Lecture 3: Dependence measures using RKHS embeddings

MLSS Tübingen, 2015

Arthur Gretton

Gatsby Unit, CSML, UCL

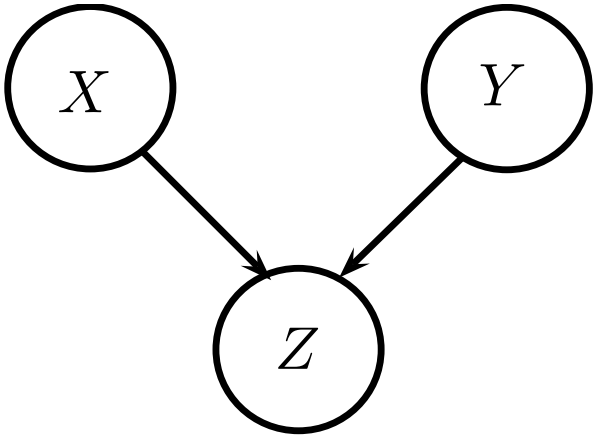
Outline

- Three or more variable interactions, comparison with conditional dependence testing [Sejdinovic et al., 2013a]
- Dependence detection in detail, covariance operators
- Bayesian inference without models, comparison with approximate Bayesian computation (ABC) [Fukumizu et al., 2013]
- Recent work (2014/2015) (not in this talk, see my webpage)
  - Testing for time series Chwialkowski and Gretton [2014], Chwialkowski et al. [2014]
  - Nonparametric adaptive expectation propagation Jitkrittum et al. [2015]
  - 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?

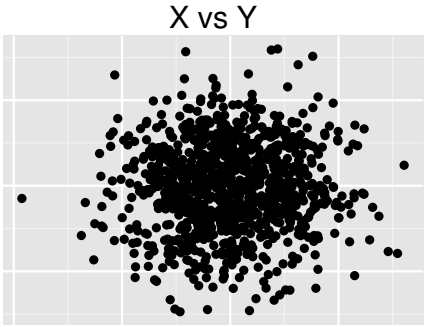


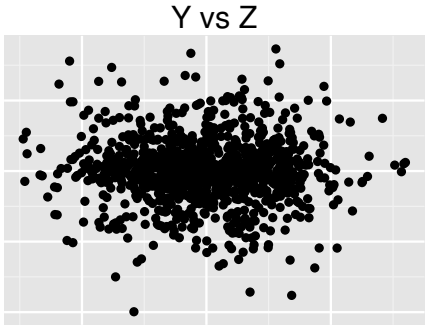
Detecting a higher order interaction

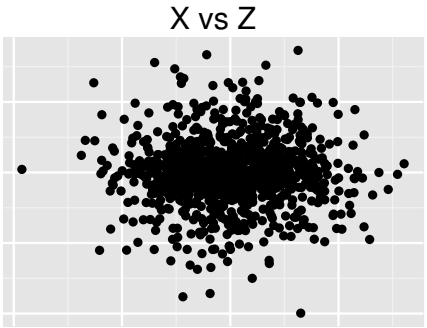
- How to detect V-structures with pairwise weak individual dependence?

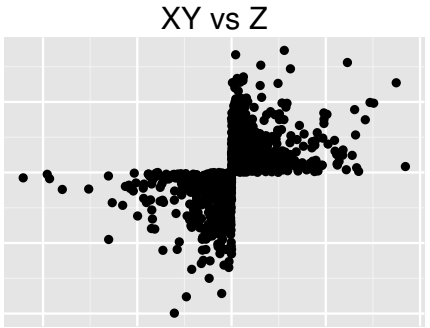
Detecting a higher order interaction

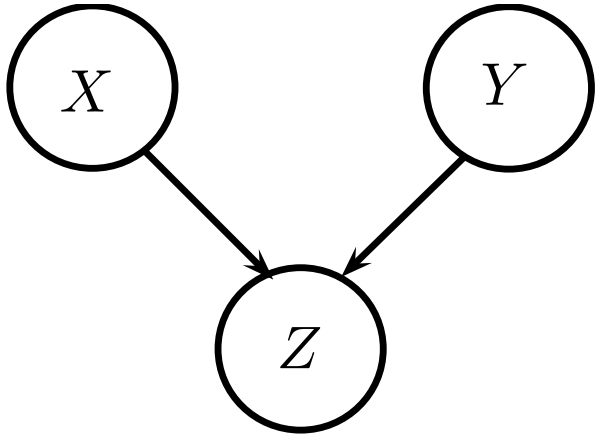
- How to detect V-structures with pairwise weak individual dependence?
- $X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$







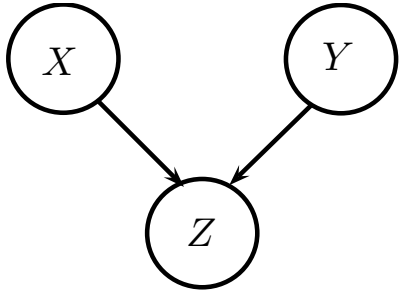




- $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,
- $Z|\ X, Y \sim \text{sign}(XY)Exp(\frac{1}{\sqrt{2}})$

Faithfulness violated here

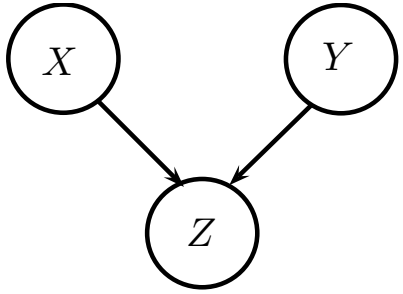
V-structure Discovery



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

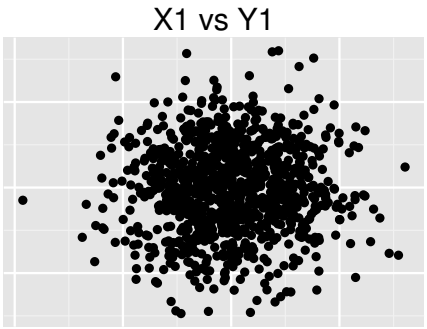


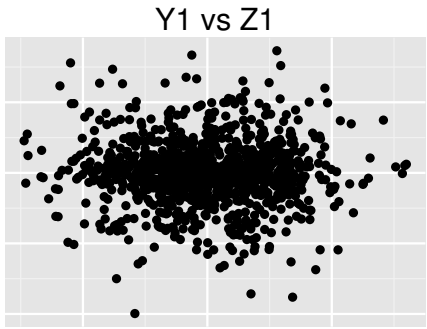
Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

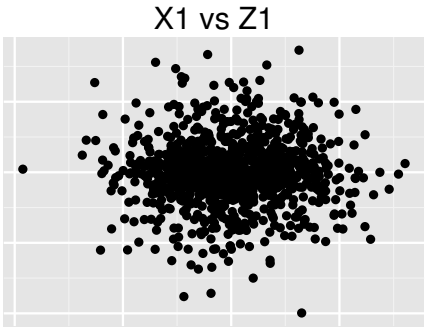
- **Consistent** CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ [Fukumizu et al., 2008, Zhang et al., 2011], or
- Factorisation test: $\mathbf{H_0} : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)
  - compute p-values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

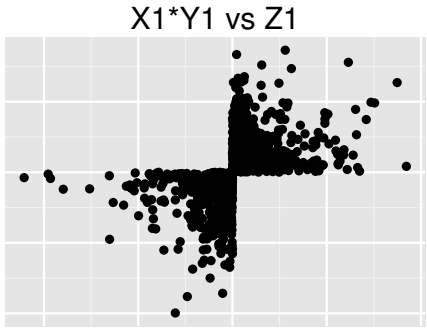
V-structure Discovery (2)

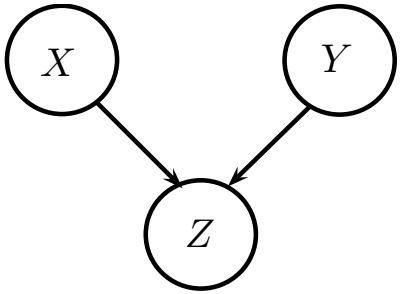
- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- $X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$











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

V-structure Discovery (3)

