^\top$?

YES:

- Given $f \in \Re^d$, $g \in \Re^{d'}$, $h \in \Re^{d'}$, define matrix $f\, g^\top$ such that $\left(f\, g^\top\right) h = f\left(g^\top h\right)$.
- Given $f \in \mathcal{F}$, $g \in \mathcal{G}$, $h \in \mathcal{G}$, define tensor product operator $f \otimes g$ such that $[f \otimes g]\, h = f \langle g, h\rangle_{\mathcal{G}}$.
- Now just set $f := \varphi(x)$, $g = \phi(y)$, to get $x\, y^\top \to \varphi(x) \otimes \phi(y)$

Does feature space covariance exist?

Given features $\varphi(x) \in \mathcal{F}$, $\phi(y) \in \mathcal{G}$

Challenge 2: Does an uncentered covariance "matrix" (operator) in feature space exist? I.e. is there some $C_{xy} : \mathcal{G} \to \mathcal{F}$ such that

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

Does "something" exist $\to$ Riesz theorem.

Does feature space covariance exist?

Given features $\varphi(x) \in \mathcal{F}$, $\phi(y) \in \mathcal{G}$

Challenge 2: Does an uncentered covariance "matrix" (operator) in feature space exist? I.e. is there some $C_{xy} : \mathcal{G} \to \mathcal{F}$ such that

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

Does "something" exist $\to$ 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.

Does feature space covariance exist?

Hints:

- In the finite dimensional case, and given basis vectors $g_j \in \Re^{d'}$ $C_{xy} \in \Re^{d \times d'}$ is in a vector space, with inner product

$$\langle C_{xy}, A \rangle_{\mathrm{HS}} = \mathrm{trace}(C_{xy}{}^\top A) = \sum_{j \in J} (C_{xy} g_j)^\top (A g_j),$$

- In particular,

$$E_{xy}\left[f(x)g(y)\right] = f^\top C_{xy} g = \mathrm{trace}(g^\top C_{xy}^\top f)$$
$$= \mathrm{trace}(C_{xy}^\top (f\ g^\top)) = \langle C_{xy}, f\ g^\top \rangle_{\mathrm{HS}}$$

Challenge 2 (reformulated via the hints): does there exist $C_{xy} : \mathcal{G} \to \mathcal{F}$ in a Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ such that:

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

and in particular,

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

Does feature space covariance exist?

Hints:

- In the finite dimensional case, and given basis vectors $g_j \in \Re^{d'}$ $C_{xy} \in \Re^{d \times d'}$ is in a vector space, with inner product

$$\langle C_{xy}, A \rangle_{\mathrm{HS}} = \mathrm{trace}(C_{xy}{}^{\top} A) = \sum_{j \in J} (C_{xy} g_j)^{\top} (A g_j),$$

- In particular,

$$\begin{aligned} E_{xy} [f(x) g(y)] &= f^{\top} C_{xy} g = \mathrm{trace}(g^{\top} C_{xy}^{\top} f) \\ &= \mathrm{trace}(C_{xy}^{\top} (f \; g^{\top})) = \langle C_{xy}, f \; g^{\top} \rangle_{\mathrm{HS}} \end{aligned}$$

Challenge 2 (reformulated via the hints): does there exist $C_{xy} \; : \; \mathcal{G} \to \mathcal{F}$ in a Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ such that:

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

and in particular,

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

The Hilbert Space 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 HS($\mathcal{G}, \mathcal{F}$)

$$\langle L, M \rangle_{\mathrm{HS}} = \sum_{j \in J} \langle L g_j, M g_j \rangle_{\mathcal{F}}$$

(independent of orthonormal basis)

- Hilbert-Schmidt norm of the operators L:

$$\|L\|^2_{\mathrm{HS}} = \sum_{j \in J} \|L g_j\|^2_{\mathcal{F}}$$

L is Hilbert-Schmidt when this norm is finite.

The Hilbert Space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

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

$$\langle L, M \rangle_{\mathrm{HS}} = \sum_{j \in J} \langle L g_j, M g_j \rangle_{\mathcal{F}}$$

(independent of orthonormal basis)

- Hilbert-Schmidt norm of the operators L:

$$\|L\|_{\mathrm{HS}}^2 = \sum_{j \in J} \|L g_j\|_{\mathcal{F}}^2$$

L is Hilbert-Schmidt when this norm is finite.

The Hilbert Space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

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

$$\langle L, M \rangle_{\mathrm{HS}} = \sum_{j \in J} \langle L g_j, M g_j \rangle_{\mathcal{F}}$$

(independent of orthonormal basis)

- Hilbert-Schmidt norm of the operators L:

$$\|L\|_{\mathrm{HS}}^2 = \sum_{j \in J} \|L g_j\|_{\mathcal{F}}^2$$

L is Hilbert-Schmidt when this norm is finite.

The tensor product $a \otimes b$ is in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, we earlier defined 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 $a \otimes b \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$?

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

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

The tensor product $a \otimes b$ is in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, we earlier defined 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 $a \otimes b \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$?

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

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

The tensor product $a \otimes b$ is in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, we earlier defined 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 $a \otimes b \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$?

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

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

The tensor product $a \otimes b$ is in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, we earlier defined 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 $a \otimes b \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$?

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

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

Inner product of $a \otimes b$ with $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given a Hilbert-Schmidt operator $L : \mathcal{G} \to \mathcal{F}$,

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}}$$

Special case:

$$\langle u \otimes v, a \otimes b \rangle_{\mathrm{HS}} = \langle u, a \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}} .$$

Proof: Use expansion

$$b = \sum_{j \in J} \langle b, g_j \rangle_{\mathcal{G}} g_j$$

Then

$$RHS = \langle a, Lb \rangle = \left\langle a, L \left(\sum_j \langle b, g_j \rangle_{\mathcal{G}} g_j \right) \right\rangle_{\mathcal{F}}$$

$$= \sum_j \langle b, g_j \rangle_{\mathcal{G}} \langle a, L g_j \rangle_{\mathcal{F}}$$

Inner product of $a \otimes b$ with $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given a Hilbert-Schmidt operator $L : \mathcal{G} \to \mathcal{F}$,

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}}$$

Special case:

$$\langle u \otimes v, a \otimes b \rangle_{\mathrm{HS}} = \langle u, a \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}} .$$

Proof: Use expansion

$$b = \sum_{j \in J} \langle b, g_j \rangle_{\mathcal{G}} \, g_j$$

Then

$$\begin{aligned} RHS = \langle a, Lb \rangle &= \left\langle a, L \left(\sum_j \langle b, g_j \rangle_{\mathcal{G}} \, g_j \right) \right\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, g_j \rangle_{\mathcal{G}} \, \langle a, Lg_j \rangle_{\mathcal{F}} \end{aligned}$$

Inner product of $a \otimes b$ with $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$

