Lecture 3: Dependence measures using RKHS embeddings

MLSS Cadiz, 2016

Arthur Gretton

Gatsby Unit, CSML, UCL

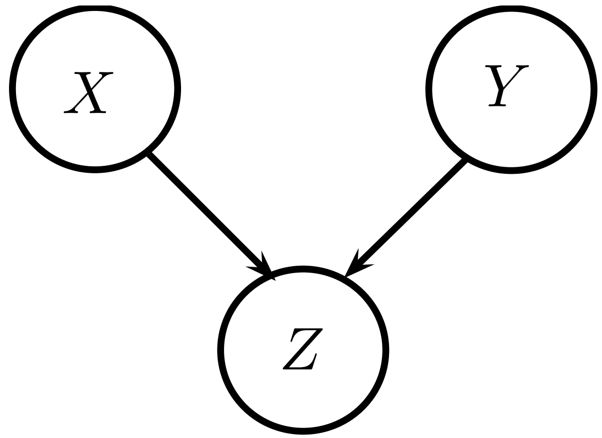
Outline

- Three or more variable interactions, comparison with conditional dependence testing [Sejdinovic et al., 2013a]
- Dependence detection in detail, covariance operators
- Choice of kernel to maximise test power Gretton et al. [2012b]
- Supervised learning with distributions as inputs Jitkrittum et al. [2015], Szabó et al. [2015]
- Recent work (2014/2015) (not in this talk, see my webpage)
  - Testing for time series Chwialkowski and Gretton [2014], Chwialkowski et al. [2014]
  - Infinite dimensional exponential families Sriperumbudur et al. [2014]
  - Adaptive MCMC, and adaptive Hamiltonian Monte Carlo Sejdinovic et al. [2014], Strathmann et al. [2015]

Lancaster (3-way) Interactions

Detecting a higher order interaction

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

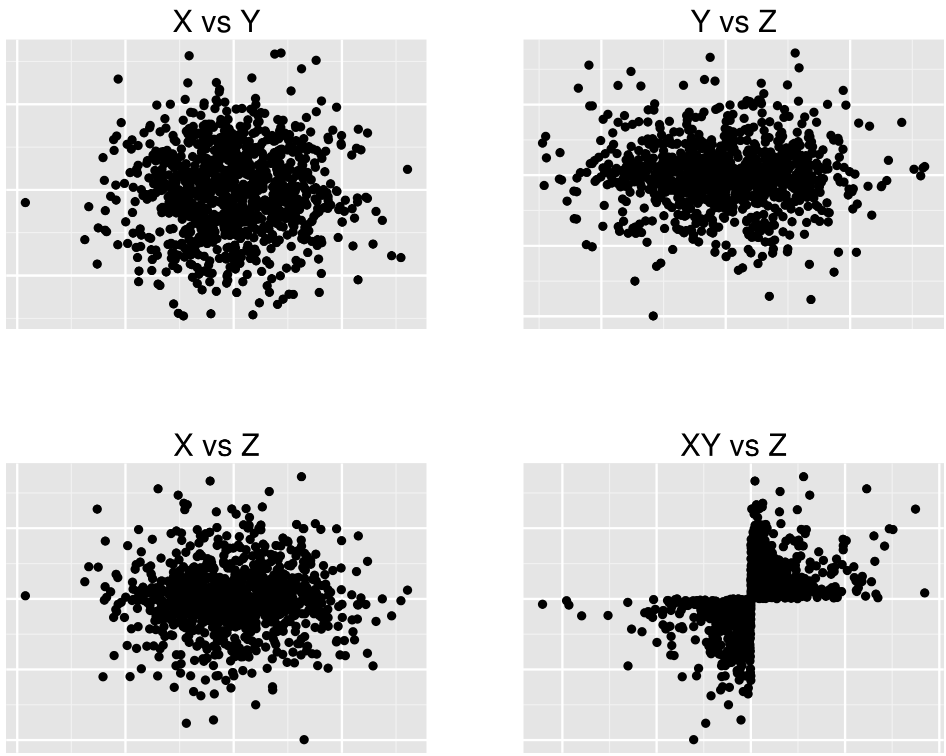


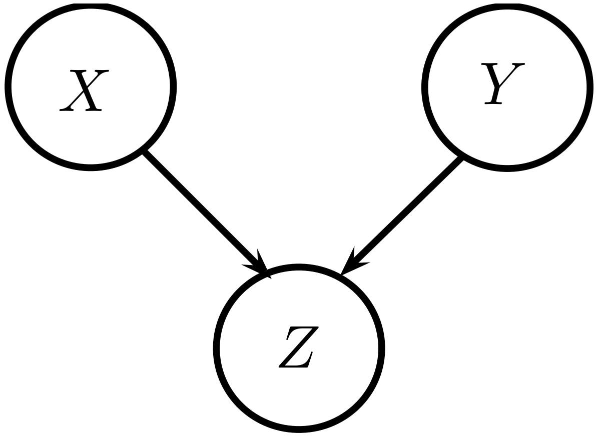
Detecting a higher order interaction

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

Detecting a higher order interaction

- How to detect V-structures with pairwise weak individual dependence?
- $X \perp\!\!\!\perp Y$, $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z$

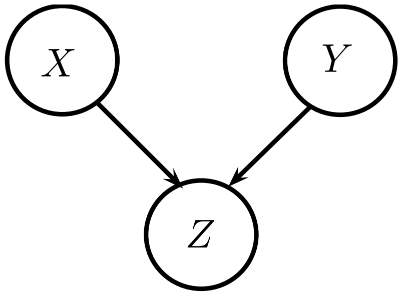




- $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$,
- $Z| \; X, Y \sim \text{sign}(XY) Exp(\frac{1}{\sqrt{2}})$

Faithfulness violated here

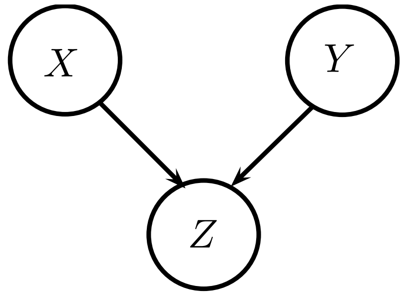
V-structure Discovery



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], or

V-structure Discovery

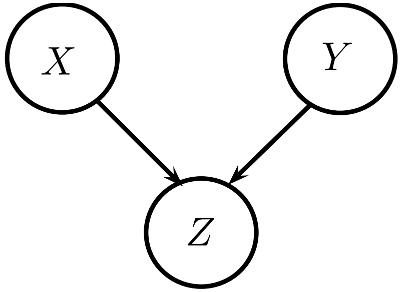


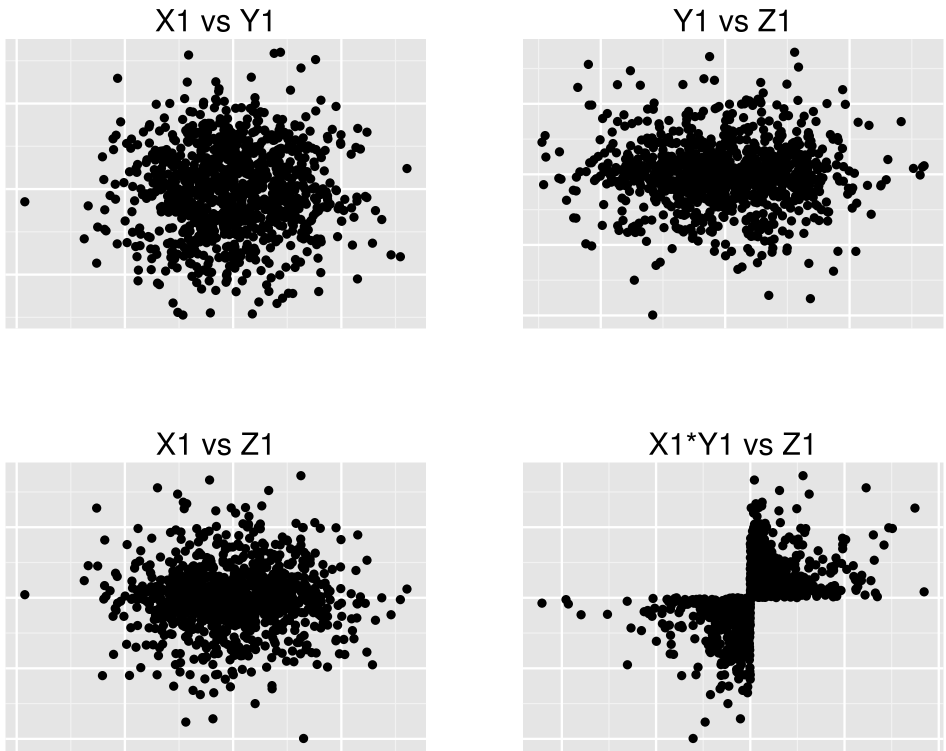
Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

- **Consistent** CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ [Fukumizu et al., 2008, Zhang et al., 2011], 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?
- $X \perp\!\!\!\perp Y$, $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z$





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

V-structure Discovery (3)

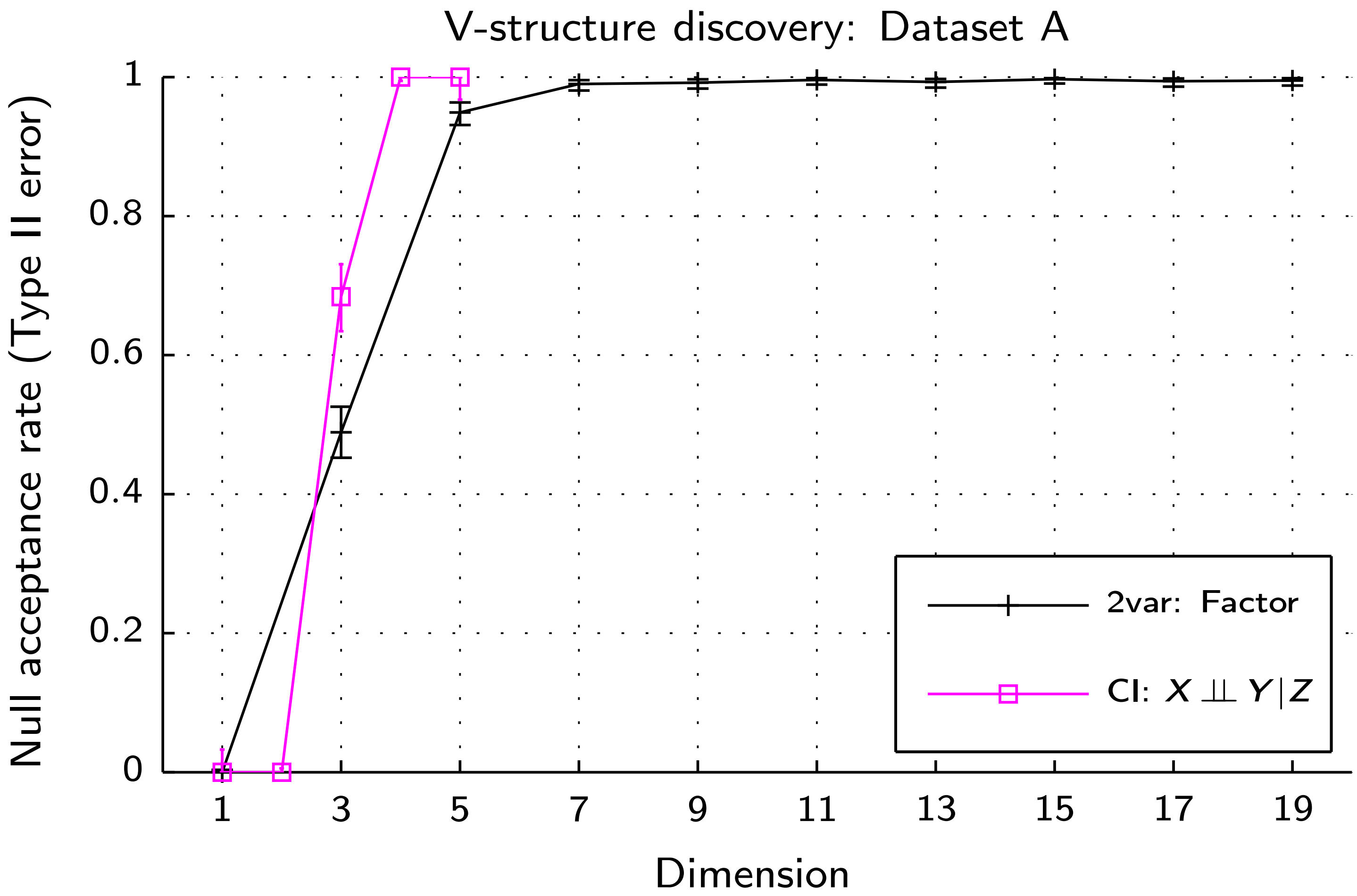


Figure 1: CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), and a factorisation test with a **HB** correction, $n = 500$

Lancaster Interaction Measure

[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

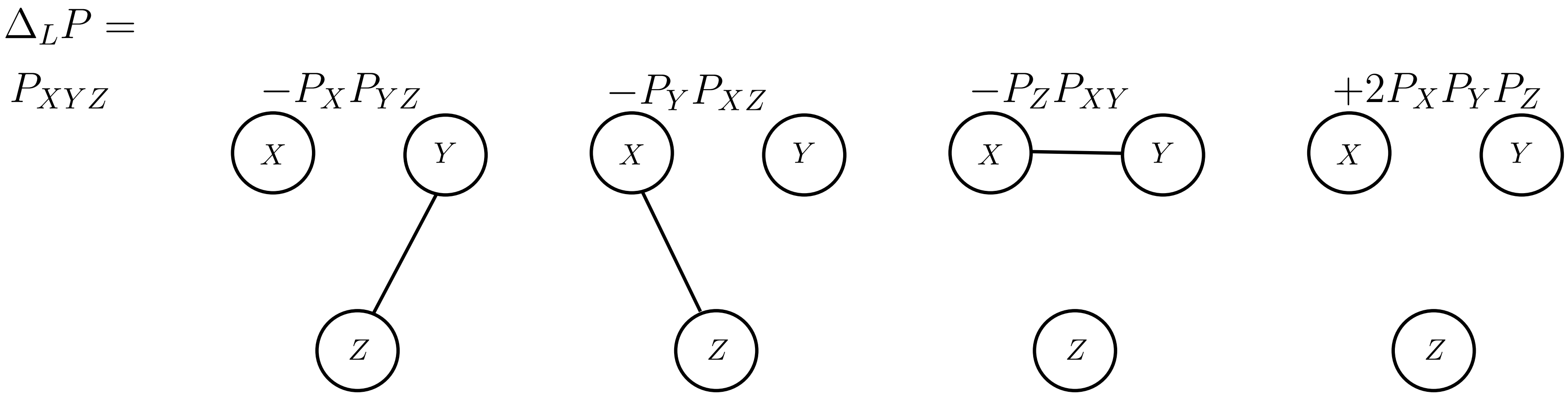
[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

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

Lancaster Interaction Measure

[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.

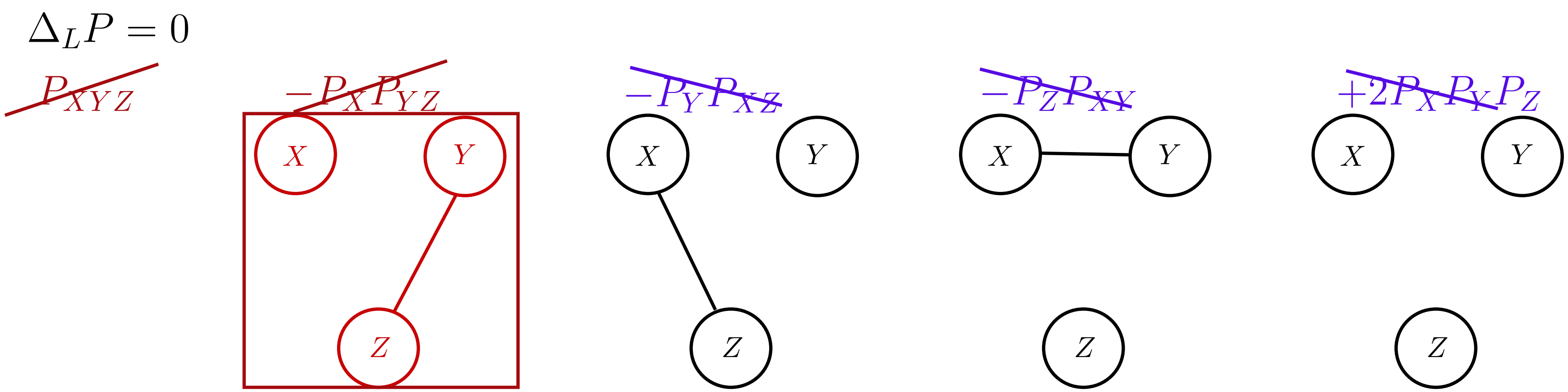
- $D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$



Lancaster Interaction Measure

[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.

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



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

Lancaster Interaction Measure

[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.

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

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

...so what might be missed?

Lancaster Interaction Measure

[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.

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

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

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 Test using Lancaster Measure

- Test statistic is empirical estimate of $\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa}$, where $\kappa = k \otimes l \otimes m$:

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

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 1: V-statistic estimators of $\left\langle \mu_\kappa \nu, \mu_\kappa \nu' \right\rangle_{\mathcal{H}_\kappa}$

Inner Product Estimators

$\nu\backslash\nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_X P_Y P_Z$
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 2: V-statistic estimators of $\left\langle \mu_\kappa\nu, \mu_\kappa\nu'\right\rangle_{\mathcal{H}_\kappa}$

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

Empirical joint central moment in the feature space

Example A: factorisation tests

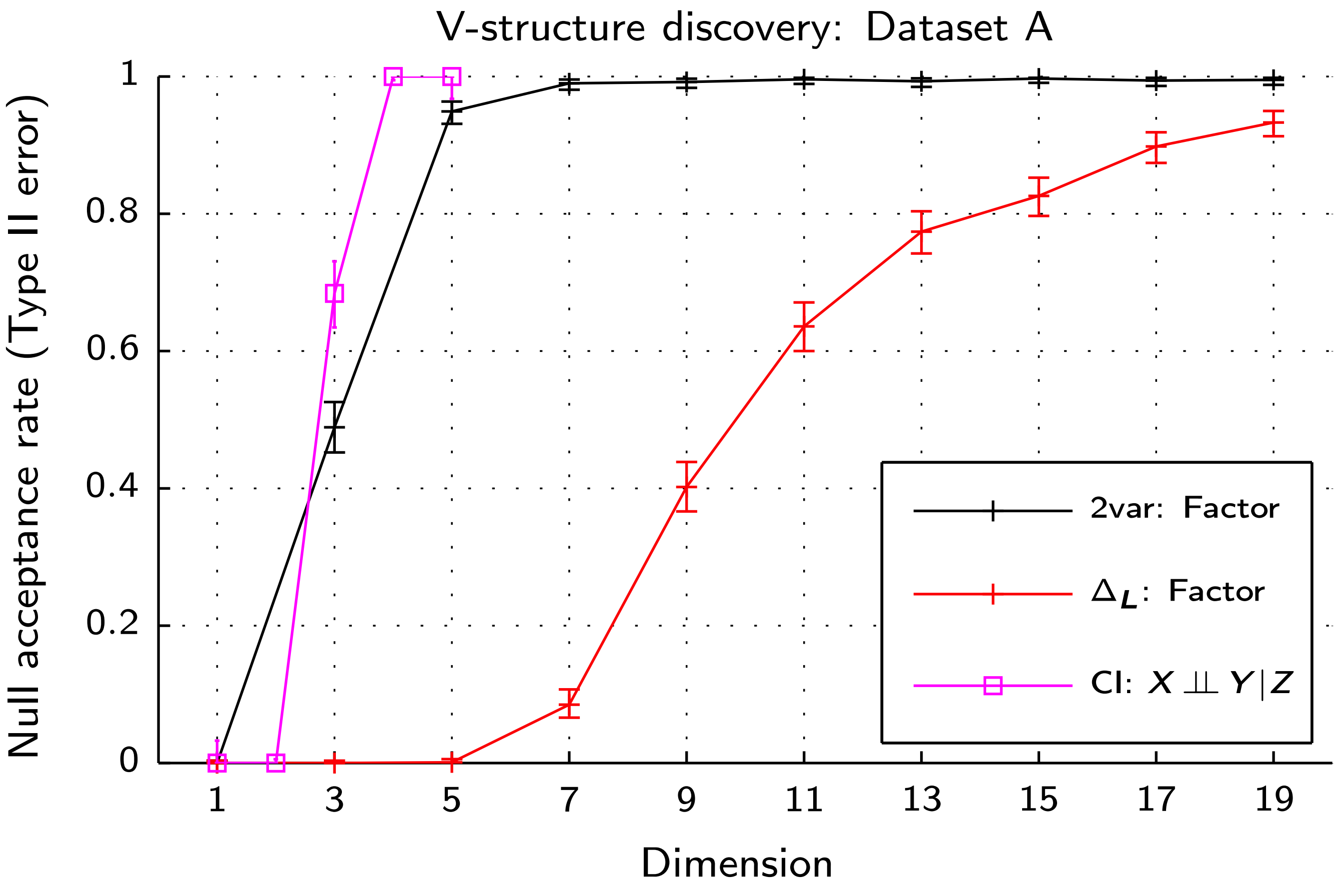


Figure 2: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Example B: Joint dependence can be easier to detect

- $X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $Z_1 = \begin{cases} X_1^2 + \epsilon, & w.p.\ 1/3, \\ Y_1^2 + \epsilon, & w.p.\ 1/3, \\ X_1 Y_1 + \epsilon, & w.p.\ 1/3, \end{cases}$ where $\epsilon \sim \mathcal{N}(0, 0.1^2)$.
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$
- dependence of Z on pair (X, Y) is stronger than on X and Y individually
- Satisfies faithfulness

Example B: factorisation tests

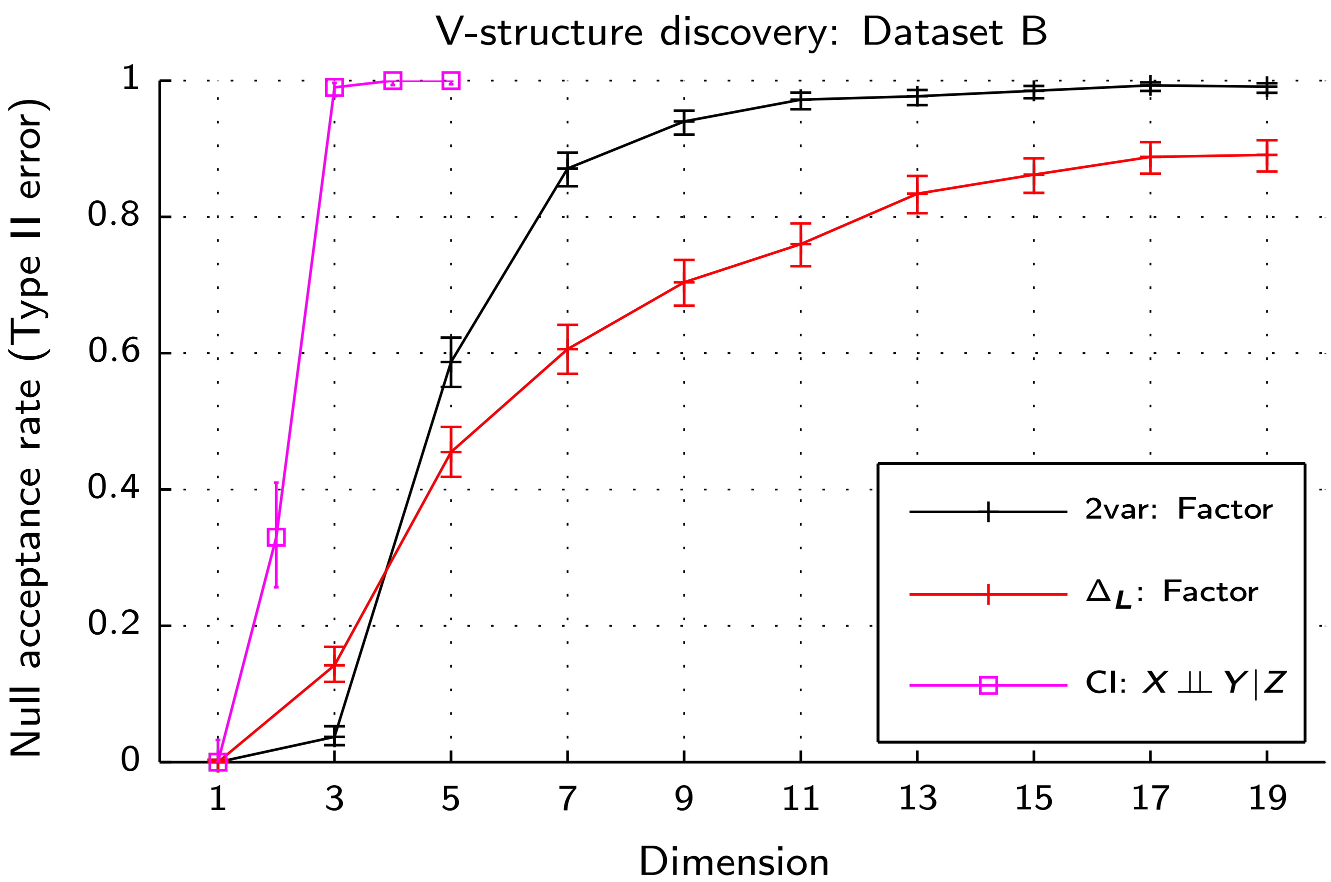


Figure 3: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Interaction for $D \geq 4$