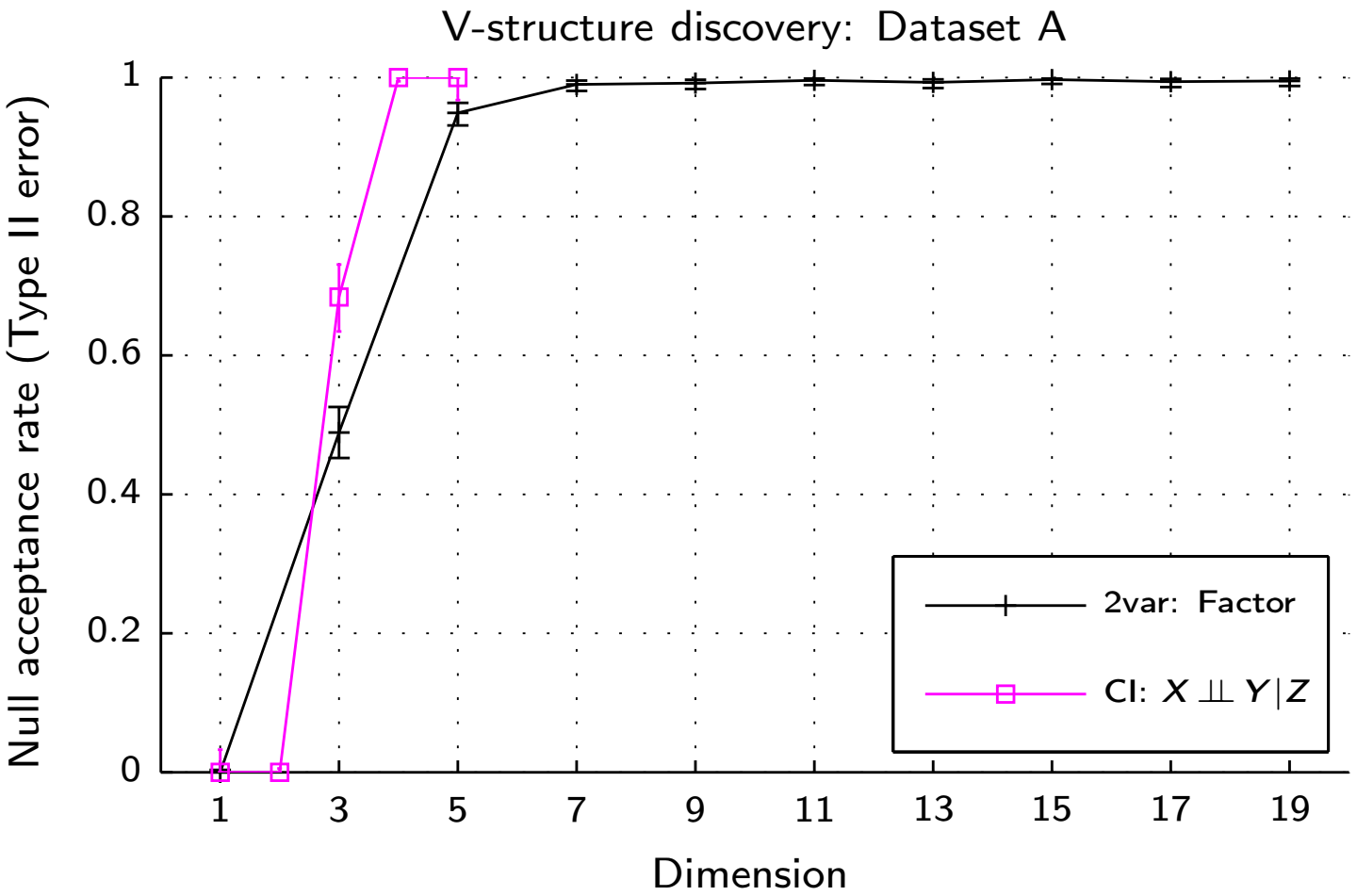


Figure 1: CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), and a factorisation test with a **HB** correction, $n = 500$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ 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.

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

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.

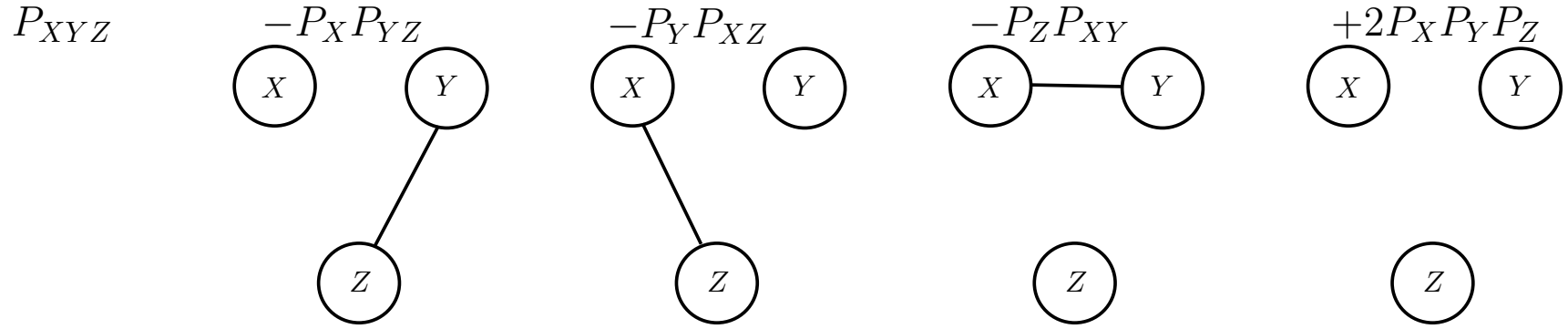
- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$

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.

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

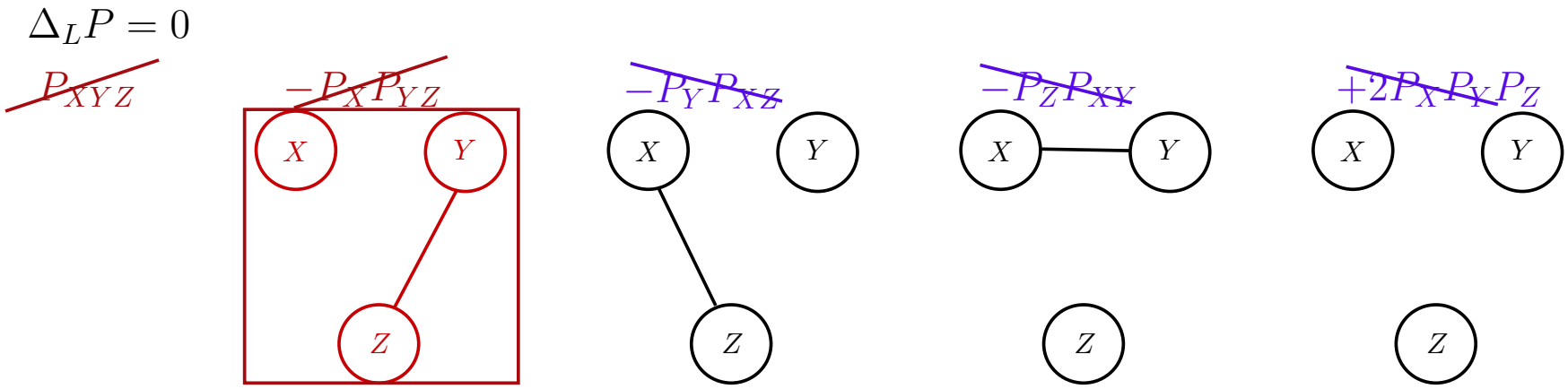
$\Delta_L P =$



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.

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



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.

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

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

...so what might be missed?

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ 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.

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

$$\Delta_L P = 0 \not\Rightarrow (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$$

Example:

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

A Test using Lancaster Measure

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

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

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 1: 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 2: V-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$