Proof (continued):

$$LHS = \langle a \otimes b, L \rangle_{\mathrm{HS}} := \sum_j \langle L g_j, (a \otimes b) g_j \rangle_{\mathcal{F}}$$
$$= \sum_j \langle b, g_j \rangle_{\mathcal{G}} \langle L g_j, a \rangle_{\mathcal{F}} .$$

Proof of special case:

$$\langle \underbrace{u \otimes v}_{L}, a \otimes b \rangle_{\mathrm{HS}} = \langle a, (u \otimes v) b \rangle_{\mathcal{F}}$$
$$= \left\langle a, u \langle v, b \rangle_{\mathcal{G}} \right\rangle_{\mathcal{F}}$$
$$= \langle u, a \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}}$$

Inner product of $a \otimes b$ with $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$

Proof (continued):

$$\begin{aligned} LHS = \langle a \otimes b, L \rangle_{\mathrm{HS}} &:= \sum_j \langle L g_j, (a \otimes b) g_j \rangle_{\mathcal{F}} \\ &= \sum_j \langle b, g_j \rangle_{\mathcal{G}} \langle L g_j, a \rangle_{\mathcal{F}} . \end{aligned}$$

Proof of special case:

$$\begin{aligned} \langle \underbrace{u \otimes v}_{L}, a \otimes b \rangle_{\mathrm{HS}} &= \langle a, (u \otimes v) b \rangle_{\mathcal{F}} \\ &= \left\langle a, u \langle v, b \rangle_{\mathcal{G}} \right\rangle_{\mathcal{F}} \\ &= \langle u, a \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}} \end{aligned}$$

Covariance operator in RKHS

Challenge 2 (reminder): does there exist $C_{xy} : \mathcal{G} \to \mathcal{F}$ in some Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ such that

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

and in particular,

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

Proof: Use Riesz representer theorem. The operator

$$\begin{aligned} C_{xy} : \mathrm{HS}(\mathcal{G}, \mathcal{F}) &\to \Re \\ A &\mapsto E_{xy} \langle \phi(x) \otimes \psi(y), A \rangle_{\mathrm{HS}} \end{aligned}$$

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

Covariance operator in RKHS

Challenge 2 (reminder): does there exist $C_{xy} : \mathcal{G} \to \mathcal{F}$ in some Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ such that

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

and in particular,

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

Proof: Use Riesz representer theorem. The operator

$$\begin{aligned} C_{xy} : \mathrm{HS}(\mathcal{G}, \mathcal{F}) &\to \Re \\ A &\mapsto E_{xy} \langle \phi(x) \otimes \psi(y), A \rangle_{\mathrm{HS}} \end{aligned}$$

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

Covariance operator in RKHS

Proof (continued): Condition comes from

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

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

Simpler condition:

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

Covariance operator in RKHS

Proof (continued): Condition comes from

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

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

Simpler condition:

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

Covariance operator in RKHS

Does the covariance do what we want? Namely,

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

Proof:

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

Covariance operator in RKHS

Does the covariance do what we want? Namely,

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

Proof:

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

(a) by definition of the covariance operator

Covariance operator in RKHS

Does the covariance do what we want? Namely,

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

Proof:

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

(a) by definition of the covariance operator

Covariance operator in RKHS

Does the covariance do what we want? Namely,

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

Proof:

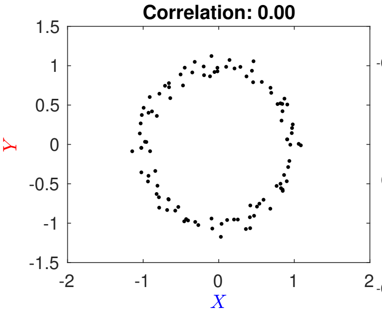
$$\begin{aligned}
\langle f, C_{xy} g \rangle_{\mathcal{F}} &= \langle C_{xy}, f \otimes g \rangle_{\mathrm{HS}} \\
&\overset{(a)}{=} E_{xy} \langle \varphi(x) \otimes \phi(y), f \otimes g \rangle_{\mathrm{HS}} \\
&= E_{xy}\left[\langle f, \varphi(x) \rangle_{\mathcal{F}} \langle g, \phi(y) \rangle_{\mathcal{F}}\right]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

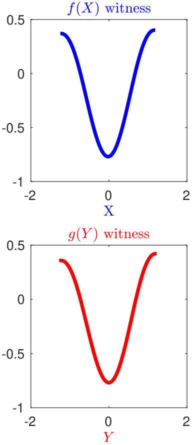
(a) by definition of the covariance operator

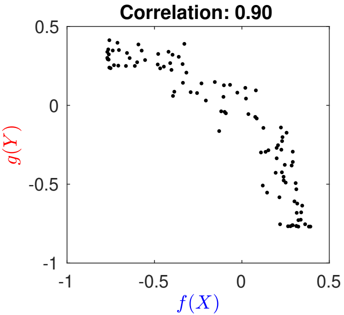
Back to the constrained covariance

The constrained covariance is

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







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

$\widehat{COCO}$ is largest eigenvalue $\gamma_{\max}$ of

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

$\widetilde{K}_{ij} = \langle \varphi(x_i) - \hat{\mu}_x, \varphi(x_j) - \hat{\mu}_x \rangle_{\mathcal{F}} =: \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{F}}$
and $\widetilde{L}_{ij} = \left\langle \tilde{\phi}(y_i), \tilde{\phi}(y_j) \right\rangle_{\mathcal{G}}$.

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

Computing COCO from finite data

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

$\widehat{COCO}$ is largest eigenvalue $\gamma_{\max}$ of

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

$\widetilde{K}_{ij} = \langle \varphi(x_i) - \hat{\mu}_x, \varphi(x_j) - \hat{\mu}_x \rangle_{\mathcal{F}} =: \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{F}}$
and $\widetilde{L}_{ij} = \left\langle \tilde{\phi}(y_i), \tilde{\phi}(y_j) \right\rangle_{\mathcal{G}}$.

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

Empirical COCO: proof

The Lagrangian is

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

$$+ \underbrace{\frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) + \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right)}_{\text{smoothness constraints}}$$

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

(Negative sign on covariance to make it a minimization problem, for consistency with later lectures).

Empirical COCO: proof

The Lagrangian is

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

$$+ \underbrace{\frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) + \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right)}_{\text{smoothness constraints}}$$

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

(Negative sign on covariance to make it a minimization problem, for consistency with later lectures).

Assume:

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

for <u>centered</u> $\tilde{\varphi}(x_i)$, $\tilde{\phi}(y_i)$.