- Interaction measure valid for all D (Streitberg, 1990):

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

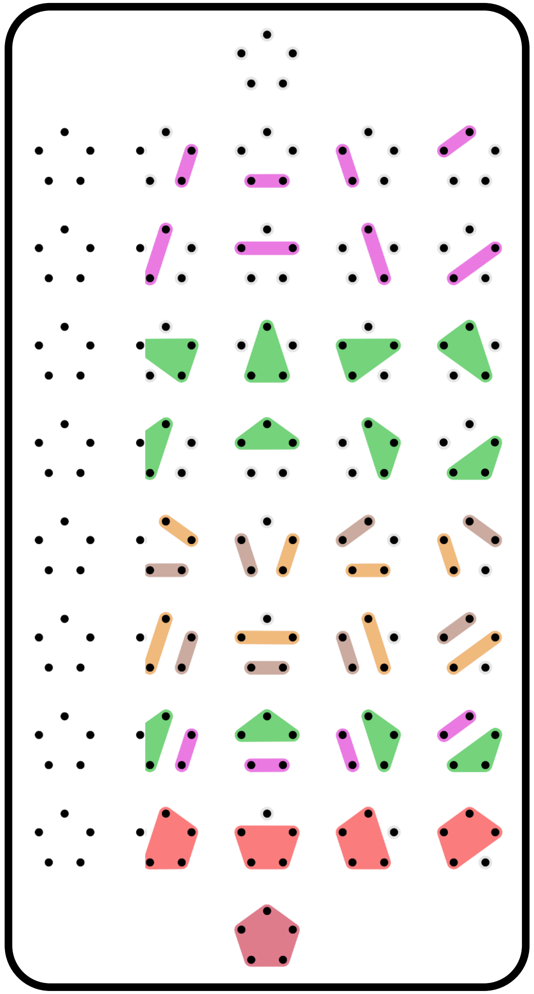
  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

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

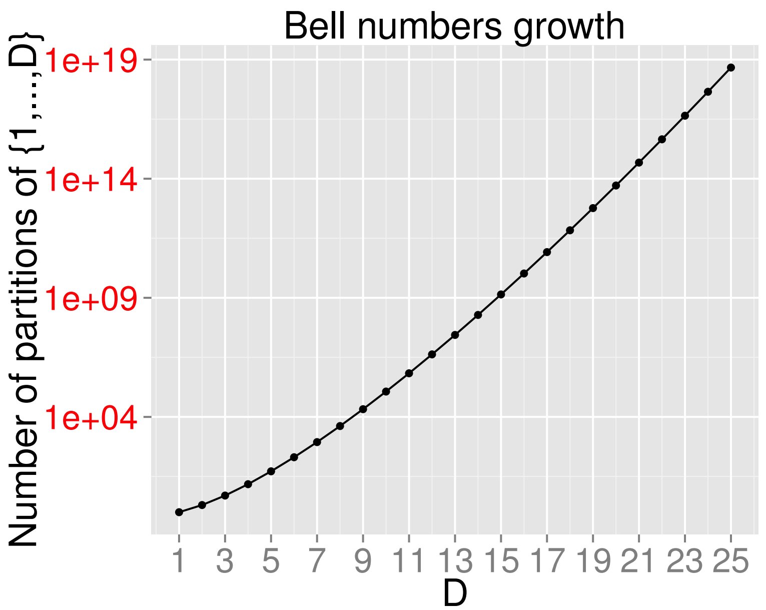


Interaction for $D \geq 4$

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



joint central moments (Lancaster interaction)

vs.

joint cumulants (Streitberg interaction)

Total independence test

- Total independence test:

  $\mathbf{H_0} : P_{XYZ} = P_X P_Y P_Z$ vs. $\mathbf{H}_1 : P_{XYZ} \neq P_X P_Y P_Z$

Total independence test

- Total independence test:
  $\mathbf{H_0}: P_{XYZ} = P_X P_Y P_Z$ vs. $\mathbf{H}_1: P_{XYZ} \neq P_X P_Y P_Z$

- For $(X_1, \ldots, X_D) \sim P_{\mathbf{X}}$, and $\kappa = \bigotimes_{i=1}^{D} k^{(i)}$:

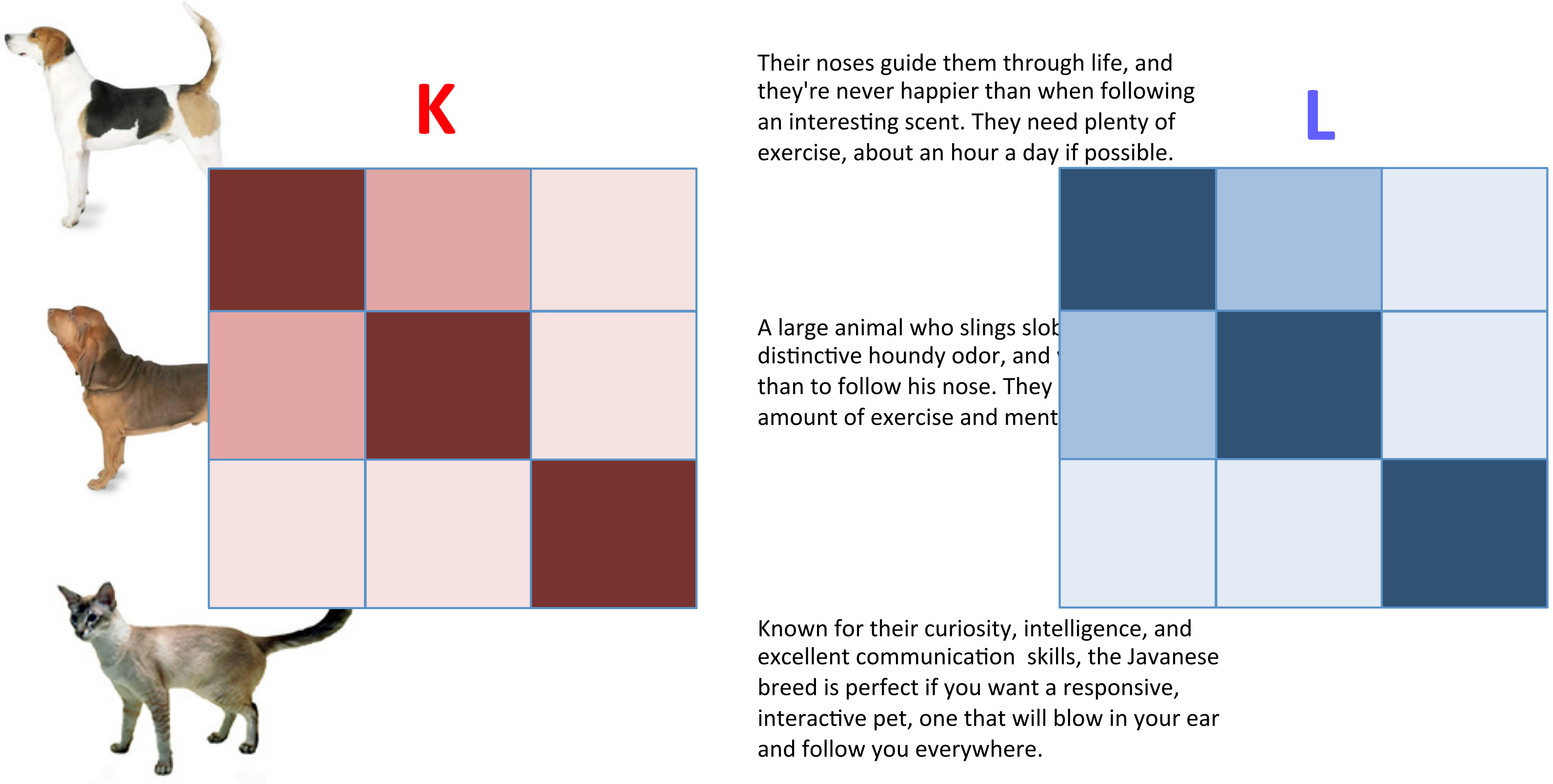
$$\Bigg\| \underbrace{\mu_\kappa \left(\hat{P}_{\mathbf{X}} - \prod_{i=1}^{D} \hat{P}_{X_i} \right)}_{\Delta_{tot} \hat{P}} \Bigg\|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} \sum_{a=1}^{n} \sum_{b=1}^{n} \prod_{i=1}^{D} K_{ab}^{(i)} - \frac{2}{n^{D+1}} \sum_{a=1}^{n} \prod_{i=1}^{D} \sum_{b=1}^{n} K_{ab}^{(i)}$$

$$+ \frac{1}{n^{2D}} \prod_{i=1}^{D} \sum_{a=1}^{n} \sum_{b=1}^{n} K_{ab}^{(i)}.$$

- Coincides with the test proposed by Kankainen (1995) using empirical characteristic functions.

Kernel dependence measures - in detail

MMD for independence: HSIC



Text from dogtime.com and petfinder.com

Empirical $HSIC(\mathbf{P}_{XY}, \mathbf{P}_X \mathbf{P}_Y)$:

$$\frac{1}{n^2} \left(HKH \circ HLH \right)_{++}$$

Covariance to reveal dependence

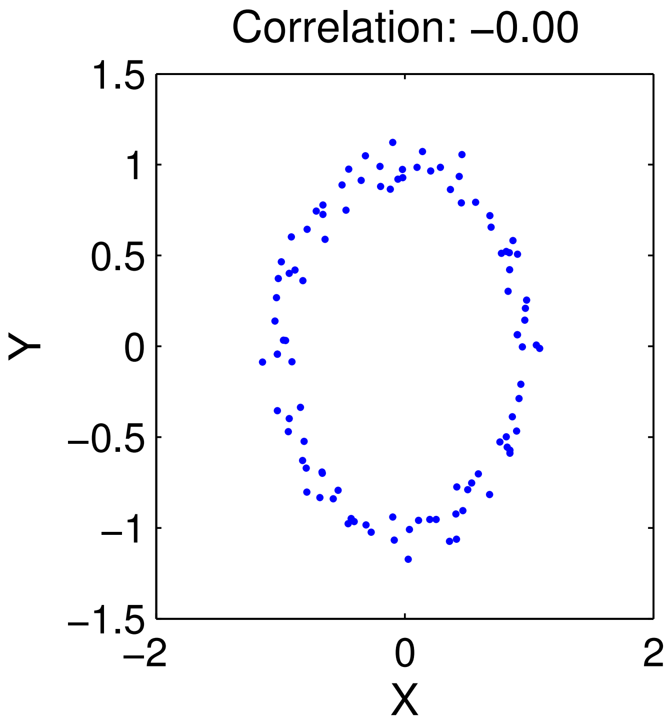
A more intuitive idea: maximize covariance of smooth mappings:

$$\mathrm{COCO}(\mathsf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$

Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

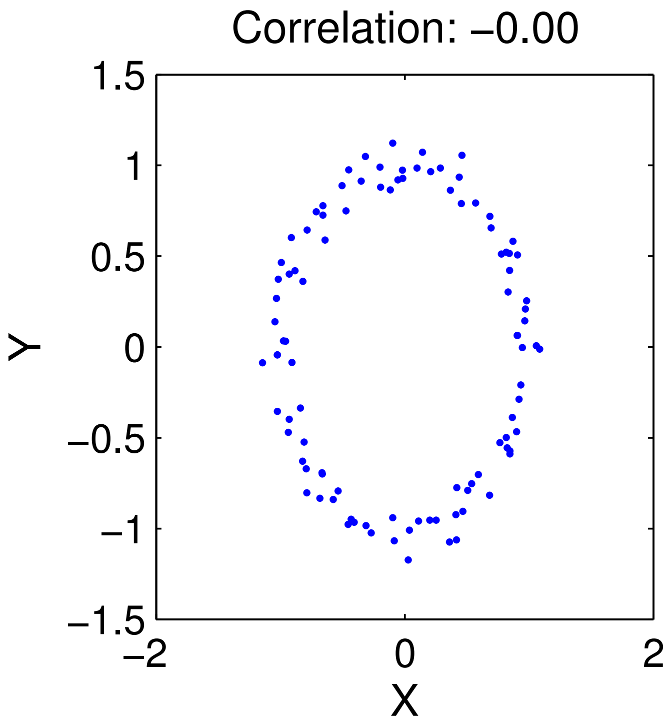
$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$

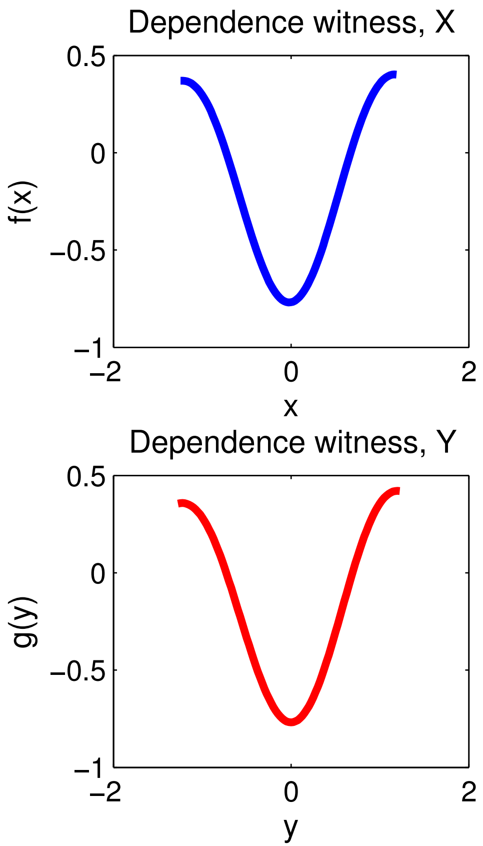


Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

$$\mathrm{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)] \right)$$

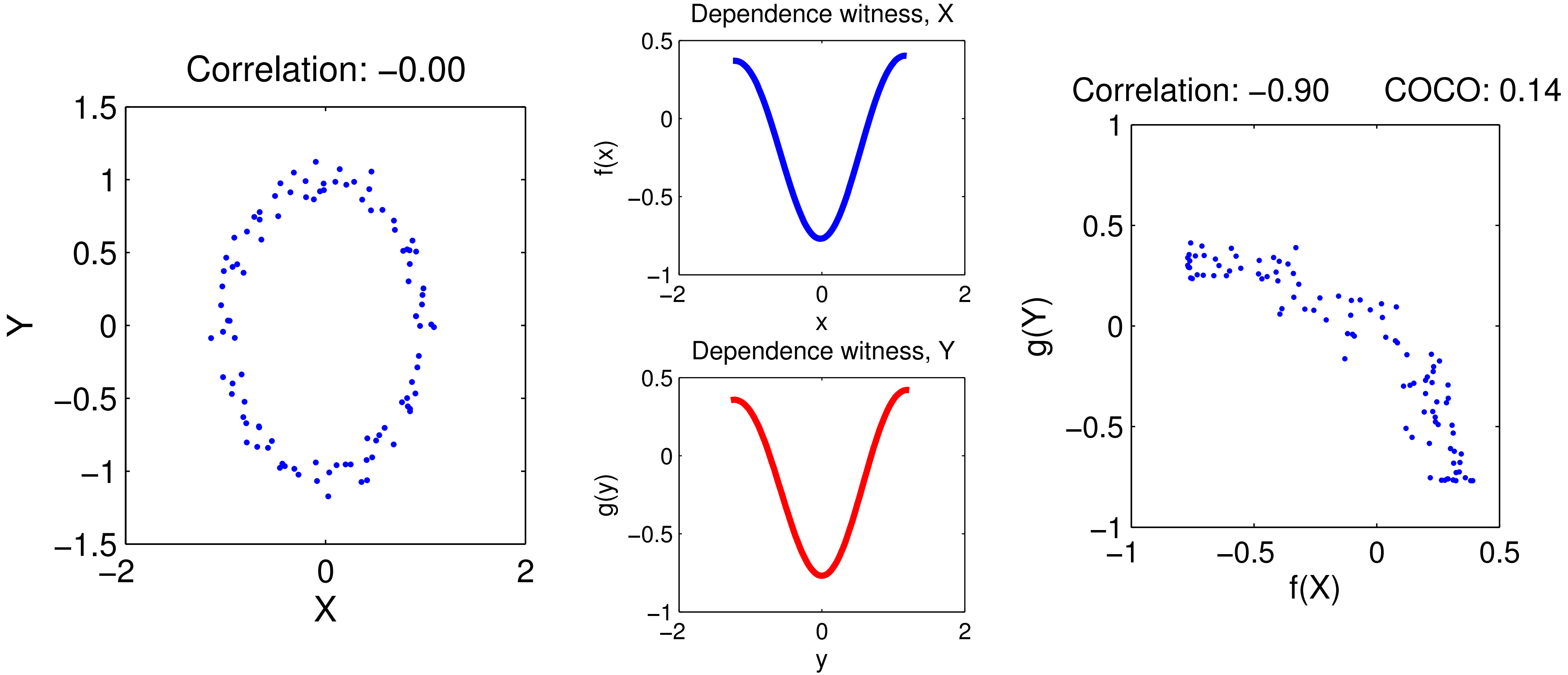




Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

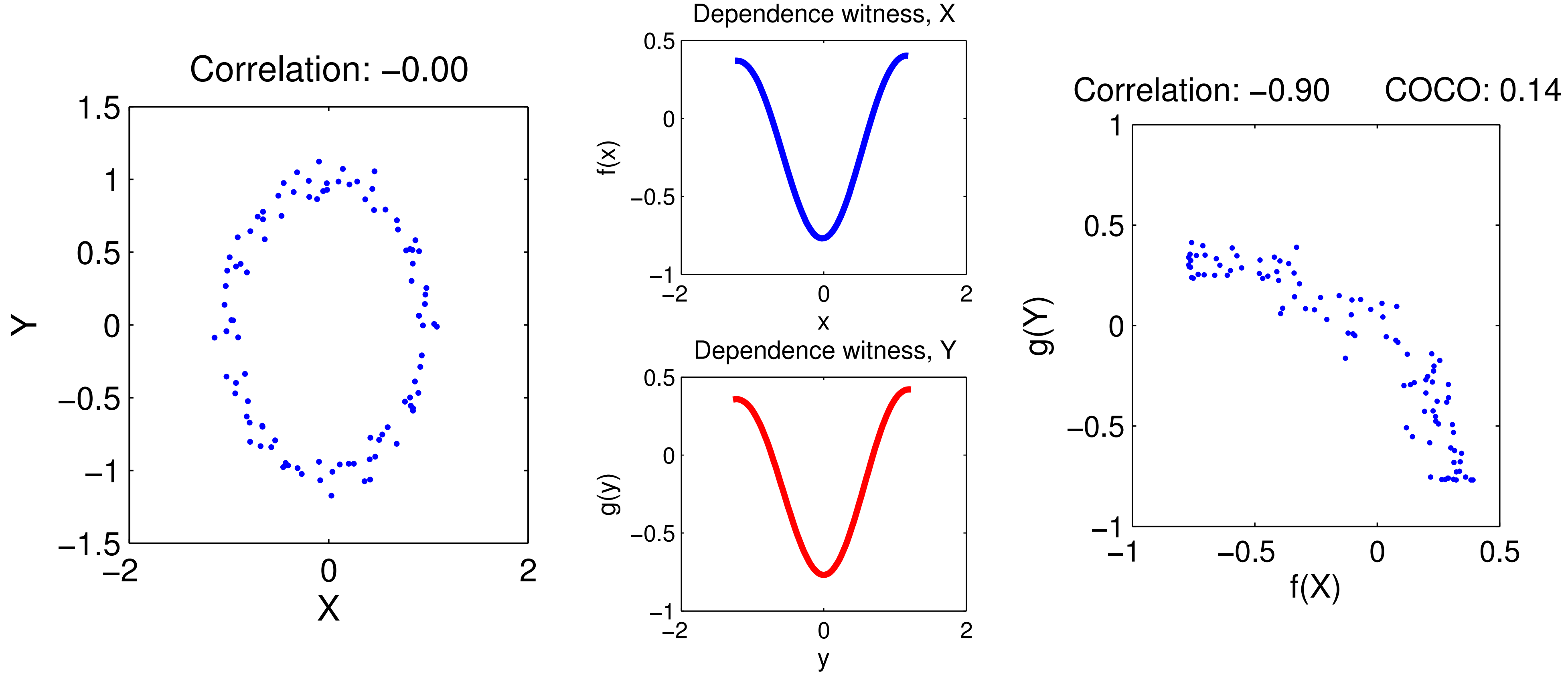
$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_{x}[f(x)]\mathbf{E}_{y}[g(y)]\right)$$



Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_{x}[f(x)]\mathbf{E}_{y}[g(y)]\right)$$



How do we define covariance in (infinite) feature spaces?

Covariance to reveal dependence

Covariance in RKHS: Let's first look at finite linear case.

We have two random vectors $\mathsf{x} \in \mathbb{R}^d$, $\mathsf{y} \in \mathbb{R}^{d'}$. Are they linearly dependent?

Covariance to reveal dependence

Covariance in RKHS: Let's first look at finite linear case.

We have two random vectors $\mathsf{x} \in \mathbb{R}^d$, $\mathsf{y} \in \mathbb{R}^{d'}$. Are they linearly dependent?

Compute their covariance matrix: (ignore centering)

$$C_{xy} = \mathbf{E}\left(\mathsf{x}\mathsf{y}^\top\right)$$

How to get a single "summary" number?

Covariance to reveal dependence

Covariance in RKHS: Let's first look at finite linear case.

We have two random vectors $\mathsf{x} \in \mathbb{R}^d$, $\mathsf{y} \in \mathbb{R}^{d'}$. Are they linearly dependent?

Compute their covariance matrix: (ignore centering)

$$C_{xy} = \mathbf{E}\left(\mathsf{x}\mathsf{y}^\top\right)$$

How to get a single "summary" number?

