Representing and comparing probabilities: Part 2

Arthur Gretton

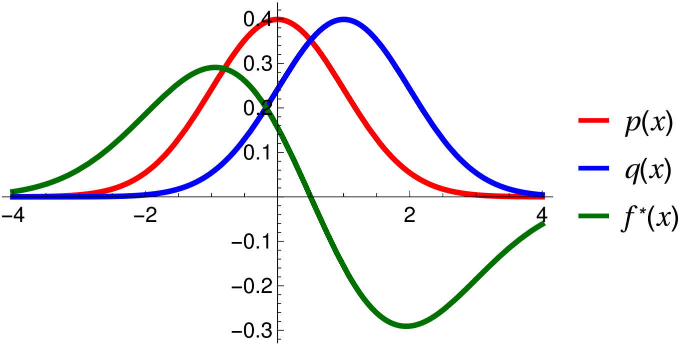
Gatsby Computational Neuroscience Unit,
University College London

UAI, 2017

Testing against a probabilistic model

Statistical model criticism

$$MMD(P, Q) = \|f^*\|^2 = \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_Q f - E_p f]$$



$f^*(x)$ is the witness function

Can we compute MMD with samples from Q and a **model** P?

Problem: usualy can't compute $E_p f$ in closed form.

Stein idea

To get rid of $E_p f$ in

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

we define the **Stein operator**

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

Then

$$E_P \, T_P f = 0$$

subject to appropriate boundary conditions. (Oates, Girolami, Chopin, 2016)

Stein idea: proof

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

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

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

$$= 0$$

Stein idea: proof

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

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

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

$$= 0$$

Stein idea: proof

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

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

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

$$= 0$$

Stein idea: proof

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

Stein idea: proof

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

Kernel Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

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

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Kernel Stein Discrepancy (KSD)

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

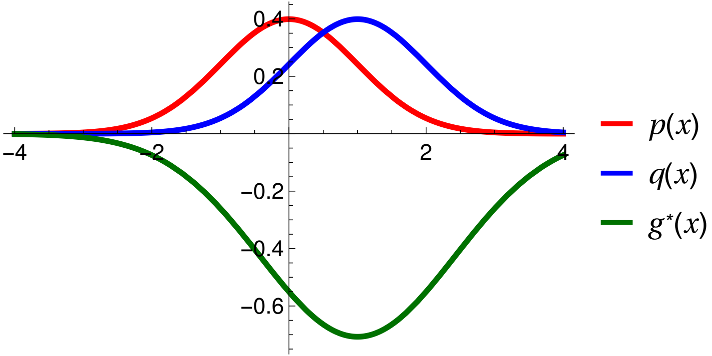
Kernel Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Kernel Stein Discrepancy (KSD)

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



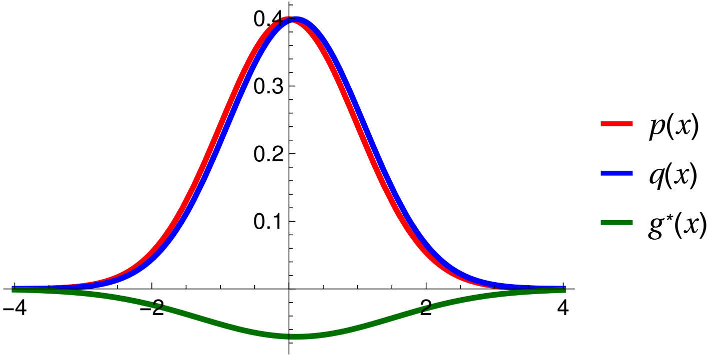
Kernel Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

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

Closed-form expression for KSD: given $Z, Z' \sim q$, then
(Chwialkowski, Strathmann, G., ICML 2016) (Liu, Lee, Jordan ICML 2016)

$$\mathrm{KSD}(p, q, \mathcal{F}) = E_q h_p(Z, Z')$$

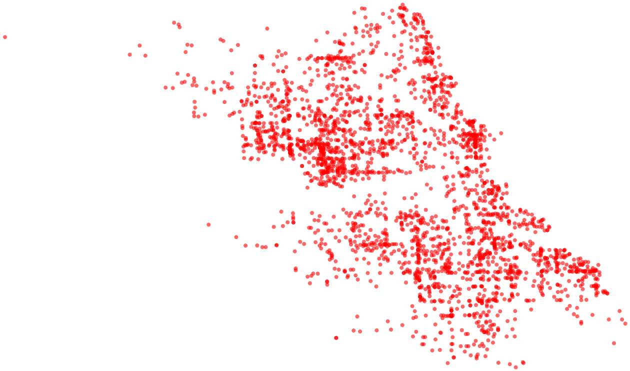
where

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

and k is RKHS kernel for $\mathcal{F}$

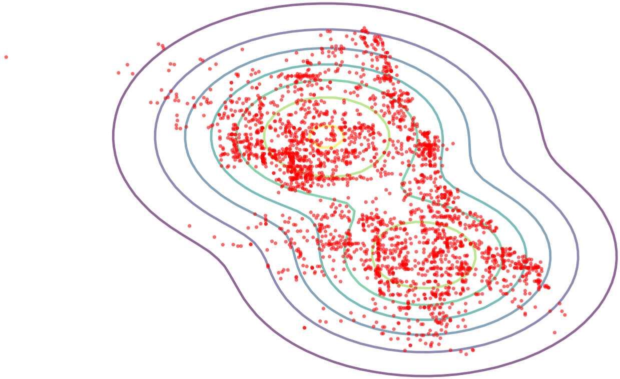
Only depends on kernel and $\partial_x \log p(x)$. Do not need to normalize p, or sample from it.