Proof (continued)

First step is smoothness constraint:

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

Proof (continued)

First step is smoothness constraint:

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

Proof (continued)

First step is smoothness constraint:

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

Proof (continued)

Second step is covariance:

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

$$= \frac{1}{n}\sum_{i=1}^{n} \langle f, \tilde{\varphi}(x_i)\rangle_{\mathcal{F}} \left\langle g, \tilde{\phi}(y_i)\right\rangle_{\mathcal{G}}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left\langle \underbrace{\sum_{\ell=1}^{n} \alpha_\ell \tilde{\varphi}(x_\ell)}_{f}, \tilde{\varphi}(x_i)\right\rangle_{\mathcal{F}} \left\langle g, \tilde{\phi}(y_i)\right\rangle_{\mathcal{G}}$$

$$= \frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta$$

Proof (continued)

Second step is covariance:

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

$$=\frac{1}{n}\sum_{i=1}^{n}\langle f,\tilde{\varphi}(x_i)\rangle_{\mathcal{F}}\left\langle g,\tilde{\phi}(y_i)\right\rangle_{\mathcal{G}}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\langle\underbrace{\sum_{\ell=1}^{n}\alpha_\ell\tilde{\varphi}(x_\ell)}_{f},\tilde{\varphi}(x_i)\right\rangle_{\mathcal{F}}\left\langle g,\tilde{\phi}(y_i)\right\rangle_{\mathcal{G}}$$

$$=\frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta$$

Proof (continued)

Second step is covariance:

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

Kernel matrices between centered variables:

$$
\widetilde{K}=HKH \qquad H=I_n-\frac{1}{n}1_n1_n^\top.
$$

Proof (continued)

Minimize Lagrangian wrt the primal variables α, β:

$$\mathcal{L}(f, g, \lambda, \gamma) = -\frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta + \frac{\lambda}{2}\left(\alpha^\top \widetilde{K}\alpha - 1\right) + \frac{\gamma}{2}\left(\beta^\top \widetilde{L}\beta - 1\right)$$

Differentiating wrt α and β and setting to zero,

$$0 = -\frac{1}{n}\widetilde{K}\widetilde{L}\beta + \lambda\widetilde{K}\alpha$$

$$0 = -\frac{1}{n}\widetilde{L}\widetilde{K}\alpha + \gamma\widetilde{L}\beta$$

Multiply the first equation by $\alpha^\top$, and the second by $\beta^\top$,

$$0 = -\frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta + \lambda\alpha^\top\widetilde{K}\alpha$$

$$0 = -\frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha + \gamma\beta^\top\widetilde{L}\beta$$

Proof (continued)

Minimize Lagrangian wrt the primal variables α, β:

$$\mathcal{L}(f, g, \lambda, \gamma) = -\frac{1}{n}\alpha^\top \widetilde{K}\tilde{L}\beta + \frac{\lambda}{2}\left(\alpha^\top \widetilde{K}\alpha - 1\right) + \frac{\gamma}{2}\left(\beta^\top \tilde{L}\beta - 1\right)$$

Differentiating wrt α and β and setting to zero,

$$0 = -\frac{1}{n}\widetilde{K}\tilde{L}\beta + \lambda\widetilde{K}\alpha$$

$$0 = -\frac{1}{n}\tilde{L}\widetilde{K}\alpha + \gamma\tilde{L}\beta$$

Multiply the first equation by $\alpha^\top$, and the second by $\beta^\top$,

$$0 = -\frac{1}{n}\alpha^\top \widetilde{K}\tilde{L}\beta + \lambda\alpha^\top \widetilde{K}\alpha$$

$$0 = -\frac{1}{n}\beta^\top \tilde{L}\widetilde{K}\alpha + \gamma\beta^\top \tilde{L}\beta$$

Proof (continued)

Minimize Lagrangian wrt the primal variables α, β:

$$\mathcal{L}(f, g, \lambda, \gamma) = -\frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta + \frac{\lambda}{2}\left(\alpha^\top \widetilde{K}\alpha - 1\right) + \frac{\gamma}{2}\left(\beta^\top \widetilde{L}\beta - 1\right)$$

Differentiating wrt α and β and setting to zero,

$$0 = -\frac{1}{n}\widetilde{K}\widetilde{L}\beta + \lambda\widetilde{K}\alpha$$

$$0 = -\frac{1}{n}\widetilde{L}\widetilde{K}\alpha + \gamma\widetilde{L}\beta$$

Multiply the first equation by $\alpha^\top$, and the second by $\beta^\top$,

$$0 = -\frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta + \lambda\alpha^\top\widetilde{K}\alpha$$

$$0 = -\frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha + \gamma\beta^\top\widetilde{L}\beta$$

Proof (continued)

Subtract second equation from first, get

$$\lambda \alpha^\top \widetilde{K} \alpha = \gamma \beta^\top \widetilde{L} \beta$$

When $\lambda \neq 0$ and $\gamma \neq 0$, then $\alpha^\top \widetilde{K} \alpha = \beta^\top \widetilde{L} \beta = 1$, hence $\lambda = \gamma$.

(Comlpementary slackness, assuming strong duality.

More later in the course!)

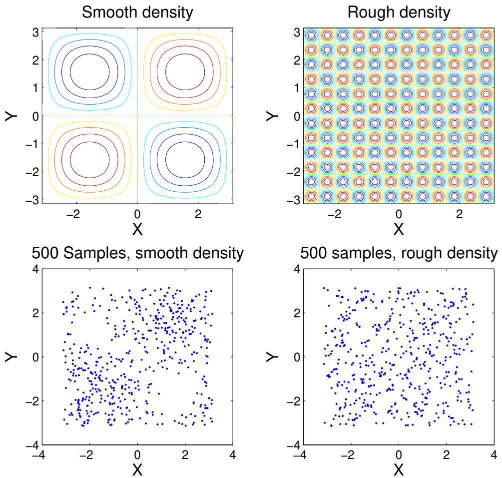
Thus $\widehat{COCO}$ is largest eigenvalue γ_{max} of

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

(Solution by maximization wrt dual variable γ).

Note: for strong duality in this case, see Appendix B.1 Boyd and Vandenberghe (2004), which is outside the scope of this lecture.

What is a large dependence with COCO?



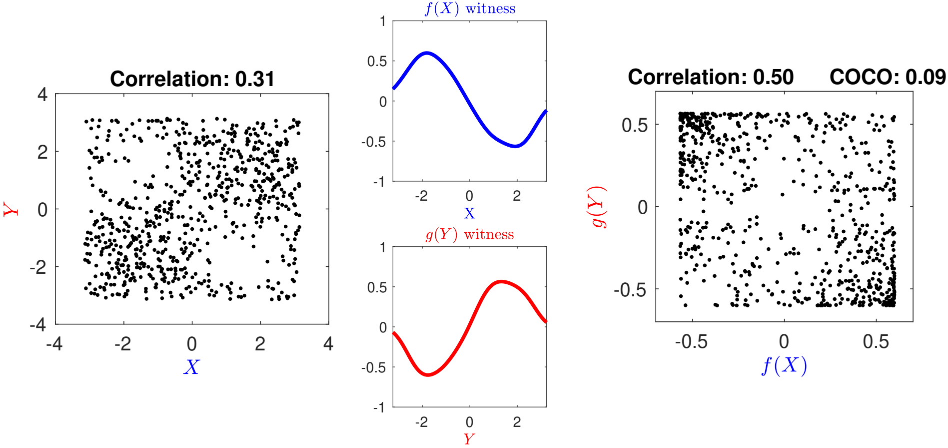
Density takes the form:

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

Which of these is the more “dependent”?