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

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

(maximum singular value)

Challenges in defining feature space covariance

Given features $\phi(x) \in \mathcal{F}$ and $\psi(y) \in \mathcal{G}$:

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

- Given $f \in \mathbb{R}^d$, $g \in \mathbb{R}^{d'}$, $h \in \mathbb{R}^{d'}$, define matrix $f\, g^\top$ such that $(f\, g^\top)h = f(g^\top h)$.
- 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 $f := \phi(x)$, $g = \psi(y)$, to get $x\, y^\top \to \phi(x) \otimes \psi(y)$
- Corresponds to the product kernel:

$$\langle \phi(x) \otimes \psi(y), \phi(x') \otimes \psi(y') \rangle = k(x, x')l(y, y')$$

Challenges in defining feature space covariance

Given features $\phi(x) \in \mathcal{F}$ and $\psi(y) \in \mathcal{G}$:

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

$$\langle f, C_{XY} g \rangle_{\mathcal{F}} = \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] = \operatorname{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)$$

Challenges in defining feature space covariance

Given features $\phi(x) \in \mathcal{F}$ and $\psi(y) \in \mathcal{G}$:

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

$$\langle f, C_{XY} g \rangle_{\mathcal{F}} = \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] = \operatorname{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)$$

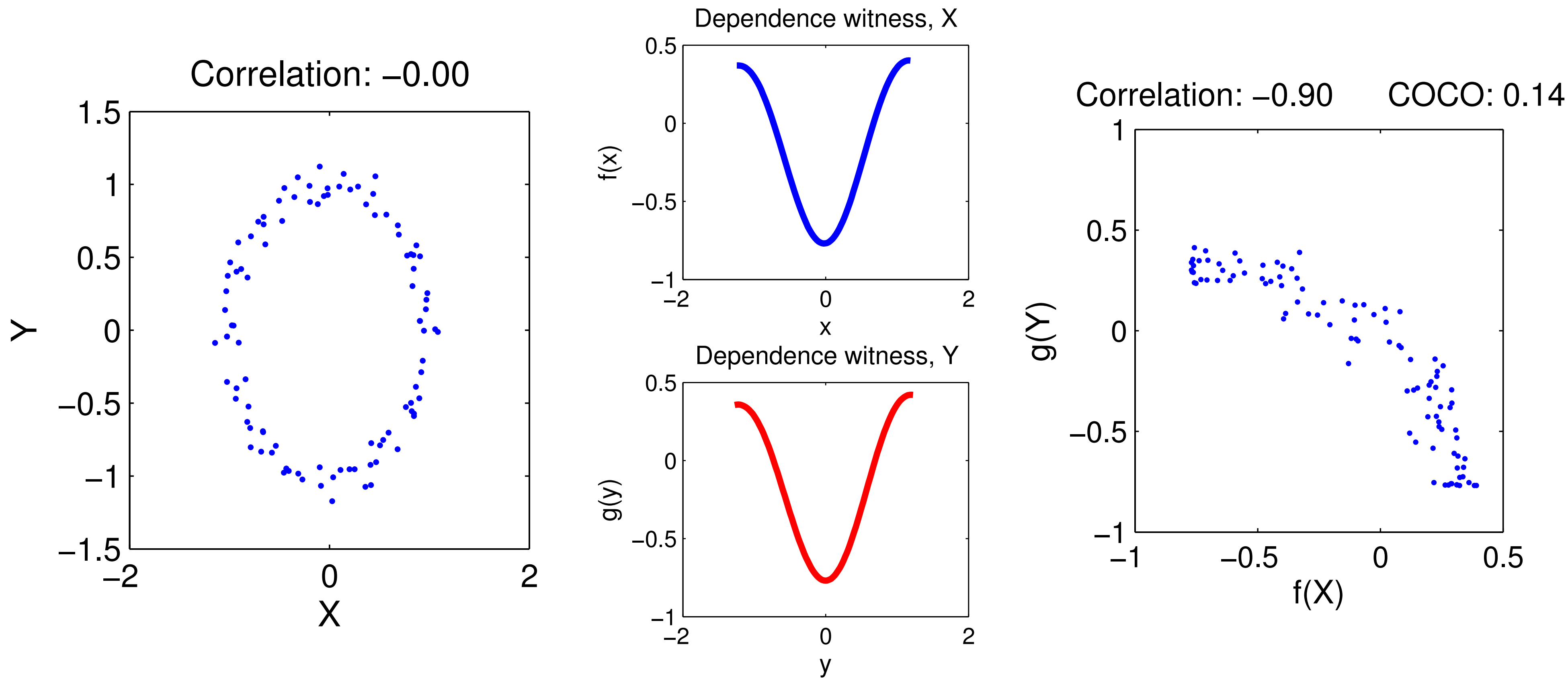
YES: via Bochner integrability argument (as with mean embedding).
Under the condition $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\sqrt{k(\mathsf{x},\mathsf{x})l(\mathsf{y},\mathsf{y})}\right) < \infty$, we can define:

$$C_{XY} := \mathbf{E}_{\mathsf{x},\mathsf{y}}\left[\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\right]$$

which is a Hilbert-Schmidt operator (sum of squared singular values is finite).

REMINDER: functions revealing dependence

$$\mathrm{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})] \right)$$



How do we compute this from finite data?

Empirical covariance operator

The empirical feature covariance given $\boldsymbol{z} := (x_i, y_i)_{i=1}^n$ (now include centering)

$$\widehat{C}_{XY} := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \otimes \hat{\mu}_y,$$

where $\hat{\mu}_x := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$.

Functions revealing dependence

Optimization problem:

$$\begin{aligned} \mathrm{COCO}(\boldsymbol{z}; \mathcal{F}, \mathcal{G}) := \max \quad & \left\langle f, \widehat{C}_{XY} g \right\rangle_{\mathcal{F}} \\ \text{subject to} \quad & \|f\|_{\mathcal{F}} \leq 1 \\ & \|g\|_{\mathcal{G}} \leq 1 \end{aligned}$$

Assume

$$f = \sum_{i=1}^{n} \alpha_i \left[\phi(x_i) - \hat{\mu}_x\right] \qquad g = \sum_{j=1}^{n} \beta_i \left[\psi(y_i) - \hat{\mu}_y\right]$$

The associated Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \left\langle f, \widehat{C}_{XY} g \right\rangle_{\mathcal{F}} - \frac{\lambda}{2} \left(\|f\|_{\mathcal{F}}^2 - 1\right) - \frac{\gamma}{2} \left(\|g\|_{\mathcal{G}}^2 - 1\right),$$

where $\lambda \geq 0$ and $\gamma \geq 0$.

Covariance to reveal dependence

- Empirical $\mathrm{COCO}(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$ largest eigenvalue of

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

- $\widetilde{K}$ and $\widetilde{L}$ are matrices of inner products between centred observations in respective feature spaces:

$$\widetilde{K} = HKH \quad \text{where} \quad K_{ij} = k(x_i, x_j) \quad \text{and} \quad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

Covariance to reveal dependence

- Empirical $\mathrm{COCO}(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$ largest eigenvalue of

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

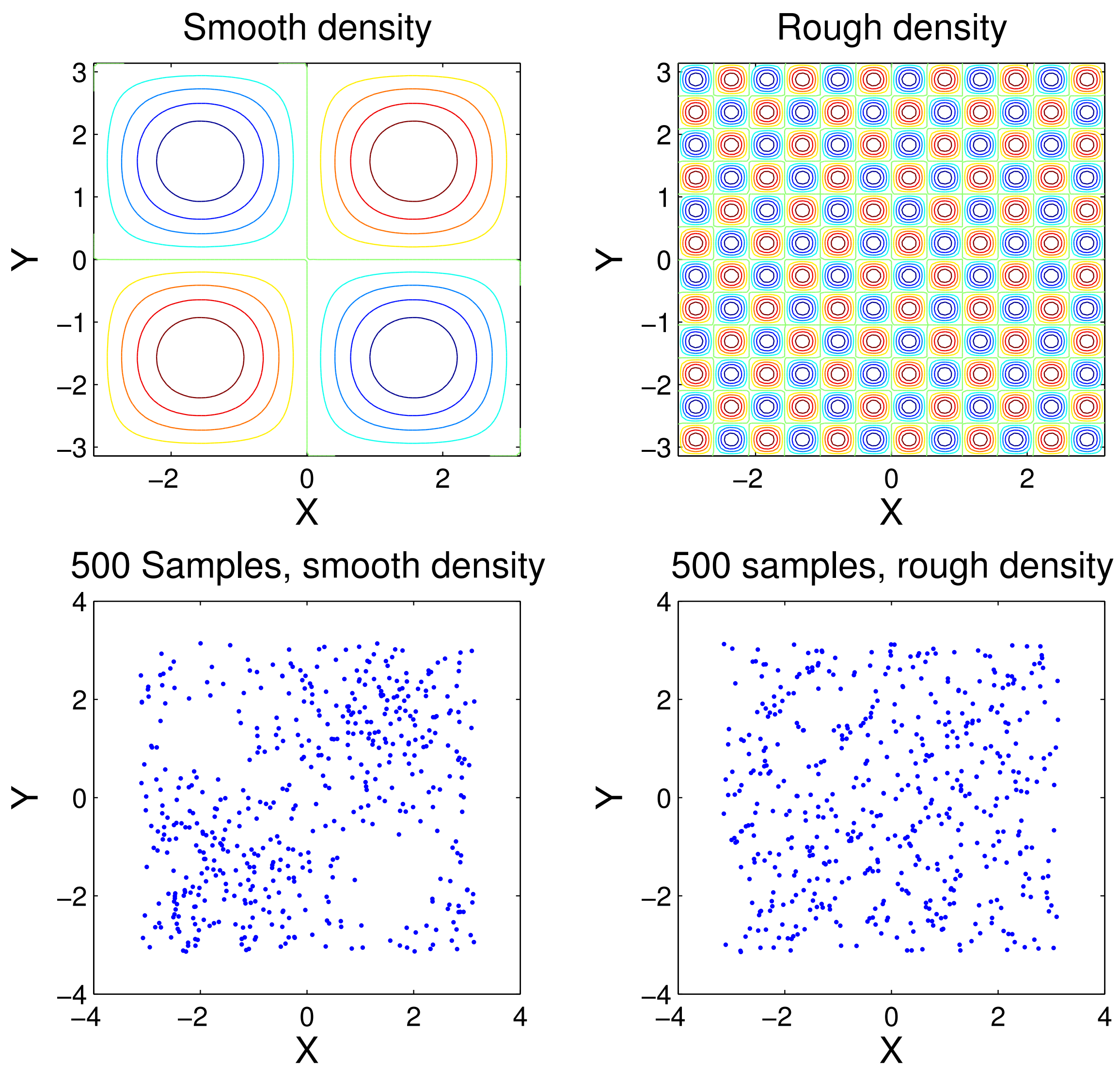
- $\widetilde{K}$ and $\widetilde{L}$ are matrices of inner products between centred observations in respective feature spaces:

$$\widetilde{K} = HKH \quad \text{where} \quad K_{ij} = k(x_i, x_j) \quad \text{and} \quad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

- Mapping function for x:

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

Hard-to-detect dependence



Density takes the form:

$$\mathbf{P}_{x,y} \propto 1 + \sin(\omega x)\sin(\omega y)$$

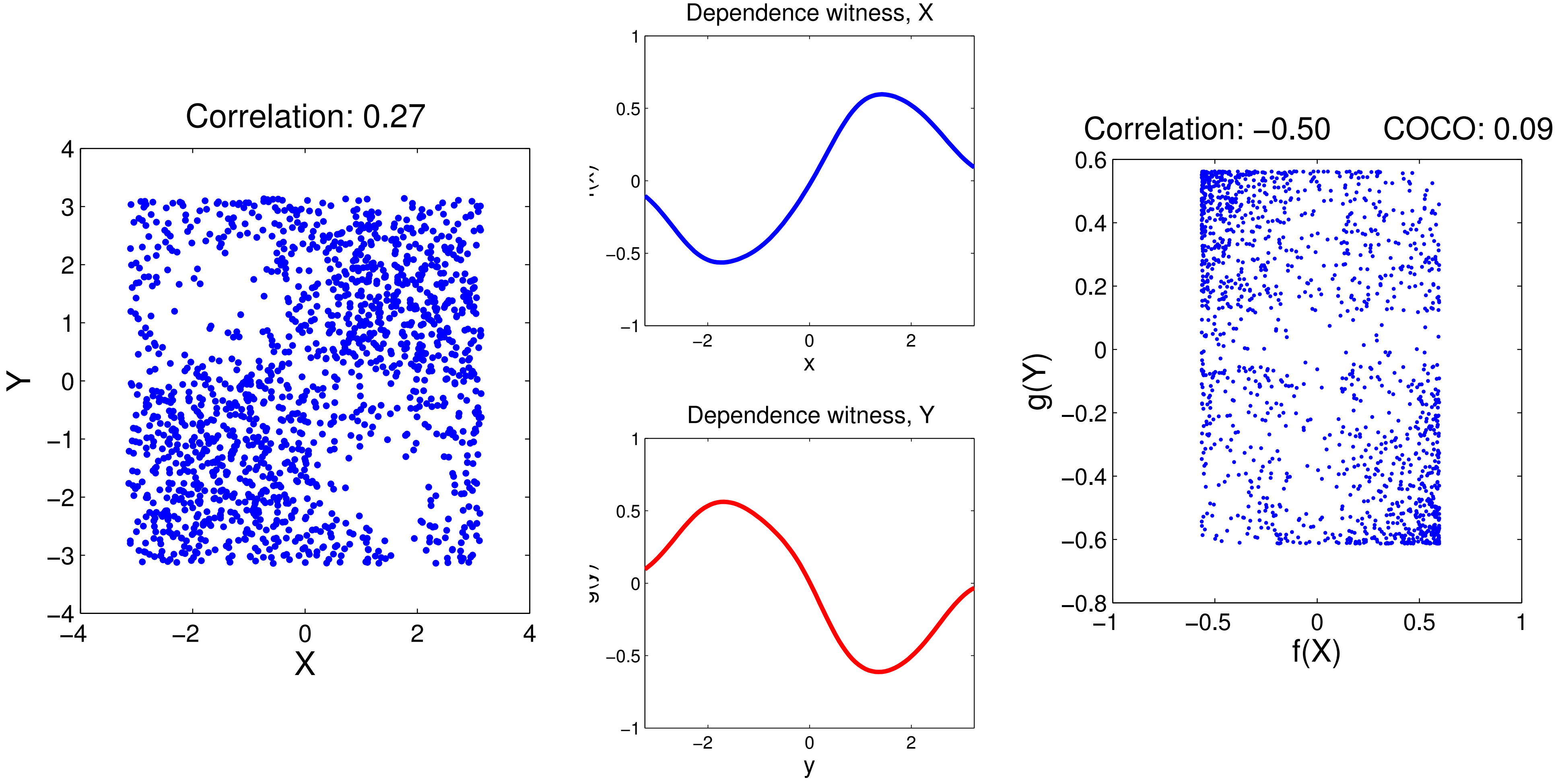
Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

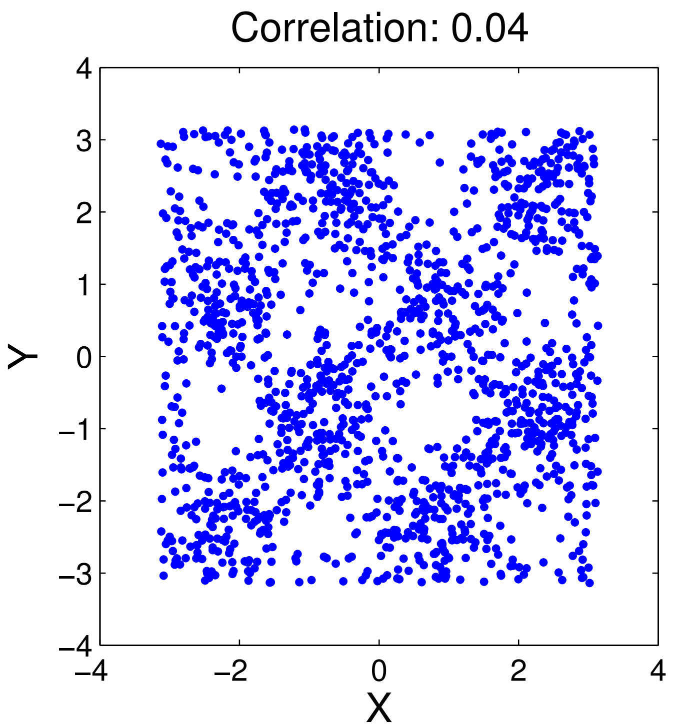
Case of $\omega = 1$

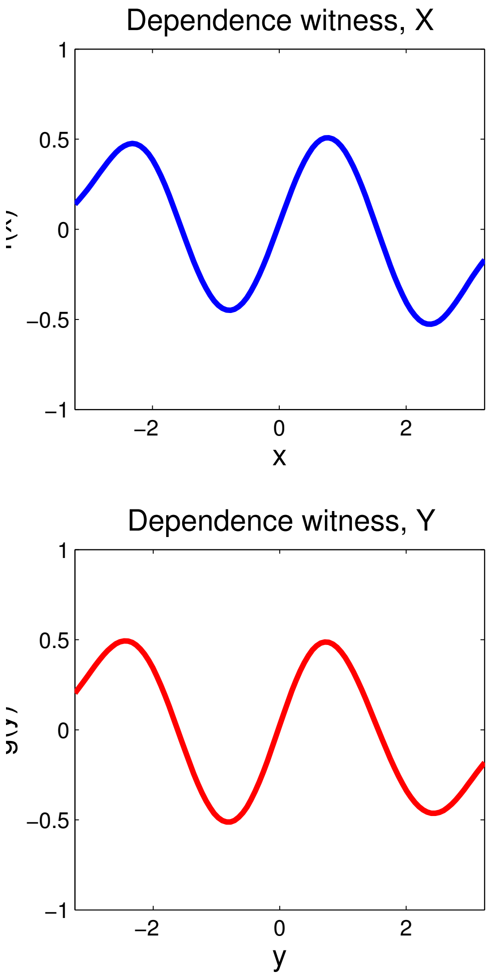


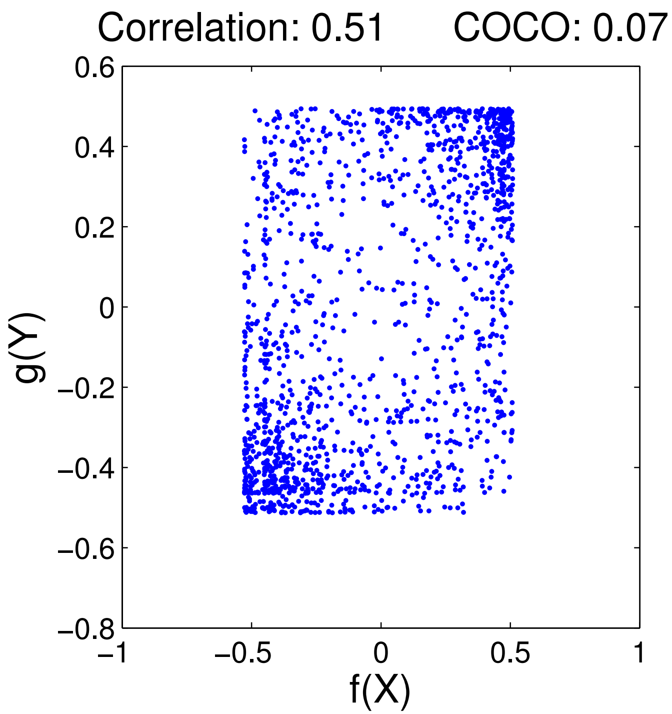
Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

Case of $\omega = 2$



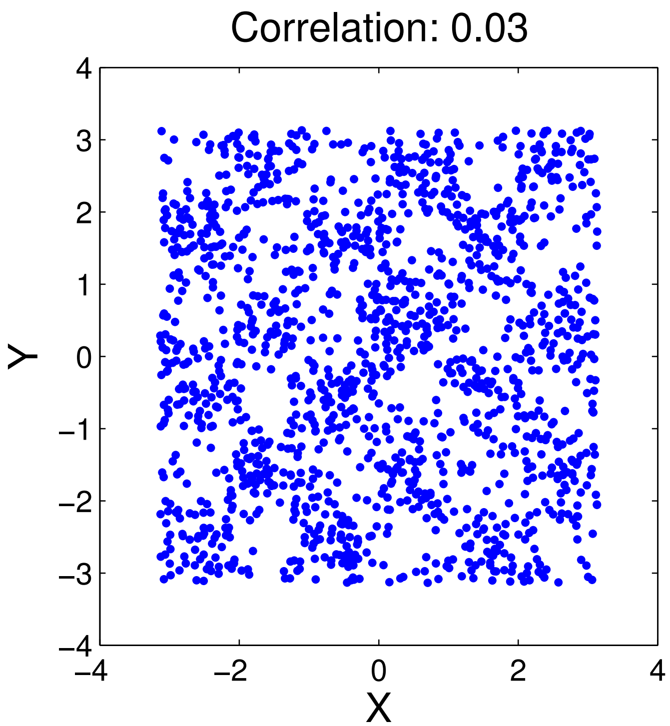


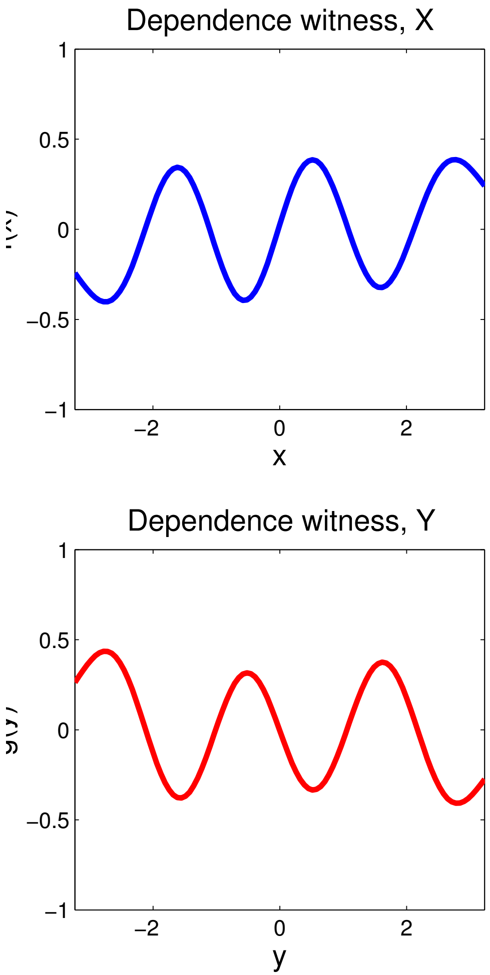


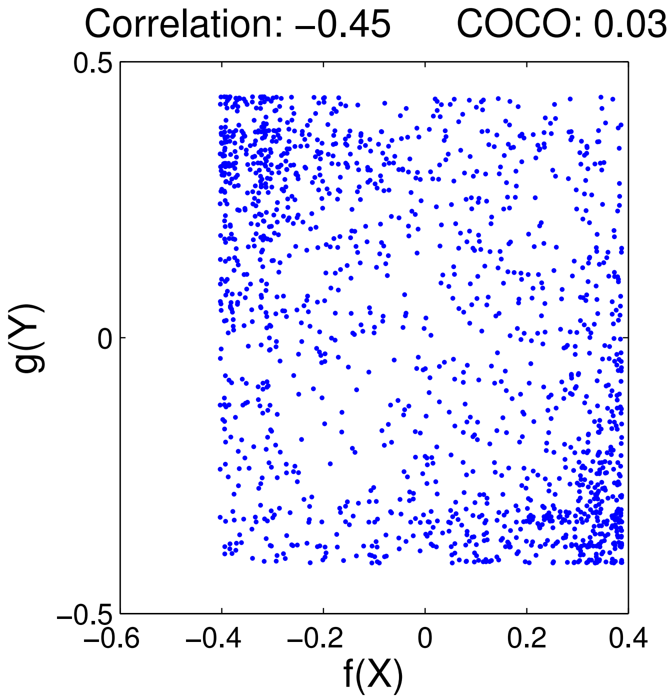
Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

