Kernel Methods and Hypothesis Testing

Dino Sejdinovic

Department of Statistics
University of Oxford

18 November 2014

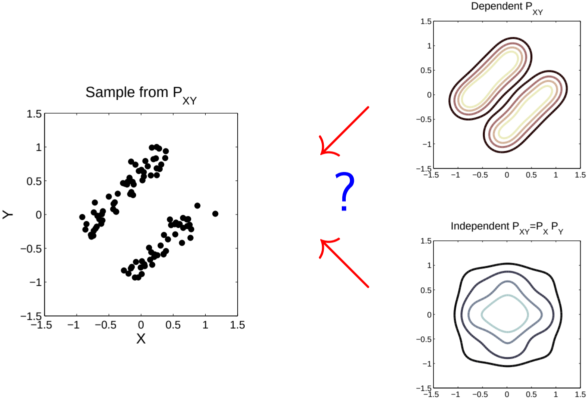
Overview

Outline

Detecting dependence

- How to detect dependence in a Euclidean space?

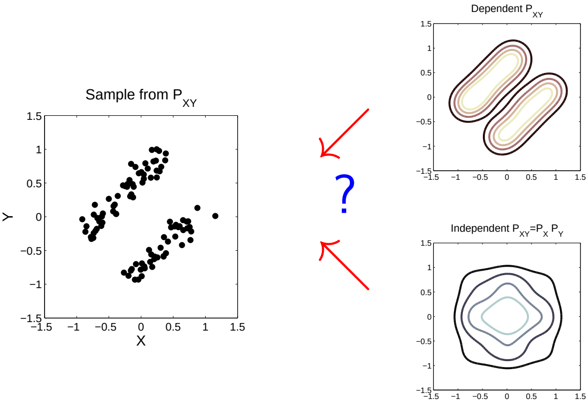
- $H_0 : X \perp\!\!\!\perp Y$
- $H_A : X \not\perp\!\!\!\perp Y$



Detecting dependence

- How to detect dependence in a Euclidean space?

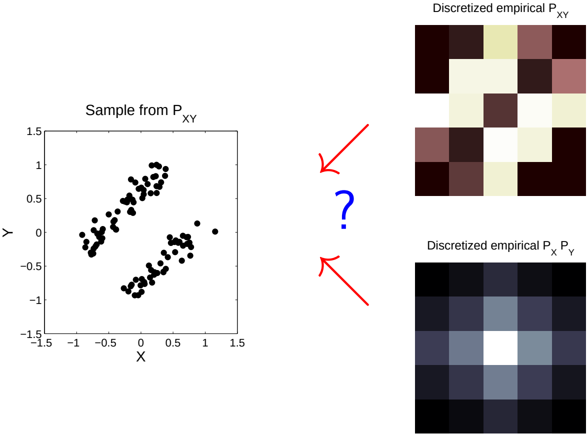
- $\mathbf{H_0}: \mathbf{P}_{XY} = \mathbf{P}_X \mathbf{P}_Y$
- $\mathbf{H_A}: \mathbf{P}_{XY} \neq \mathbf{P}_X \mathbf{P}_Y$



Detecting dependence

- How to detect dependence in a Euclidean space?

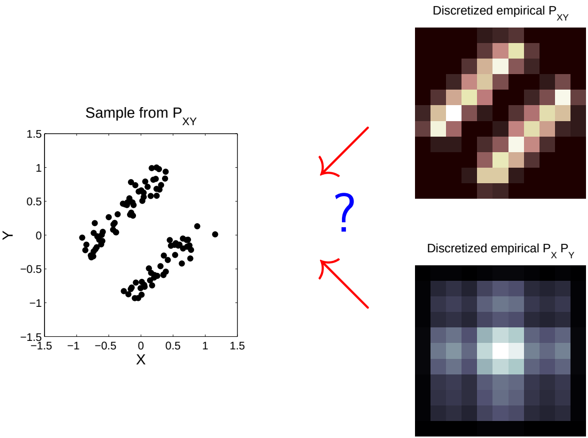
- $\mathbf{H_0}: \ P_{XY} = P_X P_Y$
- $\mathbf{H_A}: \ P_{XY} \neq P_X P_Y$



Detecting dependence

- How to detect dependence in a Euclidean space?

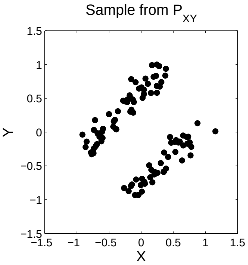
- $H_0: P_{XY} = P_X P_Y$
- $H_A: P_{XY} \neq P_X P_Y$



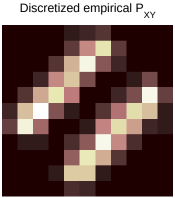
Detecting dependence

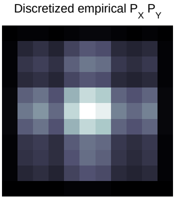
- How to detect dependence in a Euclidean space?

- $\mathrm{H}_0: \mathrm{P}_{XY} = \mathrm{P}_X \mathrm{P}_Y$
- $\mathrm{H}_A: \mathrm{P}_{XY} \neq \mathrm{P}_X \mathrm{P}_Y$



?





- $X, Y \in \mathbb{R}^4$ with dependence in a single dimension. For $n = 1024$, $\alpha = 0.05$, Type II error $\approx$**.95**. Too few points per bin!

Detecting dependence

X_1:

Y_1: The Dandie Dinmont Terrier is a sweet and hardy dog with lots of personality and pluck. He shows incredible loyalty to his owner, and is utterly devoted to his family. He is affectionate and loves to cuddle and be held in his owner's arms. He will follow you all over the house...

X_2:

$\overset{?}{\Longleftrightarrow}$

Y_2: The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps. He is a clown with a sense of humor, but he is still a true terrier: determined, keen, alert, inquisitive, and spirited....

X_3:

Y_3: Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament: inquisitive, bossy, feisty, and fearless. They are intelligent and can be a bit mischievous. Warn your flowers – many Cairns love to dig! They are not usually problem barkers, but will bark if bored or lonely...

$\cdots$ `[from justdogbreeds.com]`

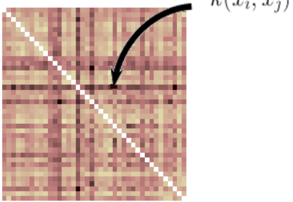
...

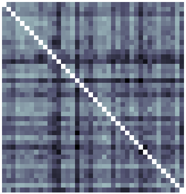
Detecting dependence

$k($  ,  $)$

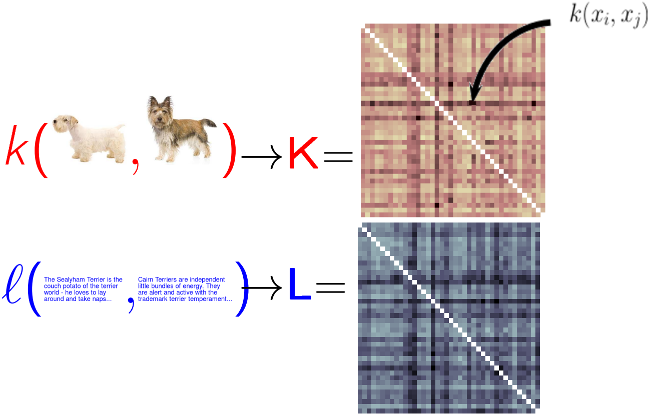
$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $)$

Detecting dependence

$k($  ,  $) \to \mathbf{K} =$ 

$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $) \to \mathbf{L} =$ 

Detecting dependence



- **Idea**: measure similarity between the kernel matrices

$$\left\langle \tilde{K}, \tilde{L} \right\rangle = \boxed{\mathrm{Tr}\left(\tilde{K}\tilde{L}\right)}$$

- $\tilde{K} = HKH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (centering matrix)

Two-sample problem

- We are given $\{x_i\}_{i=1}^m \sim \mathbf{P}$, $\{y_i\}_{i=1}^m \sim \mathbf{Q}$. Are $\mathbf{P}$ and $\mathbf{Q}$ different?

A "witness" function

$\mathbf{P} = \mathbf{Q}$ if and only if $\mathbb{E}_{X\sim\mathbf{P}} f(X) = \mathbb{E}_{Y\sim\mathbf{Q}} f(Y)$ for all *bounded continuous* functions f

Two-sample problem

- We are given $\{x_i\}_{i=1}^m \sim \mathbf{P}$, $\{y_i\}_{i=1}^m \sim \mathbf{Q}$. Are $\mathbf{P}$ and $\mathbf{Q}$ different?