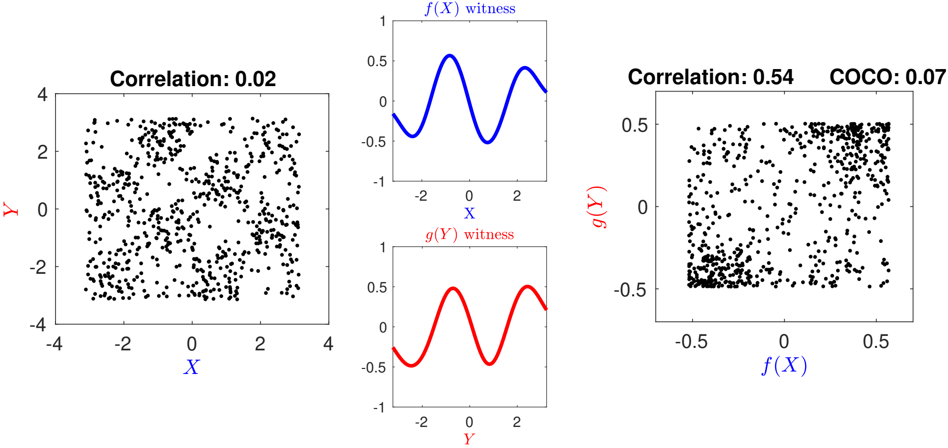
Finding covariance with smooth transformations

Case of $\omega = 1$:



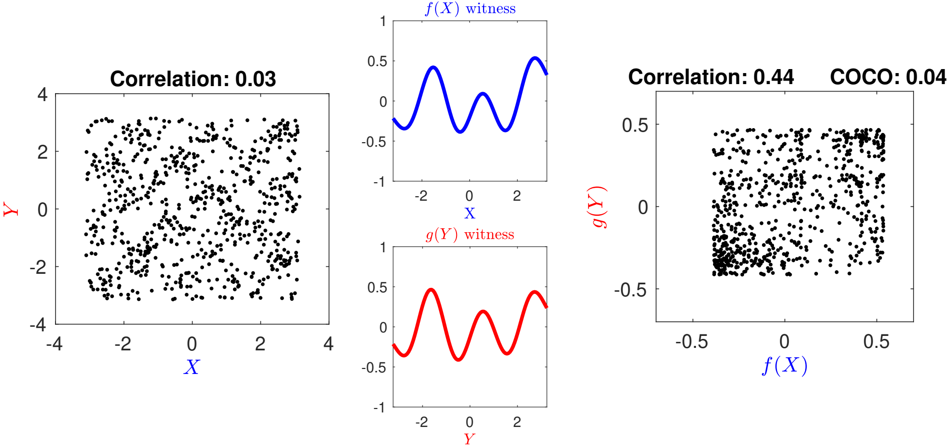
Finding covariance with smooth transformations

Case of $\omega = 2$:



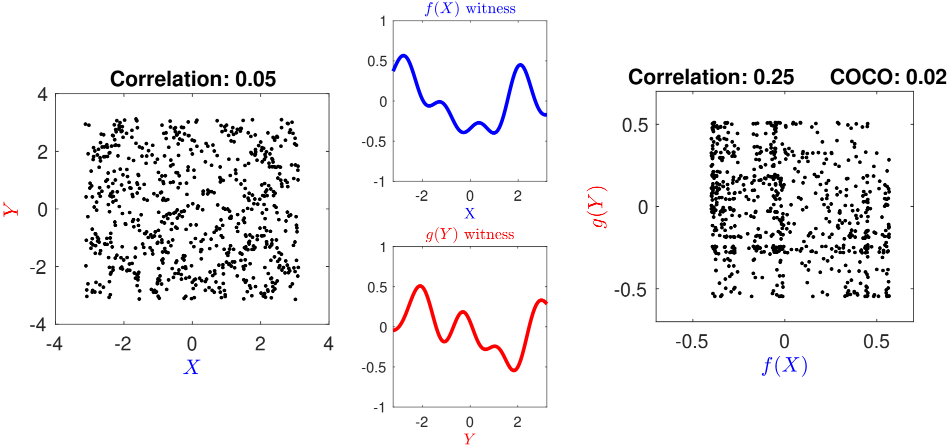
Finding covariance with smooth transformations

Case of $\omega = 3$:



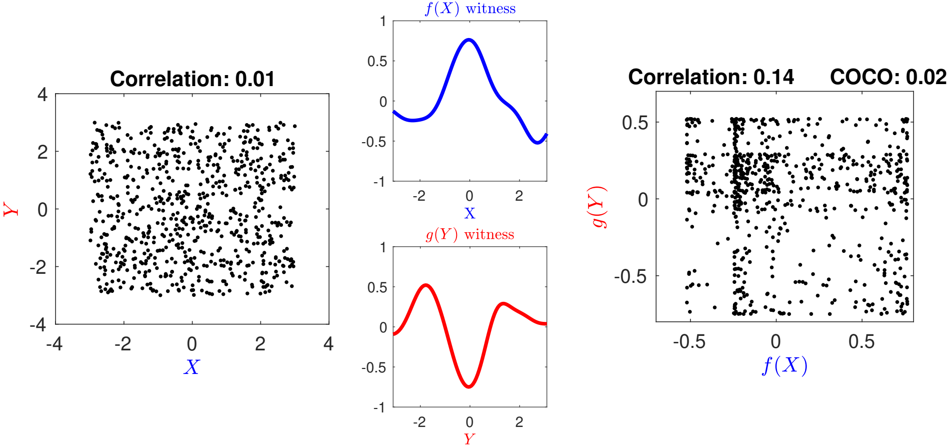
Finding covariance with smooth transformations

Case of $\omega = 4$:



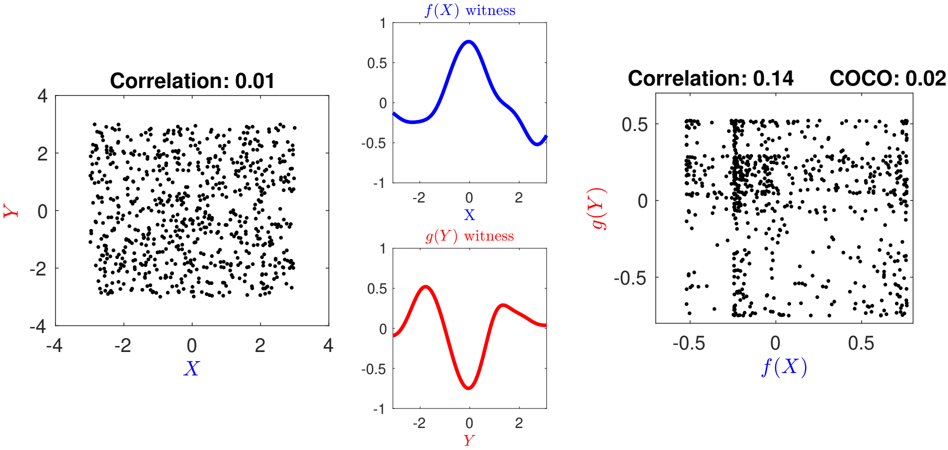
Finding covariance with smooth transformations