Case of $\omega = 3$



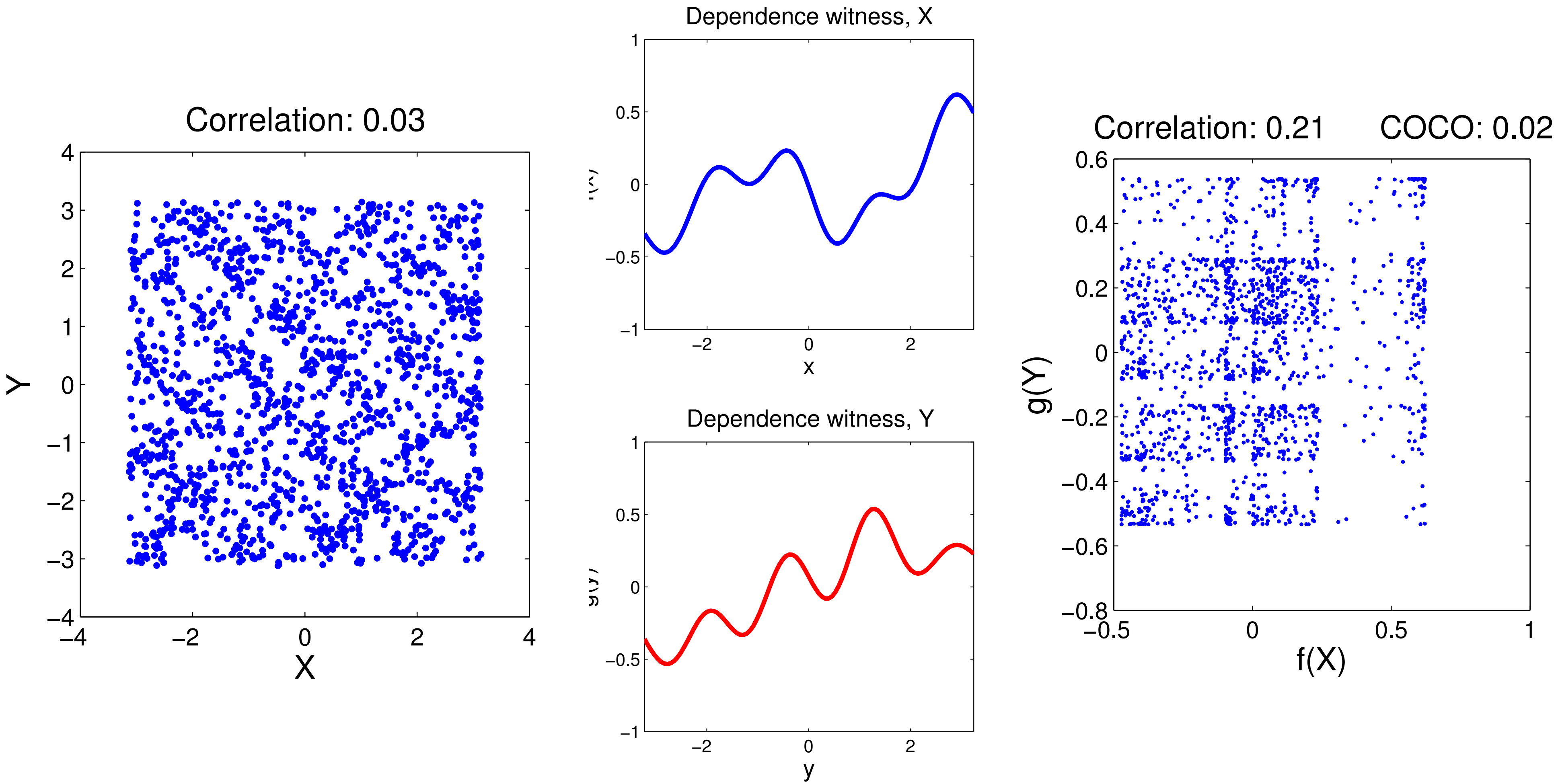




Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

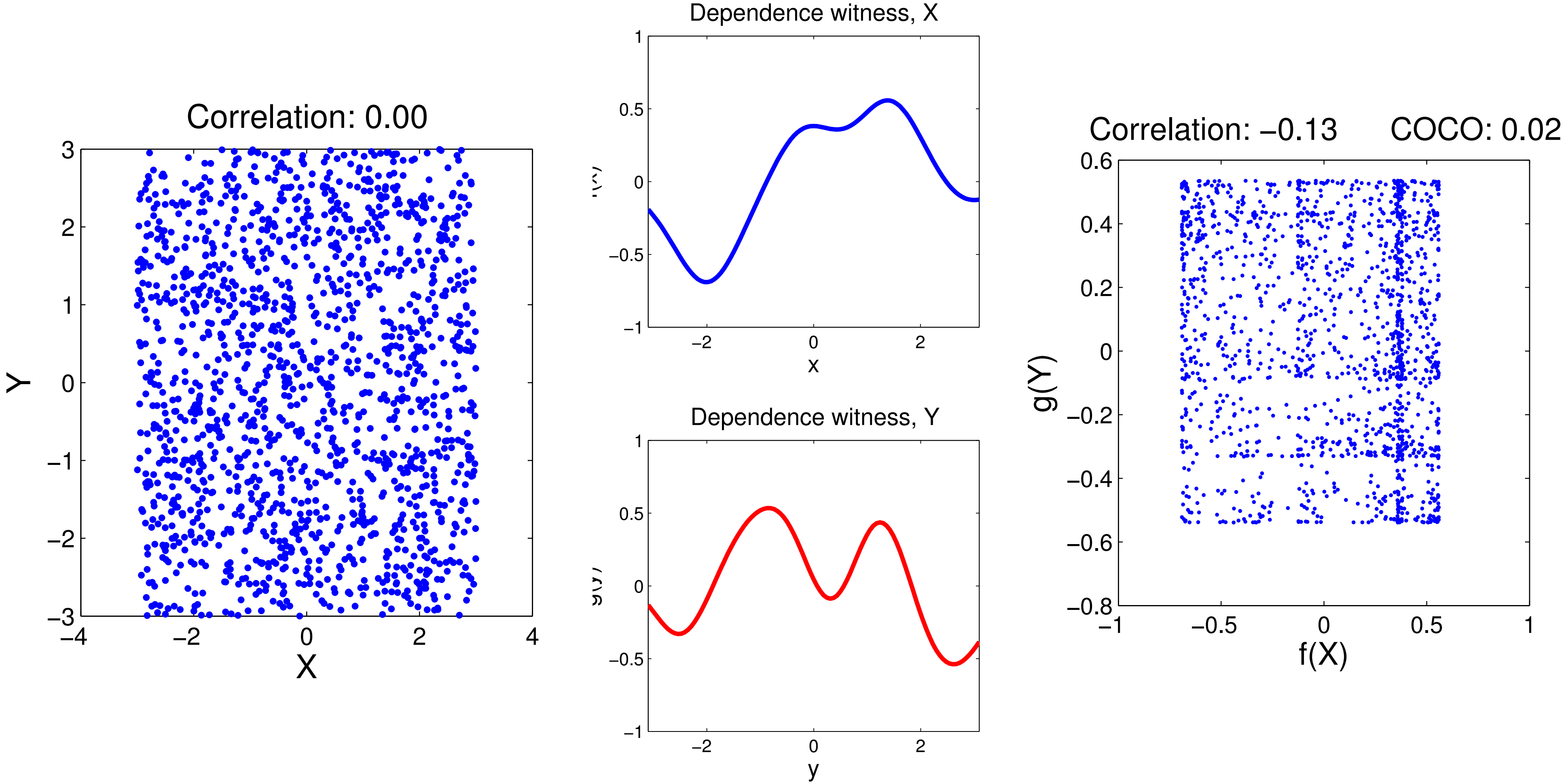
Case of $\omega = 4$



Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

Case of $\omega = ??$

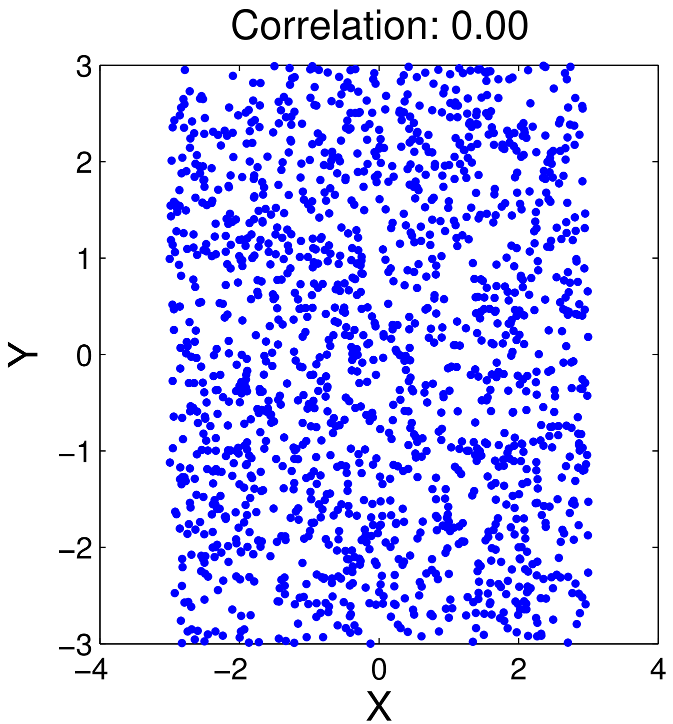


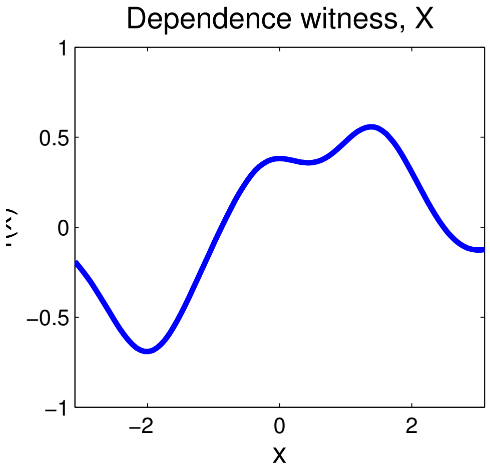
Hard-to-detect dependence

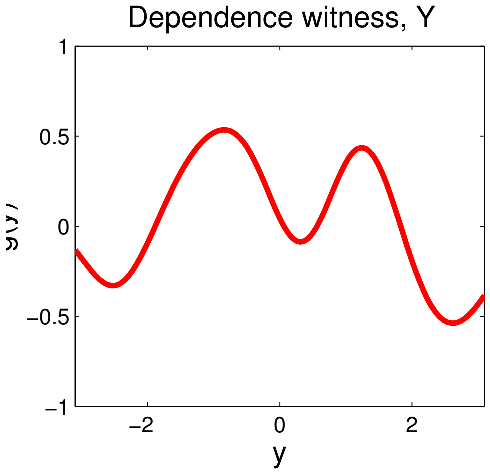
COCO vs frequency of perturbation from independence.

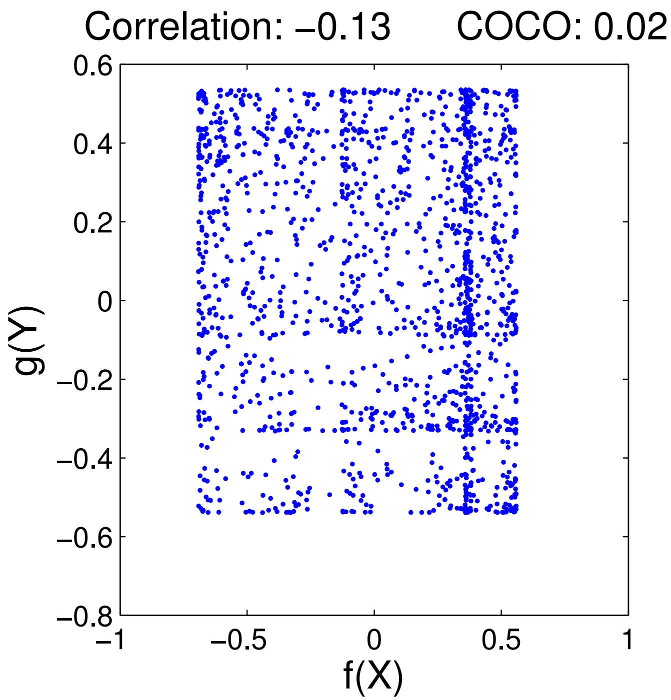
Case of uniform noise!

This bias will decrease with increasing sample size.









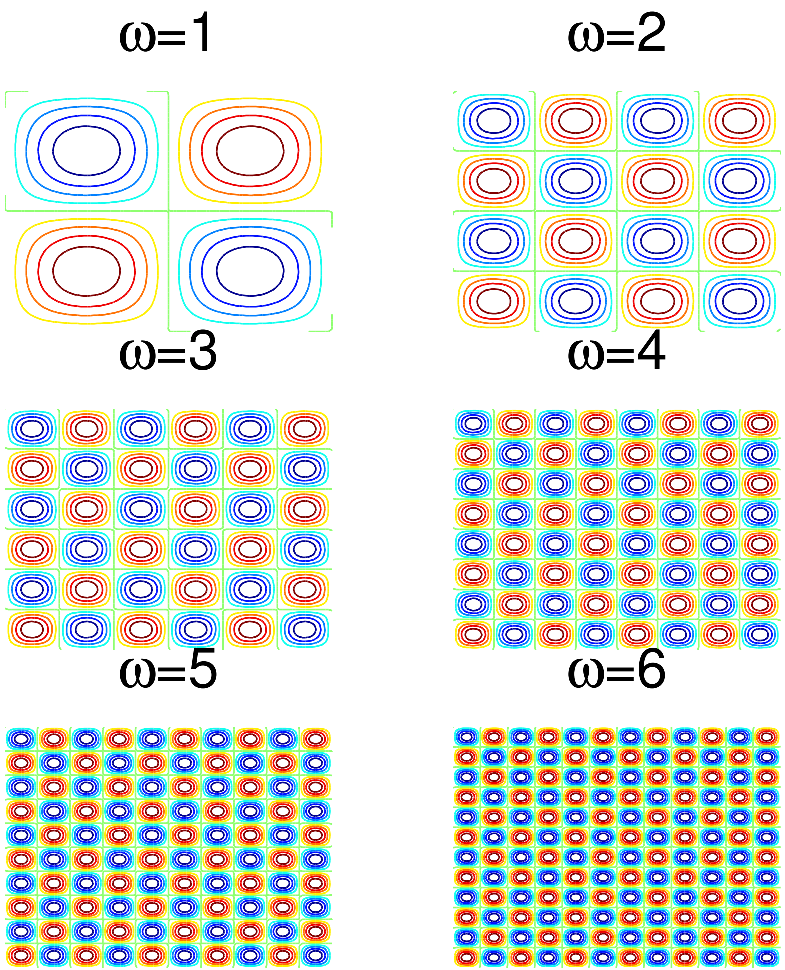
Hard-to-detect dependence

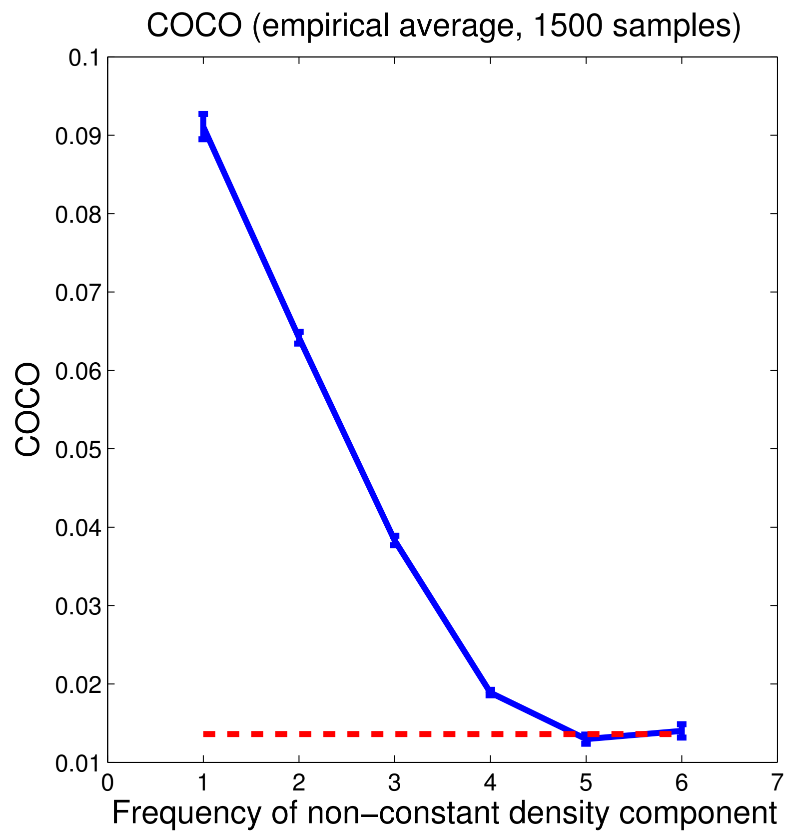
COCO vs frequency of perturbation from independence.

- As dependence is encoded at higher frequencies, the smooth mappings f, g achieve lower linear covariance.
- Even for independent variables, COCO will **not** be zero at finite sample sizes, since some mild linear dependence will be induced by f, g (bias)
- This bias will decrease with increasing sample size.

Hard-to-detect dependence

- Example: sinusoids of increasing frequency



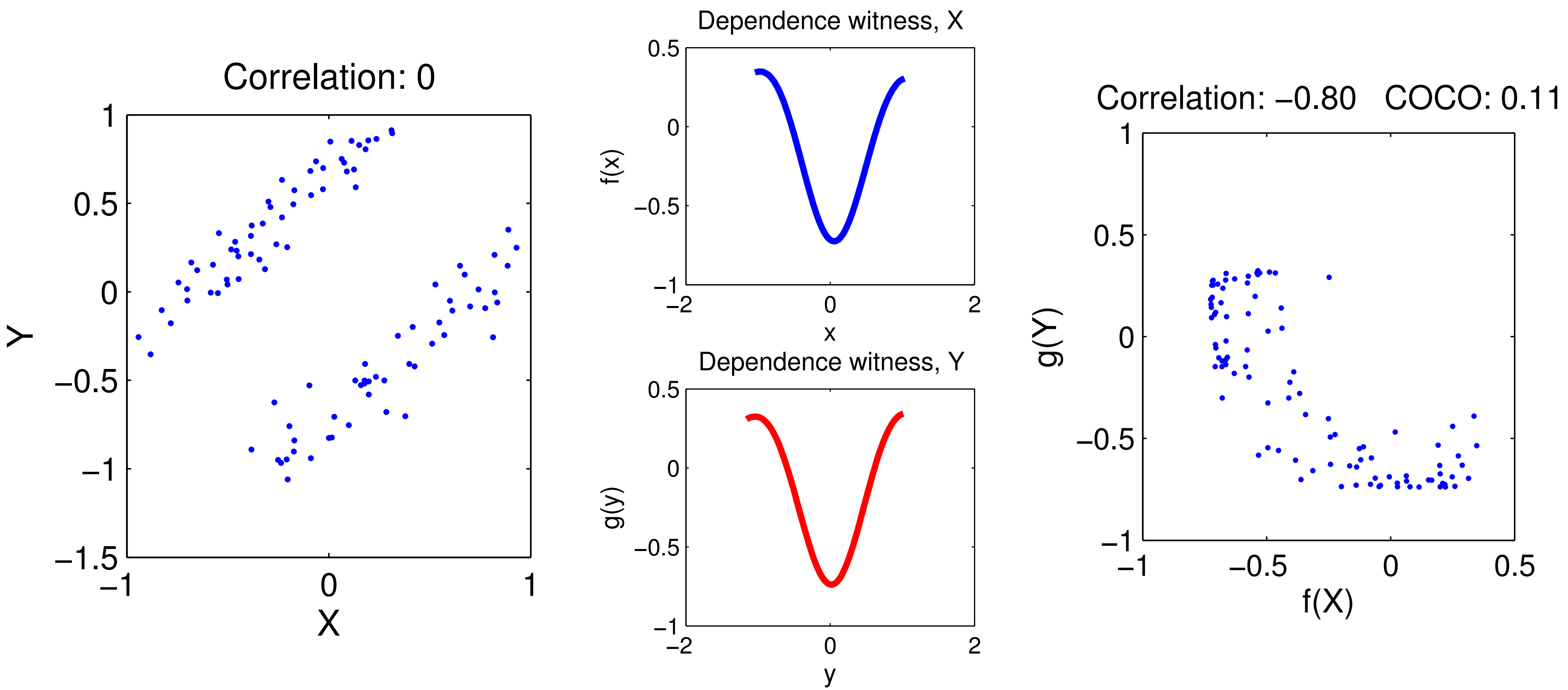


More functions revealing dependence

- Can we do better than COCO?

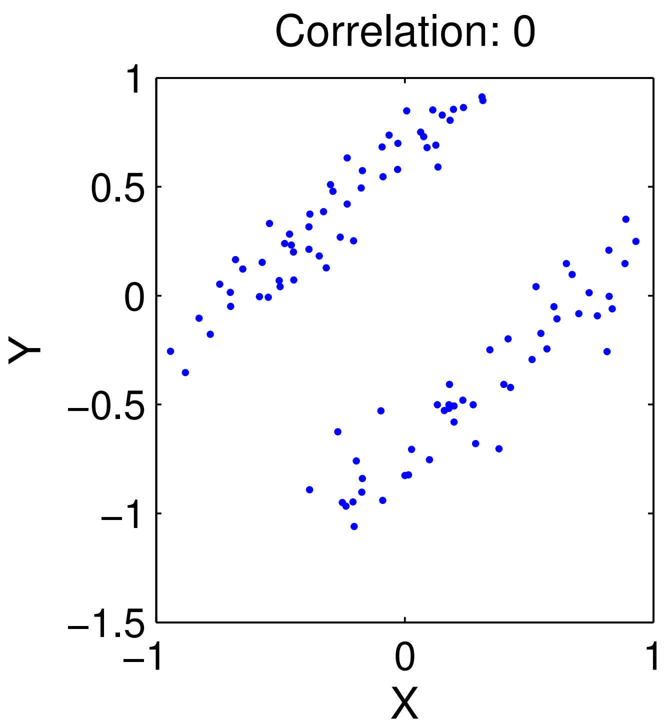
More functions revealing dependence

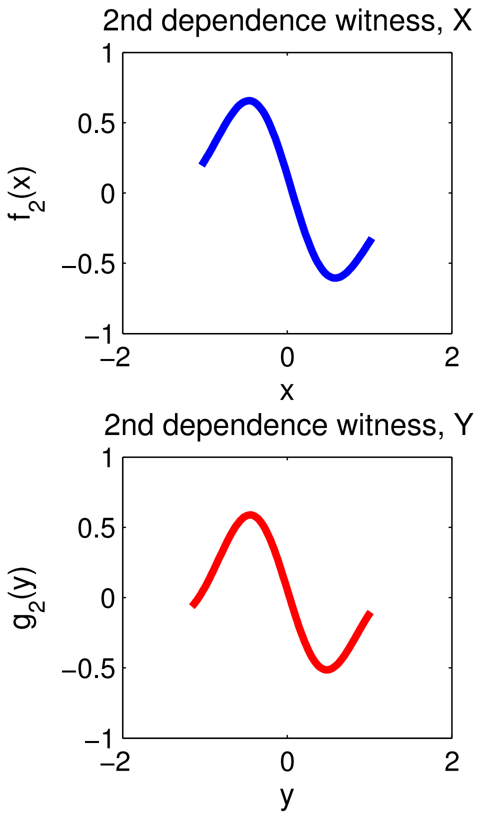
- Can we do better than COCO?
- A second example with zero correlation



More functions revealing dependence

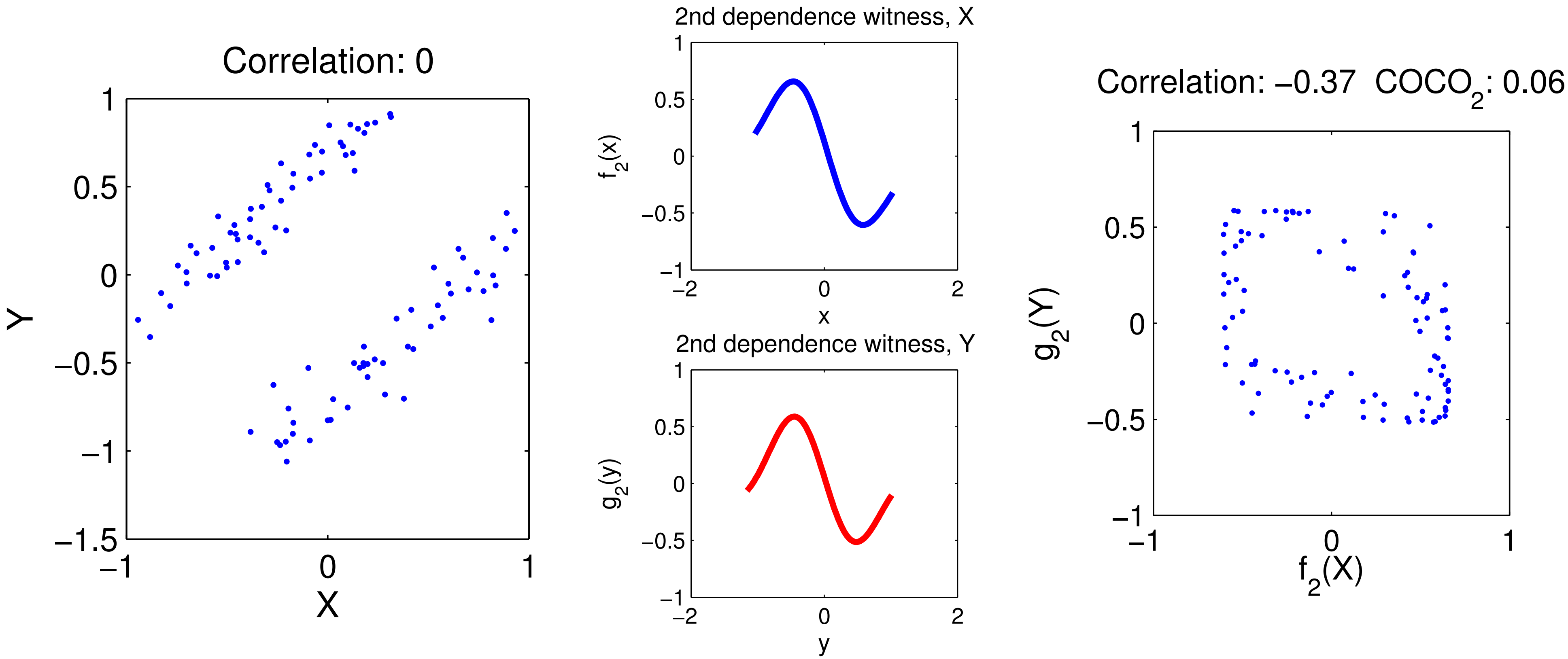
- Can we do better than COCO?
- A second example with zero correlation





More functions revealing dependence

- Can we do better than COCO?
- A second example with zero correlation



Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \mathrm{COCO}_i(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10] :

$$\mathrm{HSIC}(\boldsymbol{z}; \mathcal{F}, \mathcal{G}) := \sum_{i=1}^{n} \gamma_i^2$$

Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \mathrm{COCO}_i(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10] :

$$\mathrm{HSIC}(\boldsymbol{z}; \mathcal{F}, \mathcal{G}) := \sum_{i=1}^{n} \gamma_i^2$$

- In limit of infinite samples:

$$\begin{aligned}\mathrm{HSIC}(\mathbf{P}; F, G) &:= \|C_{xy}\|_{\mathrm{HS}}^2 \\ &= \langle C_{xy}, C_{xy} \rangle_{\mathrm{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{x}',\mathsf{y},\mathsf{y}'}[k(\mathsf{x},\mathsf{x}')l(\mathsf{y},\mathsf{y}')] + \mathbf{E}_{\mathsf{x},\mathsf{x}'}[k(\mathsf{x},\mathsf{x}')]\mathbf{E}_{\mathsf{y},\mathsf{y}'}[l(\mathsf{y},\mathsf{y}')] \\ &\quad - 2\mathbf{E}_{\mathsf{x},\mathsf{y}}\left[\mathbf{E}_{\mathsf{x}'}[k(\mathsf{x},\mathsf{x}')]\mathbf{E}_{\mathsf{y}'}[l(\mathsf{y},\mathsf{y}')]\right]\end{aligned}$$

  - x' an independent copy of x, y' a copy of y

HSIC is identical to $MMD(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y)$

When does HSIC determine independence?