Statistical model criticism



Chicago crime data

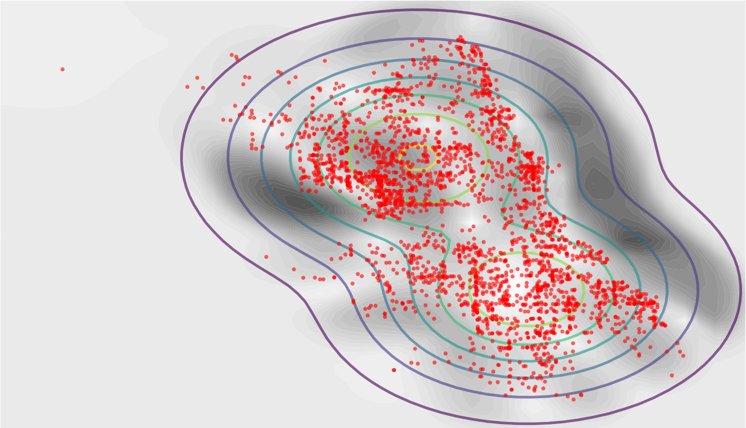
Statistical model criticism



Chicago crime data

Model is Gaussian mixture with two components.

Statistical model criticism

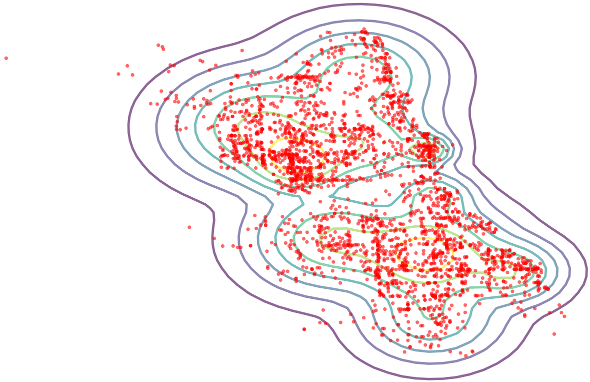


Chicago crime data

Model is Gaussian mixture with two components

Stein **witness function**

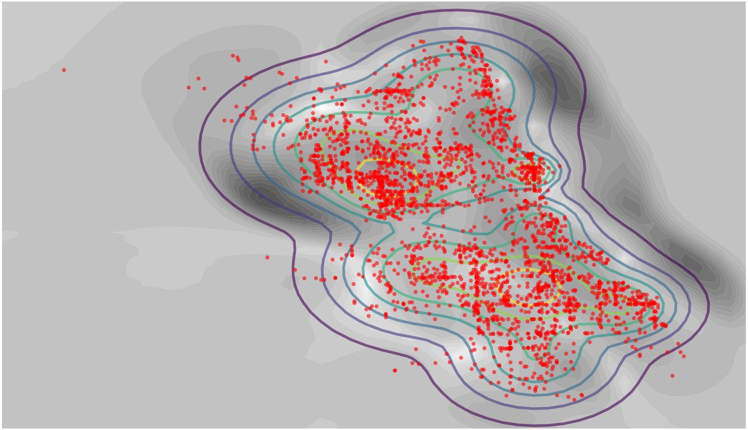
Statistical model criticism



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_{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

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 $\mathcal{F}$,

$k($  ,  $)$

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?

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

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

MMD as a dependence measure

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

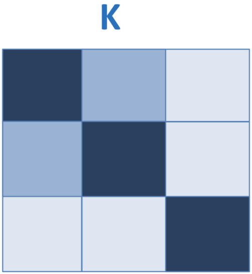
$$MMD^2(\widehat{P}_{XY}, \widehat{P}_X \widehat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2}\mathrm{trace}(K L)$$

(K, L column centered)

MMD as a dependence measure

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

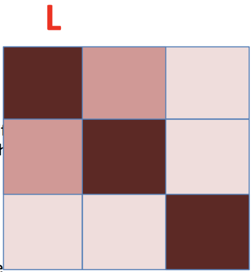
$$MMD^2(\widehat{P}_{XY}, \widehat{P}_X \widehat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2}\text{trace}(K\,L)$$



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

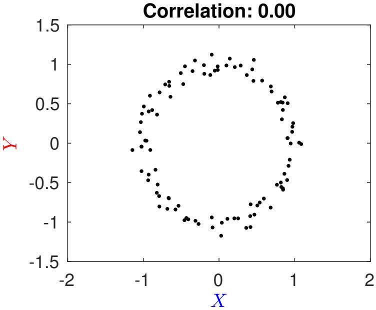
MMD as a dependence measure

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?

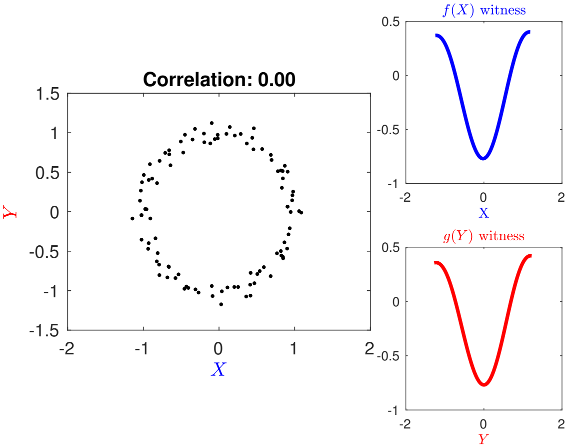
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