Case of $\omega = ??$:



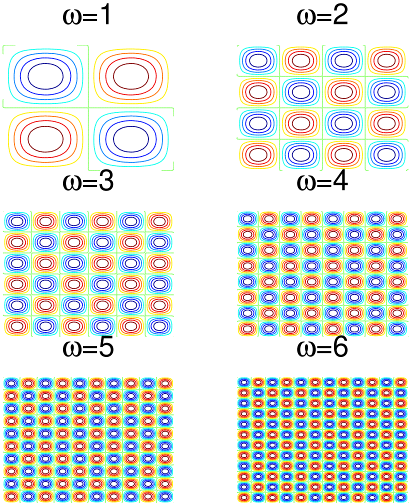
Finding covariance with smooth transformations

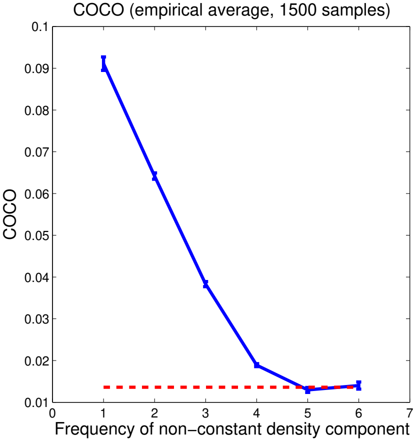
Case of $\omega = 0$: uniform noise! (shows bias)



Back to the constrained covariance

Summary: sinusoidal density $P_{XY} \propto 1 + \sin(\omega x)\sin(\omega y)$





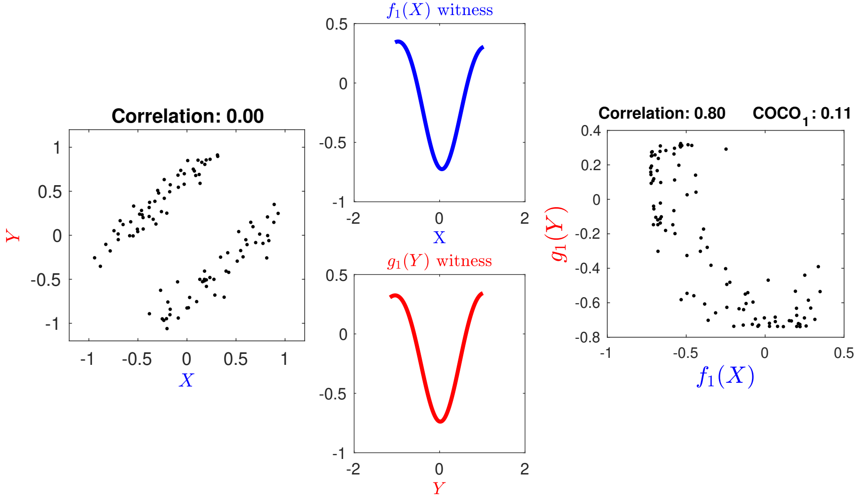
Dependence largest when at “low” frequencies

- 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.

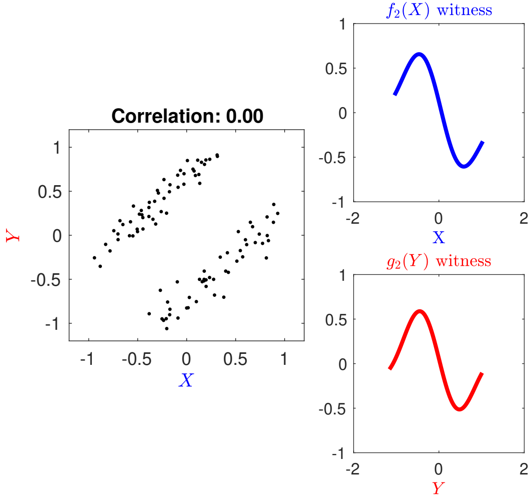
First singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



Can we do better than COCO?

A second example with zero correlation.

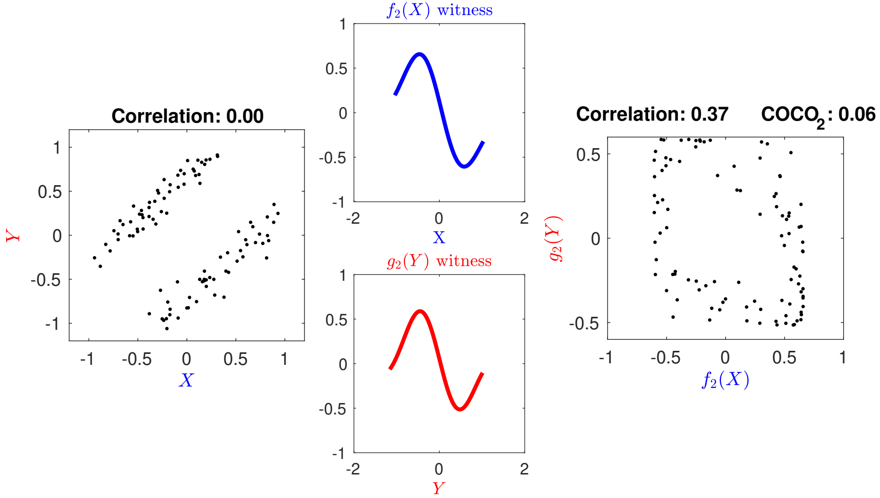
Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



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 ith 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)

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

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

The Hilbert-Schmidt Independence Criterion

Hilbert-Schmidt Independence Criterion (HSIC) in terms of HS norm:

$$
\begin{aligned}
HSIC^2(\mathrm{Pr}; \mathcal{F}, \mathcal{G}) &:= \| C_{xy} - \mu_X \otimes \mu_Y \|^2_{\mathrm{HS}} \\
&= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\mathrm{HS}} \\
&\quad - 2 \langle C_{xy}, \mu_X \otimes \mu_Y \rangle_{\mathrm{HS}} \\
&= E_{x,y} E_{x',y'} [k(x, x') l(y, y')] \\
&\quad + E_{x,x'} [k(x, x')] E_{y,y'} [l(y, y')] \\
&\quad - 2 E_{x,y} \left[E_{x'} [k(x, x')] E_{y'} [l(y, y')] \right]
\end{aligned}
$$

C_{xy} is uncentered covariance, $x, x' \sim P_x$ independent, $y, y' \sim P_y$.