Theorem: When kernels k and l are each characteristic, then $HSIC = 0$ iff $\mathbf{P}_{x,y} = \mathbf{P}_x \mathbf{P}_y$ [Gretton, 2015].

Weaker than MMD condition (which requires a kernel characteristic on $\mathcal{X} \times \mathcal{Y}$ to distinguish $\mathbf{P}_{x,y}$ from $\mathbf{Q}_{x,y}$).

Intuition: why characteristic needed on both $\mathcal{X}$ and $\mathcal{Y}$

Question: Wouldn't it be enough just to use a rich mapping from $\mathcal{X}$ to $\mathcal{Y}$, e.g. via ridge regression with characteristic $\mathcal{F}$:

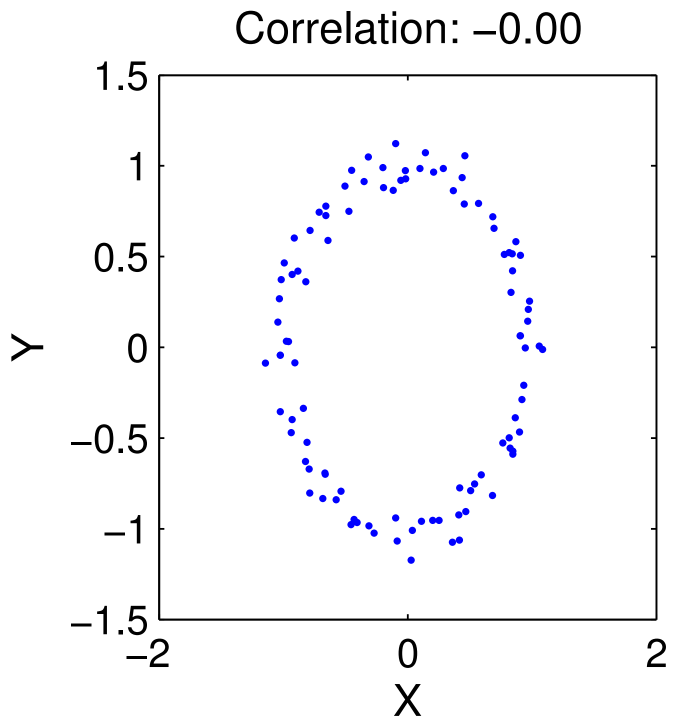
$$f^* = \arg\min_{f \in \mathcal{F}} \left(\mathbf{E}_{XY} \left(Y - \langle f, \phi(X) \rangle_{\mathcal{F}} \right)^2 + \lambda \|f\|_{\mathcal{F}}^2 \right),$$

Intuition: why characteristic needed on both $\mathcal{X}$ and $\mathcal{Y}$

Question: Wouldn't it be enough just to use a rich mapping from $\mathcal{X}$ to $\mathcal{Y}$, e.g. via ridge regression with characteristic $\mathcal{F}$:

$$f^* = \arg\min_{f \in \mathcal{F}} \left(\mathbf{E}_{XY} \left(Y - \langle f, \phi(X) \rangle_{\mathcal{F}} \right)^2 + \lambda \|f\|^2_{\mathcal{F}} \right),$$

Counterexample: density symmetric about x-axis, s.t. $p(x, y) = p(x, -y)$



Regression using distribution embeddings

Kernels on distributions in supervised learning

- Kernels have been very widely used in supervised learning
  - Support vector classification/regression, kernel ridge regression ...

Kernels on distributions in supervised learning

- Kernels have been very widely used in supervised learning
- Simple kernel on distributions (population counterpart of set kernel)

[Haussler, 1999, Gärtner et al., 2002]

$$K(\mathbf{P}, \mathbf{Q}) = \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}}$$

- Squared distance between distribution embeddings (MMD)

$$\mathrm{MMD}^2(\mu_{\mathbf{P}}, \mu_{\mathbf{Q}}) := \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_P k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_Q k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})$$

Kernels on distributions in supervised learning

- Kernels have been very widely used in supervised learning
- Simple kernel on distributions (population counterpart of set kernel)

  [Haussler, 1999, Gärtner et al., 2002]

$$K(\mathbf{P}, \mathbf{Q}) = \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}}$$

- Can define kernels on mean embedding features [Christmann, Steinwart NIPS10],[AISTATS15]

K_G	K_e	K_C	K_t	$\ldots$
$e^{-\frac{\left\|\mu_{\mathbf{P}}-\mu_{\mathbf{Q}}\right\|_{\mathcal{F}}^2}{2\theta^2}}$	$e^{-\frac{\left\|\mu_{\mathbf{P}}-\mu_{\mathbf{Q}}\right\|_{\mathcal{F}}}{2\theta^2}}$	$\left(1 + \left\|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\right\|_{\mathcal{F}}^2 / \theta^2\right)^{-1}$	$\left(1 + \left\|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\right\|_{\mathcal{F}}^{\theta}\right)^{-1}, \ \theta \leq 2$	$\ldots$

$$\left\|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\right\|_{\mathcal{F}}^2 = \mathbf{E}_P k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_Q k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})$$

Regression using *population* mean embeddings

- Samples $\mathbf{z} := \{(\mu_{\mathsf{P}_i}, y_i)\}_{i=1}^{\ell} \stackrel{\text{i.i.d.}}{\sim} \rho(\mu_{\mathsf{P}}, y) = \rho(y|\mu_{\mathsf{P}})\rho(\mu_{\mathsf{P}})$,

$$\mu_{\mathsf{P}_i} = \mathbf{E}_{\mathsf{P}_i}[\varphi_{\mathsf{x}}]$$

- Regression function

$$f_\rho(\mu_{\mathsf{P}}) = \int_{\mathbb{R}} y \, \mathrm{d}\rho(y|\mu_{\mathsf{P}}),$$

Regression using *population* mean embeddings

- Samples $\mathbf{z} := \{(\mu_{\mathsf{P}_i}, y_i)\}_{i=1}^{\ell} \overset{\text{i.i.d.}}{\sim} \rho(\mu_{\mathsf{P}}, y) = \rho(y|\mu_{\mathsf{P}})\rho(\mu_{\mathsf{P}})$,

$$\mu_{\mathsf{P}_i} = \mathbf{E}_{\mathsf{P}_i}[\varphi_{\mathsf{x}}]$$

- Regression function

$$f_\rho(\mu_{\mathsf{P}}) = \int_{\mathbb{R}} y \mathrm{d}\rho(y|\mu_{\mathsf{P}}),$$

- Ridge regression for labelled distributions

$$f_{\mathbf{z}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(f(\mu_{\mathsf{P}_i}) - y_i\right)^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad (\lambda > 0)$$

- Define RKHS $\mathcal{H}$ with kernel $K(\mu_{\mathsf{P}}, \mu_{\mathsf{Q}}) := \langle \psi_{\mu_{\mathsf{P}}}, \psi_{\mu_{\mathsf{Q}}} \rangle_{\mathcal{H}}$: functions from $F \subset \mathcal{F}$ to $\mathbb{R}$, where

$$F := \{\mu_{\mathsf{P}} \; : \; \mathsf{P} \in \mathcal{P}\} \qquad \mathcal{P} \text{ set of prob. meas. on } \mathcal{X}$$

Regression using *population* mean embeddings

- Expected risk, Excess risk

$$\mathcal{R}\left[f\right] = \mathbf{E}_{\rho(\mu_{\mathbf{P}},y)}\left(f\left(\mu_{\mathbf{P}}\right) - y\right)^2 \qquad \mathcal{E}(f_{\mathbf{z}}^{\lambda}, f_\rho) = \mathcal{R}[f_{\mathbf{z}}^{\lambda}] - \mathcal{R}[f_\rho].$$

- Minimax rate [Caponnetto and Vito, 2007]

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}, f_\rho) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right) \qquad (1 < b, c \in (1, 2]).$$

  - b size of input space, c smoothness of f_ρ

Regression using *population* mean embeddings

- Expected risk, Excess risk

$$\mathcal{R}[f] = \mathbf{E}_{\rho(\mu_{\mathsf{P}},y)}\left(f(\mu_{\mathsf{P}}) - y\right)^2 \qquad \mathcal{E}(f_{\mathbf{z}}^{\lambda}, f_{\rho}) = \mathcal{R}[f_{\mathbf{z}}^{\lambda}] - \mathcal{R}[f_{\rho}].$$

- Minimax rate [Caponnetto and Vito, 2007]

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda}, f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right) \quad (1 < b, c \in (1,2]).$$

  - b size of input space, c smoothness of f_{ρ}

- Replace μ_{P_i} with $\hat{\mu}_{\mathsf{P}_i} = N^{-1}\sum_{j=1}^{N}\varphi_{x_j} \qquad x_j \overset{\text{i.i.d.}}{\sim} \mathsf{P}_i$

- Given $N = \ell^a \log(\ell)$ and $a = 2$, (and Hölder condition on $\psi : F \to \mathcal{H}$)

$$\mathcal{E}(f_{\hat{\mathbf{z}}}^{\lambda}, f_{\rho}) = \mathcal{O}_p\left(\ell^{-\frac{bc}{bc+1}}\right) \quad (1 < b, c \in (1,2]).$$

Same rate as for population μ_{P_i} embeddings! [AISTATS15, JMLR in revision]

Kernels on distributions in supervised learning

Supervised learning applications:

- Regression: From distributions to vector spaces.[AISTATS15]
  - Atmospheric monitoring, predict aerosol value from distribution of pixel values of a multispectral satellite image over an area (performance matches engineered state-of-the-art [Wang et al., 2012])
- Expectation propagation: learn to predict outgoing messages from incoming messages, when updates would otherwise be done by numerical integration [UAI15]
- Learning causal direction with mean embeddings [Lopez-Paz et al., 2015]

Learning causal direction with mean embeddings

Additive noise model to direct an edge between random variables x and y

[Hoyer et al., 2009]

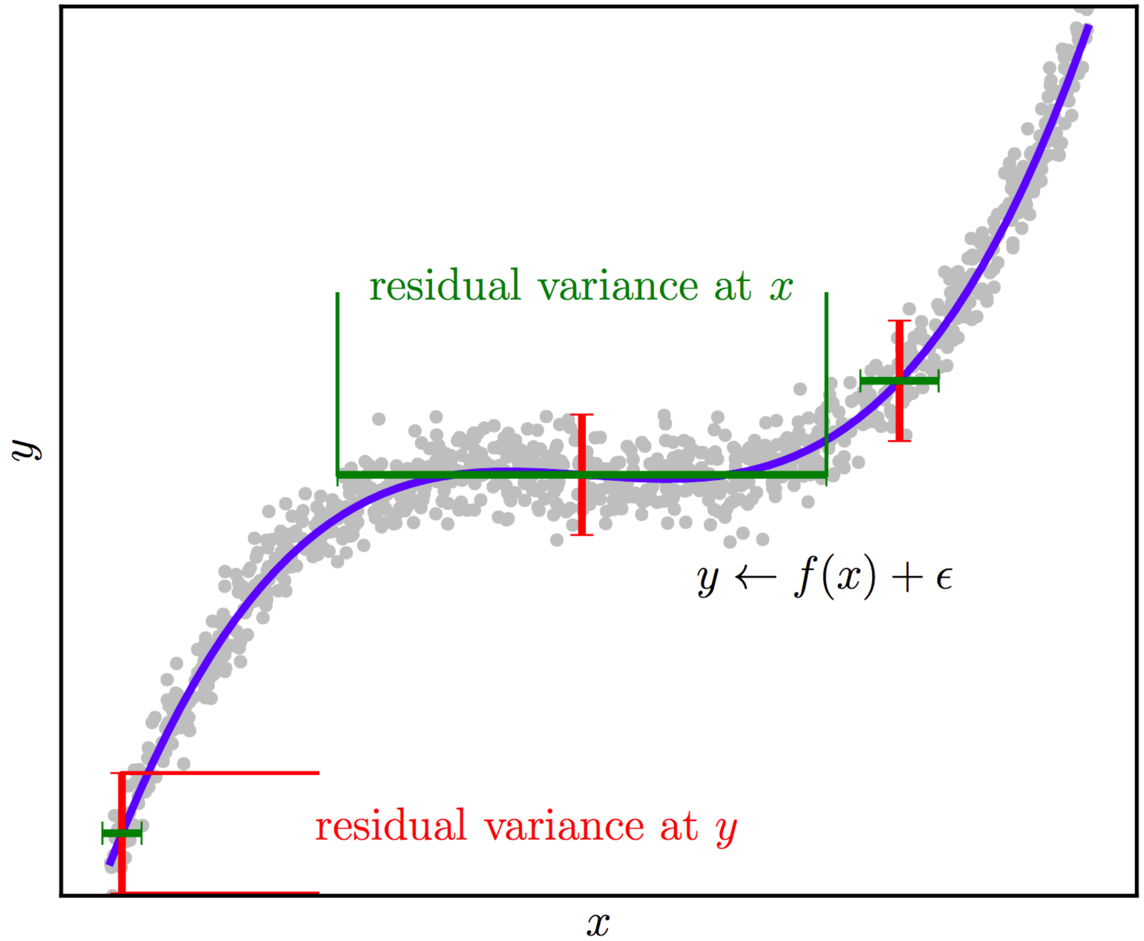


Figure: D. Lopez-Paz

Learning causal direction with mean embeddings

Classification of cause-effect relations [Lopez-Paz et al., 2015]

- Tuebingen cause-effect pairs: 82 scalar real-world examples where causes and effects known [Zscheischler, J., 2014]
- Training data: artificial, random nonlinear functions with additive gaussian noise.
- Features: $\hat{\mu}_{\mathbf{P}_x}, \hat{\mu}_{\mathbf{P}_y}, \hat{\mu}_{\mathbf{P}_{xy}}$ with labels for $x \to y$ and $y \to x$
- Performance 81% correct

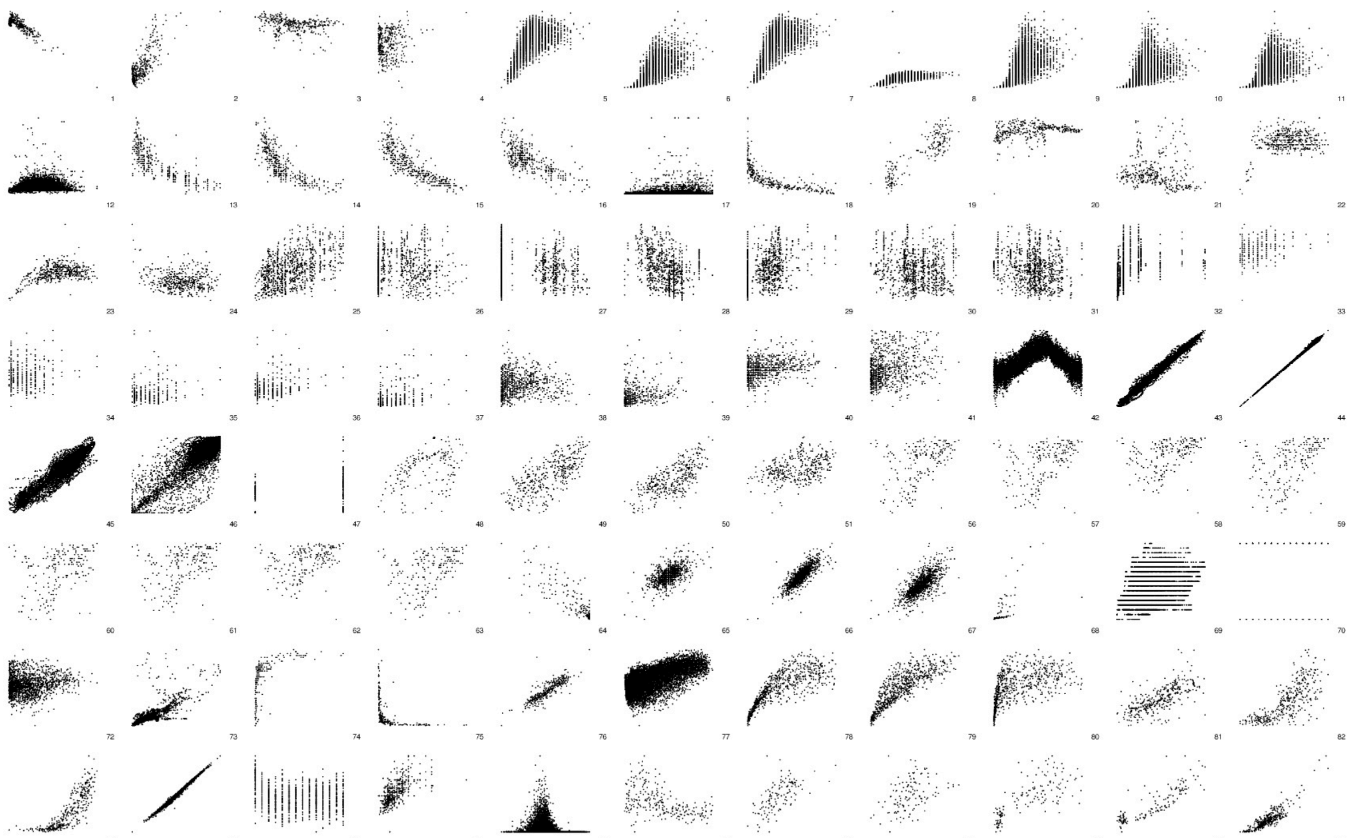


Figure:Mooij et al.(2015)

Co-authors

- **From UCL:**
  - Luca Baldasssarre
  - Steffen Grunewalder
  - Guy Lever
  - Sam Patterson
  - Massimiliano Pontil
  - Dino Sejdinovic
- **External:**
  - Karsten Borgwardt, MPI
  - Wicher Bergsma, LSE
  - Kenji Fukumizu, ISM
  - Zaid Harchaoui, INRIA
  - Bernhard Schoelkopf, MPI
  - Alex Smola, CMU/Google
  - Le Song, Georgia Tech
  - Bharath Sriperumbudur, Cambridge

Kernel two-sample tests for big data, optimal kernel choice

Quadratic time estimate of MMD

$$\text{MMD}^2 = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|^2_{\mathcal{F}} = \mathbf{E}_{\mathsf{P}} k(x, x') + \mathbf{E}_{\mathsf{Q}} k(y, y') - 2\mathbf{E}_{\mathsf{P},\mathsf{Q}} k(x, y)$$

Quadratic time estimate of MMD

$$\text{MMD}^2 = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_{\mathbf{P}} k(x, x') + \mathbf{E}_{\mathbf{Q}} k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)$$

Given i.i.d. $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_m\}$ from $\mathbf{P}, \mathbf{Q}$, respectively:

The earlier estimate: (quadratic time)

$$\widehat{\mathbb{E}}_{\mathbf{P}} k(x, x') = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j)$$

Quadratic time estimate of MMD

$$\mathrm{MMD}^2 = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_{\mathbf{P}} k(x, x') + \mathbf{E}_{\mathbf{Q}} k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)$$

Given i.i.d. $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_m\}$ from $\mathbf{P}$, $\mathbf{Q}$, respectively:

The earlier estimate: (quadratic time)

$$\widehat{\mathbb{E}}_{\mathbf{P}} k(x, x') = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j)$$

New, linear time estimate:

$$\widehat{\mathbb{E}}_{\mathbf{P}} k(x, x') = \frac{2}{m} \left[k(x_1, x_2) + k(x_3, x_4) + \ldots \right]$$

$$= \frac{2}{m} \sum_{i=1}^{m/2} k(x_{2i-1}, x_{2i})$$

Linear time MMD

Shorter expression with explicit k dependence:

$$\mathrm{MMD}^2 =: \eta_k(p,q) = \mathbb{E}_{xx'yy'} h_k(x,x',y,y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x,x',y,y') = k(x,x') + k(y,y') - k(x,y') - k(x',y),$$

and $v := [x,\ x',\ y,\ y']$.

Linear time MMD

Shorter expression with explicit k dependence:

$$\mathrm{MMD}^2 =: \eta_k(p,q) = \mathbb{E}_{xx'yy'} h_k(x,x',y,y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x,x',y,y') = k(x,x') + k(y,y') - k(x,y') - k(x',y),$$

and $v := [x,\ x',\ y,\ y']$.