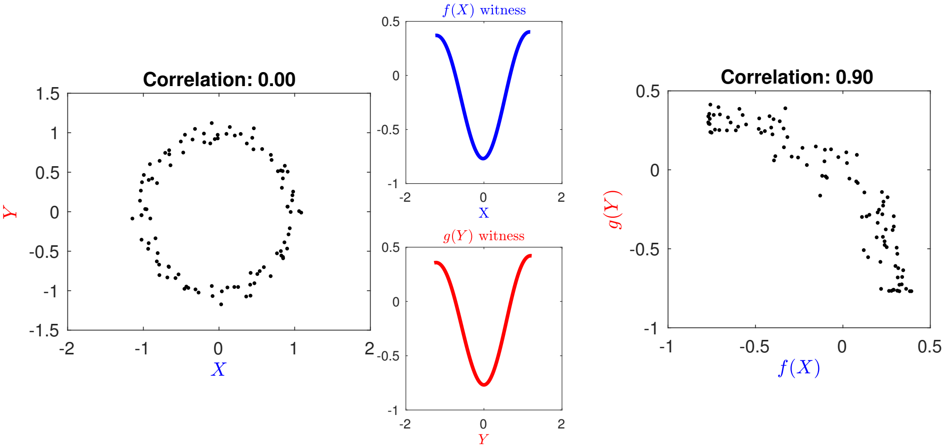
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



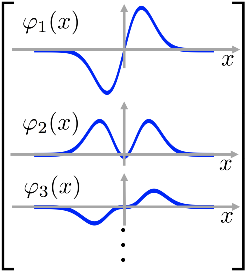
Define two spaces, one for each witness

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 $\mathcal{X}$:

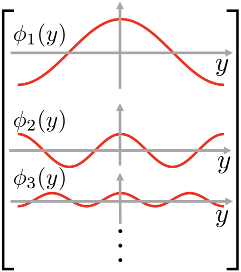
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Function in $\mathcal{G}$

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

Feature map

$$\phi(y) =$$



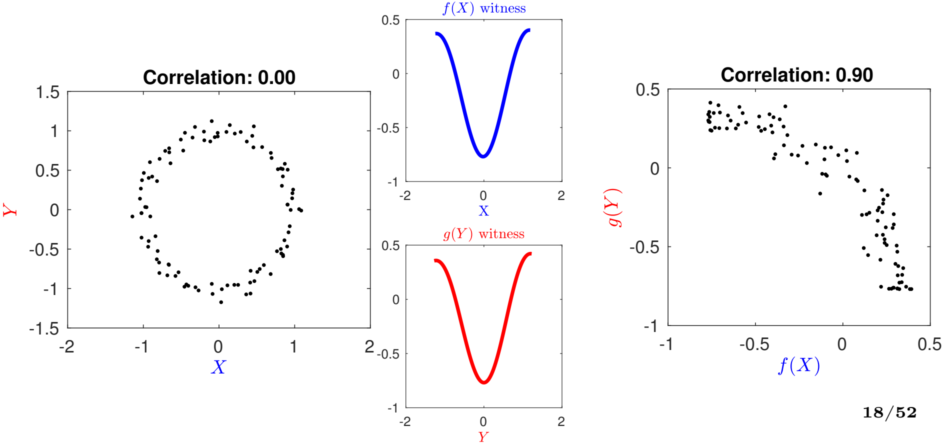
Kernel for RKHS $\mathcal{G}$ on $\mathcal{Y}$:

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

The constrained covariance

The constrained covariance is

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



The constrained covariance

The constrained covariance is

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

The constrained covariance

The constrained covariance is

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

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

The constrained covariance

The constrained covariance is

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

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

Rewriting:

$$\begin{aligned} & E_{xy}[f(x)g(y)] \\ &= \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top \underbrace{\mathbf{E}_{xy}\left(\begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \phi_1(y) & \phi_2(y) & \ldots \end{bmatrix}\right)}_{C_{\varphi(x)\phi(y)}} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix} \end{aligned}$$

The constrained covariance

The constrained covariance is

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

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

Rewriting:

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

$$= \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^{\top} \underbrace{\mathbf{E}_{xy}\left(\begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix}\begin{bmatrix} \phi_1(y) & \phi_2(y) & \dots \end{bmatrix}\right)}_{C_{\varphi(x)\phi(y)}} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}$$

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

Computing COCO in practice

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

Computing COCO in practice

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

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

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

Fine print: kernels are computed with empirically centered features $\varphi(x) - \frac{1}{n}\sum_{i=1}^n \varphi(x_i)$ and $\phi(y) - \frac{1}{n}\sum_{i=1}^n \phi(y_i)$.

AG., A. Smola., O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schoelkopf, and N. Logothetis, AISTATS'05

Computing COCO in practice

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

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

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

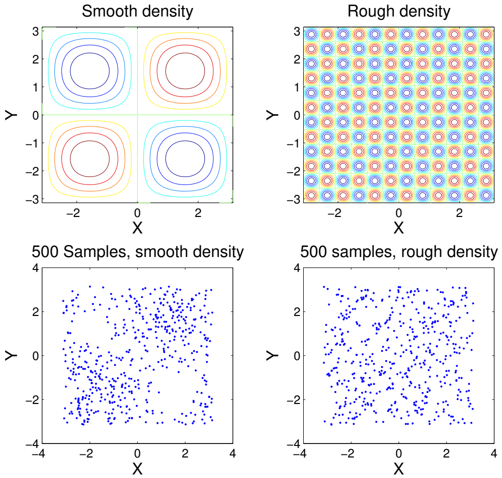
Witness functions (singular vectors):

$$f(x) \propto \sum_{i=1}^{m} \alpha_i k(x_i, x) \qquad g(y) \propto \sum_{i=1}^{n} \beta_i l(y_i, y)$$

Fine print: kernels are computed with empirically centered features $\varphi(x) - \frac{1}{n}\sum_{i=1}^n \varphi(x_i)$ and $\phi(y) - \frac{1}{n}\sum_{i=1}^n \phi(y_i)$.