The Hilbert-Schmidt Independence Criterion

Hilbert-Schmidt Independence Criterion (HSIC) in terms of HS norm:

$$
\begin{aligned}
HSIC^2(\mathrm{Pr}; \mathcal{F}, \mathcal{G}) &:= \| C_{xy} - \mu_X \otimes \mu_Y \|_{\mathrm{HS}}^2 \\
&= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\mathrm{HS}} \\
&\quad - 2 \langle C_{xy}, \mu_X \otimes \mu_Y \rangle_{\mathrm{HS}} \\
&= E_{x,y} E_{x',y'} [k(x, x') l(y, y')] \\
&\quad + E_{x,x'} [k(x, x')] E_{y,y'} [l(y, y')] \\
&\quad - 2 E_{x,y} \left[E_{x'} [k(x, x')] E_{y'} [l(y, y')] \right]
\end{aligned}
$$

C_{xy} is uncentered covariance, $x, x' \sim P_x$ independent, $y, y' \sim P_y$.

The Hilbert-Schmidt Independence Criterion

Proof: Recall

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \qquad \langle C_{xy}, A \rangle_{\mathrm{HS}} = E_{x,y} \langle \phi(x) \otimes \psi(y), A \rangle_{\mathrm{HS}}$$

and

$$[a \otimes b]c = \langle b, c \rangle a$$

Applying the (uncentered) covariance operator definition twice,

$$\begin{aligned}
\| C_{xy} \|_{\mathrm{HS}}^2 &= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} \langle \varphi(x) \otimes \phi(y), C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y') \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x), [\varphi(x') \otimes \phi(y')] \phi(y) \rangle_{\mathcal{F}} \\
&= E_{xy} E_{x'y'} \left[\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \langle \phi(y'), \phi(y) \rangle_{\mathcal{G}} \right] \\
&= E_{xy} E_{x'y'} \left[k(x, x') l(y, y') \right]
\end{aligned}$$

The Hilbert-Schmidt Independence Criterion

Proof: Recall

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \qquad \langle C_{xy}, A \rangle_{\mathrm{HS}} = E_{x,y} \langle \phi(x) \otimes \psi(y), A \rangle_{\mathrm{HS}}$$

and

$$[a \otimes b]c = \langle b, c \rangle a$$

Applying the (uncentered) covariance operator definition twice,

$$\begin{aligned}
\| C_{xy} \|_{\mathrm{HS}}^2 &= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} \langle \varphi(x) \otimes \phi(y), C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y') \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x), [\varphi(x') \otimes \phi(y')] \phi(y) \rangle_{\mathcal{F}} \\
&= E_{xy} E_{x'y'} \left[\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \langle \phi(y'), \phi(y) \rangle_{\mathcal{G}} \right] \\
&= E_{xy} E_{x'y'} \left[k(x, x') l(y, y') \right]
\end{aligned}$$

The Hilbert-Schmidt Independence Criterion

Proof: Recall

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \qquad \langle C_{xy}, A \rangle_{\mathrm{HS}} = E_{x,y} \langle \phi(x) \otimes \psi(y), A \rangle_{\mathrm{HS}}$$

and

$$[a \otimes b]c = \langle b, c \rangle a$$

Applying the (uncentered) covariance operator definition twice,

$$\begin{aligned}
\| C_{xy} \|_{\mathrm{HS}}^2 &= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} \langle \varphi(x) \otimes \phi(y), C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y') \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x), [\varphi(x') \otimes \phi(y')] \phi(y) \rangle_{\mathcal{F}} \\
&= E_{xy} E_{x'y'} \left[\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \langle \phi(y'), \phi(y) \rangle_{\mathcal{G}} \right] \\
&= E_{xy} E_{x'y'} \left[k(x, x') l(y, y') \right]
\end{aligned}$$

The Hilbert-Schmidt Independence Criterion

Proof: Recall

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \qquad \langle C_{xy}, A \rangle_{\mathrm{HS}} = E_{x,y} \langle \phi(x) \otimes \psi(y), A \rangle_{\mathrm{HS}}$$

and

$$[a \otimes b]c = \langle b, c \rangle a$$

Applying the (uncentered) covariance operator definition twice,

$$\begin{aligned}
\| C_{xy} \|_{\mathrm{HS}}^2 &= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} \langle \varphi(x) \otimes \phi(y), C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y') \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x), [\varphi(x') \otimes \phi(y')] \phi(y) \rangle_{\mathcal{F}} \\
&= E_{xy} E_{x'y'} \left[\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \langle \phi(y'), \phi(y) \rangle_{\mathcal{G}} \right] \\
&= E_{xy} E_{x'y'} [k(x, x') l(y, y')]
\end{aligned}$$

The Hilbert-Schmidt Independence Criterion

Proof: Recall

$$\langle L, a \otimes b \rangle_{\text{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \qquad \langle C_{xy}, A \rangle_{\text{HS}} = E_{x,y} \langle \phi(x) \otimes \psi(y), A \rangle_{\text{HS}}$$

and

$$[a \otimes b] c = \langle b, c \rangle a$$

Applying the (uncentered) covariance operator definition twice,

$$\begin{aligned}
\| C_{xy} \|^2_{\text{HS}} &= \langle C_{xy}, C_{xy} \rangle_{\text{HS}} \\
&= E_{xy} \langle \varphi(x) \otimes \phi(y), C_{xy} \rangle_{\text{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y') \rangle_{\text{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x), [\varphi(x') \otimes \phi(y')] \phi(y) \rangle_{\mathcal{F}} \\
&= E_{xy} E_{x'y'} \left[\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \langle \phi(y'), \phi(y) \rangle_{\mathcal{G}} \right] \\
&= E_{xy} E_{x'y'} \left[k(x, x') l(y, y') \right]
\end{aligned}$$

The Hilbert-Schmidt Independence Criterion

Proof: Recall

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \qquad \langle C_{xy}, A \rangle_{\mathrm{HS}} = E_{x,y} \langle \phi(x) \otimes \psi(y), A \rangle_{\mathrm{HS}}$$

and

$$[a \otimes b] c = \langle b, c \rangle a$$

Applying the (uncentered) covariance operator definition twice,

$$\begin{aligned}
\| C_{xy} \|_{\mathrm{HS}}^2 &= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} \langle \varphi(x) \otimes \phi(y), C_{xy} \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x) \otimes \phi(y), \varphi(x') \otimes \phi(y') \rangle_{\mathrm{HS}} \\
&= E_{xy} E_{x'y'} \langle \varphi(x), [\varphi(x') \otimes \phi(y')] \phi(y) \rangle_{\mathcal{F}} \\
&= E_{xy} E_{x'y'} \left[\langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \langle \phi(y'), \phi(y) \rangle_{\mathcal{G}} \right] \\
&= E_{xy} E_{x'y'} \left[k(x, x') l(y, y') \right]
\end{aligned}$$