The linear time estimate again:

$$\check{\eta}_k = \frac{2}{m} \sum_{i=1}^{m/2} h_k(v_i),$$

where $v_i := [x_{2i-1},\ x_{2i},\ y_{2i-1},\ y_{2i}]$ and

$h_k(v_i) := k(x_{2i-1}, x_{2i}) + k(y_{2i-1}, y_{2i}) - k(x_{2i-1}, y_{2i}) - k(x_{2i}, y_{2i-1})$

Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given m, hence…
- …a much less powerful test for a given m

Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given m, hence...
- ...a much less powerful test for a given m

Advantages of the linear time MMD vs quadratic time MMD

- Very simple asymptotic null distribution (a Gaussian, vs an infinite weighted sum of χ^2)
- Both test statistic and threshold computable in $O(m)$, with storage $O(1)$.
- Given unlimited data, a given Type II error can be attained with less computation

Asymptotics of linear time MMD

By central limit theorem,

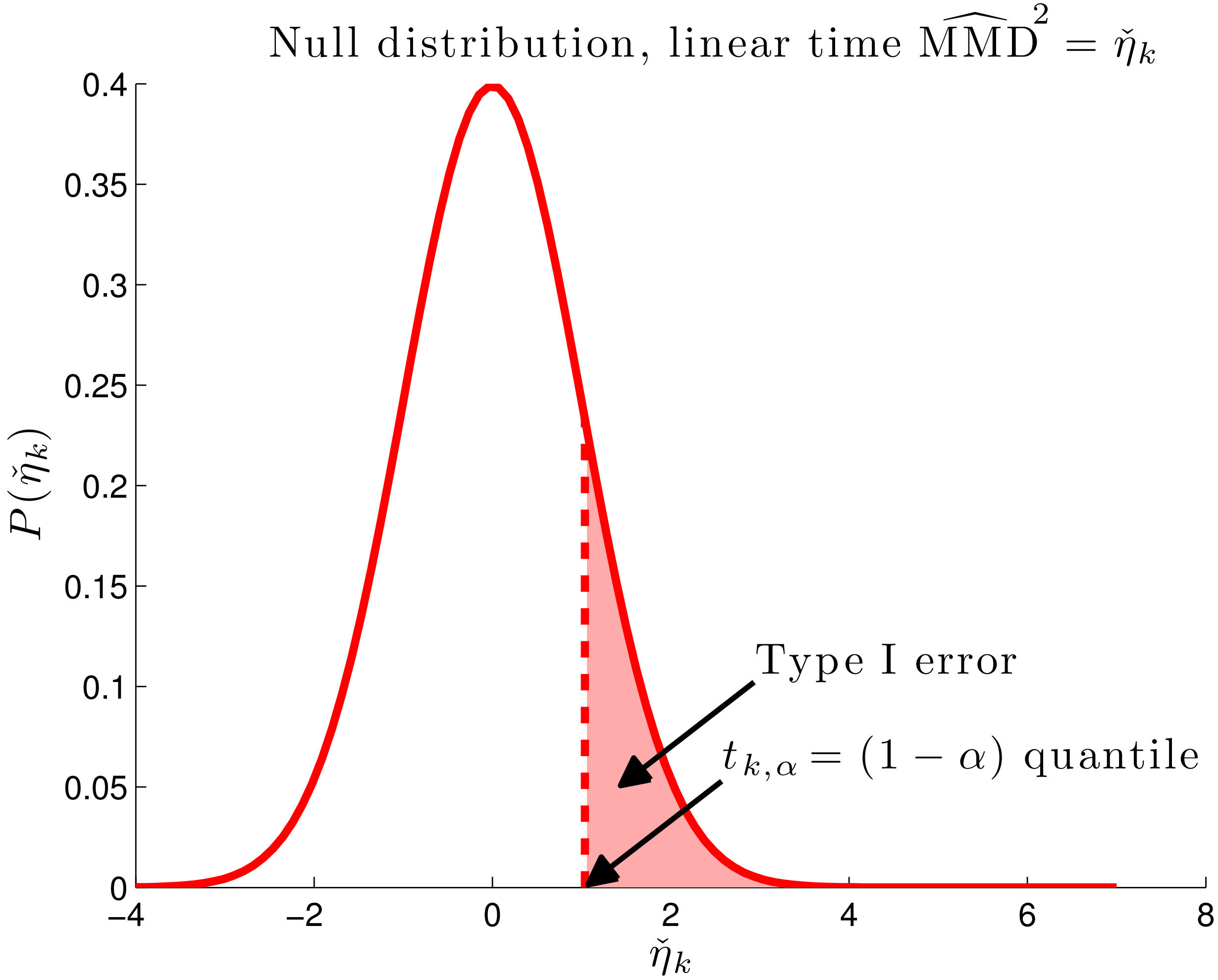
$$m^{1/2}\left(\check{\eta}_k - \eta_k(p,q)\right) \xrightarrow{D} \mathcal{N}(0, 2\sigma_k^2)$$

- assuming $0 < \mathbb{E}(h_k^2) < \infty$ (true for bounded k)
- $\sigma_k^2 = \mathbb{E}_v h_k^2(v) - \left[\mathbb{E}_v(h_k(v))\right]^2.$

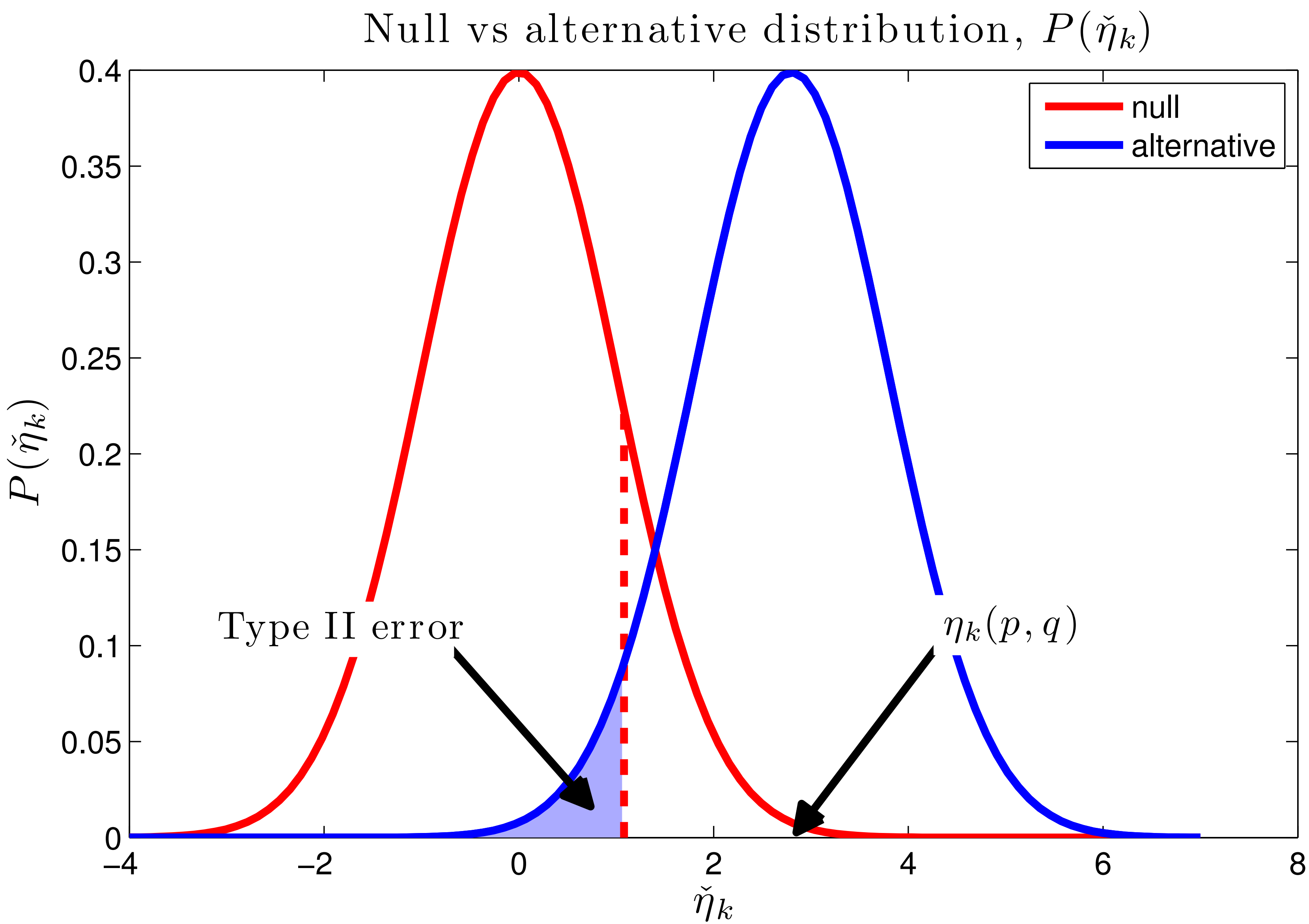
Hypothesis test

Hypothesis test of asymptotic level α:

$t_{k,\alpha} = m^{-1/2}\sigma_k\sqrt{2}\Phi^{-1}(1-\alpha)$ where Φ^{-1} is inverse CDF of $\mathcal{N}(0,1)$.



Type II error



The best kernel: minimizes Type II error

Type II error: $\check{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p,q) > 0$.

Prob. of a Type II error:

$$P(\check{\eta}_k < t_{k,\alpha}) = \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\eta_k(p,q)\sqrt{m}}{\sigma_k\sqrt{2}}\right)$$

where Φ is a Normal CDF.

The best kernel: minimizes Type II error

Type II error: $\check{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p,q) > 0$.

Prob. of a Type II error:

$$P(\check{\eta}_k < t_{k,\alpha}) = \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\eta_k(p,q)\sqrt{m}}{\sigma_k\sqrt{2}}\right)$$

where Φ is a Normal CDF.

Since Φ monotonic, best kernel choice to minimize Type II error prob. is:

$$k_* = \arg\max_{k\in\mathcal{K}} \eta_k(p,q)\sigma_k^{-1},$$

where $\mathcal{K}$ is the family of kernels under consideration.

Learning the best kernel in a family

Define the family of kernels as follows:

$$\mathcal{K} := \left\{ k \, : \, k = \sum_{u=1}^{d} \beta_u k_u, \; \|\beta\|_1 = D, \; \beta_u \geq 0, \; \forall u \in \{1, \ldots, d\} \right\}.$$

Properties: if at least one $\beta_u > 0$

- all $k \in \mathcal{K}$ are valid kernels,
- If all k_u charateristic then k characteristic

Test statistic

The squared MMD becomes

$$\eta_k(p,q) = \|\mu_k(p) - \mu_k(q)\|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p,q),$$

where $\eta_u(p,q) := \mathbb{E}_v h_u(v)$.

Test statistic

The squared MMD becomes

$$\eta_k(p,q) = \|\mu_k(p) - \mu_k(q)\|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p,q),$$

where $\eta_u(p,q) := \mathbb{E}_v h_u(v)$.

Denote:

- $\beta = (\beta_1, \beta_2, \ldots, \beta_d)^\top \in \mathbb{R}^d$,
- $h = (h_1, h_2, \ldots, h_d)^\top \in \mathbb{R}^d$,
  - $h_u(x, x', y, y') = k_u(x, x') + k_u(y, y') - k_u(x, y') - k_u(x', y)$
- $\eta = \mathbb{E}_v(h) = (\eta_1, \eta_2, \ldots, \eta_d)^\top \in \mathbb{R}^d$.

Quantities for test:

$$\eta_k(p,q) = \mathbb{E}(\beta^\top h) = \beta^\top \eta \qquad \sigma_k^2 := \beta^\top \mathrm{cov}(h) \beta.$$

Optimization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \qquad\qquad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta},$$

$\hat{Q}$ is empirical estimate of $\mathrm{cov}(h)$.

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\check{\eta}_k, \check{\sigma}_k$ on data to be tested (why?)

Optimization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \qquad\qquad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top \left(\hat{Q} + \lambda_m I\right)\beta},$$

$\hat{Q}$ is empirical estimate of $\mathrm{cov}(h)$.

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\check{\eta}_k, \check{\sigma}_k$ on data to be tested (why?)

Objective:

$$\begin{aligned}\hat{\beta}^* &= \arg\max_{\beta \succeq 0} \; \hat{\eta}_k(p,q)\hat{\sigma}_{k,\lambda}^{-1} \\ &= \arg\max_{\beta \succeq 0} \; \left(\beta^\top \hat{\eta}\right)\left(\beta^\top \left(\hat{Q} + \lambda_m I\right)\beta\right)^{-1/2} \\ &=: \alpha(\beta; \hat{\eta}, \hat{Q})\end{aligned}$$

Optmization of ratio $\eta_k(p, q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\quad \beta \succeq 0 \quad$ s.t. $\quad \alpha(\beta; \hat{\eta}, \hat{Q}) > 0.$

Thus: $\quad \alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\quad \beta \succeq 0 \quad$ s.t. $\quad \alpha(\beta; \hat{\eta}, \hat{Q}) > 0.$

Thus: $\quad \alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier problem: $\quad \hat{\beta}^* = \arg\max_{\beta \succeq 0} \alpha^2(\beta; \hat{\eta}, \hat{Q}).$

Quadratic program:

$$\min\{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta : \beta^\top \hat{\eta} = 1,\ \beta \succeq 0\}$$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\quad \beta \succeq 0 \quad$ s.t. $\quad \alpha(\beta; \hat{\eta}, \hat{Q}) > 0.$

Thus: $\quad \alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier problem: $\quad \hat{\beta}^* = \arg\max_{\beta \succeq 0} \alpha^2(\beta; \hat{\eta}, \hat{Q}).$

Quadratic program:

$$\min\{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta : \beta^\top \hat{\eta} = 1,\ \beta \succeq 0\}$$

What if $\hat{\eta}$ has no positive entries?

Test procedure

1. Split the data into testing and training.
2. On the training data:
   (a) Compute $\hat{\eta}_u$ for all $k_u \in \mathcal{K}$
   (b) If at least one $\hat{\eta}_u > 0$, solve the QP to get β^*, else choose random kernel from $\mathcal{K}$
3. On the test data:
   (a) Compute $\check{\eta}_{k^*}$ using $k^* = \sum_{u=1}^{d} \beta^* k_u$
   (b) Compute test threshold $\check{t}_{\alpha,k^*}$ using $\check{\sigma}_{k^*}$
4. Reject null if $\check{\eta}_{k^*} > \check{t}_{\alpha,k^*}$

Convergence bounds

Assume bounded kernel, σ_k, bounded away from 0.
If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| = O_P \left(m^{-1/3} \right).$$

Convergence bounds

Assume bounded kernel, σ_k, bounded away from 0.

If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| = O_P\left(m^{-1/3}\right).$$

Idea:

$$\begin{aligned}
&\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| \\
&\leq \sup_{k \in \mathcal{K}} \left| \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \eta_k \sigma_{k,\lambda}^{-1} \right| + \sup_{k \in \mathcal{K}} \left| \eta_k \sigma_{k,\lambda}^{-1} - \eta_k \sigma_k^{-1} \right| \\
&\leq \frac{\sqrt{d}}{D\sqrt{\lambda_m}} \left(C_1 \sup_{k \in \mathcal{K}} |\hat{\eta}_k - \eta_k| + C_2 \sup_{k \in \mathcal{K}} |\hat{\sigma}_{k,\lambda} - \sigma_{k,\lambda}| \right) + C_3 D^2 \lambda_m,
\end{aligned}$$

Experiments

Competing approaches

- Median heuristic
- Max. MMD: choose $k_u \in \mathcal{K}$ with the largest $\hat{\eta}_u$
  - same as maximizing $\beta^\top \hat{\eta}$ subject to $\|\beta\|_1 \leq 1$
- ℓ_2 statistic: maximize $\beta^\top \hat{\eta}$ subject to $\|\beta\|_2 \leq 1$
- Cross validation on training set

Also compare with:

- Single kernel that maximizes ratio $\eta_k(p, q)\sigma_k^{-1}$

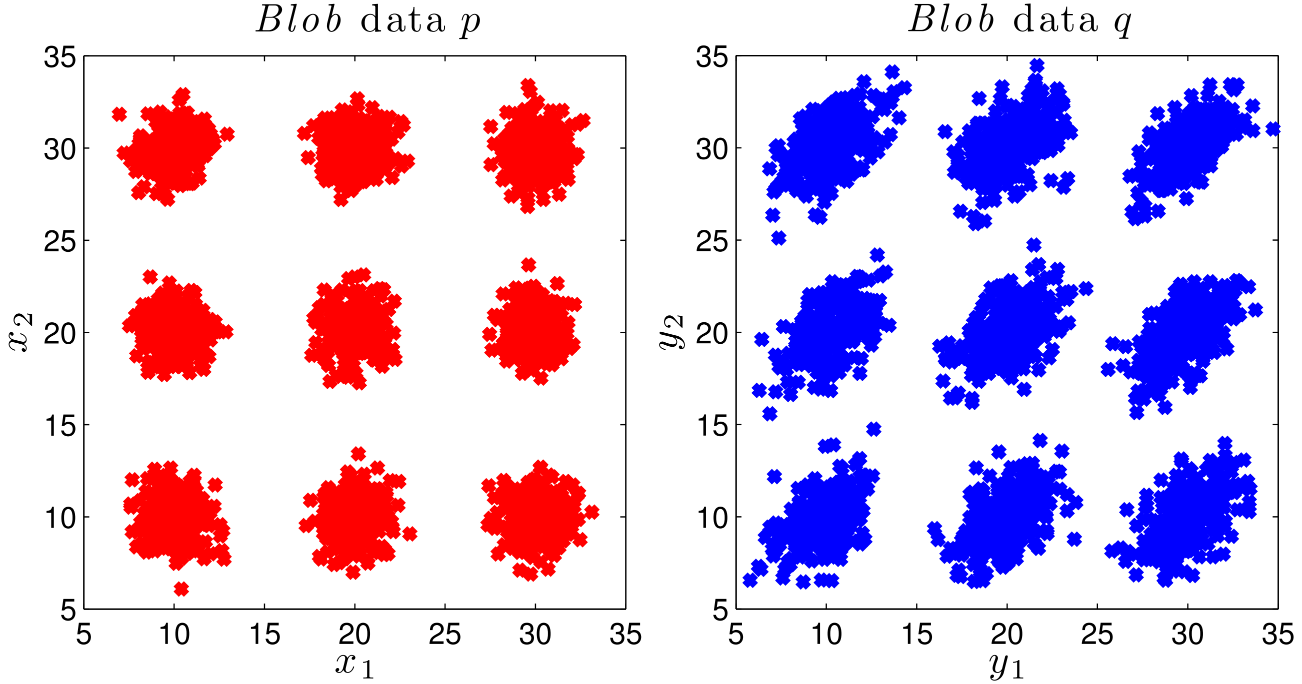
Blobs: data

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

Blobs: data

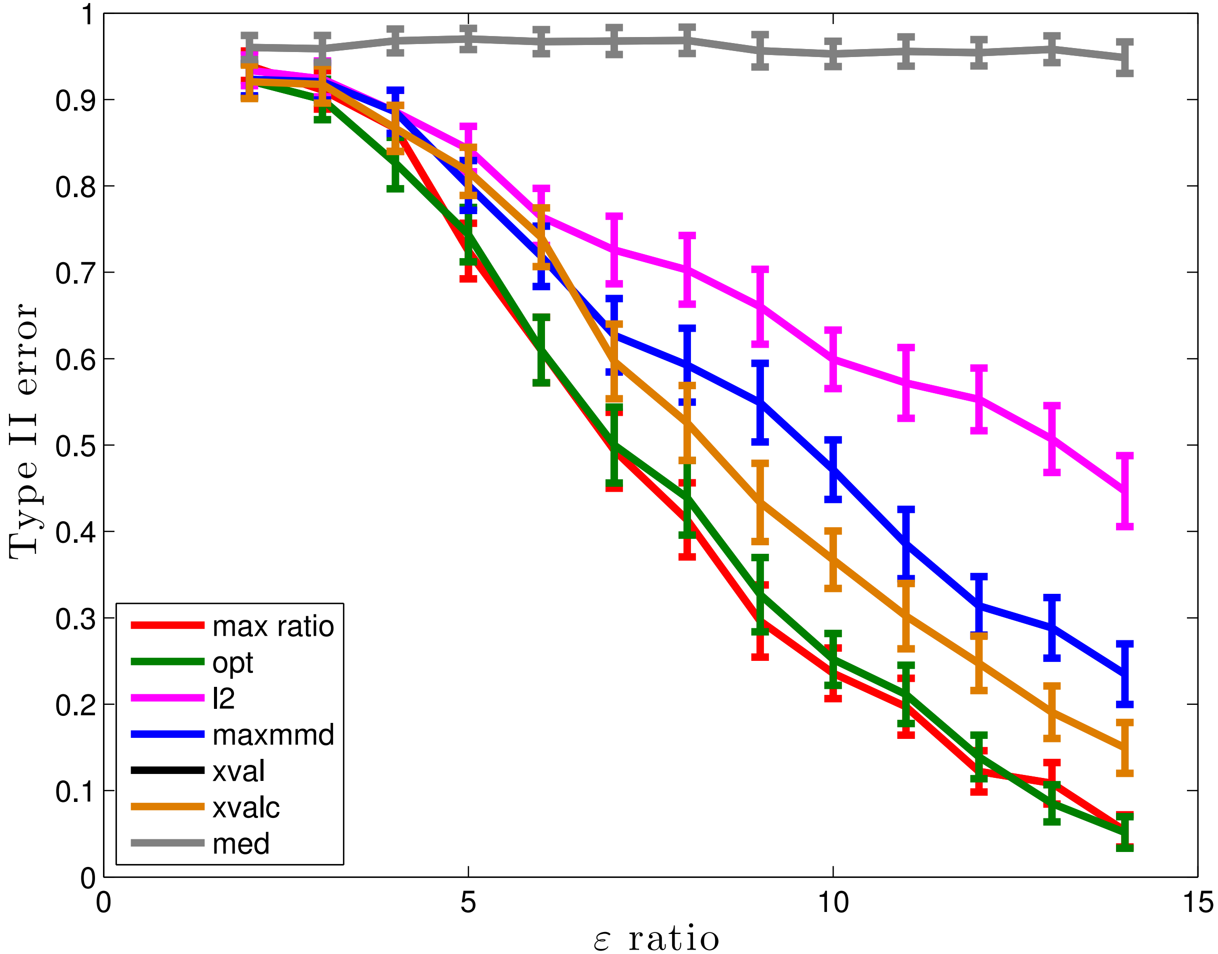
Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

We distinguish a field of Gaussian blobs with different covariances.



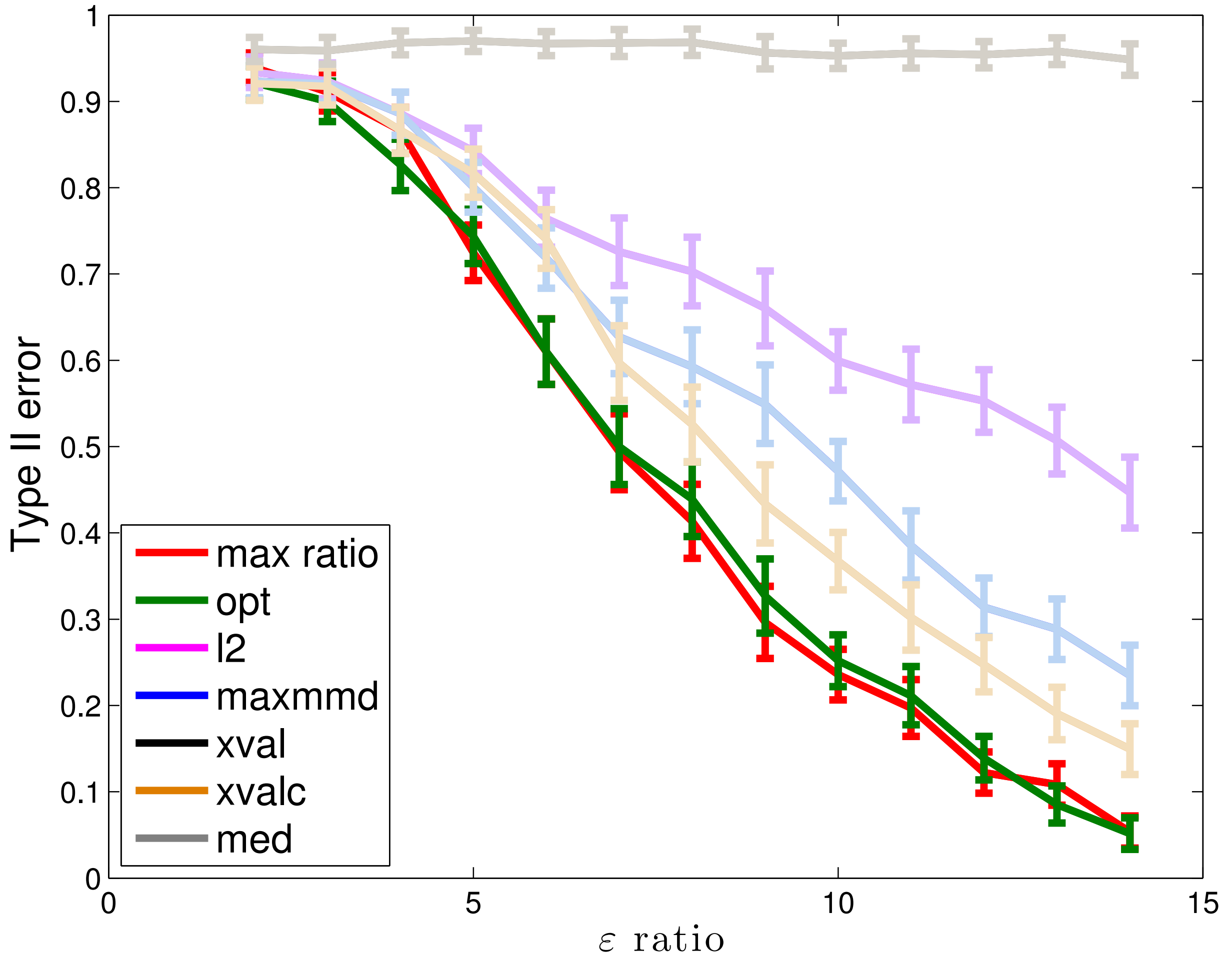
Ratio $\varepsilon = 3.2$ of largest to smallest eigenvalues of blobs in q.