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

Empirical joint central moment in the feature space

Example A: factorisation tests

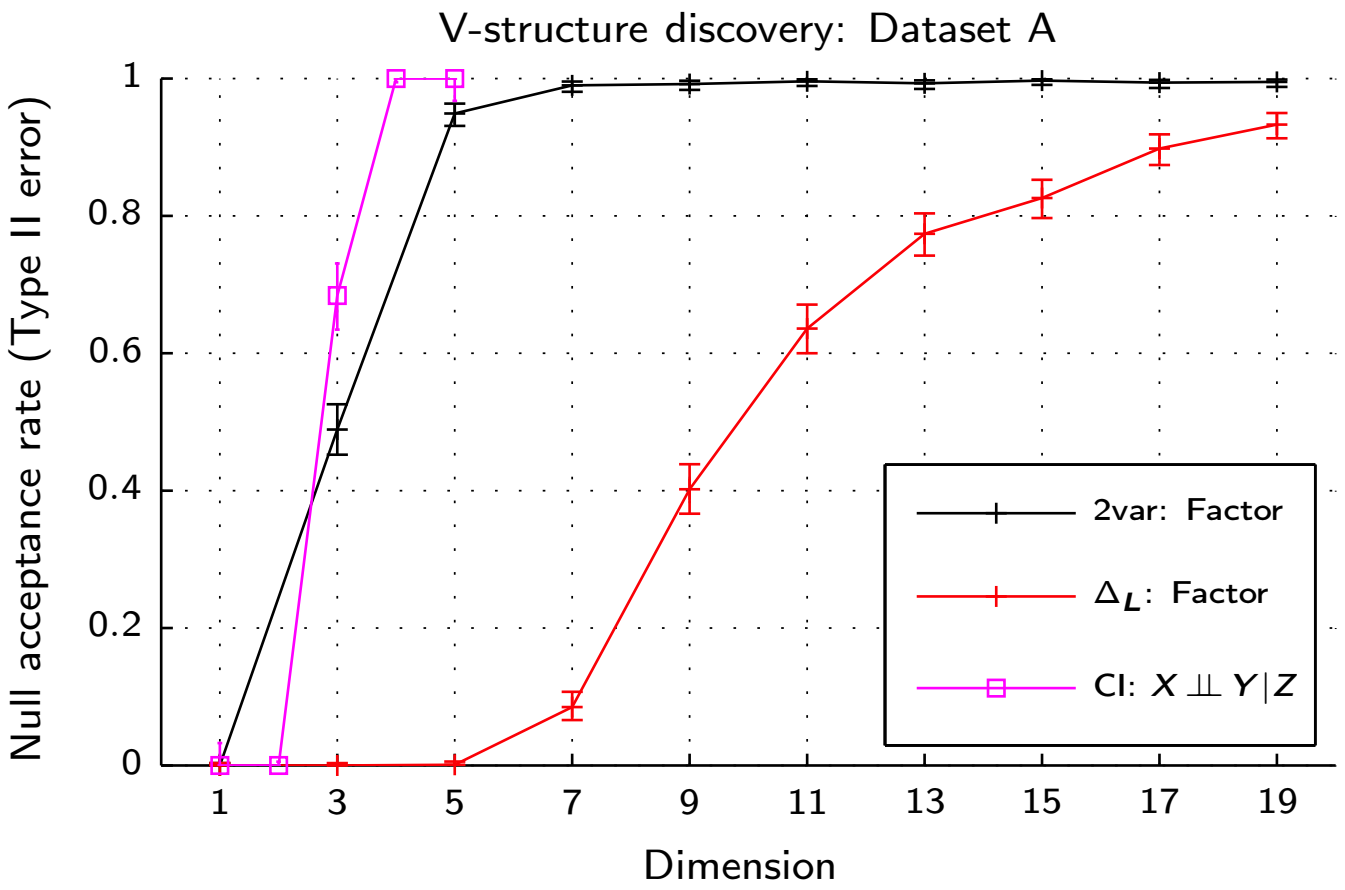


Figure 2: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

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

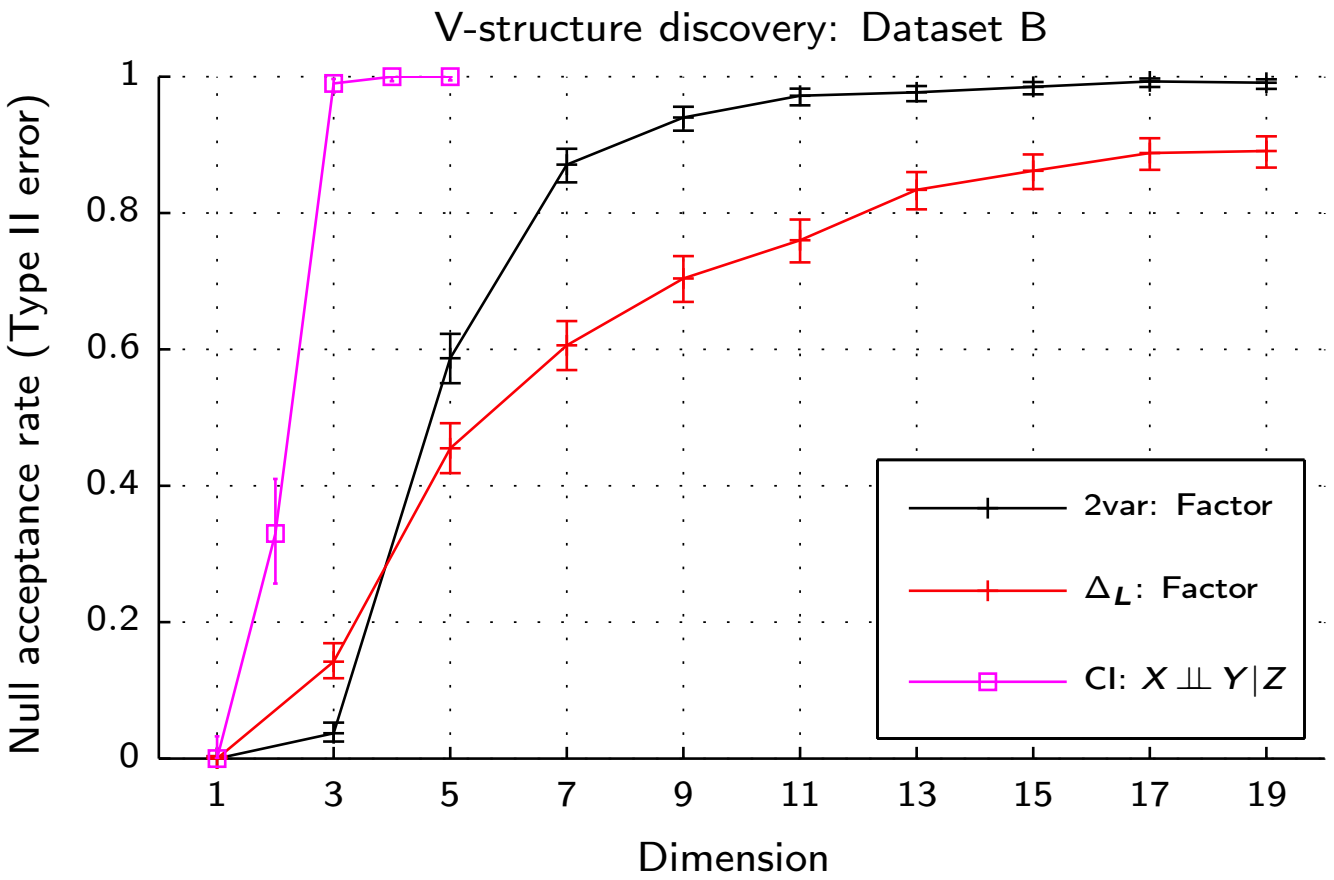


Figure 3: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Interaction for $D \geq 4$

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

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

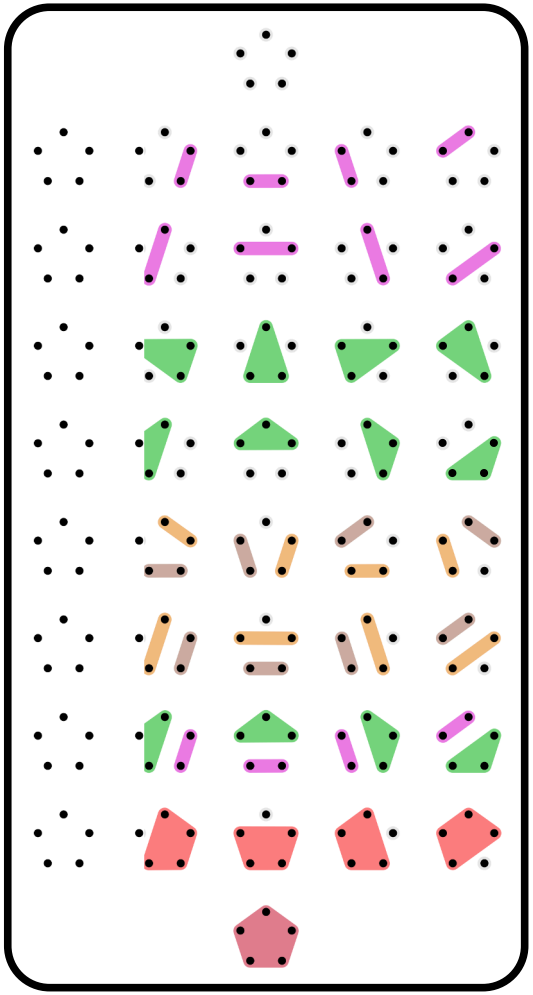
  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

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

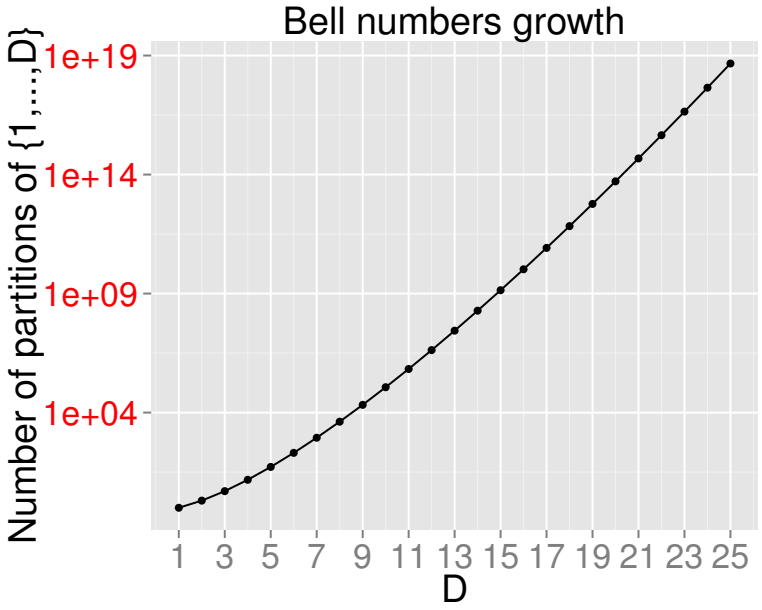


Interaction for $D \geq 4$

- 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 $\kappa = \bigotimes_{i=1}^{D} k^{(i)}$:

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

- Coincides with the test proposed by Kankainen (1995) using empirical characteristic functions: similar relationship to that between dCov and HSIC (DS et al, 2013)