AG., A. Smola., O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schoelkopf, and N. Logothetis, AISTATS'05

What is a large dependence with COCO?



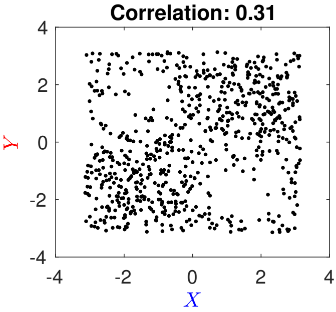
Density takes the form:

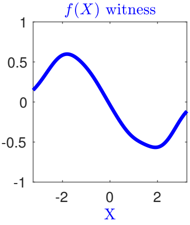
$$P_{XY} \propto 1+\sin(\omega x)\sin(\omega y)$$

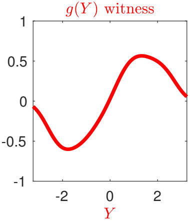
Which of these is the more “dependent”?

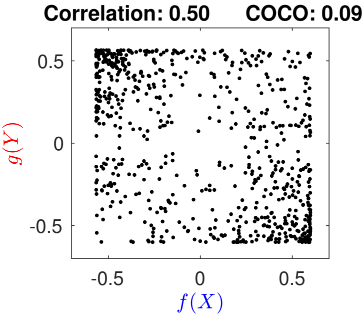
Finding covariance with smooth transformations

Case of $\omega = 1$:



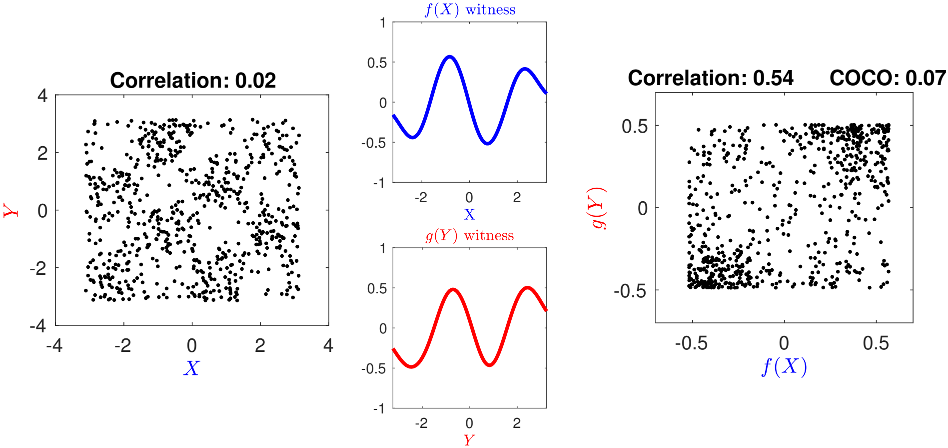






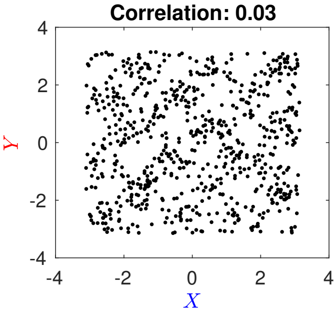
Finding covariance with smooth transformations

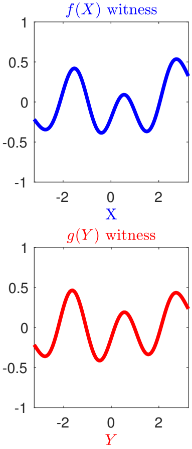
Case of $\omega = 2$:

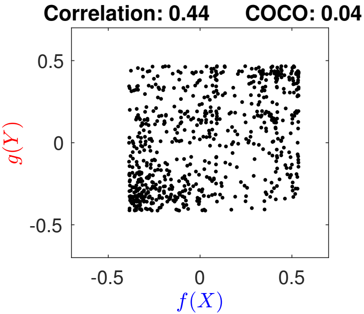


Finding covariance with smooth transformations

Case of $\omega = 3$:

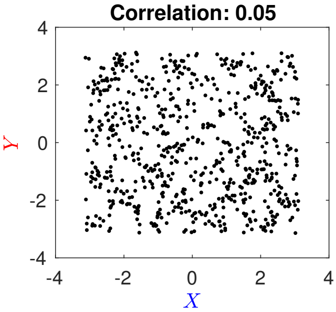


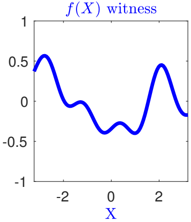


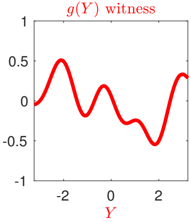


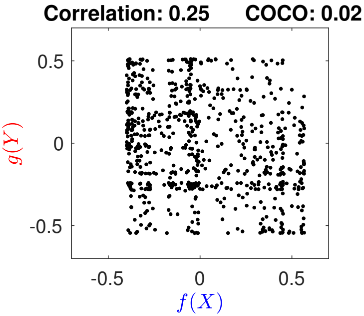
Finding covariance with smooth transformations

Case of $\omega = 4$:



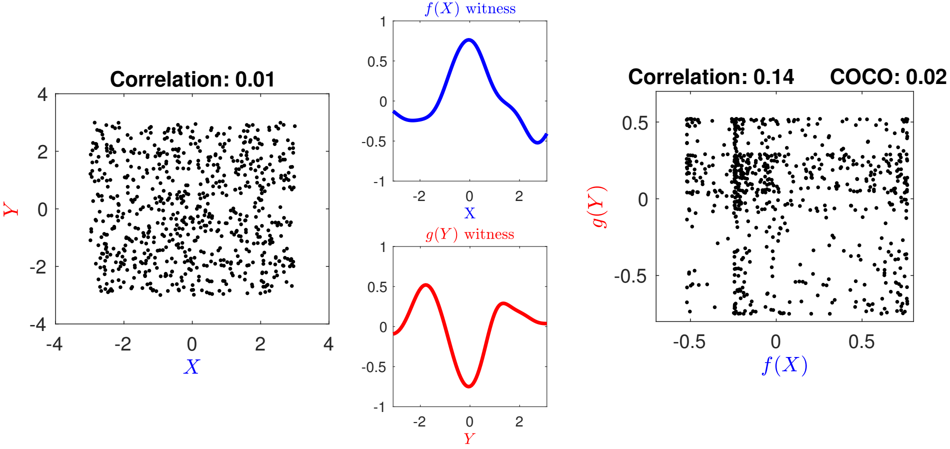






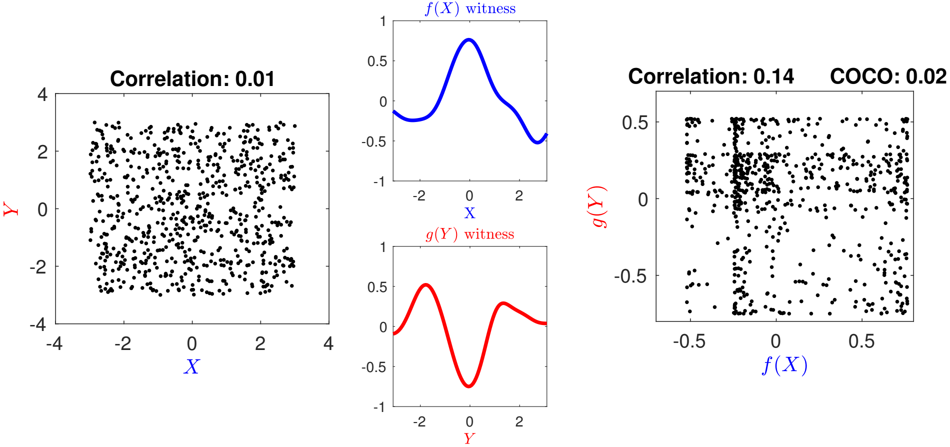
Finding covariance with smooth transformations

Case of $\omega = ??$:



Finding covariance with smooth transformations

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



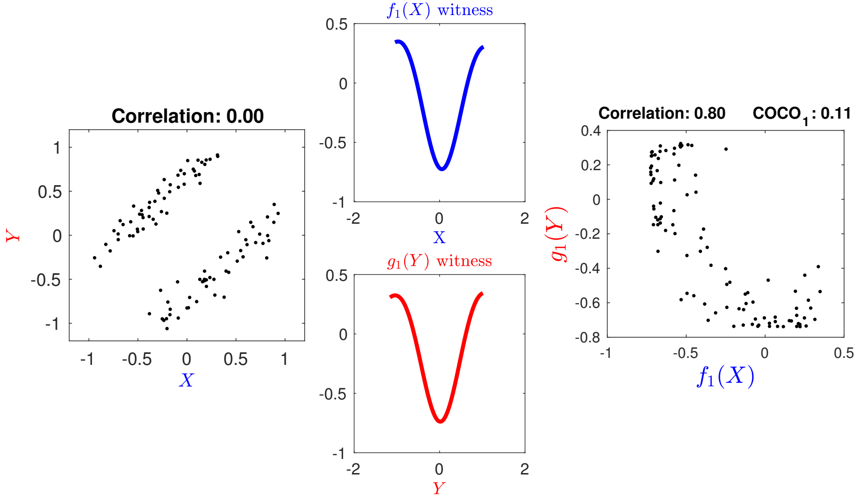
Dependence largest when at “low” frequencies

- 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.

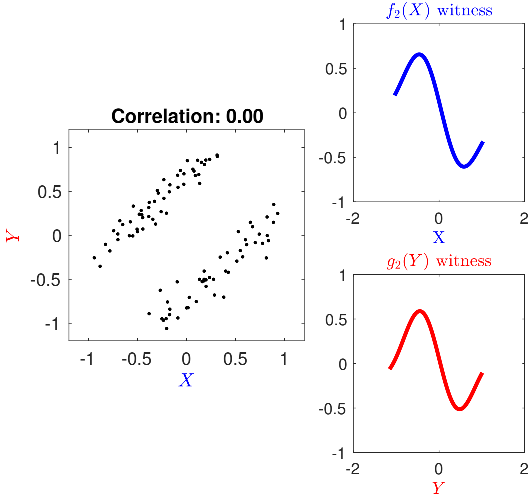
First singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



Can we do better than COCO?

A second example with zero correlation.

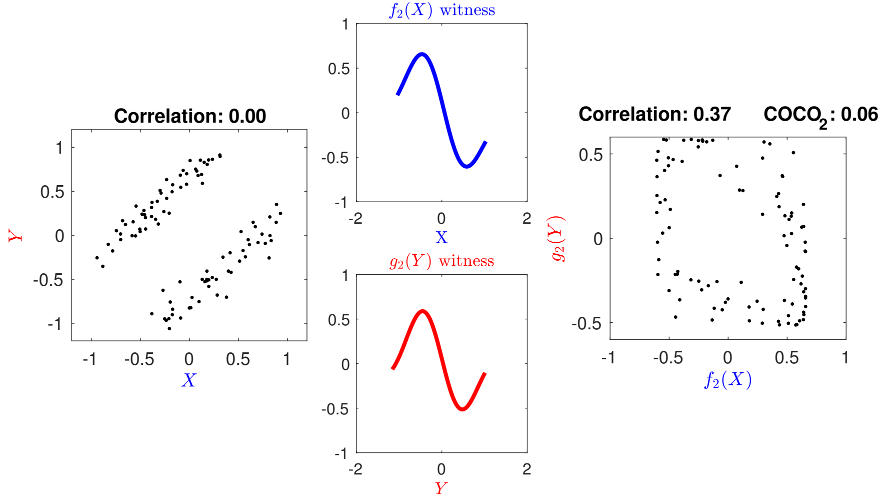
Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



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)

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

AG, O. Bousquet , A. Smola., and B. Schoelkopf, ALT2005; AG,.,K. Fukumizu,,C.H. Teo., L. Song., B. Schoelkopf., and A. Smola, NIPS 2007,.