Blobs: results



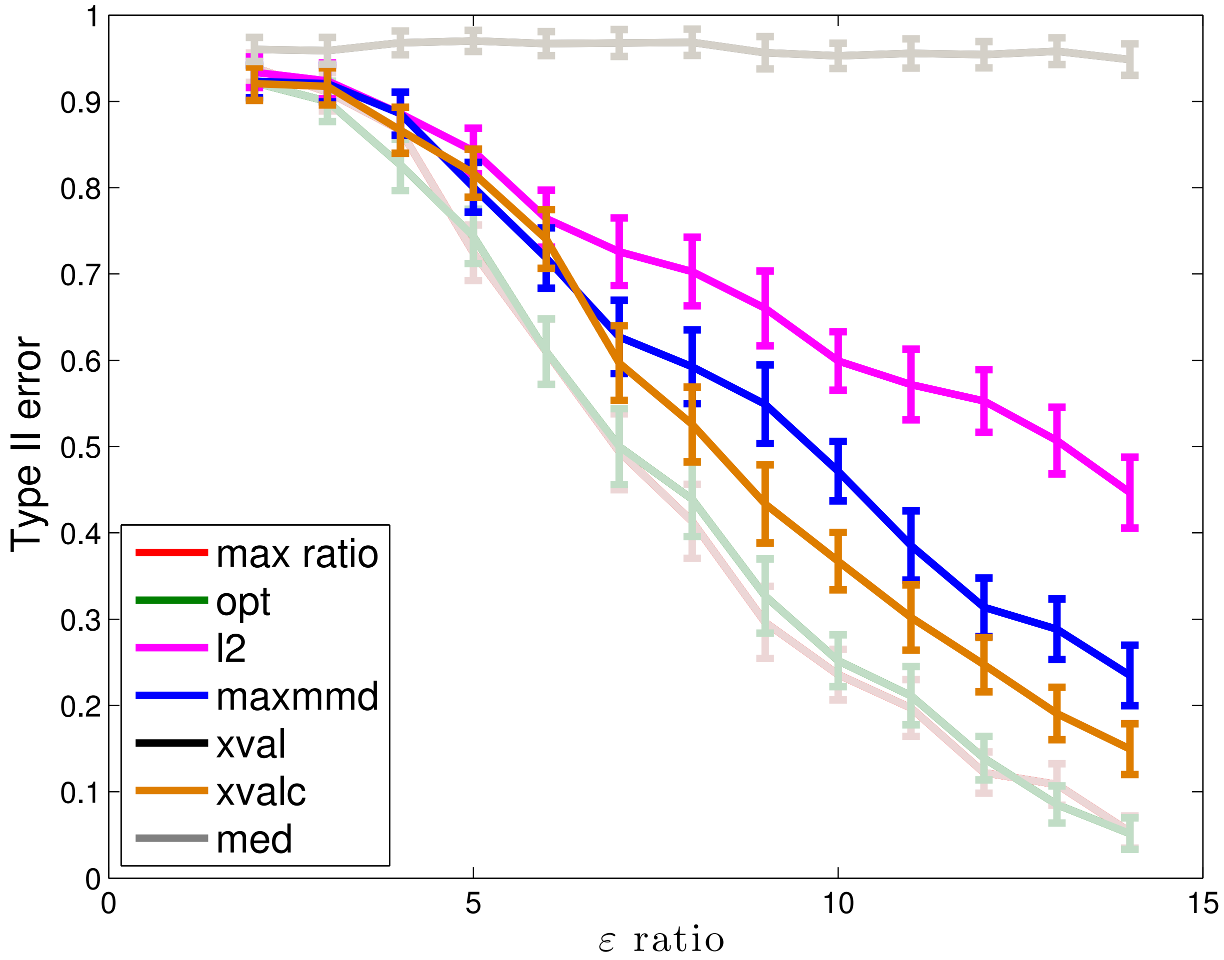
Parameters: $m = 10,000$ (for training and test). Ratio ε of largest to smallest eigenvalues of blobs in q. Results are average over 617 trials.

Blobs: results



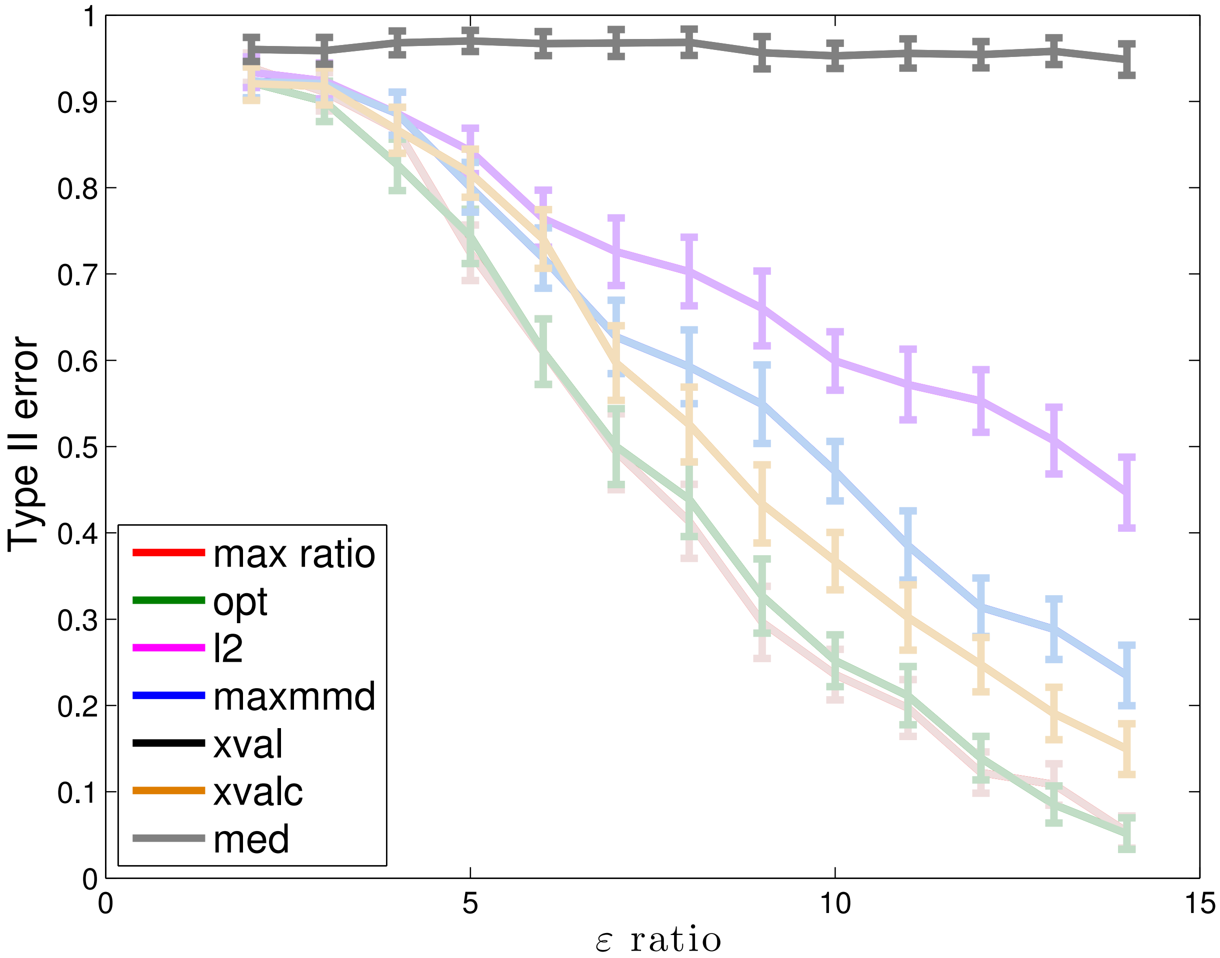
Optimize ratio $\eta_k(p,q)\sigma_k^{-1}$

Blobs: results



Maximize $\eta_k(p, q)$ with β constraint

Blobs: results



Median heuristic

Feature selection: data

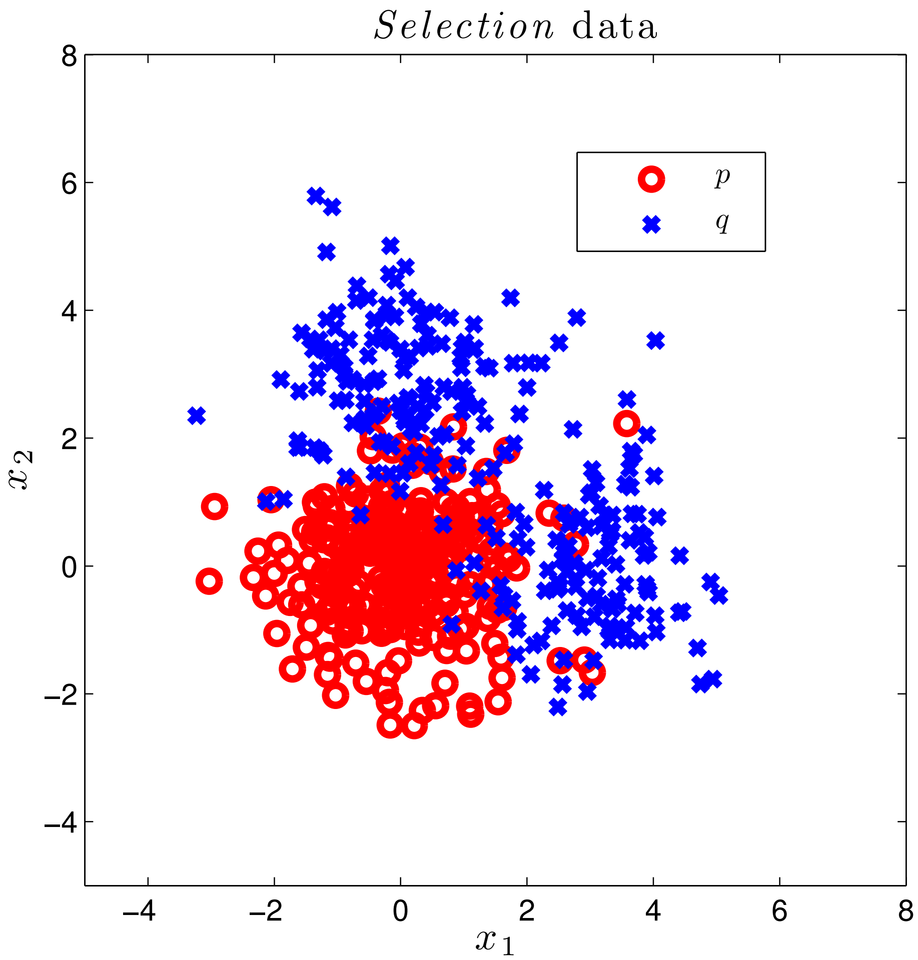
Idea: no single best kernel.

Each of the k_u are univariate (along a single coordinate)

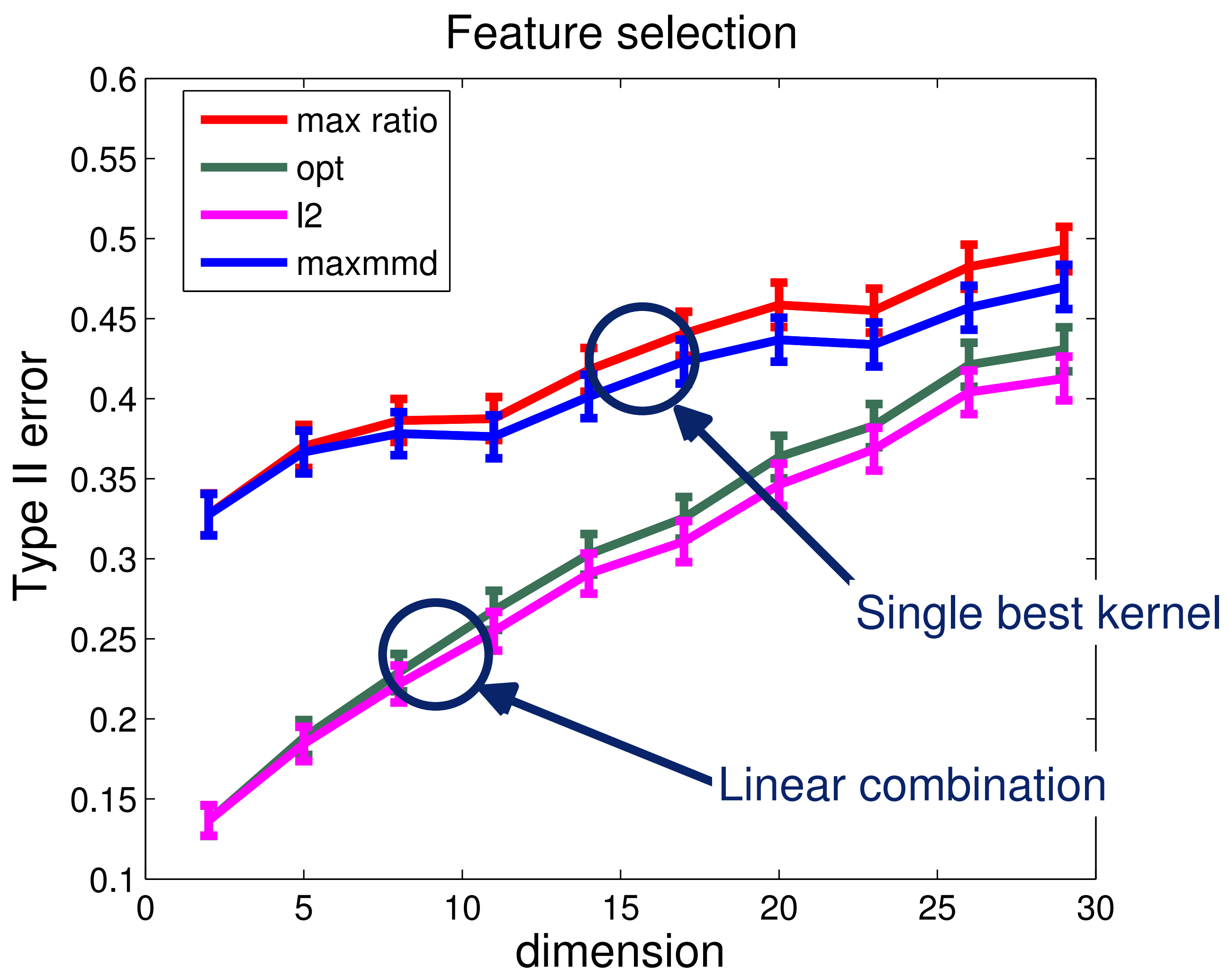
Feature selection: data

Idea: no single best kernel.

Each of the k_u are univariate (along a single coordinate)



Feature selection: results



$m = 10,000$, average over 5000 trials

Amplitude modulated signals

Given an audio signal $s(t)$, an amplitude modulated signal can be defined

$$u(t) = \sin(\omega_c t)\,[a\,s(t) + l]$$

- ω_c: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset

Amplitude modulated signals

Given an audio signal $s(t)$, an amplitude modulated signal can be defined

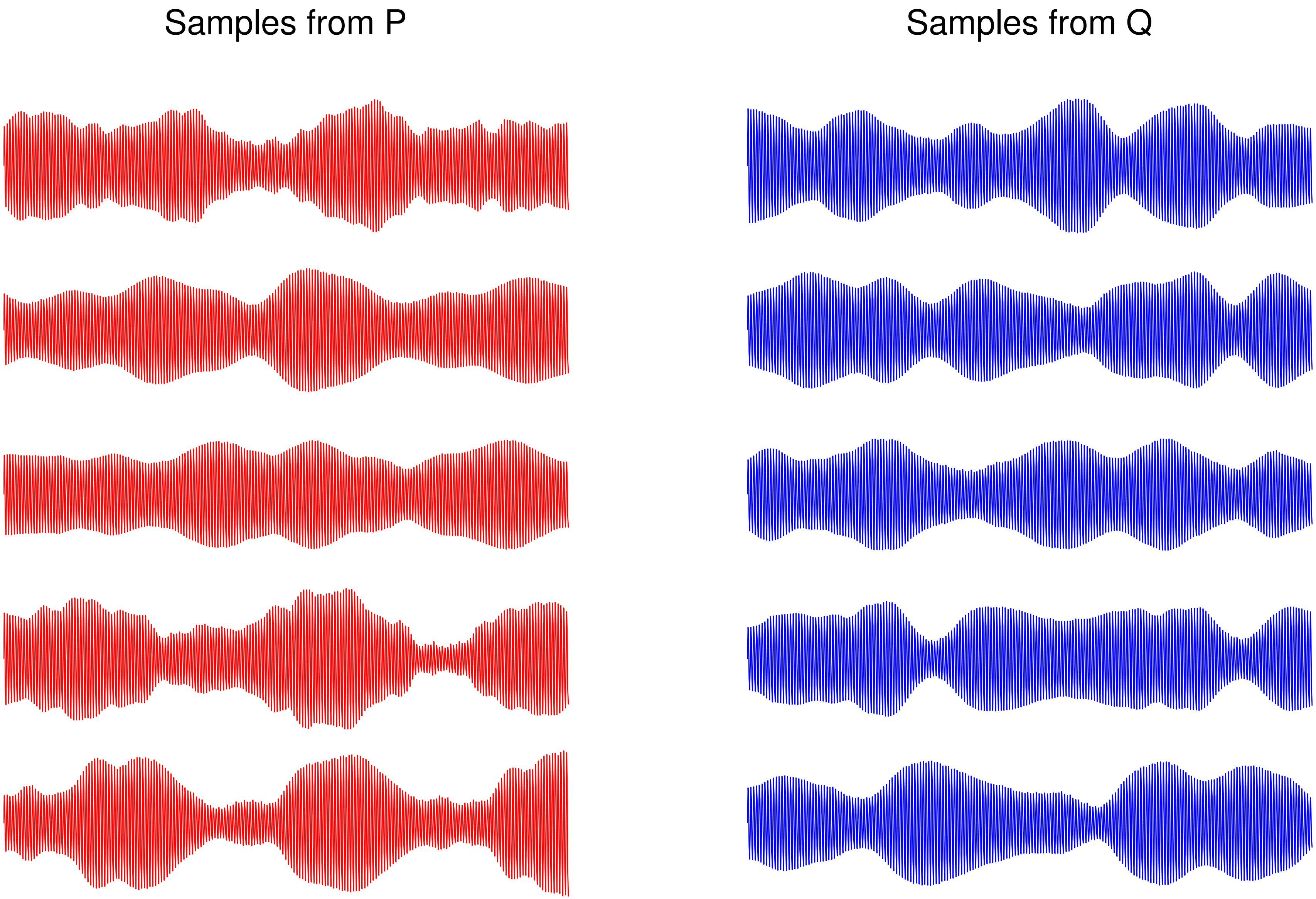
$$u(t) = \sin(\omega_c t)\,[a\,s(t) + l]$$

- ω_c: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset

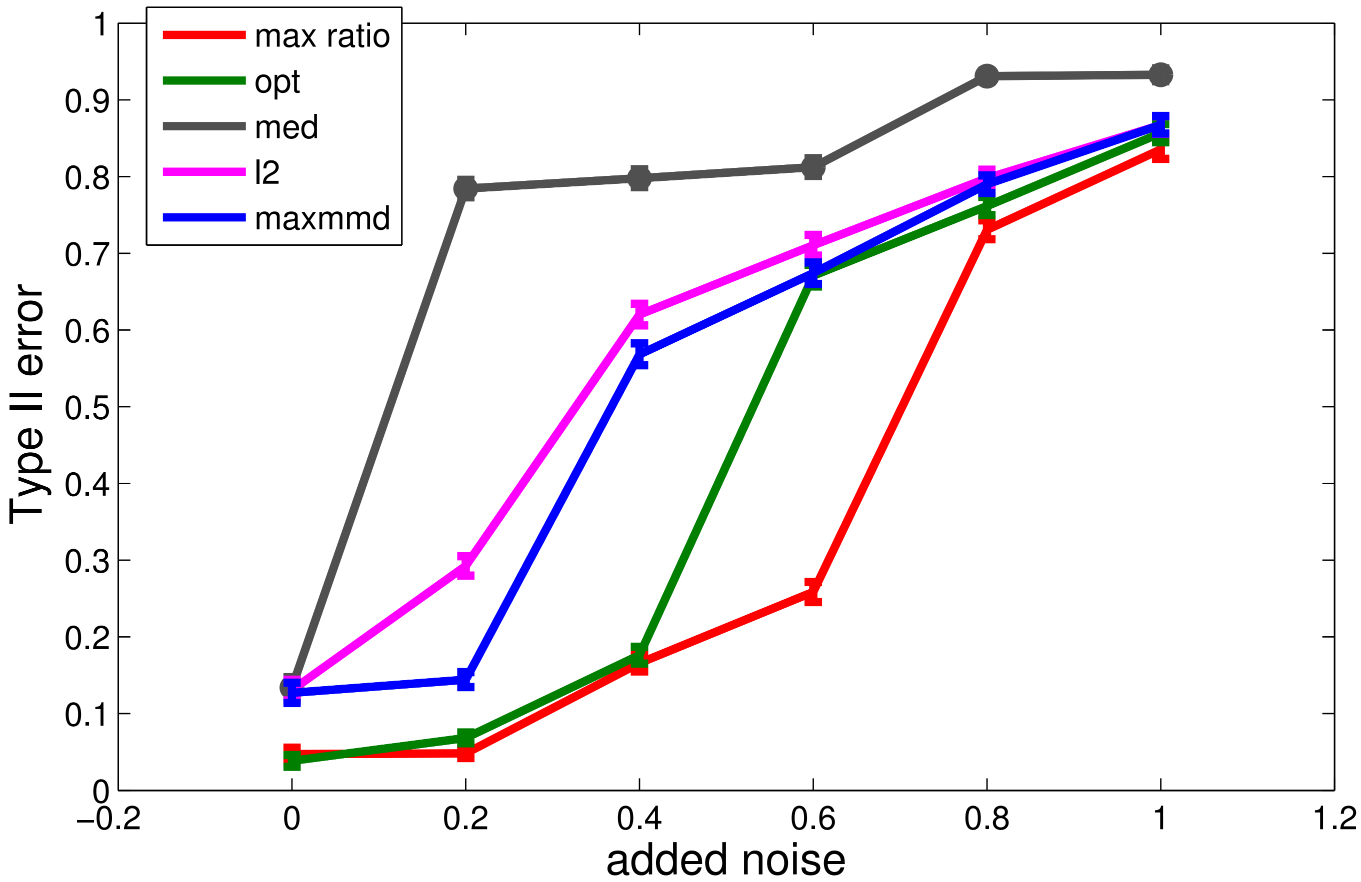
Two amplitude modulated signals from same artist (in this case, Magnetic Fields).

- Music sampled at 8KHz (very low)
- Carrier frequency is 24kHz
- AM signal observed at 120kHz
- Samples are extracts of length $N = 1000$, approx. 0.01 sec (very short).
- Total dataset size is 30,000 samples from each of p, q.

Amplitude modulated signals



Results: AM signals



$m = 10,000$ (for training and test) and scaling $a = 0.5$. Average over 4124 trials. Gaussian noise added.

Observations on kernel choice

- It is possible to choose the best kernel for a kernel two-sample test
- Kernel choice matters for "difficult" problems, where the distributions differ on a lengthscale different to that of the data.
- Ongoing work:
  - quadratic time statistic
  - avoid training/test split

Energy Distance and the MMD

Energy distance and MMD

Distance between probability distributions:

Energy distance: [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, 2005]

$$D_E(\mathbf{P}, \mathbf{Q}) = \mathbf{E}_{\mathbf{P}}\|X - X'\|^q + \mathbf{E}_{\mathbf{Q}}\|Y - Y'\|^q - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}}\|X - Y\|^q$$

$$0 < q \leq 2$$

Maximum mean discrepancy [Gretton et al., 2007, Smola et al., 2007, Gretton et al., 2012a]

$$\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}; F) = \mathbf{E}_{\mathbf{P}} k(X, X') + \mathbf{E}_{\mathbf{Q}} k(Y, Y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(X, Y)$$

Energy distance and MMD

Distance between probability distributions:

Energy distance: [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, 2005]

$$D_E(\mathbf{P}, \mathbf{Q}) = \mathbf{E}_{\mathbf{P}} \|X - X'\|^q + \mathbf{E}_{\mathbf{Q}} \|Y - Y'\|^q - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} \|X - Y\|^q$$

$$0 < q \leq 2$$

Maximum mean discrepancy [Gretton et al., 2007, Smola et al., 2007, Gretton et al., 2012a]

$$\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}; F) = \mathbf{E}_{\mathbf{P}} k(X, X') + \mathbf{E}_{\mathbf{Q}} k(Y, Y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(X, Y)$$

Energy distance is MMD with a particular kernel! [Sejdinovic et al., 2013b]

Distance covariance and HSIC

Distance covariance $(0 < q, r \leq 2)$ [Feuerverger, 1993, Székely et al., 2007]

$$\mathcal{V}^2(X,Y) = \mathbf{E}_{XY}\mathbf{E}_{X'Y'}\left[\|X - X'\|^q \|Y - Y'\|^r\right] + \mathbf{E}_X\mathbf{E}_{X'}\|X - X'\|^q \mathbf{E}_Y\mathbf{E}_{Y'}\|Y - Y'\|^r - 2\mathbf{E}_{XY}\left[\mathbf{E}_{X'}\|X - X'\|^q \mathbf{E}_{Y'}\|Y - Y'\|^r\right]$$

Hilbert-Schmdit Indepence Criterion [Gretton et al., 2005, Smola et al., 2007, Gretton et al., 2008, Gretton and Gyorfi, 2010]Define RKHS $\mathcal{F}$ on $\mathcal{X}$ with kernel k, RKHS $\mathcal{G}$ on $\mathcal{Y}$ with kernel l. Then

$$\begin{aligned}&\mathrm{HSIC}(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y)\\ &= \mathbf{E}_{XY}\mathbf{E}_{X'Y'}k(X,X')l(Y,Y') + \mathbf{E}_X\mathbf{E}_{X'}k(X,X')\mathbf{E}_Y\mathbf{E}_{Y'}l(Y,Y')\\ &\quad - 2\mathbf{E}_{X'Y'}\left[\mathbf{E}_X k(X,X')\mathbf{E}_Y l(Y,Y')\right].\end{aligned}$$

Distance covariance and HSIC

Distance covariance $(0 < q, r \leq 2)$ [Feuerverger, 1993, Székely et al., 2007]

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

Hilbert-Schmdit Indepence Criterion [Gretton et al., 2005, Smola et al., 2007, Gretton et al., 2008, Gretton and Gyorfi, 2010]Define RKHS $\mathcal{F}$ on $\mathcal{X}$ with kernel k, RKHS $\mathcal{G}$ on $\mathcal{Y}$ with kernel l. Then

$$\begin{aligned}&\mathrm{HSIC}(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y) \\ &= \mathbf{E}_{XY}\mathbf{E}_{X'Y'}k(X,X')l(Y,Y') + \mathbf{E}_X\mathbf{E}_{X'}k(X,X')\mathbf{E}_Y\mathbf{E}_{Y'}l(Y,Y') \\ &\quad - 2\mathbf{E}_{X'Y'}\left[\mathbf{E}_X k(X,X')\mathbf{E}_Y l(Y,Y')\right].\end{aligned}$$

Distance covariance is HSIC with particular kernels! [Sejdinovic et al., 2013b]

Semimetrics and Hilbert spaces

Theorem [Berg et al., 1984, Lemma 2.1, p. 74]

$\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semimetric (no triangle inequality) on $\mathcal{X}$. Let $z_0 \in \mathcal{X}$, and denote

$$k_\rho(z, z') = \frac{1}{2}(\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')).$$

Then k is positive definite (via Moore-Arnonsajn, defines a unique RKHS) iff ρ is of negative type.