Example B: total independence tests

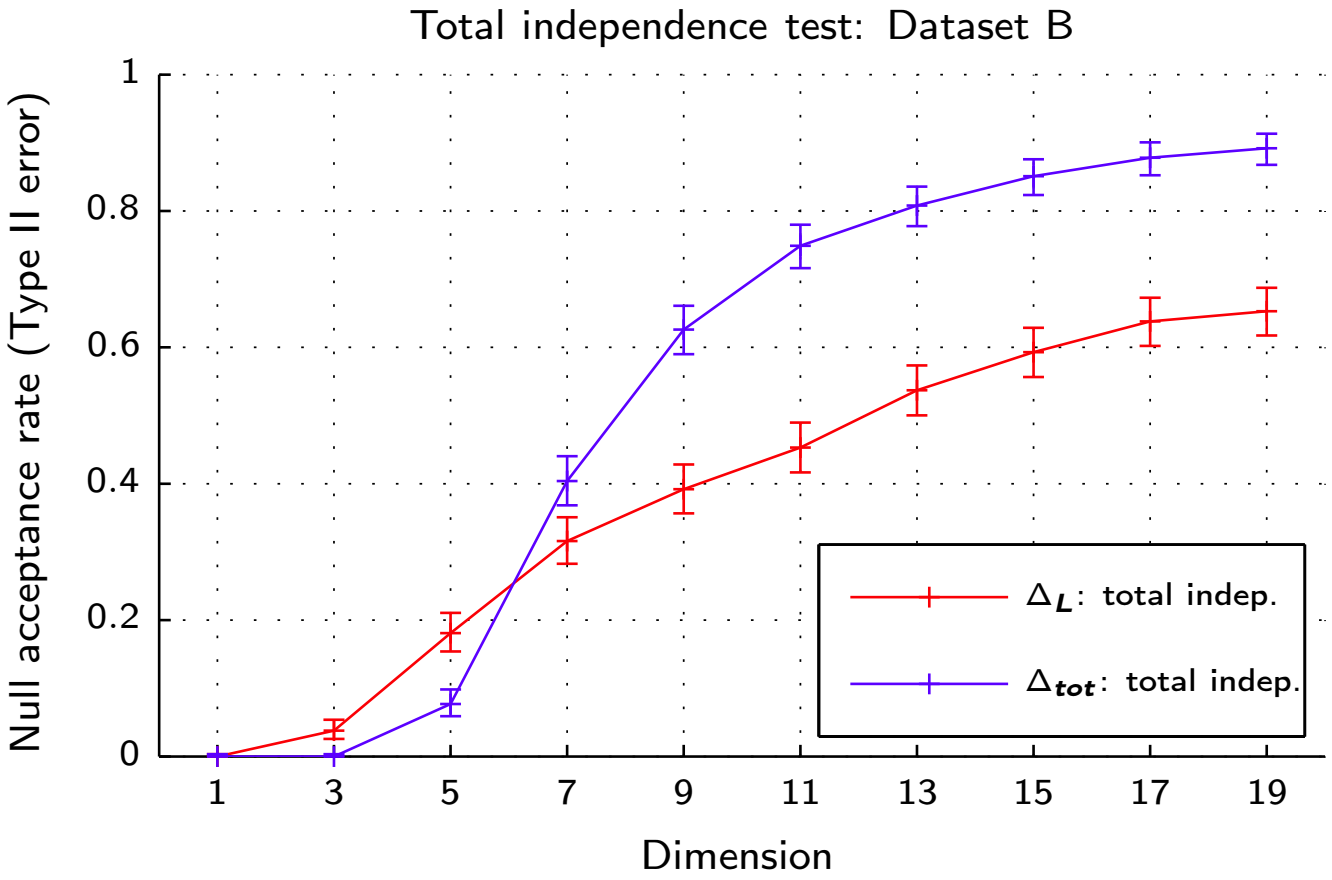
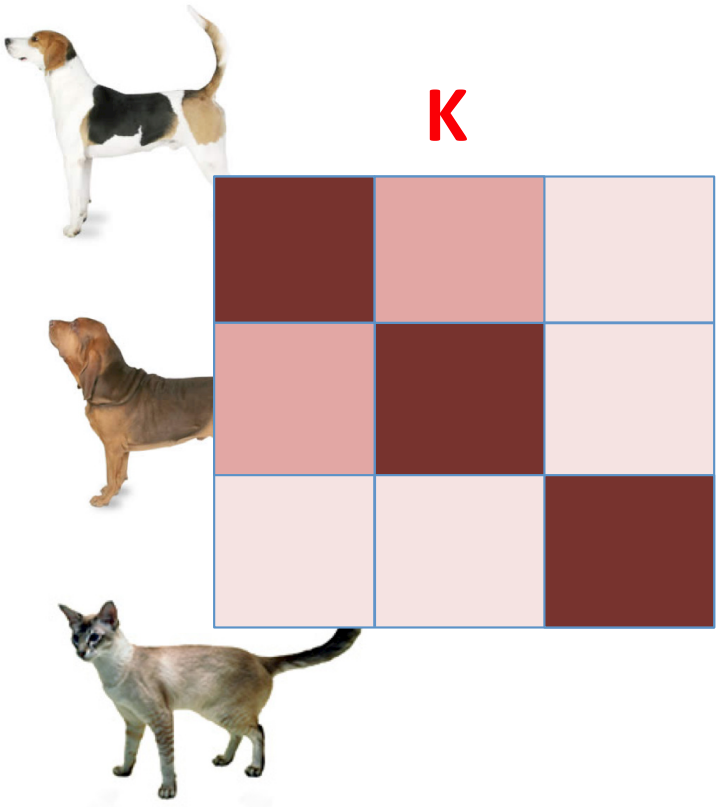


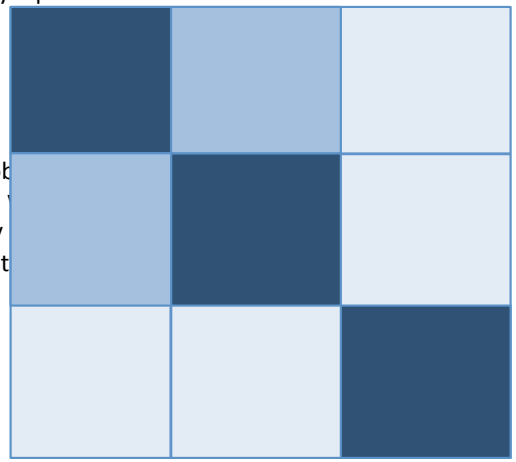
Figure 4: Total independence: $\Delta_{tot}\hat{P}$ vs. $\Delta_L\hat{P}$, $n = 500$

Kernel dependence measures - in detail

MMD for independence: HSIC



Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.



A large animal who slings slob
distinctive houndy odor, and
than to follow his nose. They
amount of exercise and ment

Known for their curiosity, intelligence, and excellent communication skills, the Javanese breed is perfect if you want a responsive, interactive pet, one that will blow in your ear and follow you everywhere.

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

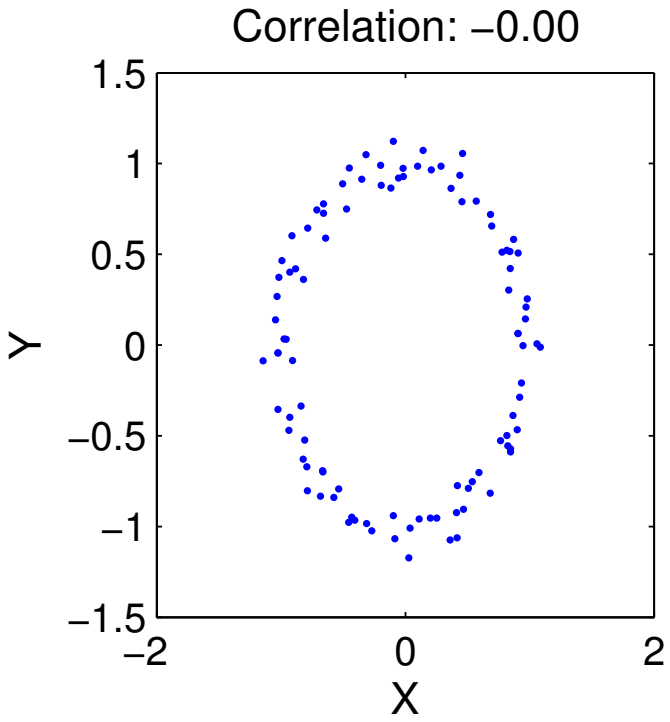
A more intuitive idea: maximize covariance of smooth mappings:

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

Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

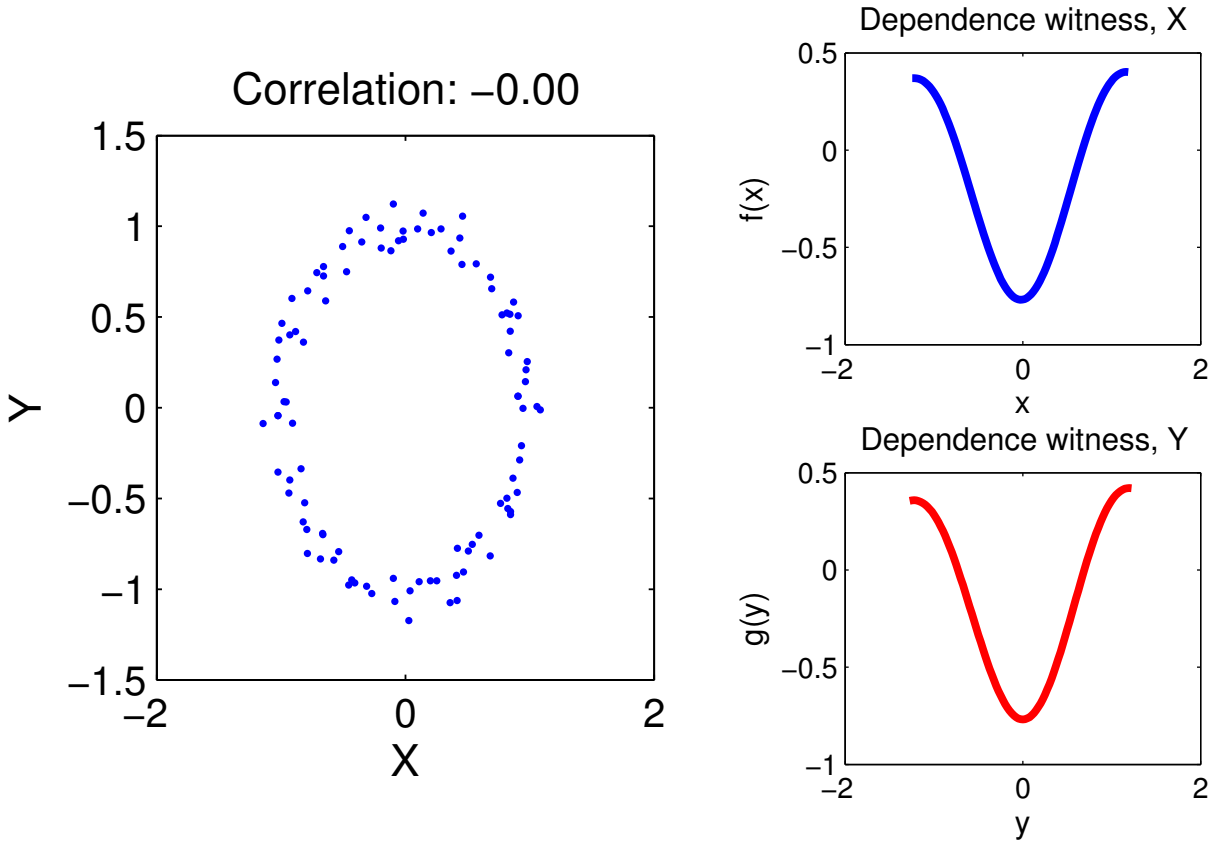
$$\mathrm{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$



Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

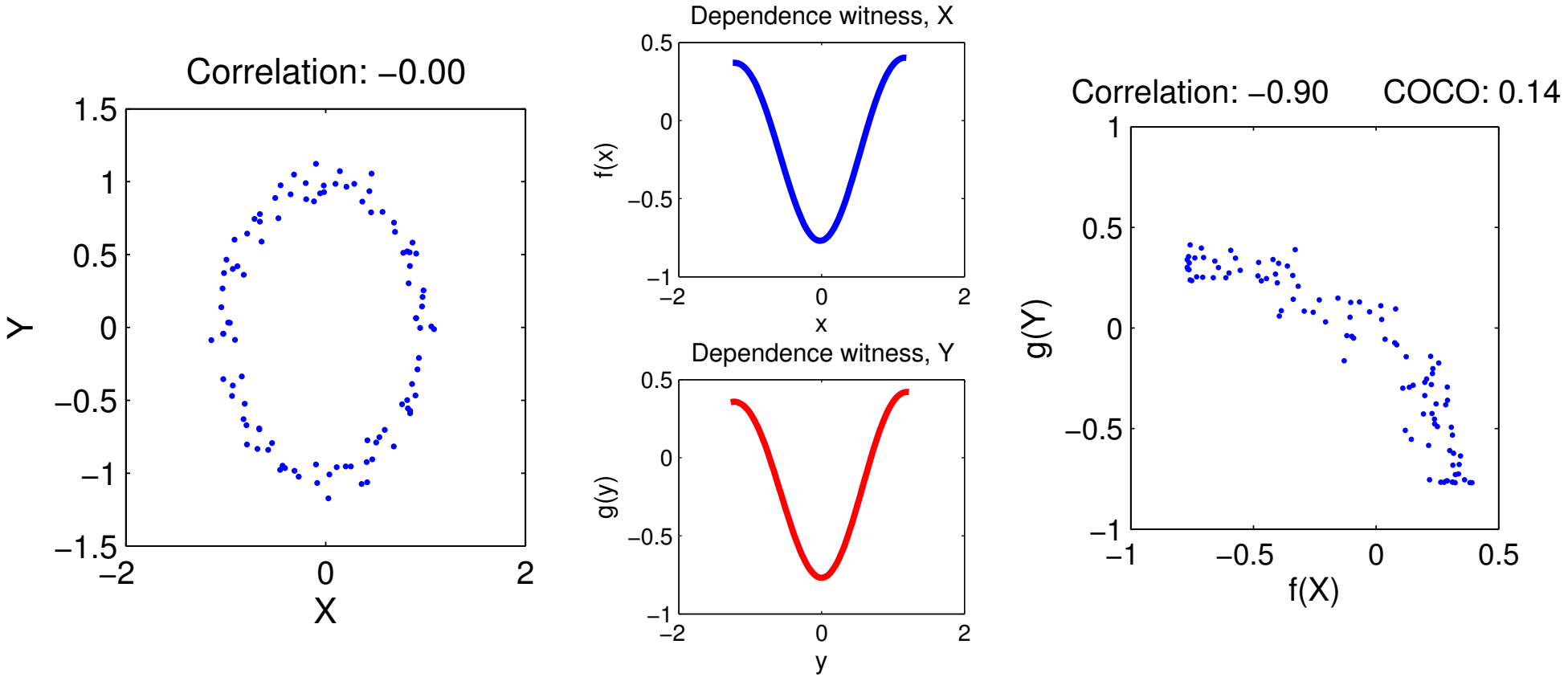
$$\mathrm{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$



Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

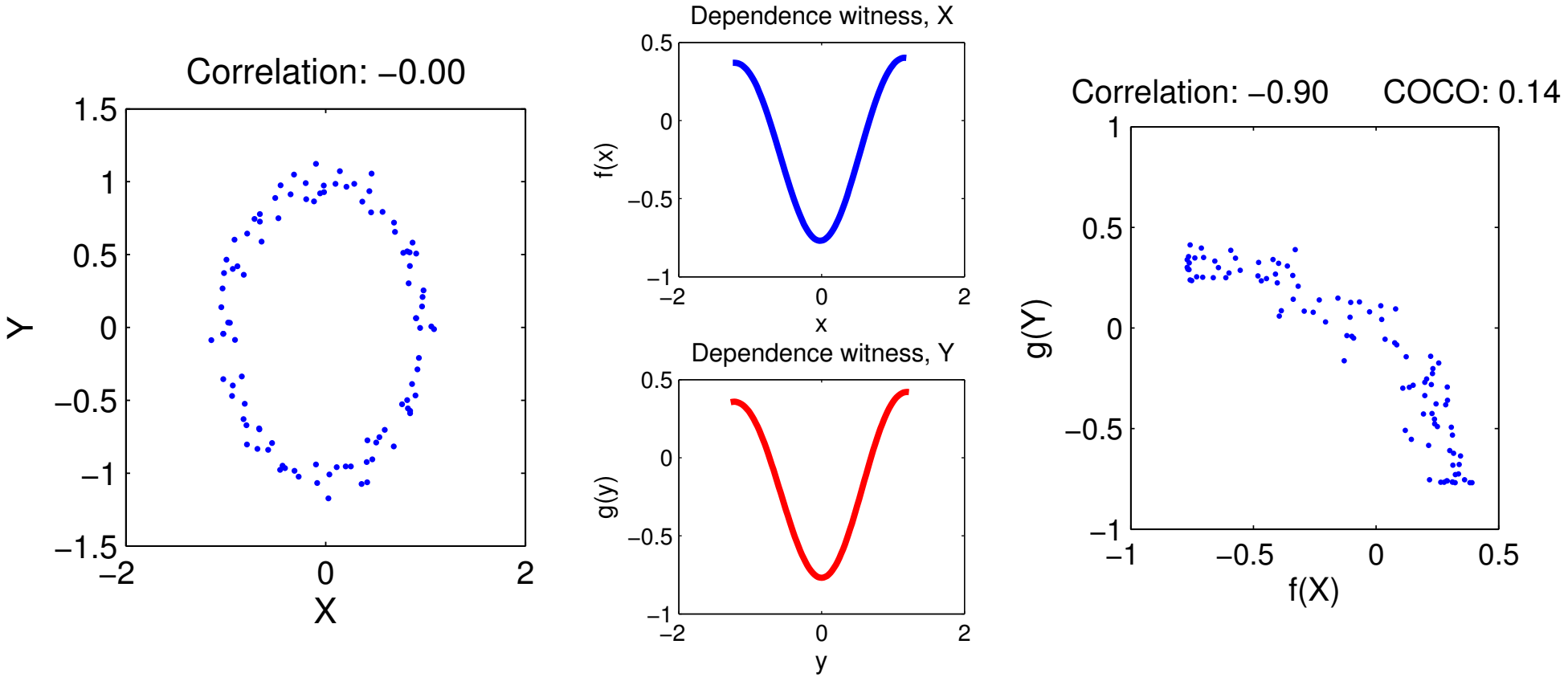
$$\mathrm{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$



Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

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



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

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

(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 $x\, y^\top \to \phi(x) \otimes \psi(y)$

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

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

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

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

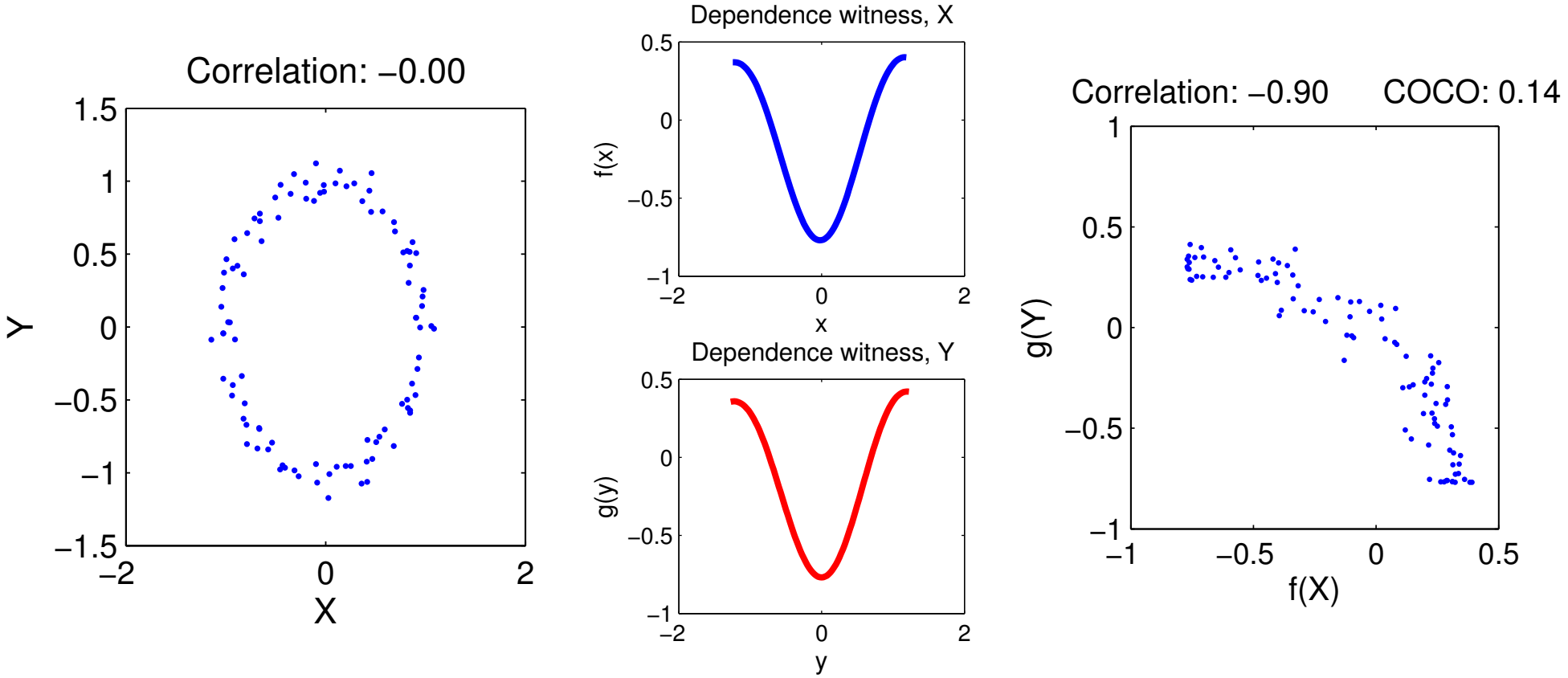
YES: via Bochner integrability argument (as with mean embedding).
Under the condition $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\sqrt{k(\mathsf{x},\mathsf{x})l(\mathsf{y},\mathsf{y})}\right) < \infty$, we can define:

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

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

REMINDER: functions revealing dependence

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



How do we compute this from finite data?

Empirical covariance operator

The empirical 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 $\hat{\mu}_x := \frac{1}{n} \sum_{i=1}^n \phi(x_i)$. More concisely,

$$\widehat{C}_{XY} = \frac{1}{n} X H Y^\top,$$

where $H = I_n - n^{-1} \mathbf{1}_n$, and $\mathbf{1}_n$ is an $n \times n$ matrix of ones, and

$$X = \begin{bmatrix} \phi(x_1) & \ldots & \phi(x_n) \end{bmatrix} \qquad Y = \begin{bmatrix} \psi(y_1) & \ldots & \psi(y_n) \end{bmatrix}.$$

Define the kernel matrices

$$K_{ij} = \left(X^\top X \right)_{ij} = k(x_i, x_j) \qquad L_{ij} = l(y_i, y_j),$$

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

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

The associated Lagrangian is

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

Covariance to reveal dependence

- Empirical 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:

$$\widetilde{K} = HKH \qquad \text{where} \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

Covariance to reveal dependence

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

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

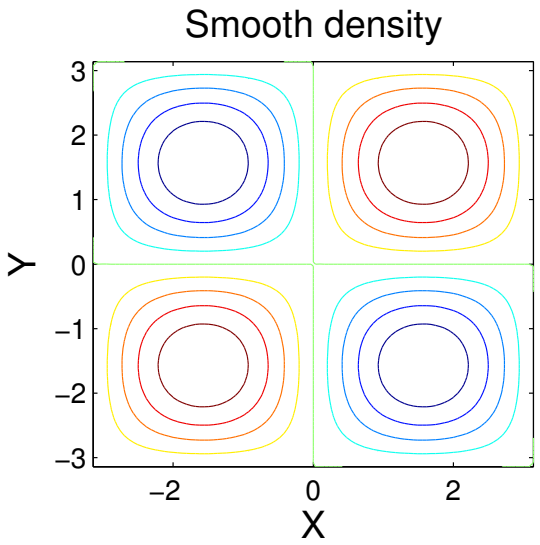
- $\widetilde{K}$ and $\widetilde{L}$ are matrices of inner products between centred observations in respective feature spaces:

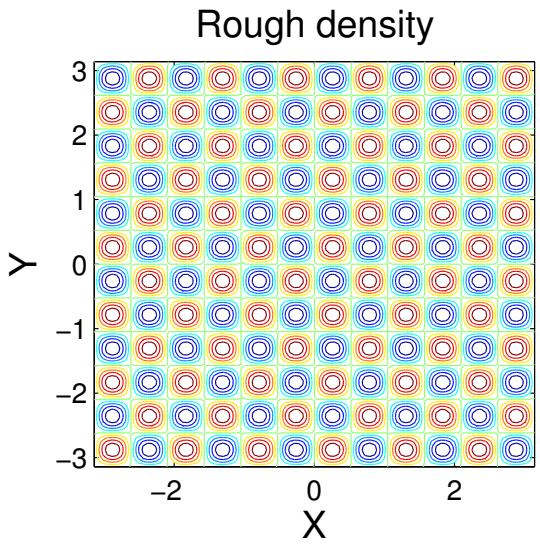
$$\widetilde{K} = HKH \qquad \text{where} \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

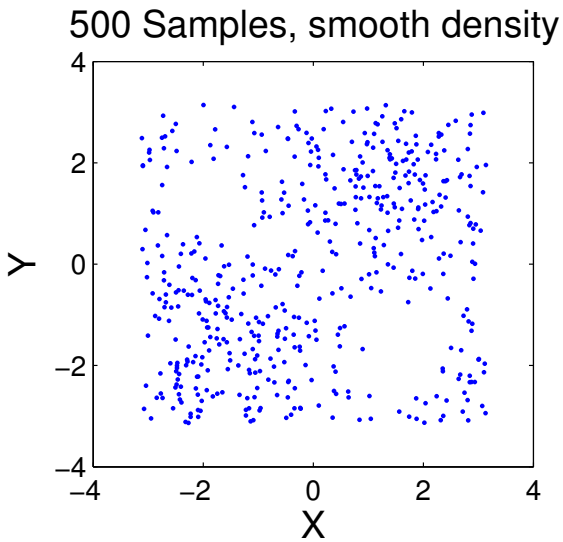
- Mapping function for x:

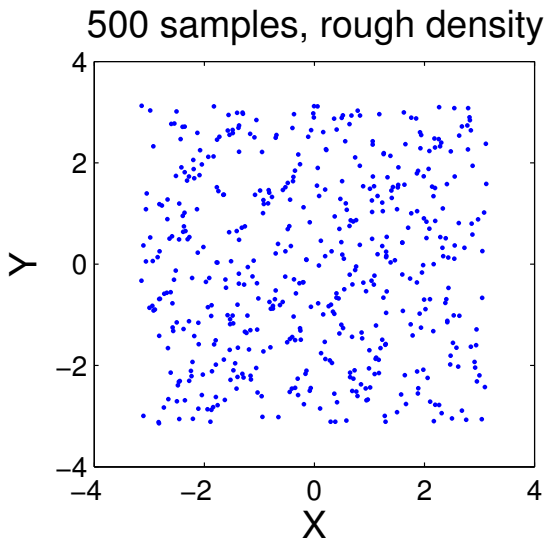
$$f(x) = \sum_{i=1}^{n} \alpha_i \left(k(x_i, x) - \frac{1}{n} \sum_{j=1}^{n} k(x_j, x) \right)$$