Empirical estimates of HSIC

Unbiased estimate: define $\widehat{A}$ as the empirical estimator of $\|C_{xy}\|^2_{\mathrm{HS}} = E_{xy}E_{x'y'}\left[k(x,x')l(y,y')\right]$,

$$\widehat{A} := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j)l(y_i,y_j)$$

Alternative: plug in empirical covariance operator (uncentered),

$$\check{C}_{xy} = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\psi(y_i),$$

Biased estimate:

$$\widehat{A}_b = \left\|\check{C}_{xy}\right\|^2_{\mathrm{HS}} = \left\langle \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\phi(y_i), \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\phi(y_i)\right\rangle_{\mathrm{HS}}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i,x_j)l(y_i,y_j) = \frac{1}{n^2}\mathrm{tr}(KL),$$

Empirical estimates of HSIC

Unbiased estimate: define $\widehat{A}$ as the empirical estimator of $\|C_{xy}\|^2_{\mathrm{HS}} = E_{xy}E_{x'y'}\left[k(x,x')l(y,y')\right]$,

$$\widehat{A} := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j)l(y_i,y_j)$$

Alternative: plug in empirical covariance operator (uncentered),

$$\check{C}_{xy} = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\psi(y_i),$$

Biased estimate:

$$\widehat{A}_b = \left\|\check{C}_{xy}\right\|^2_{\mathrm{HS}} = \left\langle \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\phi(y_i), \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\phi(y_i)\right\rangle_{\mathrm{HS}}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i,x_j)l(y_i,y_j) = \frac{1}{n^2}\mathrm{tr}(KL),$$

Empirical estimates of HSIC

Unbiased estimate: define $\widehat{A}$ as the empirical estimator of $\|C_{xy}\|^2_{\text{HS}} = E_{xy}E_{x'y'}\left[k(x,x')l(y,y')\right]$,

$$\widehat{A} := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j)l(y_i,y_j)$$

Alternative: plug in empirical covariance operator (uncentered),

$$\check{C}_{xy} = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\psi(y_i),$$

Biased estimate:

$$\widehat{A}_b = \left\|\check{C}_{xy}\right\|^2_{\text{HS}} = \left\langle \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\phi(y_i), \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\otimes\phi(y_i)\right\rangle_{\text{HS}}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i,x_j)l(y_i,y_j) = \frac{1}{n^2}\text{tr}(KL),$$

How large is the bias?

Difference is:

$$
\begin{aligned}
\widehat{A}_b - \widehat{A} &= \frac{1}{n^2} \sum_{i,j=1}^{n} k_{ij} l_{ij} - \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k_{ij} l_{ij} \\
&= \frac{1}{n^2} \sum_{i=1}^{n} k_{ii} l_{ii} + \left(\frac{1}{n^2} - \frac{1}{n(n-1)} \right) \left(\sum_{i \neq j}^{n} k_{ij} l_{ij} \right) \\
&= \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^{n} k_{ii} l_{ii} - \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k_{ij} l_{ij} \right),
\end{aligned}
$$

where $k_{ij} = k(x_i, x_j)$.

The expectation of this difference (i.e., the bias) is of $O(n^{-1})$.

Remaining terms covered in lecture notes.

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 $\widehat{HSIC}^2$?

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 $\widehat{HSIC}^2$?
- Empirical HSIC (biased)

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

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ $\qquad (H = I_n - \frac{1}{n}1_n 1_n^\top)$

HSIC is MMD with product kernel!

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

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

Proof: exercise!

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 $\widehat{HSIC}^2$?
- Empirical HSIC (biased)

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

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ $\quad (H = I_n - \frac{1}{n}1_n 1_n^\top)$

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α?

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 $\widehat{HSIC}^2$?
- Empirical HSIC (biased)

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

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ $\qquad (H = I_n - \frac{1}{n}1_n 1_n^\top)$

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α?
- Asymptotics of $\widehat{HSIC}$ when $P_{XY} = P_X P_Y$:

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

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

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α (prob. of false positive)?

- Original sample:

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

- Permutation:

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

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. 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 $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α (prob. of false positive)?
- Original sample:

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

- Permutation:

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

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. 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 $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α (prob. of false positive)?
- Original sample:

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

- Permutation:

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

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. 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 $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α (prob. of false positive)?
- Null distribution via moment matching

$$n\mathrm{HSIC}_b^2(Z) \sim \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$$

where

$$\alpha = \frac{(E(\mathrm{HSIC}_b^2))^2}{\mathrm{var}(\mathrm{HSIC}_b^2)}, \quad \beta = \frac{\mathrm{var}(\mathrm{HSIC}_b^2)}{nE(\mathrm{HSIC}_b^2)}.$$

- Purely a heuristic, no guarantees

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
⋮	⋮

Text from the aligned hansards of the 36th parliament of canada,
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 $n = 10$

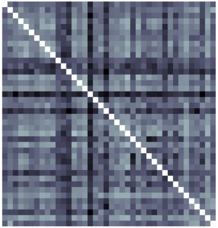
X

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

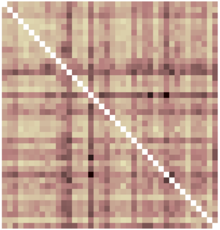
Y

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

$$\widehat{HSIC}^2 = \frac{1}{n^2}\operatorname{trace}(KHLH)$$

$H = I_n - \frac{1}{n}1_n 1_n^\top$

Application:Dependence detection across languages

Results (for $\alpha = 0.05$)

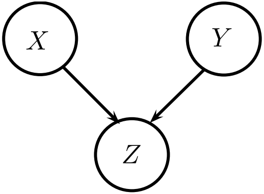
- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

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

Testing higher order interactions

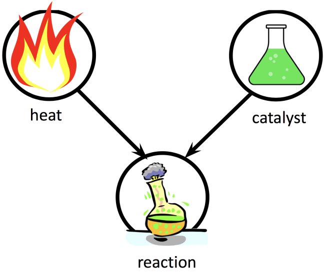
Detecting higher order interaction

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



Detecting higher order interaction

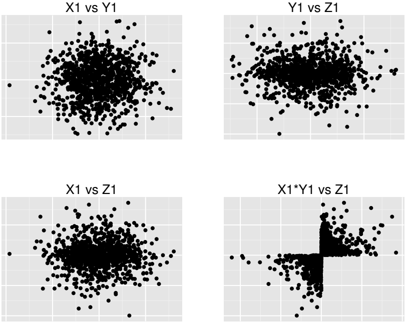
How to detect V-structures with pairwise weak individual dependence?