Call k_ρ a distance induced kernel

Negative type: The semimetric space $(\mathcal{Z}, \rho)$ is said to have negative type if $\forall n \geq 2$, $z_1, \ldots, z_n \in \mathcal{Z}$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^{n} \alpha_i = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0. \tag{1}$$

Semimetrics and Hilbert spaces

Theorem [Berg et al., 1984, Lemma 2.1, p. 74]

$\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semimetric (no triangle inequality) on $\mathcal{X}$. Let $z_0 \in \mathcal{X}$, and denote

$$k_\rho(z, z') = \frac{1}{2}(\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')).$$

Then k is positive definite (via Moore-Arnonsajn, defines a unique RKHS) iff ρ is of negative type.

Call k_ρ a distance induced kernel

Special case: $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\rho_q(z, z') = \|z - z'\|^q$. Then ρ_q is a valid semimetric of negative type for $0 < q \leq 2$.

Semimetrics and Hilbert spaces

Theorem [Berg et al., 1984, Lemma 2.1, p. 74]

$\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a semimetric (no triangle inequality) on $\mathcal{X}$. Let $z_0 \in \mathcal{X}$, and denote

$$k_\rho(z, z') = \frac{1}{2}(\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')).$$

Then k is positive definite (via Moore-Arnonsajn, defines a unique RKHS) iff ρ is of negative type.

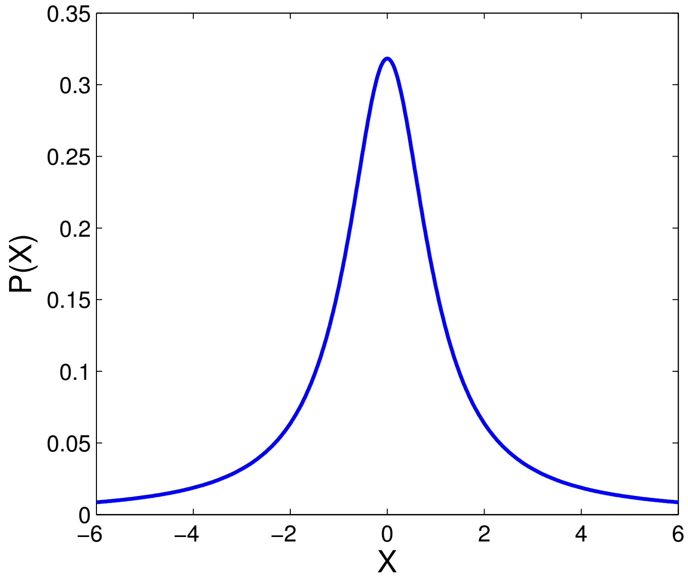
Call k_ρ a distance induced kernel

Special case: $\mathcal{Z} \subseteq \mathbb{R}^d$ and $\rho_q(z, z') = \|z - z'\|^q$. Then ρ_q is a valid semimetric of negative type for $0 < q \leq 2$.

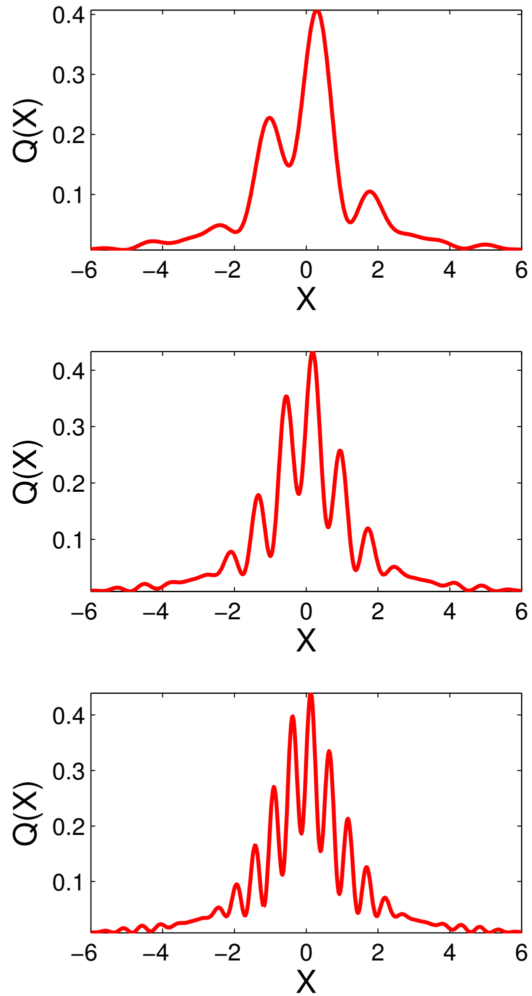
Energy distance is MMD with a distance induced kernel
Distance covariance is HSIC with distance induced kernels

Two-sample testing benchmark

Two-sample testing example in 1-D:

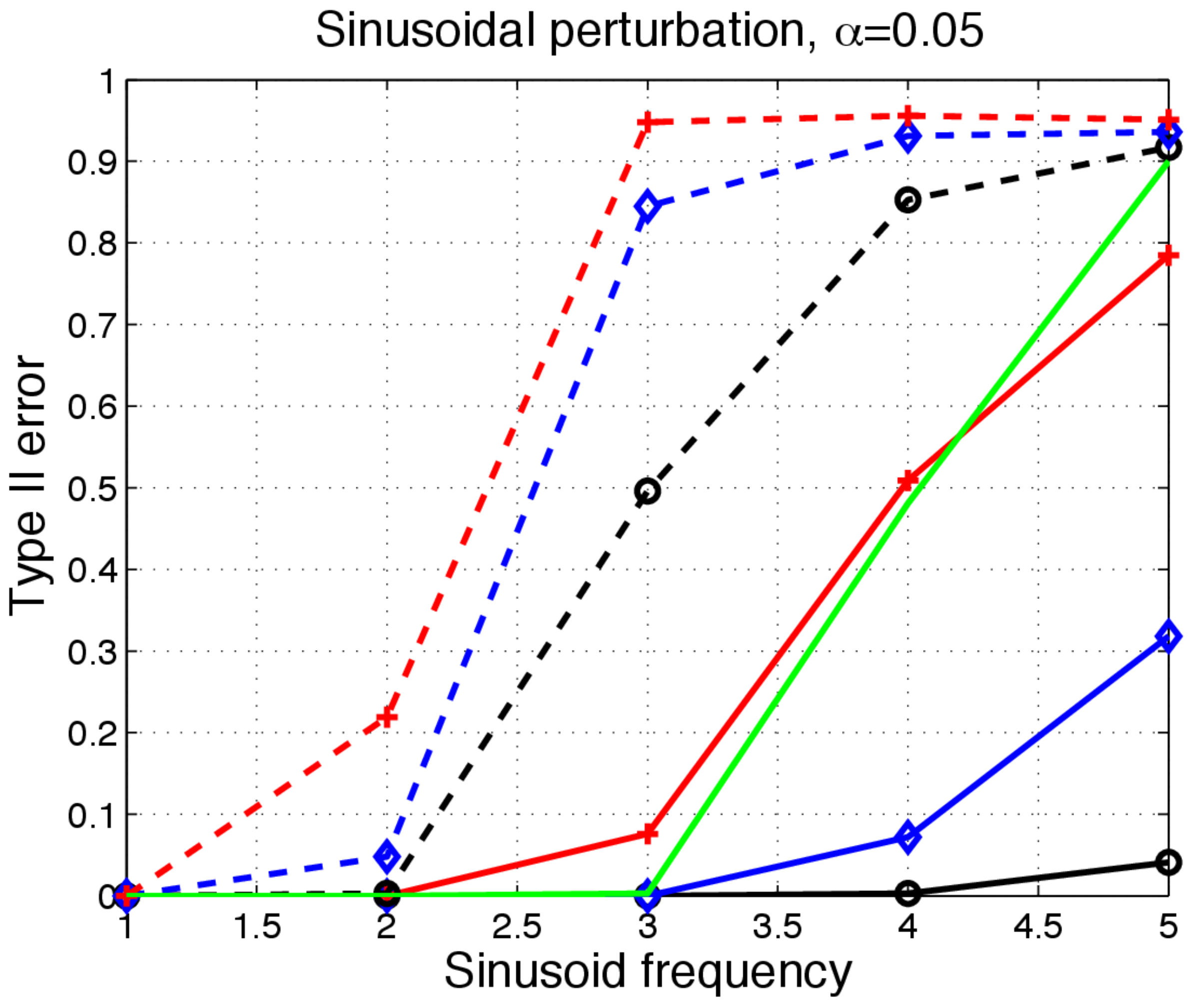


VS



Two-sample test, MMD with distance kernel

Obtain more powerful tests on this problem when $q \neq 1$ (exponent of distance)



Key:

- Gaussian kernel
- $q = 1$
- Best: $q = 1/3$
- Worst: $q = 2$

Selected references

Characteristic kernels and mean embeddings:

- Smola, A., Gretton, A., Song, L., Schoelkopf, B. (2007). A hilbert space embedding for distributions. ALT.
- Sriperumbudur, B., Gretton, A., Fukumizu, K., Schoelkopf, B., Lanckriet, G. (2010). Hilbert space embeddings and metrics on probability measures. JMLR.
- Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., Smola, A. (2012). A kernel two- sample test. JMLR.

Two-sample, independence, conditional independence tests:

- Gretton, A., Fukumizu, K., Teo, C., Song, L., Schoelkopf, B., Smola, A. (2008). A kernel statistical test of independence. NIPS
- Fukumizu, K., Gretton, A., Sun, X., Schoelkopf, B. (2008). Kernel measures of conditional dependence.
- Gretton, A., Fukumizu, K., Harchaoui, Z., Sriperumbudur, B. (2009). A fast, consistent kernel two-sample test. NIPS.
- Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., Smola, A. (2012). A kernel two- sample test. JMLR

Energy distance, relation to kernel distances

- Sejdinovic, D., Sriperumbudur, B., Gretton, A.,, Fukumizu, K., (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. Annals of Statistics.

Three way interaction

- Sejdinovic, D., Gretton, A., and Bergsma, W. (2013). A Kernel Test for Three-Variable Interactions. NIPS.

Selected references (continued)

Conditional mean embedding, RKHS-valued regression:

- Weston, J., Chapelle, O., Elisseeff, A., Schölkopf, B., and Vapnik, V., (2003). Kernel Dependency Estimation, NIPS.
- Micchelli, C., and Pontil, M., (2005). On Learning Vector-Valued Functions. Neural Computation.
- Caponnetto, A., and De Vito, E. (2007). Optimal Rates for the Regularized Least-Squares Algorithm. Foundations of Computational Mathematics.
- Song, L., and Huang, J., and Smola, A., Fukumizu, K., (2009). Hilbert Space Embeddings of Conditional Distributions. ICML.
- Grunewalder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., Pontil, M. (2012). Conditional mean embeddings as regressors. ICML.
- Grunewalder, S., Gretton, A., Shawe-Taylor, J. (2013). Smooth operators. ICML.

Kernel Bayes rule:

- Song, L., Fukumizu, K., Gretton, A. (2013). Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. IEEE Signal Processing Magazine.
- Fukumizu, K., Song, L., Gretton, A. (2013). Kernel Bayes rule: Bayesian inference with positive definite kernels, JMLR

Kernel CCA: Definition

- There exists a factorization of C_{xy} such that [Baker, 1973]

$$C_{xy} = C_{xx}^{1/2} V_{xy} C_{YY}^{1/2} \qquad \|V_{xy}\|_S \leq 1$$

- Regularized empirical estimate of spectral norm: [JMLR07]

$$\|\hat{V}_{xy}\|_S := \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \langle f, \hat{C}_{xy} g \rangle_{\mathcal{F}} \quad \text{subject to} \quad \begin{cases} \langle f, (\hat{C}_{xx} + \epsilon_n I) f \rangle_{\mathcal{F}} = 1, \\ \langle g, (\hat{C}_{yy} + \epsilon_n I) g \rangle_{\mathcal{G}} = 1, \end{cases}$$

  – First canonical correlate

Kernel CCA: Definition

- There exists a factorization of C_{xy} such that [Baker, 1973]

$$C_{xy} = C_{xx}^{1/2} V_{xy} C_{YY}^{1/2} \qquad \|V_{xy}\|_S \leq 1$$

- Regularized empirical estimate of spectral norm: [JMLR07]

$$\|\hat{V}_{xy}\|_S := \sup_{f\in\mathcal{F}, g\in\mathcal{G}} \langle f, \hat{C}_{xy} g\rangle_{\mathcal{F}} \quad \text{subject to} \quad \begin{cases} \langle f, (\hat{C}_{xx} + \epsilon_n I) f\rangle_{\mathcal{F}} = 1, \\ \langle g, (\hat{C}_{yy} + \epsilon_n I) g\rangle_{\mathcal{G}} = 1, \end{cases}$$

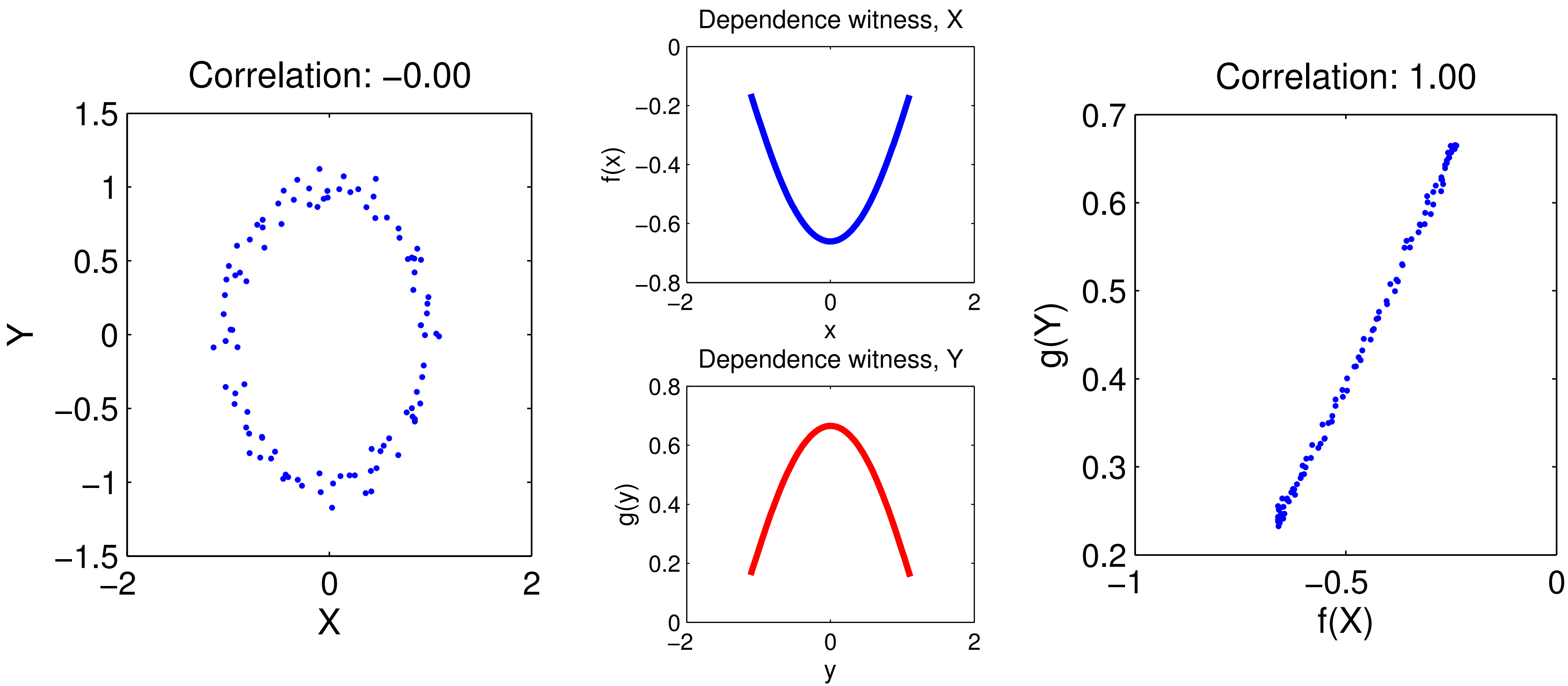
  – First canonical correlate

- Regularized empirical estimate of HS norm: [NIPS07b]

$$\mathrm{NOCCO}(\boldsymbol{z}; F, G) := \|\hat{V}_{xy}\|_{HS}^2 = \mathrm{tr}\left[\mathbf{R}_y \mathbf{R}_x\right], \qquad R_x := \widetilde{\mathbf{K}}_x (\widetilde{\mathbf{K}}_x + n\epsilon_n I_n)^{-1}$$

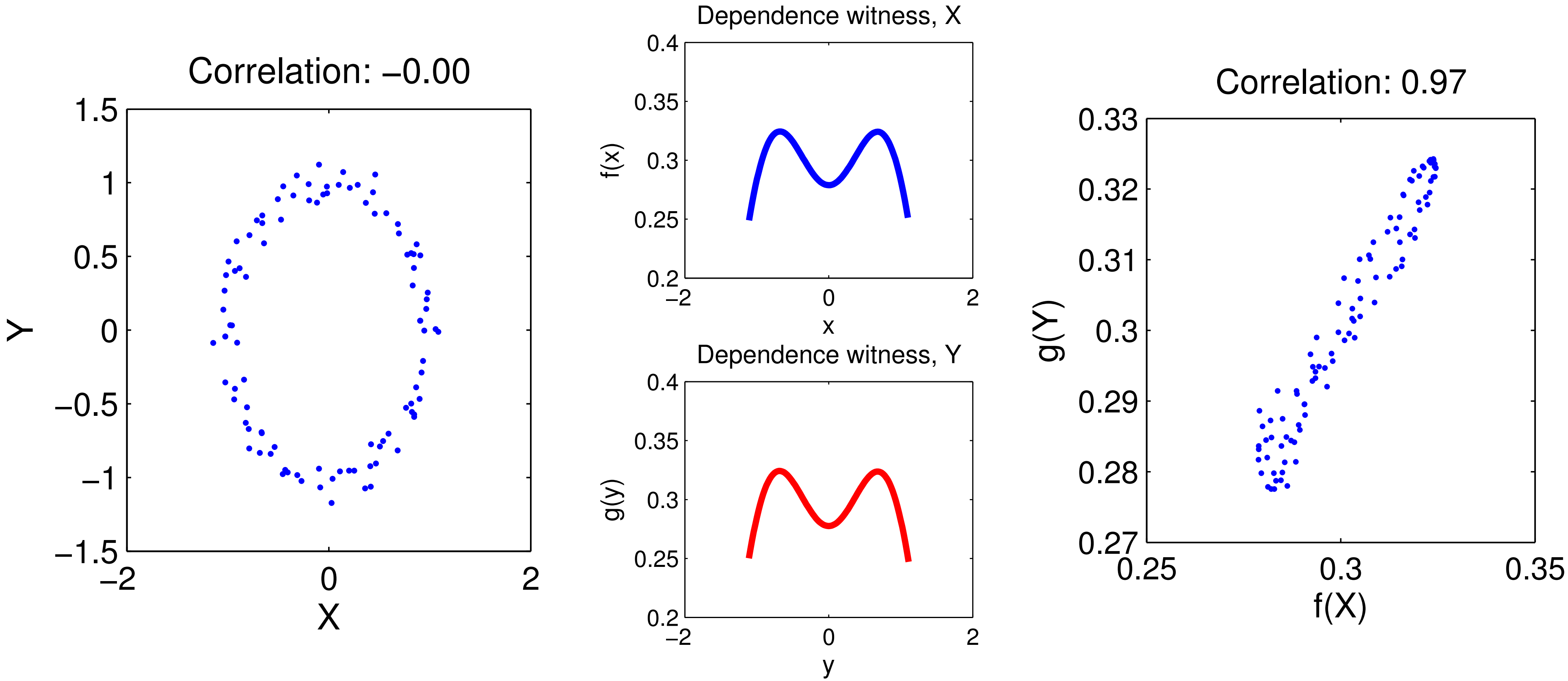
Kernel CCA: Illustration

- Ring-shaped density, first eigenvalue



Kernel CCA: Illustration

- Ring-shaped density, third eigenvalue



NOCCO: HS Norm of Normalized Cross Covariance

- Define NOCCO as

$$\text{NOCCO} := \|V_{xy}\|_{HS}^2$$

- Characteristic kernels: population NOCCO is mean-square contingency, indep. of RKHS

$$\text{NOCCO} = \int\int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{p_{xy}(x,y)}{p_x(x)p_y(y)} - 1 \right)^2 p_x(x)p_y(y)d\mu(x)d\mu(y).$$

  - $\mu(x)$ and $\mu(y)$ Lebesgue measures on $\mathcal{X}$ and $\mathcal{Y}$; P_{xy} absolutely continuous w.r.t. $\mu(x) \times \mu(y)$, density p_{xy}, marginal densities p_x and p_y

- Convergence result: assume regularization ϵ_n satisfies $\epsilon_n \to 0$ and $\epsilon_n^3 n \to \infty$, Then

$$\|\hat{V}_{xy} - V_{xy}\|_{HS} \;\to\; 0$$

in probability

References

C. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *J. Multivariate Anal.*, 88:190–206, 2004.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, New York, 1984.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. *ICML*, 2014.

K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. *NIPS*, 2014.

Andrey Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3):419–433, 1993.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proceedings of the International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann Publishers Inc., 2002.

A. Gretton and L. Gyorfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 63–77. Springer-Verlag, 2005.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems 15*, pages 513–520, Cambridge, MA, 2007. MIT Press.

A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, Cambridge, MA, 2008. MIT Press.

A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012a.

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 *NIPS*, 2012b.

Arthur Gretton. A simpler condition for consistency of a kernel independence test. Technical Report 1501.06103, arXiv, 2015.

David Haussler. Convolution kernels on discrete structures. Technical Report UCS-CRL-99-10, UC Santa Cruz, 1999.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, 2009.

Wittawat Jitkrittum, Arthur Gretton, Nicolas Heess, S. M. Ali Eslami, Balaji Lakshminarayanan, Dino Sejdinovic, and Zoltán Szabó. Kernel-based just-in-time learning for passing expectation propagation messages. *UAI*, 2015.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *ICML*, 2015.

D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *NIPS*, 2013a.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2702, 2013b.

D. Sejdinovic, H. Strathmann, M. Lomeli Garcia, C. Andrieu, and A. Gretton. Kernel adaptive Metropolis-Hastings. *ICML*, 2014.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 4754, pages 13–31. Springer, 2007.

B. Sriperumbudur, K. Fukumizu, A. Gretton, and A. Hyvärinen. Density estimation in infinite dimensional exponential families. Technical Report 1312.3516, ArXiv e-prints, 2014.

Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltán Szabó, and Arthur Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. *arxiv*, 2015.

Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. *AISTATS*, 2015.

G. Székely and M. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

G. Székely and M. Rizzo. A new test for multivariate normality. *J. Multivariate Anal.*, 93:58–80, 2005.

G. Székely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Stat.*, 35(6):2769–2794, 2007.

K. Zhang, J. Peters, D. Janzing, B., and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.