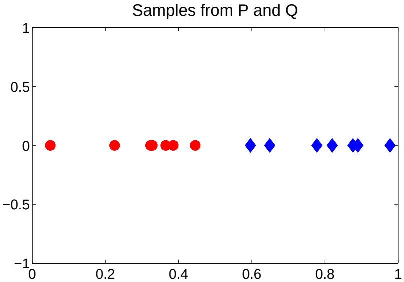
A "witness" function

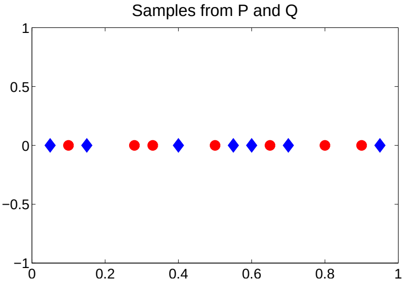
$\mathbf{P} = \mathbf{Q}$ if and only if $\mathbb{E}_{X \sim \mathbf{P}} f(X) = \mathbb{E}_{Y \sim \mathbf{Q}} f(Y)$ for all *bounded continuous* functions f

- true for many other (sufficiently rich) classes of functions

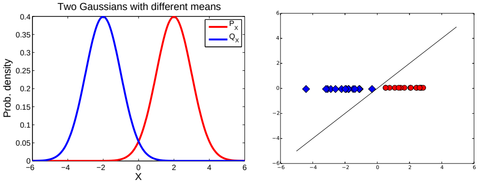
Two-sample problem

- We are given $\{x_i\}_{i=1}^m \sim \mathbf{P}$, $\{y_i\}_{i=1}^m \sim \mathbf{Q}$. Are $\mathbf{P}$ and $\mathbf{Q}$ different?

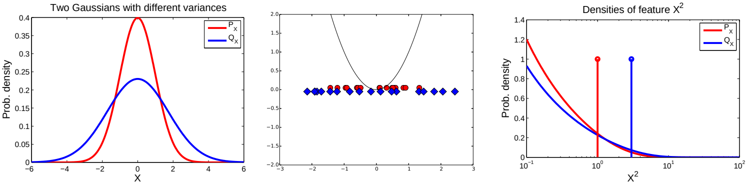




Difference in means of features

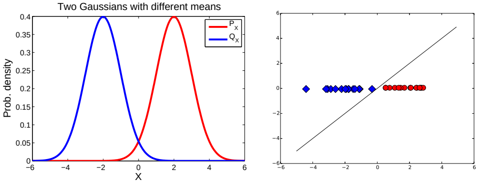


feature $\varphi(x) = x$

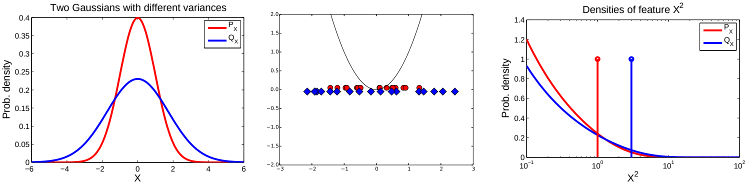


feature $\varphi(x) = x^2$

Difference in means of features



feature $\varphi(x) = x$

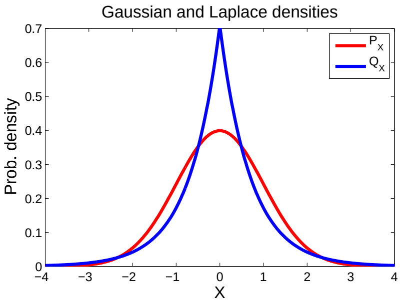


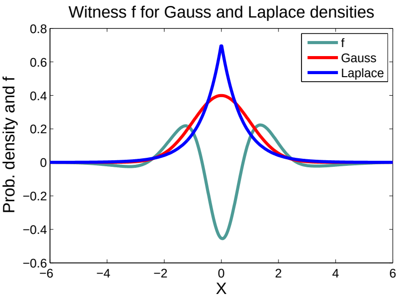
feature $\varphi(x) = x^2$

In general, given a feature $\varphi(x)$, find $\|\mathbb{E}_P\varphi(X) - \mathbb{E}_Q\varphi(X)\|$.

Difference in means of features

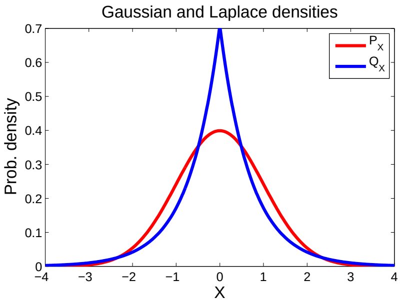
- Difference in means of higher order features

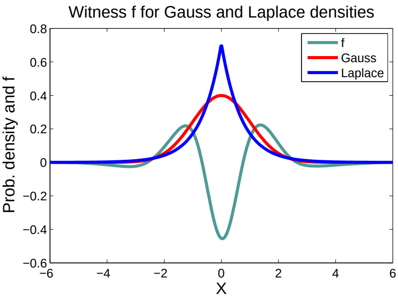




Difference in means of features

- Difference in means of higher order features



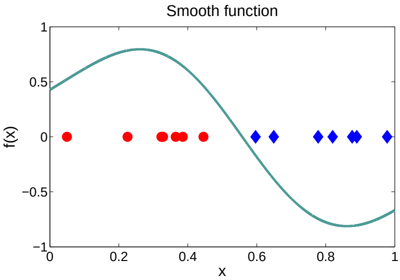


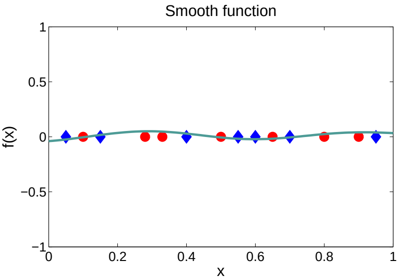
A systematic way to discover the appropriate features which distinguish distributions?

Functions Showing Difference in Distributions

- Maximum mean discrepancy: find a **smooth function** that best distinguishes **P** vs. **Q**:

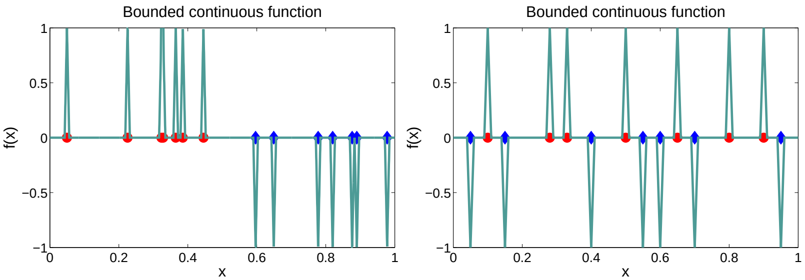
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbb{E}_{X \sim \mathbf{P}} f(X) - \mathbb{E}_{Y \sim \mathbf{Q}} f(Y) \right]$$





Function Showing Difference in Distributions

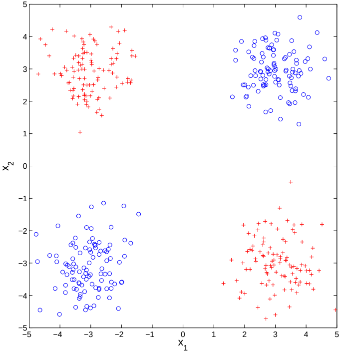
- What if the "witness" is not smooth?

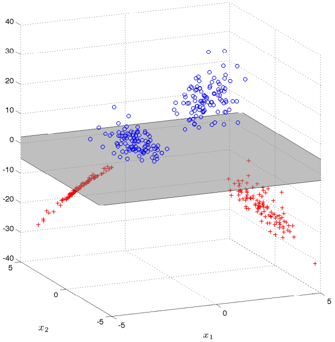


- Not useful for distinguishing distributions on the basis of samples! A smoothness constraint is required.