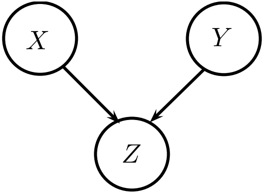


Detecting higher order interaction

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

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

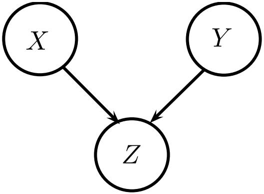




- $X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- $Z|\, X, Y \sim \text{sign}(XY)Exp(\frac{1}{\sqrt{2}})$

Fine print: Faithfulness violated here!

V-structure discovery



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

V-structure can then be detected by:

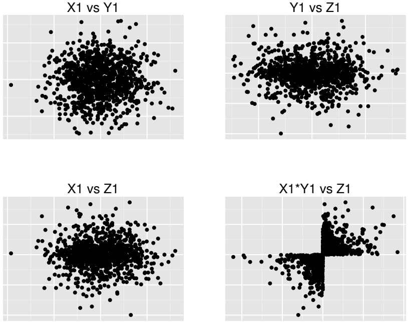
- Consistent CI test: $\mathrm{H}_0 : X \perp\!\!\!\perp Y | Z$ [Fukumizu et al. 2008, Zhang et al. 2011]
- Factorisation test: $\mathrm{H}_0 : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)

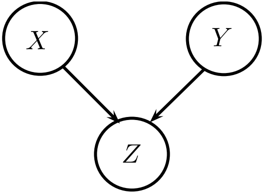
How well do these work?

Detecting higher order interaction

Generalise earlier example to p dimensions

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

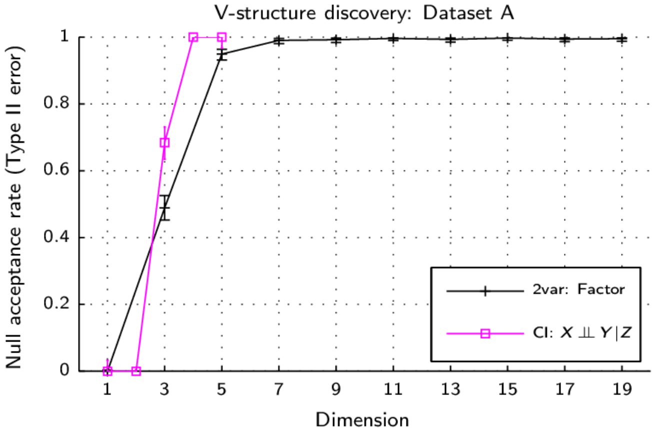




- $X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- $Z|\, X, Y \sim \text{sign}(XY)Exp(\frac{1}{\sqrt{2}})$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathrm{I}_{p-1})$

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

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

Lancaster interaction measure

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

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

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

Lancaster interaction measure

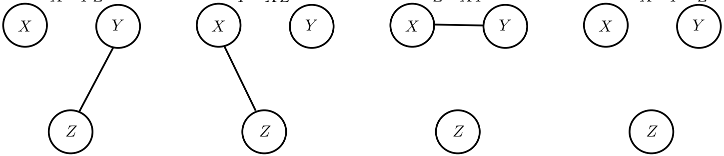
Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

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

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

$\Delta_L P =$

P_{XYZ} $\quad -P_X P_{YZ}$ $\quad -P_Y P_{XZ}$ $\quad -P_Z P_{XY}$ $\quad +2P_X P_Y P_Z$

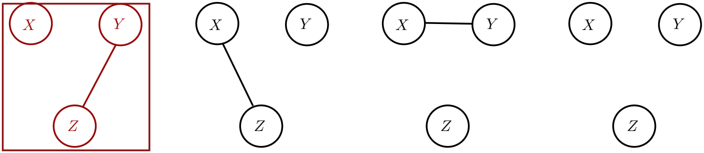


Lancaster interaction measure

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.

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

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



$\Delta_L P = 0$

$\cancel{P_{XYZ}}$ $\cancel{-P_X P_{YZ}}$ $\cancel{-P_{XZ} P_Y}$ $\cancel{-P_{XY} P_Z}$ $\cancel{+2 P_X P_Y P_Z}$

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

Lancaster interaction measure

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

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

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

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

...so what might be missed?

Lancaster interaction measure

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

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

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

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

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 $\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa}$, 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

$\nu \backslash \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_X P_Y P_Z$
P_{XYZ}	$(\mathbf{K}\circ\mathbf{L}\circ\mathbf{M})_{++}$	$((\mathbf{K}\circ\mathbf{L})\,\mathbf{M})_{++}$	$((\mathbf{K}\circ\mathbf{M})\,\mathbf{L})_{++}$	$((\mathbf{M}\circ\mathbf{L})\,\mathbf{K})_{++}$	$tr(\mathbf{K}_+\circ\mathbf{L}_+\circ\mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K}\circ\mathbf{L})_{++}\,\mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K}\circ\mathbf{M})_{++}\,\mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L}\circ\mathbf{M})_{++}\,\mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_X P_Y P_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

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

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

Empirical joint central moment in the feature space

A kernel test statistic using Lancaster Measure

$\nu \backslash \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_X P_Y P_Z$
P_{XYZ}	$(\mathbf{K} \circ \mathbf{L} \circ \mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{L})\, \mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{M})\, \mathbf{L})_{++}$	$((\mathbf{M} \circ \mathbf{L})\, \mathbf{K})_{++}$	$tr(\mathbf{K}_+ \circ \mathbf{L}_+ \circ \mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K} \circ \mathbf{L})_{++}\, \mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K} \circ \mathbf{M})_{++}\, \mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L} \circ \mathbf{M})_{++}\, \mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_X P_Y P_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

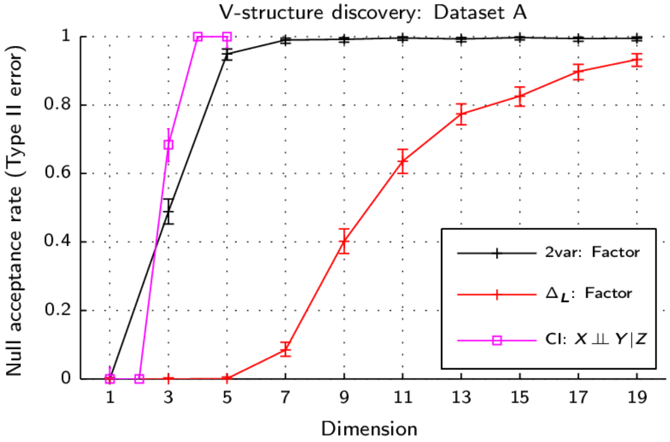
Table: V-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_X P_Y P_Z$). H is centering matrix $I - n^{-1}$

Lancaster interaction statistic: Sejdinovic, G, Bergsma, NIPS13

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

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., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

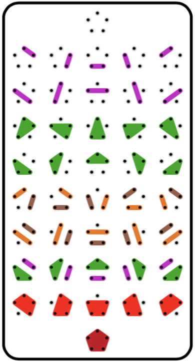
Interaction for $D \geq 4$

- 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., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.