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)

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

AG, O. Bousquet , A. Smola., and B. Schoelkopf, ALT2005; AG,.,K. Fukumizu,,C.H. Teo., L. Song., B. Schoelkopf., and A. 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 $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$, 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 $P_{XY} = P_X P_Y$:

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

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

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

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

$K_{ij} = k(x_i, x_j)$ 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 $P_{XY} = P_X P_Y$:

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

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

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 $\widehat{HSIC}$?
- 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 $P_{XY} = P_X P_Y$:

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

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

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 $\widehat{HSIC}$?
- 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 $P_{XY} = P_X P_Y$:

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

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 $P(\widehat{HSIC} > c_\alpha) < \alpha$ 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 $\{1, \ldots, n\}$. 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 $P(\widehat{HSIC} > c_\alpha) < \alpha$ 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 $\{1, \ldots, n\}$. 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 $P(\widehat{HSIC} > c_\alpha) < \alpha$ 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
⋮	⋮

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 $n = 10$

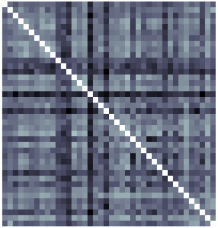
X

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.

$\vdots$



K

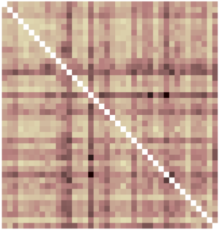
Y

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

$\vdots$



L

$$\widehat{HSIC} = \frac{1}{n^2}\mathrm{trace}(K\,L)$$

(K and L column centered)

Application:Dependence detection across languages

Results (for $\alpha = 0.05$)

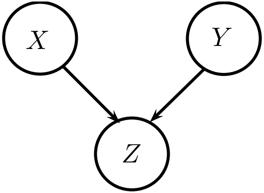
- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

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

Testing higher order interactions

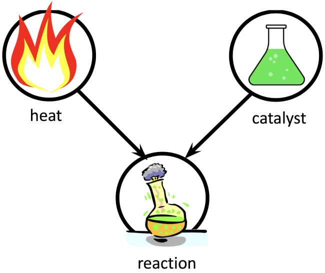
Detecting higher order interaction

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



Detecting higher order interaction

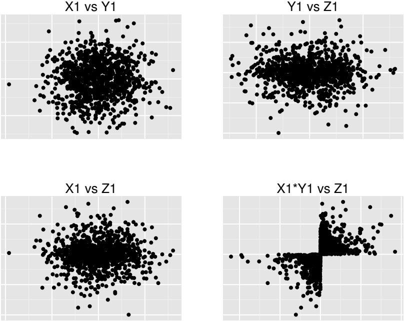
How to detect V-structures with pairwise weak individual dependence?

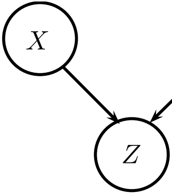


Detecting higher order interaction

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

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

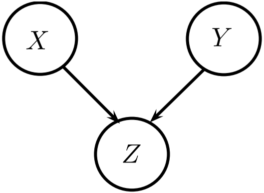




- $X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- $Z|\ X, Y \sim \text{sign}(XY)Exp(\frac{1}{\sqrt{2}})$

Fine print: Faithfulness violated here!

V-structure discovery



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

V-structure can then be detected by:

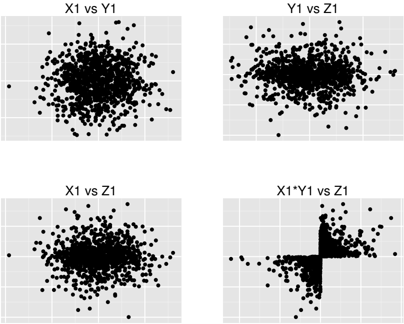
- Consistent CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ [Fukumizu et al. 2008, Zhang et al. 2011]
- Factorisation test: $\mathbf{H_0} : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)

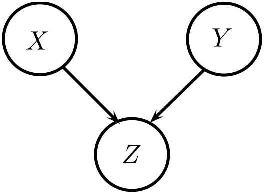
How well do these work?

Detecting higher order interaction

Generalise earlier example to p dimensions

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

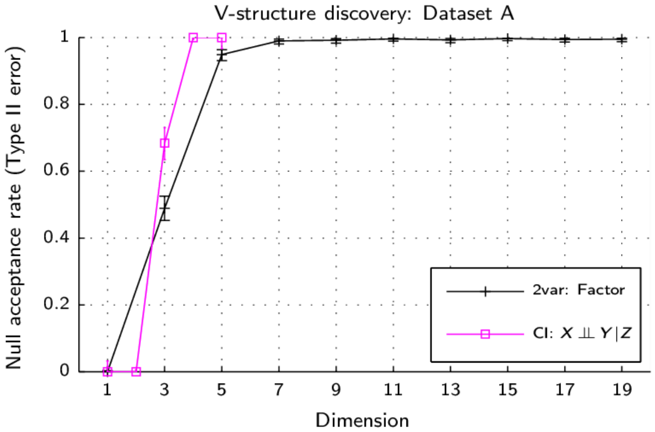




- $X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- $Z|\, X, Y \sim \text{sign}(XY) Exp(\frac{1}{\sqrt{2}})$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$