Outline

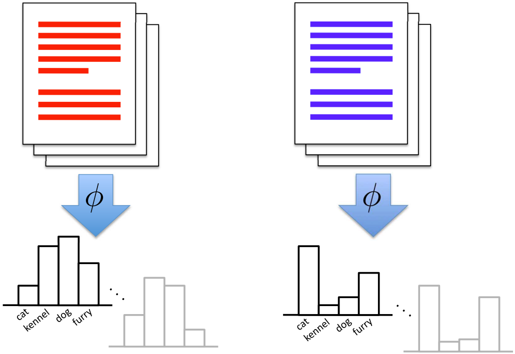
Why kernel methods (1): XOR example





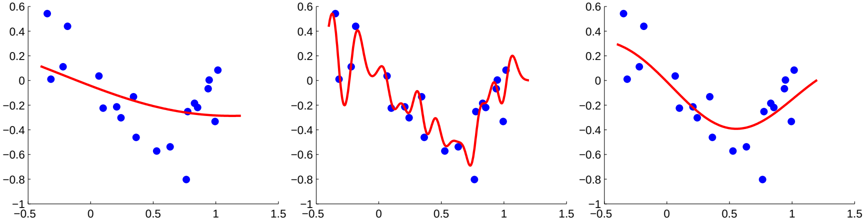
- No linear separation exists in the original space $\mathbb{R}^2$, but it does after feature map $\phi(x) = \begin{bmatrix} x_1 & x_2 & x_1 x_2 \end{bmatrix} \in \mathbb{R}^3$
- kernel methods allow **rich feature space representations.**

Why kernel methods (2): document classification



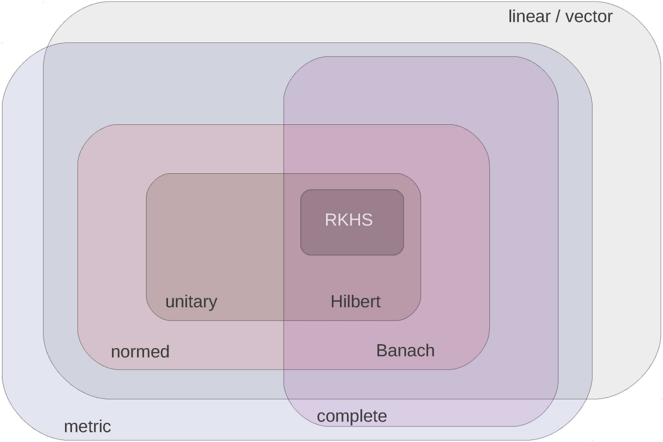
Kernels let us compare **complex data objects** on the basis of **features.**

Why kernel methods (3): smoothing



Kernel methods can control **smoothness** and **avoid overfitting/underfitting**.

RKHS: a function space with a very special structure



Evaluation functional

Definition (Evaluation functional)

Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, map $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x : f \mapsto f(x)$ is called the (Dirac) evaluation functional at x.

- Are evaluation functionals continuous?

Discontinuous evaluation

$\mathcal{F}$: the space of polynomials over $[0,1]$, endowed with the L_p norm, i.e.,

$$\|f_1 - f_2\|_p = \left(\int_0^1 |f_1(x) - f_2(x)|^p \, dx \right)^{1/p}.$$

Consider the sequence of functions $\{q_n\}_{n=1}^{\infty}$, where $q_n = x^n$. Then: $\lim_{n\to\infty} \|q_n - 0\|_p = 0$, i.e., $\{q_n\}$ converges to "zero function" in L_p norm, but does not get close to zero function everywhere:

$$1 = \lim_{n\to\infty} \delta_1(q_n) \neq \delta_1(\lim_{n\to\infty} q_n) = 0.$$

Discontinuous evaluation

$\mathcal{F}$: the space of polynomials over $[0,1]$, endowed with the L_p norm, i.e.,

$$\|f_1 - f_2\|_p = \left(\int_0^1 |f_1(x) - f_2(x)|^p \, dx \right)^{1/p}.$$

Consider the sequence of functions $\{q_n\}_{n=1}^{\infty}$, where $q_n = x^n$. Then: $\lim_{n\to\infty} \|q_n - 0\|_p = 0$, i.e., $\{q_n\}$ converges to "zero function" in L_p norm, but does not get close to zero function everywhere:

$$1 = \lim_{n\to\infty} \delta_1(q_n) \neq \delta_1(\lim_{n\to\infty} q_n) = 0.$$

$$\delta_1 : f \mapsto f(1) \text{ is not continuous!}$$

RKHS

Definition (Reproducing kernel Hilbert space)

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if all evaluation functionals are continuous on $\mathcal{H}$.

RKHS

Definition (Reproducing kernel Hilbert space)

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if all evaluation functionals are continuous on $\mathcal{H}$.

Theorem (Norm convergence implies pointwise convergence)

If $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$, *then* $\lim_{n\to\infty} f_n(x) = f(x)$, $\forall x \in \mathcal{X}$.

RKHS

Definition (Reproducing kernel Hilbert space)

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if all evaluation functionals are continuous on $\mathcal{H}$.

Theorem (Norm convergence implies pointwise convergence)

If $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$, *then* $\lim_{n\to\infty} f_n(x) = f(x)$, $\forall x \in \mathcal{X}$.

If two functions $f, g \in \mathcal{H}$ are close in the norm of $\mathcal{H}$, then $f(x)$ and $g(x)$ are close for all $x \in \mathcal{X}$

RKHS

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, and
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space* (RKHS).

RKHS

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if:

1. $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$, and
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.

If $\mathcal{H}$ has a reproducing kernel, it is said to be *a reproducing kernel Hilbert space* (RKHS).

In particular, for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle k(\cdot, y), k(\cdot, x)\rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}}.$$

Feature map and kernel trick

- **A "nonlinear method"**: a linear method in a transformed space: $z \mapsto (\varphi_1(z), \ldots \varphi_s(z)) \in \mathbb{R}^s$

Feature map and kernel trick

- **A "nonlinear method"**: a linear method in a transformed space: $z \mapsto (\varphi_1(z), \ldots \varphi_s(z)) \in \mathbb{R}^s$
  - only feature-space inner product computations: $k(z, z') = \sum_{i=1}^{s} \varphi_i(z)\varphi_i(z')$

Feature map and kernel trick

- **A "nonlinear method"**: a linear method in a transformed space: $z \mapsto (\varphi_1(z), \ldots \varphi_s(z)) \in \mathbb{R}^s$
  - only feature-space inner product computations: $k(z, z') = \sum_{i=1}^{s} \varphi_i(z)\varphi_i(z')$
- **A kernel method**: Hilbert-space valued features: $\varphi(z) \in \mathcal{H}$

Feature map and kernel trick

- **A "nonlinear method"**: a linear method in a transformed space: $z \mapsto (\varphi_1(z), \ldots \varphi_s(z)) \in \mathbb{R}^s$
  - only feature-space inner product computations: $k(z, z') = \sum_{i=1}^{s} \varphi_i(z)\varphi_i(z')$
- **A kernel method**: Hilbert-space valued features: $\varphi(z) \in \mathcal{H}$
  - canonical feature: $\varphi : z \mapsto k(\cdot, z)$, $k(z, z') = \langle k(\cdot, z), k(\cdot, z') \rangle_{\mathcal{H}}$ (**kernel trick**: no need for explicit coordinates)

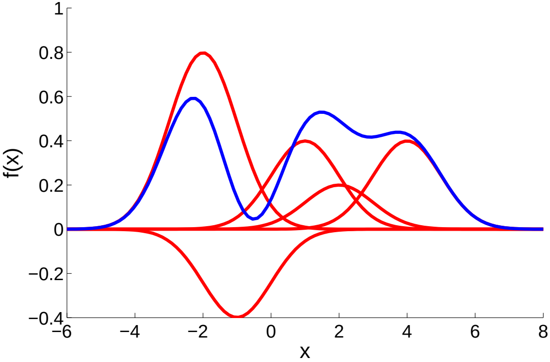
Feature map and kernel trick

- **A "nonlinear method"**: a linear method in a transformed space: $z \mapsto (\varphi_1(z), \ldots \varphi_s(z)) \in \mathbb{R}^s$
  - only feature-space inner product computations: $k(z, z') = \sum_{i=1}^{s} \varphi_i(z)\varphi_i(z')$
- **A kernel method**: Hilbert-space valued features: $\varphi(z) \in \mathcal{H}$
  - canonical feature: $\varphi : z \mapsto k(\cdot, z)$, $k(z, z') = \langle k(\cdot, z), k(\cdot, z') \rangle_{\mathcal{H}}$ (**kernel trick**: no need for explicit coordinates)
  - **Moore-Aronszajn Theorem**: every symmetric psd $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is a reproducing kernel and has a unique RKHS $\mathcal{H}_k$.