Hard-to-detect dependence







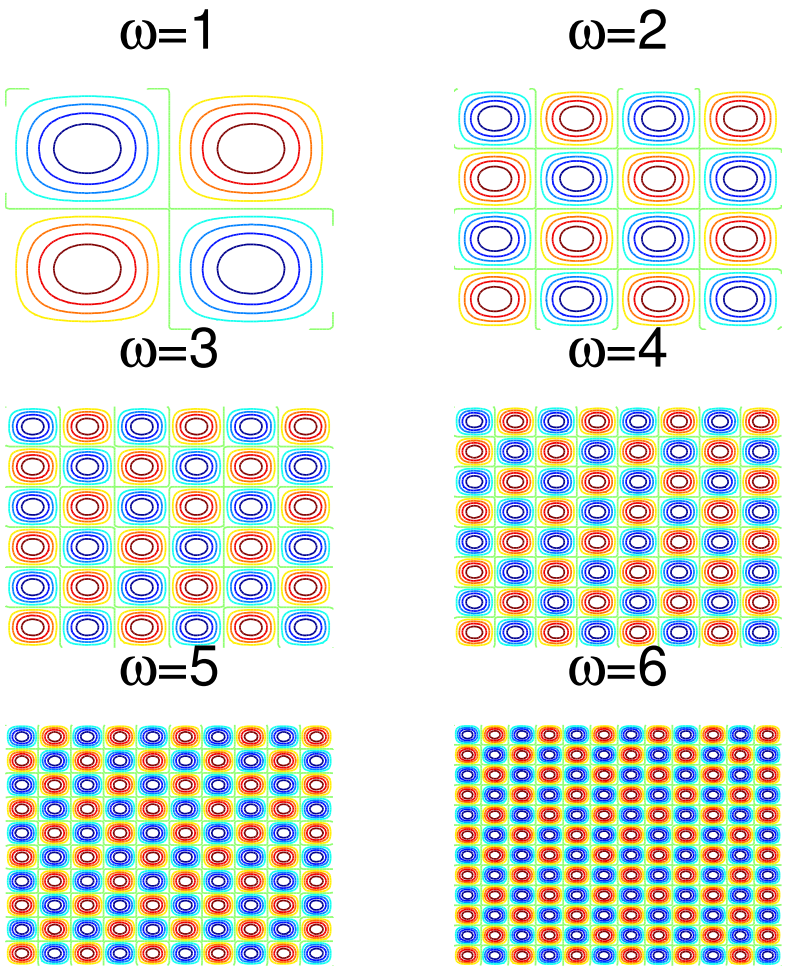


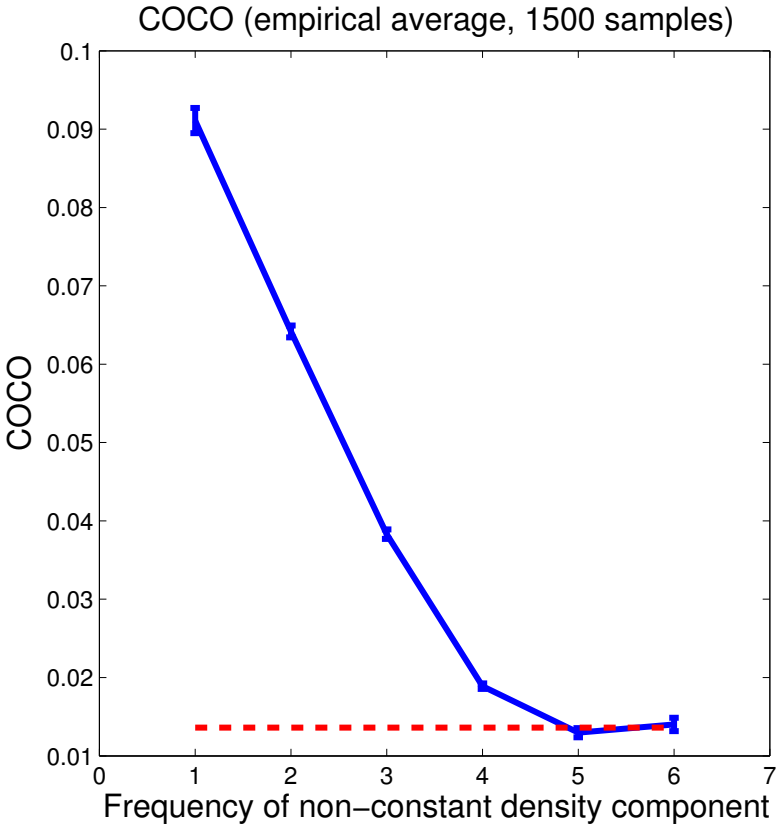
Density takes the form:

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

Hard-to-detect dependence

- Example: sinusoids of increasing frequency





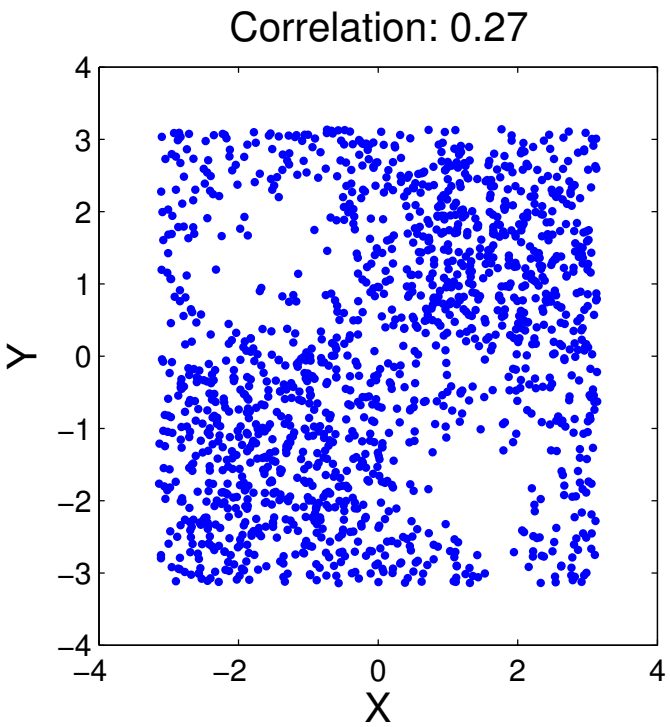
Hard-to-detect dependence

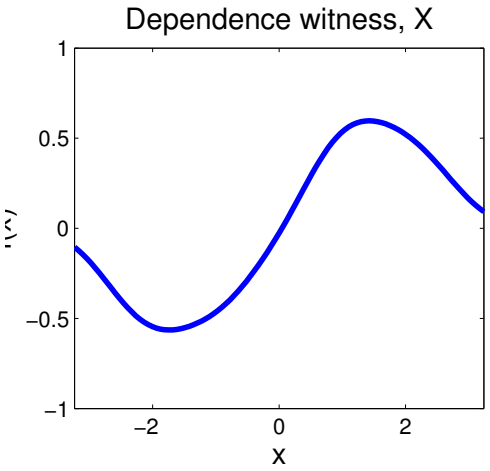
COCO vs frequency of perturbation from independence.

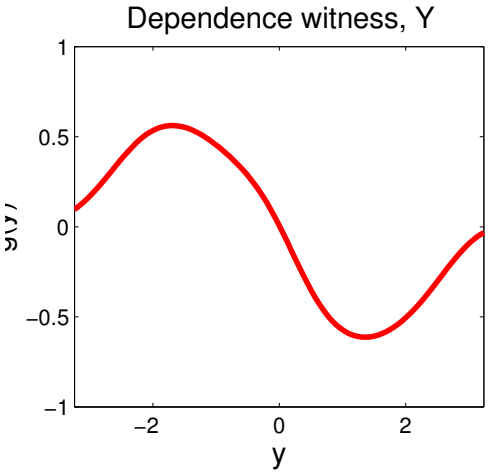
Hard-to-detect dependence

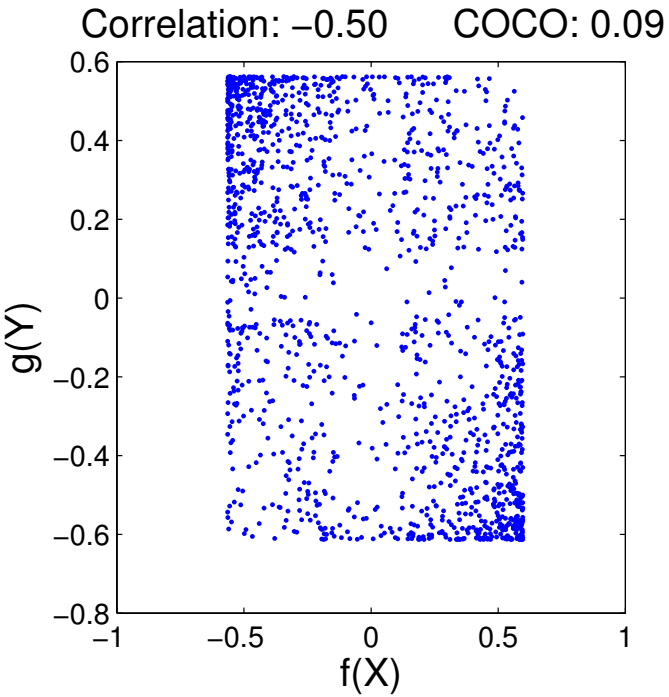
COCO vs frequency of perturbation from independence.