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 $(X_1, \ldots, X_D) \sim P$ 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

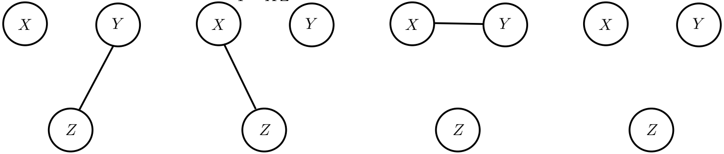
Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

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

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

$\Delta_L P =$

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

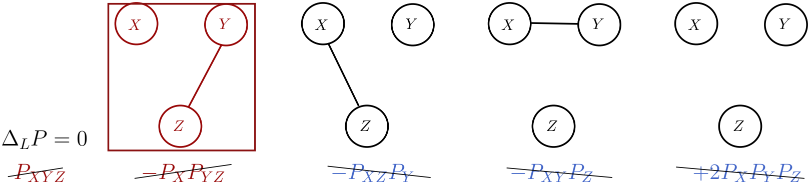


Lancaster interaction measure

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.

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

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



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: D. Sejdinovic, AG, W. 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}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(\mathbf{K}\circ\mathbf{L}\circ\mathbf{M})_{++}$	$((\mathbf{K}\circ\mathbf{L})\,\mathbf{M})_{++}$	$((\mathbf{K}\circ\mathbf{M})\,\mathbf{L})_{++}$	$((\mathbf{M}\circ\mathbf{L})\,\mathbf{K})_{++}$	$tr(\mathbf{K}_+\circ\mathbf{L}_+\circ\mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K}\circ\mathbf{L})_{++}\,\mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K}\circ\mathbf{M})_{++}\,\mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L}\circ\mathbf{M})_{++}\,\mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_XP_YP_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

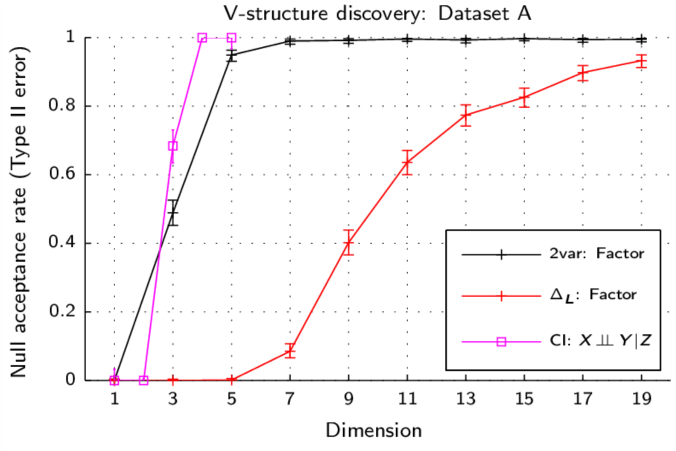
Table: V-statistic estimators of $\langle\mu_\kappa\nu,\mu_\kappa\nu'\rangle_{\mathcal{H}_\kappa}$ (without terms $P_XP_YP_Z$). H is centering matrix $I-n^{-1}$