Moore-Aronszajn Theorem

$\mathcal{H}_k = \overline{span\left\{k(\cdot,x) \mid x \in \mathcal{X}\right\}}$ includes functions of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$



Potentially infinite-dimensional feature space

Under certain conditions (cf. Mercer's theorem), we can write

$$k(x,x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x'), \qquad \int_{\mathcal{X}} e_i(x) e_j(x) d\mu(x) = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

where this sum is guaranteed to converge whatever the x and x'.
Infinite-dimensional feature map can then be identified with a sequence:

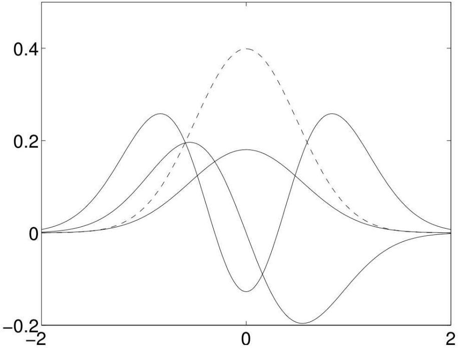
$$\varphi(x) = \begin{bmatrix} \vdots \\ \sqrt{\lambda_i} e_i(x) \\ \vdots \end{bmatrix} \in \ell_2$$

Smoothness interpretation

Gaussian kernel, $k(x,y) = \exp\left(-\sigma \|x-y\|^2\right)$,

$$\begin{aligned} \lambda_j &\propto b^j \qquad b < 1 \\ e_j(x) &\propto \exp(-(c-a)x^2)H_j(x\sqrt{2c}), \end{aligned}$$

a, b, c are functions of σ, and H_j is jth order Hermite polynomial.



NOTE that $\|f\|_{\mathcal{H}_k} < \infty$ is a "smoothness" constraint: λ_j decay as e_j become "rougher" and

$$\|f\|^2_{\mathcal{H}_k} = \sum_{j \in J} \frac{a_j^2}{\lambda_j}$$

(Figure from Rasmussen and Williams)

Kernel Embedding

Definition (Kernel embedding)

Let k be a kernel on $\mathcal{Z}$, and $P \in \mathcal{M}_+^1(\mathcal{Z})$ a probability measure. The *kernel embedding* of P into the RKHS $\mathcal{H}_k$ is $\mu_k(P) \in \mathcal{H}_k$ such that $\int f(z)dP(z) = \langle f, \mu_k(P)\rangle_{\mathcal{H}_k}$ **for all** $f \in \mathcal{H}_k$.

Kernel Embedding

Definition (Kernel embedding)

Let k be a kernel on $\mathcal{Z}$, and $P \in \mathcal{M}_+^1(\mathcal{Z})$ a probability measure. The *kernel embedding* of P into the RKHS $\mathcal{H}_k$ is $\mu_k(P) \in \mathcal{H}_k$ such that $\int f(z)dP(z) = \langle f, \mu_k(P) \rangle_{\mathcal{H}_k}$ **for all** $f \in \mathcal{H}_k$.

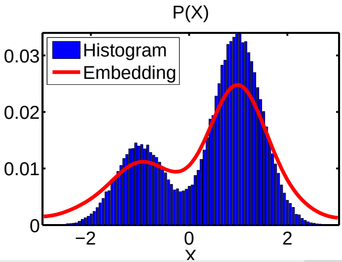
- If k is measurable and bounded, $\mu_k(P)$ exists for every P

Kernel Embedding

Definition (Kernel embedding)

Let k be a kernel on $\mathcal{Z}$, and $P \in \mathcal{M}_+^1(\mathcal{Z})$ a probability measure. The *kernel embedding* of P into the RKHS $\mathcal{H}_k$ is $\mu_k(P) \in \mathcal{H}_k$ such that $\int f(z)dP(z) = \langle f, \mu_k(P)\rangle_{\mathcal{H}_k}$ **for all** $f \in \mathcal{H}_k$.

- If k is measurable and bounded, $\mu_k(P)$ exists for every P
- Alternatively, can be defined as $\mu_k(P) = \mathbb{E}k(\cdot, Z) \in \mathcal{H}_k$ ("expected canonical feature").

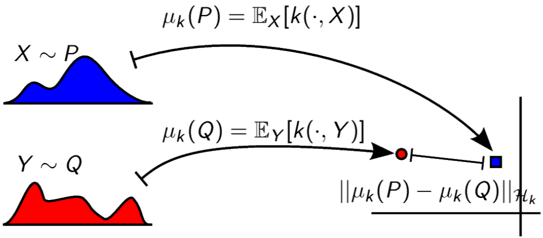


Kernel MMD

Definition

Kernel metric (MMD) between P and Q:

$$\begin{aligned}\mathrm{MMD}_k^2(P,Q) &= \|\mathbb{E}k(\cdot,X)-\mathbb{E}k(\cdot,Y)\|_{\mathcal{H}_k}^2\\ &= \mathbb{E}_{XX'}k(X,X')+\mathbb{E}_{YY'}k(Y,Y')-2\mathbb{E}_{XY}k(X,Y)\end{aligned}$$

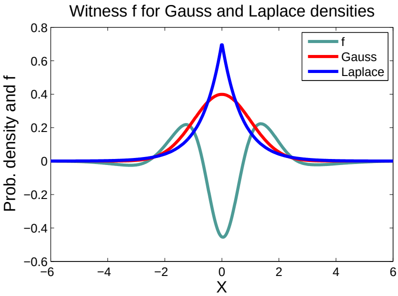


Kernel MMD

- An alternative interpretation of MMD is as an integral probability metric (Müller, 1997), i.e.,

$$\mathrm{MMD}_k(P, Q) = \sup_{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1} \left[\mathbb{E}_{Z \sim P} f(Z) - \mathbb{E}_{W \sim Q} f(W)\right].$$

- Supremum acheived at the "witness function" $f = (\mu_k(P) - \mu_k(Q)) \,/\, \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}$.



Kernel MMD

- A polynomial kernel $k(z,z') = \left(1 + z^\top z'\right)^s$ captures the difference in first s moments only

Kernel MMD

- A polynomial kernel $k(z, z') = \left(1 + z^\top z'\right)^s$ captures the difference in first s moments only
- For a certain family of kernels (**characteristic**): $\mathrm{MMD}_k(P, Q) = 0$ if and only if $P = Q$.

Kernel MMD

- A polynomial kernel $k(z,z')=\left(1+z^\top z'\right)^s$ captures the difference in first s moments only
- For a certain family of kernels (**characteristic**): $\mathrm{MMD}_k(P,Q)=0$ if and only if $P=Q$.
- Gaussian $\exp(-\gamma\left\|z-z'\right\|_2^2)$, Laplacian, inverse multiquadratics, B_{2n+1}- splines are all characteristic.

Kernel MMD

- A polynomial kernel $k(z, z') = \left(1 + z^\top z'\right)^s$ captures the difference in first s moments only
- For a certain family of kernels (**characteristic**): $\text{MMD}_k(P, Q) = 0$ if and only if $P = Q$.
- Gaussian $\exp(-\gamma \|z - z'\|_2^2)$, Laplacian, inverse multiquadratics, B_{2n+1}- splines are all characteristic.
- Under weak assumptions, k-MMD metrizes weak* topology on probability measures (Sriperumbudur, 2010):

$$\text{MMD}_k(P_n, P) \to 0 \Leftrightarrow P_n \xrightarrow{w} P$$

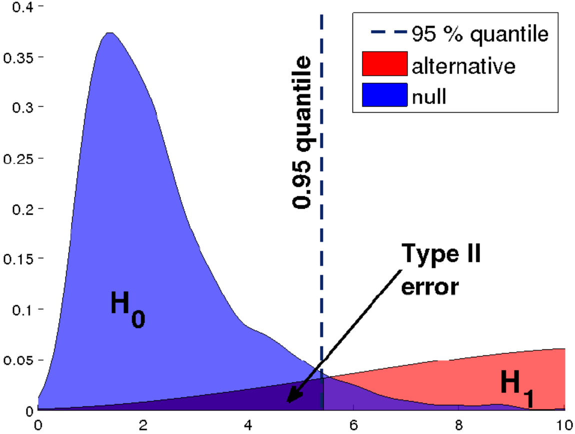
Nonparametric two-sample tests

- $\mathbf{H_0}$: $\mathbf{P} = \mathbf{Q}$ vs. $\mathbf{H_A}$: $\mathbf{P} \neq \mathbf{Q}$ based on samples $\{x_i\}_{i=1}^{n_x} \sim \mathbf{P}$, $\{y_i\}_{i=1}^{n_y} \sim \mathbf{Q}$.
- Test statistic (estimate of $\mathrm{MMD}_k^2(P, Q) = \mathbb{E}_{XX'} k(X, X') + \mathbb{E}_{YY'} k(Y, Y') - 2\mathbb{E}_{XY} k(X, Y)$)

$$\begin{aligned}
\mathrm{MMD}_k^2(\hat{P}, \hat{Q}) \quad = \quad & \frac{1}{n_x(n_x-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n_y(n_y-1)} \sum_{i \neq j} k(y_i, y_j) \\
& -\frac{2}{n_x n_y} \sum_{i,j} k(x_i, y_j).
\end{aligned}$$

Test threshold

- distribution under $\mathbf{H_0}$: $\mathbf{P} = \mathbf{Q}$:
$\frac{n_x n_y}{n_x + n_y} \mathrm{MMD}_k^2(\hat{P}, \hat{Q}) \xrightarrow{d} \sum_{r=1}^{\infty} \lambda_r \left(Z_r^2 - 1 \right), \quad \{Z_r\} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $\{\lambda_r\}$ depend on the kernel k and the underlying distribution $\mathbf{P}$



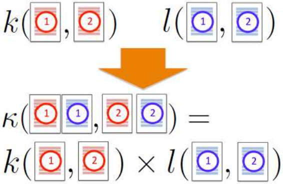
Non-parametric independence tests

- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y$ (null hypothesis)
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y$ (alternative hypothesis)

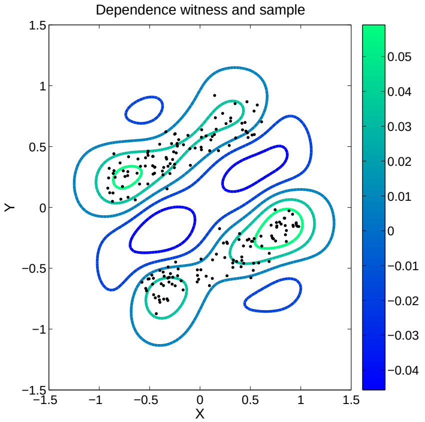
Non-parametric independence tests

- $\mathbf{H_0}$: $X \perp\!\!\!\perp Y \Leftrightarrow \mathrm{P}_{XY} = \mathrm{P}_X \mathrm{P}_Y$ **(null hypothesis)**
- $\mathbf{H_A}$: $X \not\perp\!\!\!\perp Y \Leftrightarrow \mathrm{P}_{XY} \neq \mathrm{P}_X \mathrm{P}_Y$ **(alternative hypothesis)**

- Test statistic:
  $$\mathrm{HSIC}(X, Y) = \left\| \mu_\kappa(\hat{P}_{XY}) - \mu_\kappa(\hat{P}_X \hat{P}_Y) \right\|^2_{\mathcal{H}_\kappa},$$
  with $\kappa = k \otimes l$
  Gretton et al (2005, 2008); Smola et al (2007)



HSIC as integral probability metric



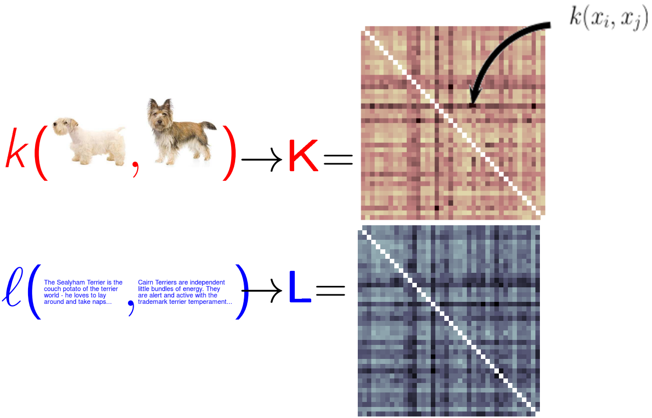
- $\|\mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y)\|_{\mathcal{H}_\kappa} = \sup_f \left[\mathbb{E}_{X,Y} f(X,Y) - \mathbb{E}_X \mathbb{E}_Y f(X,\right.$
- witness lies in the unit ball of $\mathcal{H}_\kappa$, the RKHS of functions on $\mathcal{X} \times \mathcal{Y}$

HSIC computation

$k($  ,  $)$

$\ell($ The Sealyham Terrier is the couch potato of the terrier world - he loves to lay around and take naps... , Cairn Terriers are independent little bundles of energy. They are alert and active with the trademark terrier temperament... $)$

HSIC computation

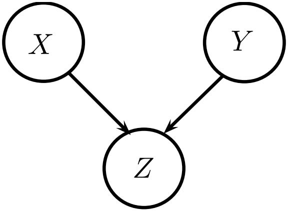


- **HSIC** measures similarity between the kernel matrices: $\mathrm{HSIC}(X, Y) = \frac{1}{n^2} \langle H\mathbf{K}H, H\mathbf{L}H \rangle$

Outline

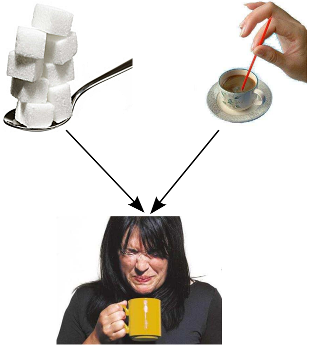
Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?



Detecting a higher order interaction

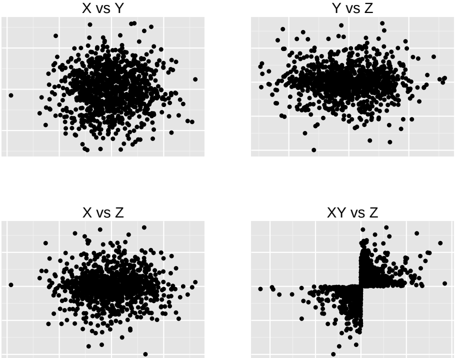
- How to detect V-structures with pairwise weak (or nonexistent) dependence?

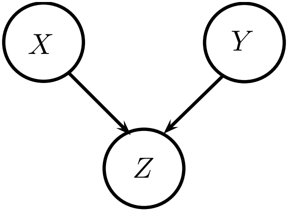


Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?

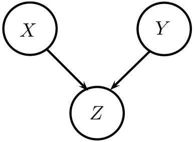
- $X \perp\!\!\!\perp Y$, $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z$





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

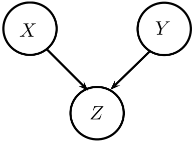
V-structure Discovery



Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

- CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ (Zhang et al 2011) or

V-structure Discovery



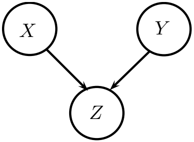
Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

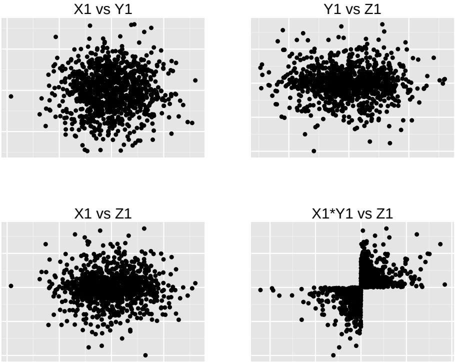
- CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ (Zhang et al 2011) or
- Factorisation test: $\mathbf{H_0} : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)
  - compute p-values for each of the marginal tests for $(Y, Z) \perp\!\!\!\perp X$, $(X, Z) \perp\!\!\!\perp Y$, or $(X, Y) \perp\!\!\!\perp Z$
  - apply Holm-Bonferroni (**HB**) sequentially rejective correction (Holm 1979)

V-structure Discovery (2)

- How to detect V-structures with pairwise weak (or nonexistent) dependence?

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



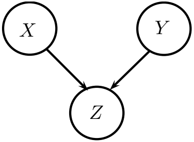


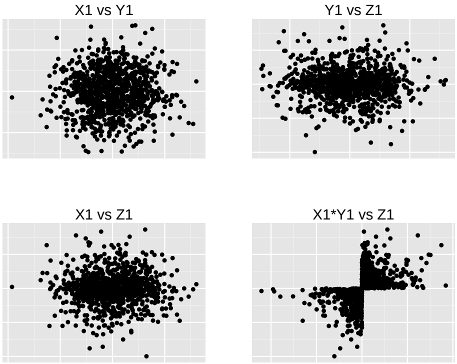
- $X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,
- $Z_1 | X_1, Y_1 \sim \mathrm{sign}(X_1 Y_1) Exp(\frac{1}{\sqrt{2}})$

V-structure Discovery (2)

- How to detect V-structures with pairwise weak (or nonexistent) dependence?