Lancaster interaction statistic: D. Sejdinovic, AG, W. Bergsma, NIPS13

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

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

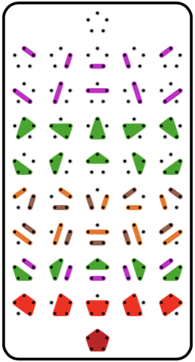
Interaction for $D \geq 4$

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



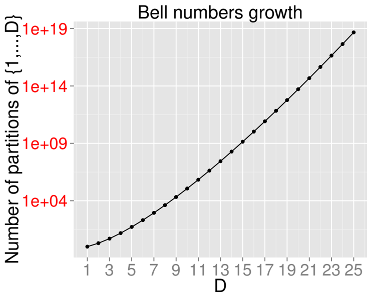
Interaction for $D \geq 4$

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

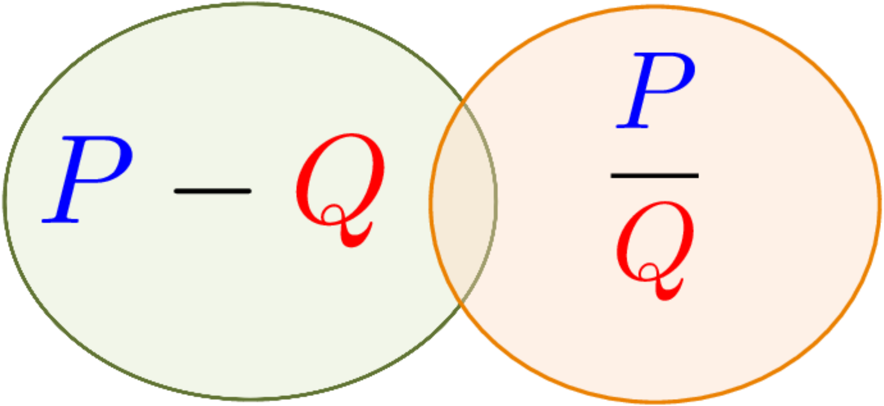


Part 4: Advanced topics

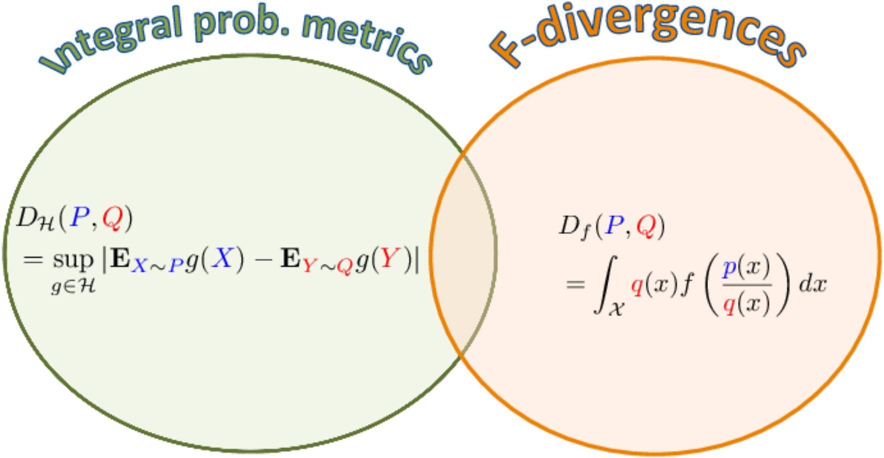
Advanced topics

- testing on time series
- testing for conditional dependence
- regression and conditional mean embedding

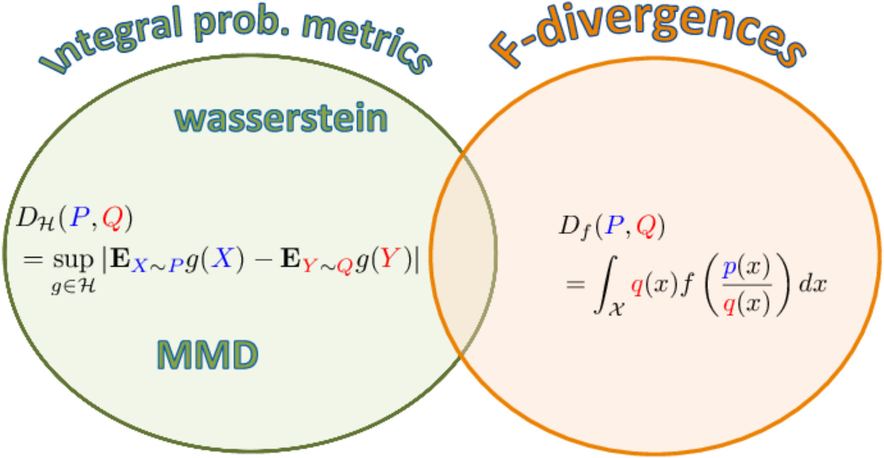
Measures of divergence



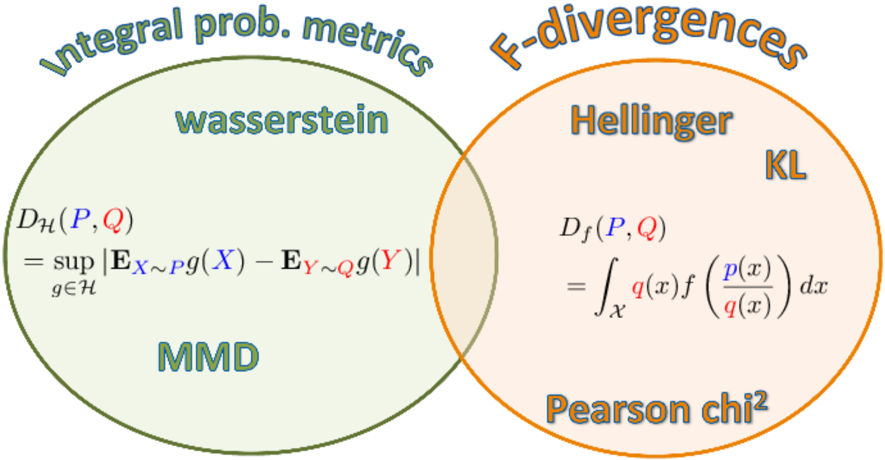
Measures of divergence



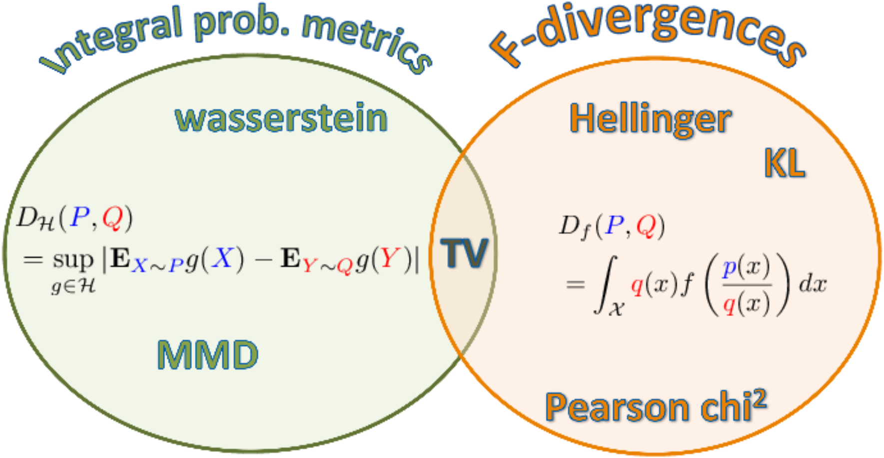
Measures of divergence



Measures of divergence



Measures of divergence



Sriperumbudur, Fukumizu, G, Schoelkopf, Lanckriet (2012)

Co-authors

From Gatsby:

- Kacper Chwialkowski
- Wittawat Jitkrittum
- Bharath Sriperumbudur
- Heiko Strathmann
- Dougal Sutherland
- Zoltan Szabo
- Wenkai Xu

External collaborators:

- Kenji Fukumizu
- Bernhard Schoelkopf
- Alex Smola

Questions?