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





- $X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,
- $Z_1 | X_1, Y_1 \sim \text{sign}(X_1 Y_1) Exp(\frac{1}{\sqrt{2}})$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$

V-structure Discovery (3)

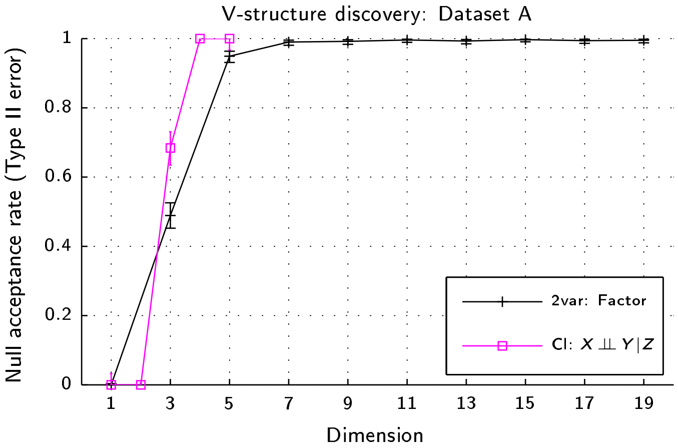


Figure : CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), and a factorisation test with a **HB** correction, $n = 500$

Lancaster Interaction Measure

Definition (Bahadur (1961); Lancaster (1969))

Interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

Lancaster Interaction Measure

Definition (Bahadur (1961); Lancaster (1969))

Interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

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

Lancaster Interaction Measure

Definition (Bahadur (1961); Lancaster (1969))

Interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

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

Lancaster Interaction Measure

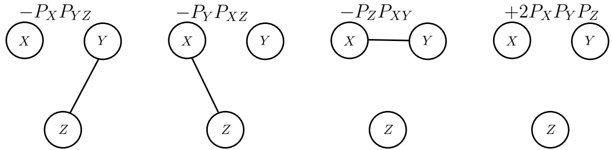
Definition (Bahadur (1961); Lancaster (1969))

Interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

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

$\Delta_L P =$

P_{XYZ}

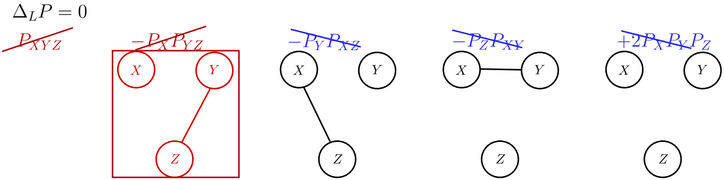


Lancaster Interaction Measure

Definition (Bahadur (1961); Lancaster (1969))

Interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

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



A Test using Lancaster Measure

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

$$\left\|\mu_\kappa(P_{XYZ} - P_{XY}P_Z - \cdots)\right\|^2_{\mathcal{H}_\kappa} =$$
$$\left\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ}\right\rangle_{\mathcal{H}_\kappa} - 2\left\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY}P_Z\right\rangle_{\mathcal{H}_\kappa} \cdots$$

Inner Product Estimators

$\nu\backslash\nu'$	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 $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$

Inner Product Estimators

$\nu\backslash\nu'$	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 $\langle\mu_\kappa\nu,\mu_\kappa\nu'\rangle_{\mathcal{H}_\kappa}$

Proposition (Lancaster interaction statistic)

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

Empirical joint central moment in the feature space

Example A: factorisation tests

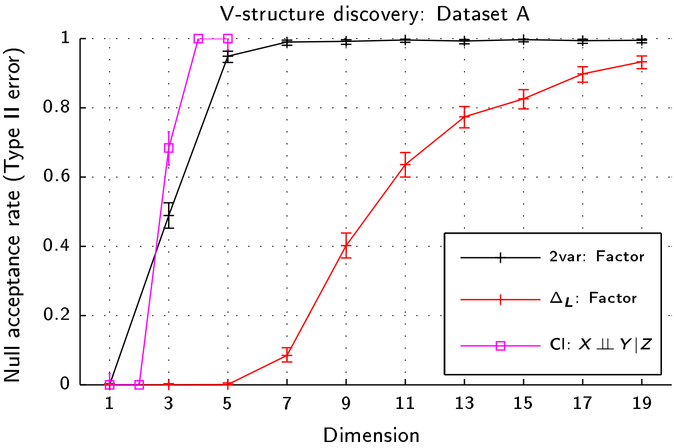


Figure : Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Outline

Computing estimates of MMD

- Write MMD as $\mathbb{E}_{XX'YY'} h(X, X', Y, Y')$, where
  $h(X, X', Y, Y') = k(X, X') + k(Y, Y') - k(X, Y') - k(X', Y)$
- Given i.i.d. samples $\mathbf{x} = \{x_i\}_{i=1}^m \sim P$ and $\mathbf{y} = \{y_i\}_{i=1}^m \sim Q$,
  - an estimator that needs $O(m^2)$ time to compute: U- or V-statistic
  - an estimator that needs $O(m)$ time to compute: a running average $\frac{2}{m} \sum_{i=1}^{m/2} h(x_{2i-1}, x_{2i}, y_{2i-1}, y_{2i})$

Computing estimates of MMD

- Write MMD as $\mathbb{E}_{XX'YY'}h(X, X', Y, Y')$, where
  $h(X, X', Y, Y') = k(X, X') + k(Y, Y') - k(X, Y') - k(X', Y)$
- Given i.i.d. samples $\mathbf{x} = \{x_i\}_{i=1}^{m} \sim P$ and $\mathbf{y} = \{y_i\}_{i=1}^{m} \sim Q$,
  - an estimator that needs $O(m^2)$ time to compute: U- or V-statistic
  - an estimator that needs $O(m)$ time to compute: a running average $\frac{2}{m}\sum_{i=1}^{m/2} h(x_{2i-1}, x_{2i}, y_{2i-1}, y_{2i})$

	U- or V-statistic	running average
time	$O(m^2)$	$O(m)$
storage	$O(m^2)$	$O(1)$
null distribution	infinite sum of chi-squares	normal
convergence rate	$1/m$	$1/\sqrt{m}$

Experiment: Gaussian blobs

Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

Experiment: Gaussian blobs

Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.
We distinguish grids of Gaussian blobs with different covariances.

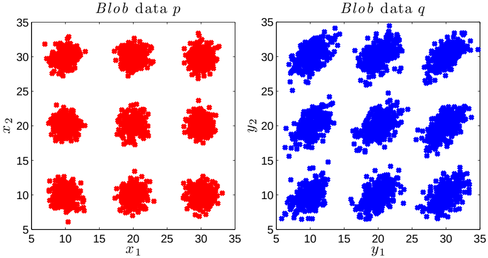


Figure : 3×3 blobs, ratio $\varepsilon = 3.2$ of largest-to-smallest eigenvalues of blobs in Q.

Experiment: Gaussian blobs (2)

12×12 blobs with $\varepsilon = 1.4$. Linear time statistic vs. Quadratic time statistic. Fixed kernel.

Experiment: Gaussian blobs (2)

12×12 blobs with $\varepsilon = 1.4$. Linear time statistic vs. Quadratic time statistic. Fixed kernel.

	m **per trial**	**Type II error**	**Trials**
Quadratic	5,000	$[0.7996, 0.8516]$	820
	10,000	$[0.5161, 0.6175]$	367
	$> 10,000$	Buy more RAM!	

Experiment: Gaussian blobs (2)

12×12 blobs with $\varepsilon = 1.4$. Linear time statistic vs. Quadratic time statistic. Fixed kernel.

	m **per trial**	**Type II error**	**Trials**
Quadratic	5,000	$[0.7996, 0.8516]$	820
	10,000	$[0.5161, 0.6175]$	367
	$> 10,000$	Buy more RAM!	
Linear	$\sim 100,000,000$	$[0.2250, 0.3049]$	468
	$\sim 200,000,000$	$[0.1873, 0.2829]$	302
	$\vdots$	$\vdots$	$\vdots$
	$\sim 500,000,000$	$\mathbf{0.0270 \pm 0.0302}$	111

Asymptotic efficiency criterion

- A. Gretton, B. Sriperumbudur, DS, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, in *Advances in Neural Information Processing Systems (NIPS) 25*, 2012.

Proposition

For given P and Q. Let $\eta_k = MMD_k^2(P, Q)$, *and let* σ_k^2 *be the asymptotic variance of the linear-time statistic. Then*

$$k_* = \arg\max_{k \in \mathcal{K}} \eta_k \sigma_k^{-1}$$

minimizes the asymptotic Type II error probability on $\mathcal{K}$.

Asymptotic efficiency criterion

- A. Gretton, B. Sriperumbudur, DS, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, in *Advances in Neural Information Processing Systems (NIPS) 25*, 2012.

Proposition

For given P and Q. Let $\eta_k = MMD_k^2(P, Q)$*, and let* σ_k^2 *be the asymptotic variance of the linear-time statistic. Then*

$$k_* = \arg\max_{k \in \mathcal{K}} \eta_k \sigma_k^{-1}$$

minimizes the asymptotic Type II error probability on $\mathcal{K}$*.*

- We only have estimates of η_k and σ_k (If we knew η_k, our problem would have been solved)!
  - Will the kernel optimization using these esimates be consistent? yes!
  - Over what families of kernels can we perform such optimization *efficiently*? linear combinations (MKL)

Experiment: Gaussian blobs (3)

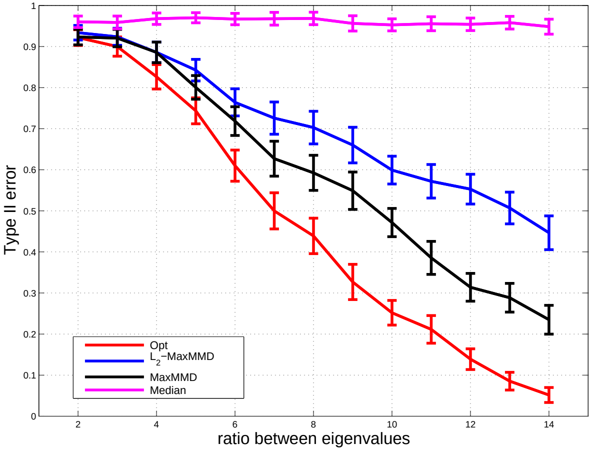


Figure : $m = 10,000$; family generated by gaussian kernels with bandwiths $\{2^{-5}, \ldots, 2^{15}\}$.

Summary

- A kernel selection criterion to explicitly optimize the (Hodges and Lehmann) asymptotic relative efficiency

Summary

- A kernel selection criterion to explicitly optimize the (Hodges and Lehmann) asymptotic relative efficiency
- Shown consistency of a regularized empirical criterion, which can be solved by a quadratic program

Summary

- A kernel selection criterion to explicitly optimize the (Hodges and Lehmann) asymptotic relative efficiency
- Shown consistency of a regularized empirical criterion, which can be solved by a quadratic program
- Both optimization and testing are performed with computational cost linear in the sample size

Outline

Distance covariance (dCov)

(Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009; Lyons 2011)

- For random vectors X and Y, dCov $\mathcal{V}^2(X,Y)$ is the **weighted L_2-norm of $f_{XY} - f_X f_Y$**:

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|X-X'\|_2\|Y-Y'\|_2 \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\|X-X'\|_2\,\mathbb{E}_Y\mathbb{E}_{Y'}\|Y-Y'\|_2 \\ &\quad - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\|X-X'\|_2\,\mathbb{E}_{Y'}\|Y-Y'\|_2\right],\end{aligned}$$

where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$.

Distance covariance (dCov)

(Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009; Lyons 2011)

- For random vectors X and Y, dCov $\mathcal{V}^2(X,Y)$ is the **weighted L_2-norm of $f_{XY} - f_X f_Y$**:

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\left\|X-X'\right\|_2\left\|Y-Y'\right\|_2 \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_Y\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2 \\ &\quad - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2\right],\end{aligned}$$

  where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$.
- $\mathcal{V}^2(X,Y) = 0$ if and only if X and Y are independent

Distance covariance (dCov)

(Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009; Lyons 2011)

- For random vectors X and Y, dCov $\mathcal{V}^2(X,Y)$ is the **weighted L_2-norm of $f_{XY} - f_X f_Y$:**

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\left\|X-X'\right\|_2\left\|Y-Y'\right\|_2 \\ &\quad + \mathbb{E}_X\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_Y\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2 \\ &\quad - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\left\|X-X'\right\|_2\mathbb{E}_{Y'}\left\|Y-Y'\right\|_2\right],\end{aligned}$$

  where (X,Y) and (X',Y') are $\overset{i.i.d.}{\sim} P_{XY}$.
- $\mathcal{V}^2(X,Y) = 0$ if and only if X and Y are independent
- Simon and Tibshirani (2011) show that dCov results in more powerful tests than MIC in 1D

dCov vs. MMD

- DS, A. Gretton, B. Sriperumbudur and K. Fukumizu, **Hypothesis testing using pairwise distances and associated kernels**, in *Proc. International Conference on Machine Learning ICML*, 2012

Theorem

dCov is MMD with $k(x,x') = \frac{1}{2}\left[\|x\|_2 + \|x'\|_2 - \|x - x'\|_2\right]$*, and* $l(y,y') = \frac{1}{2}\left[\|y\|_2 + \|y'\|_2 - \|y - y'\|_2\right]$.

dCov vs. MMD

- DS, A. Gretton, B. Sriperumbudur and K. Fukumizu, **Hypothesis testing using pairwise distances and associated kernels**, in *Proc. International Conference on Machine Learning ICML*, 2012

Theorem

dCov is MMD with $k(x,x') = \frac{1}{2}\left[\|x\|_2 + \|x'\|_2 - \|x-x'\|_2\right]$*, and* $l(y,y') = \frac{1}{2}\left[\|y\|_2 + \|y'\|_2 - \|y-y'\|_2\right]$.

- Series of examples that demonstrate that a more powerful test can be achieved with $k(x,x') = \frac{1}{2}\left[\|x\|_2^\alpha + \|x'\|_2^\alpha - \|x-x'\|_2^\alpha\right]$.

dCov vs. MMD (2)

- DS, B. Sriperumbudur, A. Gretton and K. Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**, *Annals of Statistics*, 2013
- When generalized to semimetric spaces of negative type (ensures $\mathcal{V}^2(X, Y) \geq 0$), dCov and MMD approaches are equivalent.

dCov vs. MMD (2)

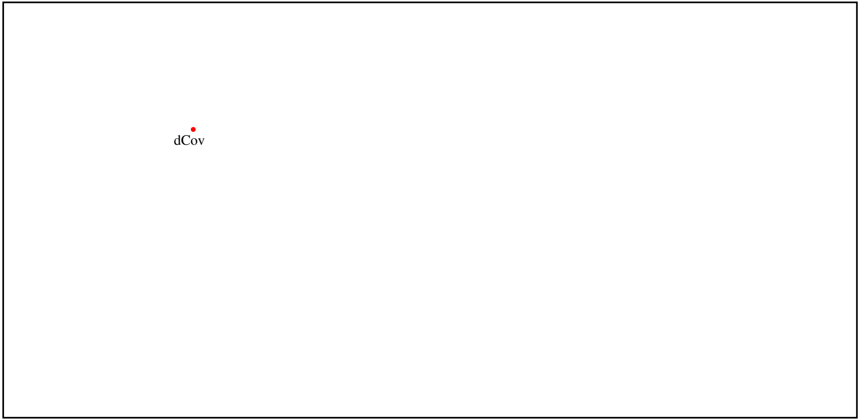
- DS, B. Sriperumbudur, A. Gretton and K. Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**, *Annals of Statistics*, 2013
- When generalized to semimetric spaces of negative type (ensures $\mathcal{V}^2(X, Y) \geq 0$), dCov and MMD approaches are equivalent.

Theorem

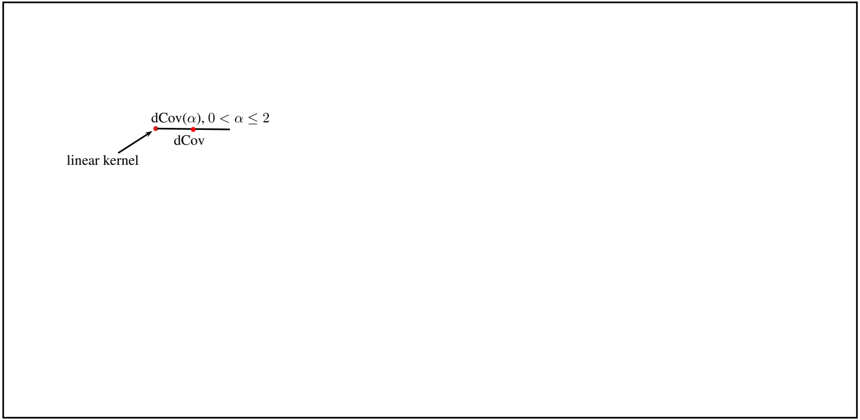
Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semimetric spaces of negative type, and let k and l be any two kernels on $\mathcal{X}$ and $\mathcal{Y}$ that generate $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$, respectively, and define $\kappa = k \otimes l$. Then, if $(X, Y) \sim P_{XY}$, with marginals $P_X \in \mathcal{M}_k^2(\mathcal{X})$, $P_Y \in \mathcal{M}_l^2(\mathcal{Y})$

$$\mathcal{V}^2_{\rho_{\mathcal{X}}, \rho_{\mathcal{Y}}}(X, Y) = 4 \left\| \mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y) \right\|^2_{\mathcal{H}_\kappa}.$$

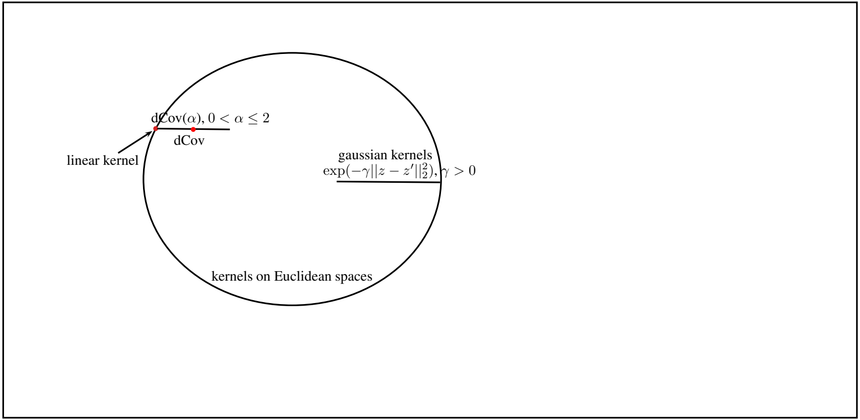
Dependence measure Zoo



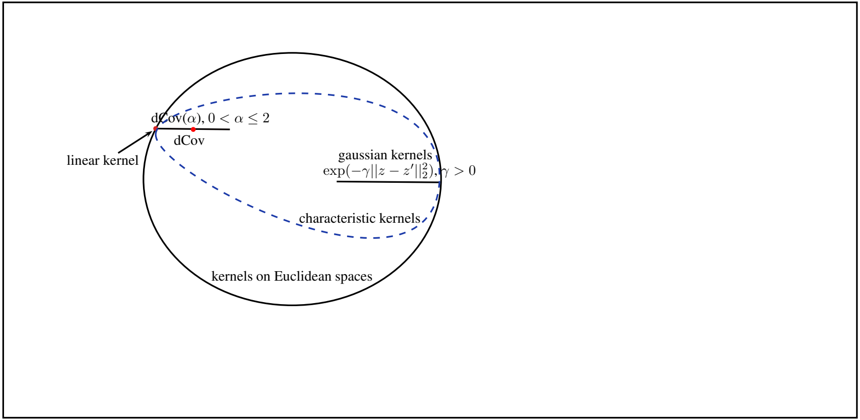
Dependence measure Zoo



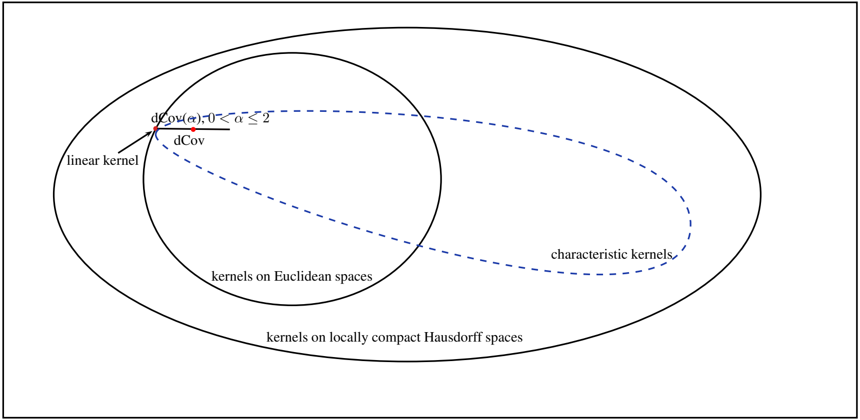
Dependence measure Zoo



Dependence measure Zoo



Dependence measure Zoo



Summary

- Distance-based statistics of Szekely et al are a special case of the RKHS framework.

Summary

- Distance-based statistics of Szekely et al are a special case of the RKHS framework.
- Conversely, RKHS-based statistics have a clear interpretation in terms of implicitly imposing a negative type (semi)metric onto the original space.

Summary

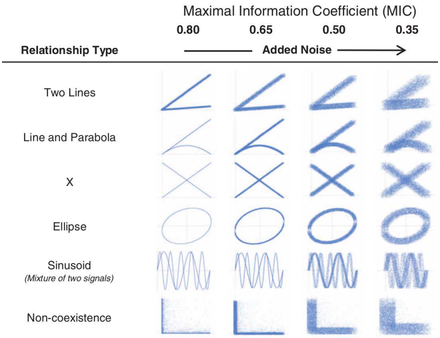
- Distance-based statistics of Szekely et al are a special case of the RKHS framework.
- Conversely, RKHS-based statistics have a clear interpretation in terms of implicitly imposing a negative type (semi)metric onto the original space.
- A new way to estimate the null distribution of distance-based statistics through the link with kernels. A new way to construct characteristic kernels.

References

- A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf and A. Smola, **A kernel two-sample test**, *Journal of Machine Learning Research*, 13:723-773, 2012.
- A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil and K. Fukumizu, **Optimal kernel choice for large-scale two-sample tests**, in *Advances in Neural Information Processing Systems (NIPS) 25*, 2012.
- D. Sejdinovic, A. Gretton and W. Bergsma, **A kernel test for three-variable interactions**, *Advances in Neural Information Processing Systems (NIPS)* 26, 2013.
- D. Sejdinovic, B. Sriperumbudur, A. Gretton and K. Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**, *Ann. Statist.* 41(5), pp 2263-2291, Oct. 2013.
- G. Székely and M. Rizzo, **Brownian distance covariance**. *Ann. Appl. Statist.*, 4(3):1233–1303, 2009.

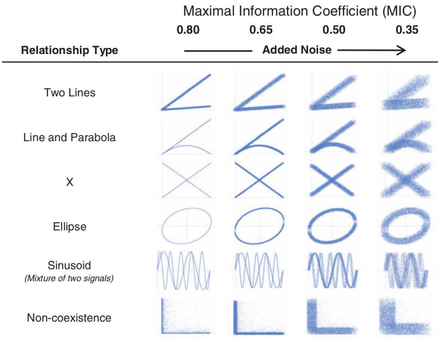
1D case: MIC, dCov

- Reshef et al, **Detecting Novel Associations in Large Data Sets**. *Science*, 334: 1518–1524, 2011.



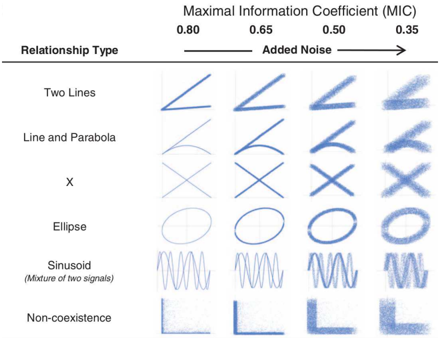
1D case: MIC, dCov

- Reshef et al, **Detecting Novel Associations in Large Data Sets**. *Science*, 334: 1518–1524, 2011.
- MIC vs. dCov: Simon and Tibshirani, **Comment on *Reshef et. al.*** (*Science*)



1D case: MIC, dCov

- Reshef et al, **Detecting Novel Associations in Large Data Sets**. *Science*, 334: 1518–1524, 2011.
- MIC vs. dCov: Simon and Tibshirani, **Comment on *Reshef et. al.*** (*Science*)
- Székely and Rizzo, **Brownian distance covariance** (with discussion). *Ann. Appl. Statist.*, 4:1233–1303, 2009.



1D case: MIC, dCov

- Reshef et al, **Detecting Novel Associations in Large Data Sets**. *Science*, 334: 1518–1524, 2011.
- MIC vs. dCov: Simon and Tibshirani, **Comment on *Reshef et. al.*** (*Science*)
- Székely and Rizzo, **Brownian distance covariance** (with discussion). *Ann. Appl. Statist.*, 4:1233–1303, 2009.
- dCov - a special case of a kernel dependence measure: DS, Sriperumbudur, Gretton and Fukumizu, **Equivalence of distance-based and RKHS-based statistics in hypothesis testing**, *Ann. Statist.*, 2013.