Case of $\omega = 1$





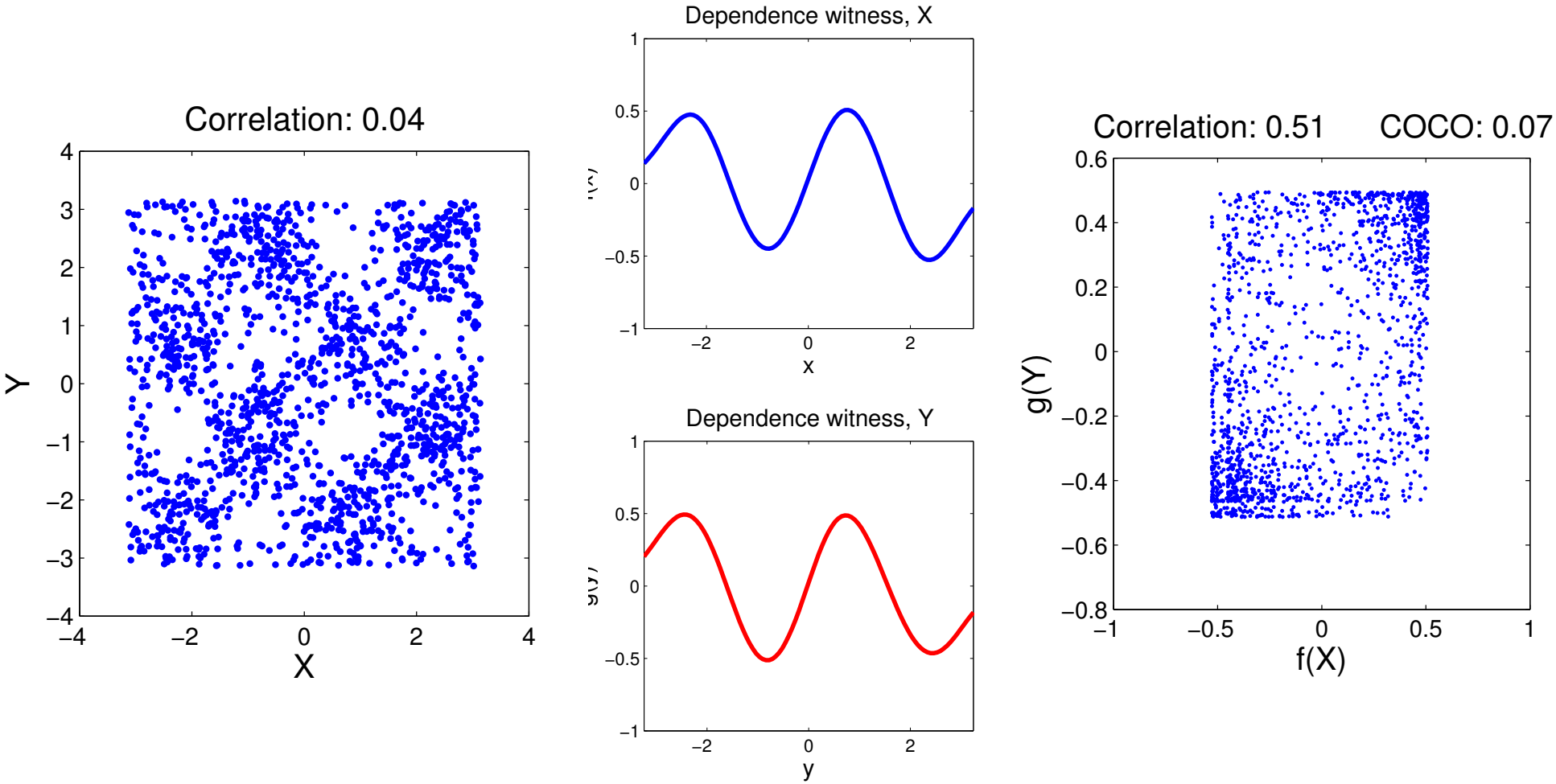




Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

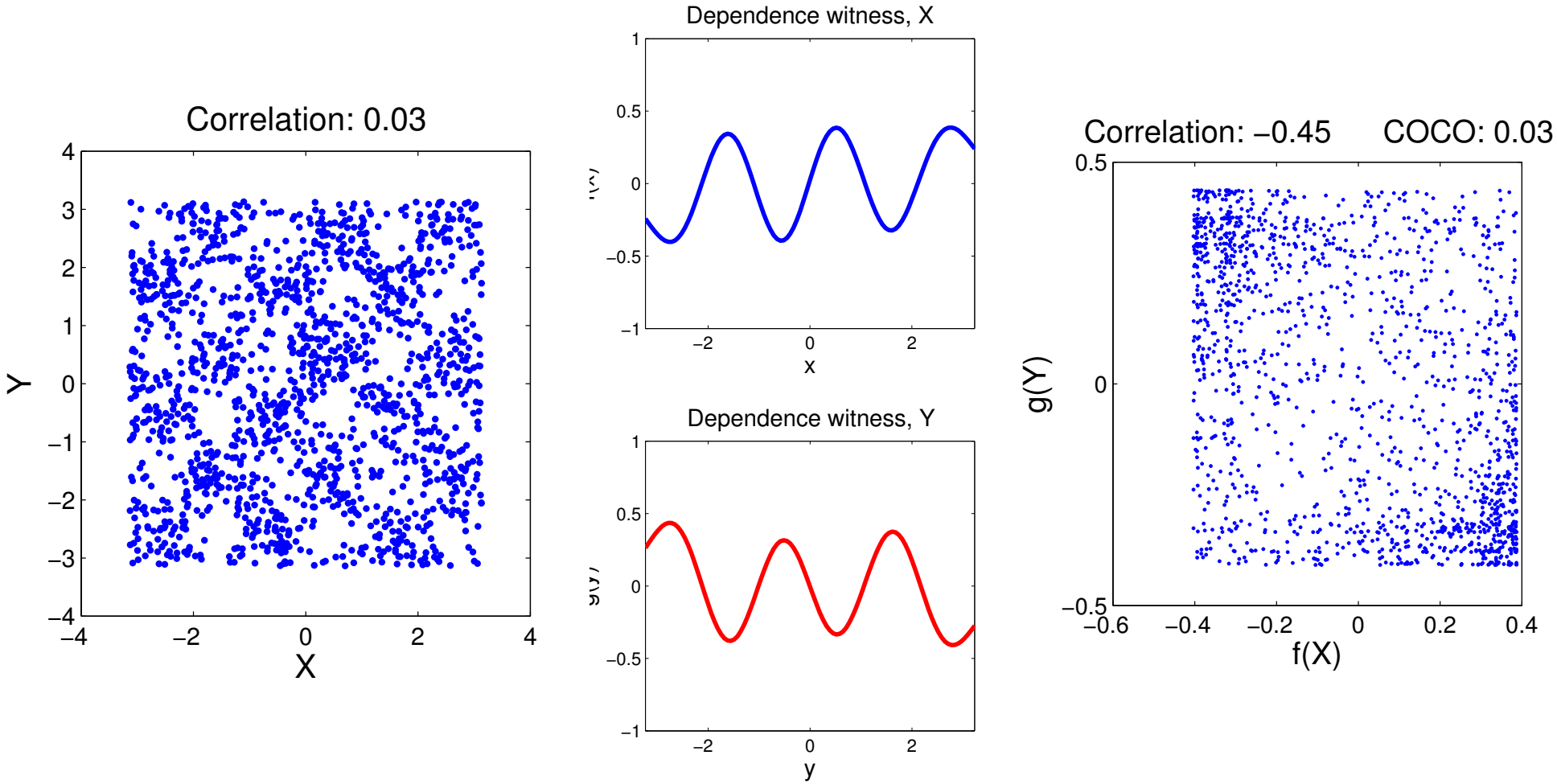
Case of $\omega = 2$



Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

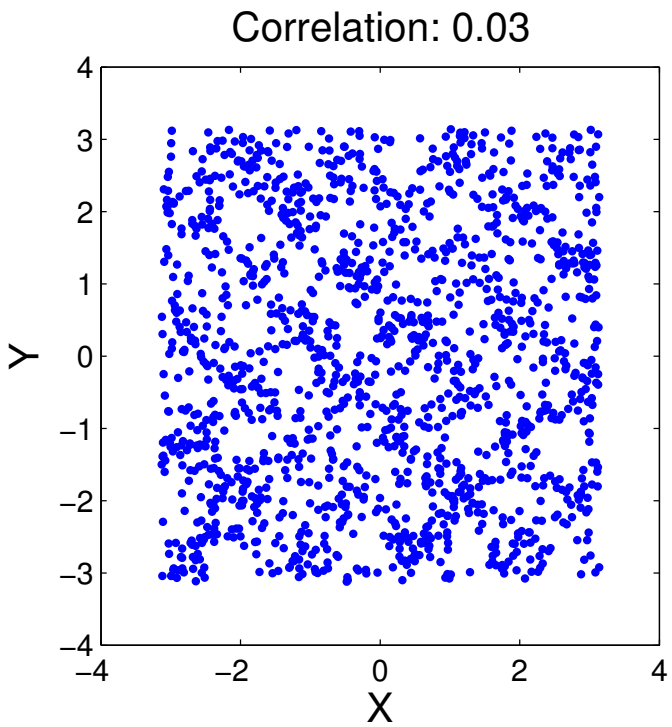
Case of $\omega = 3$

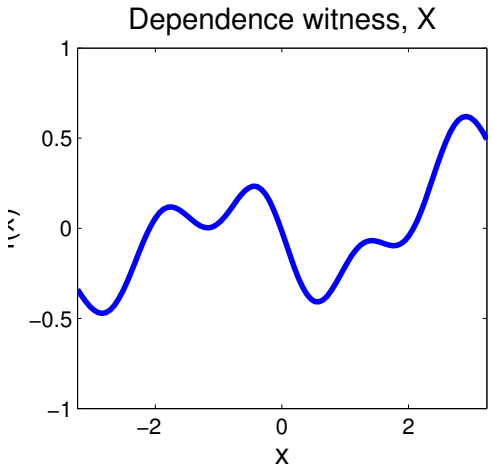


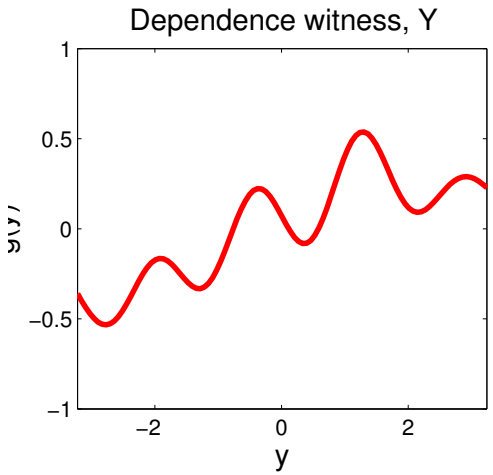
Hard-to-detect dependence

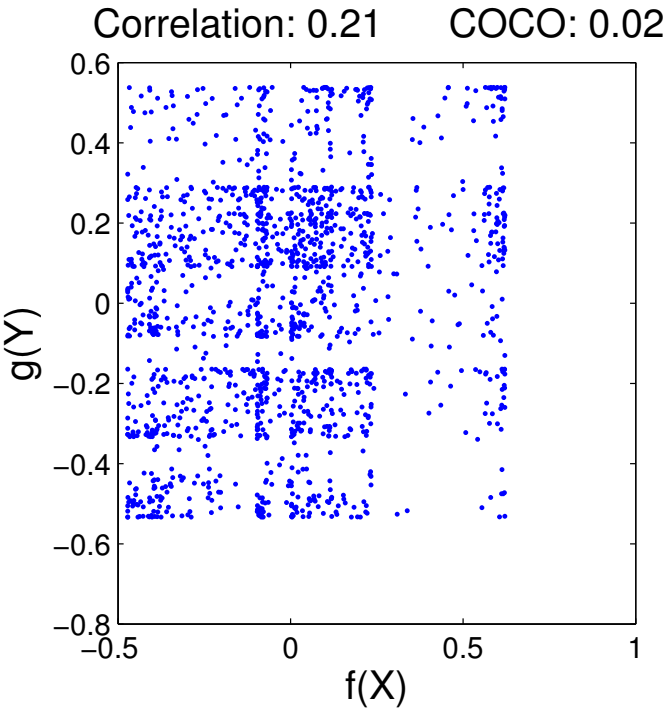
COCO vs frequency of perturbation from independence.

Case of $\omega = 4$





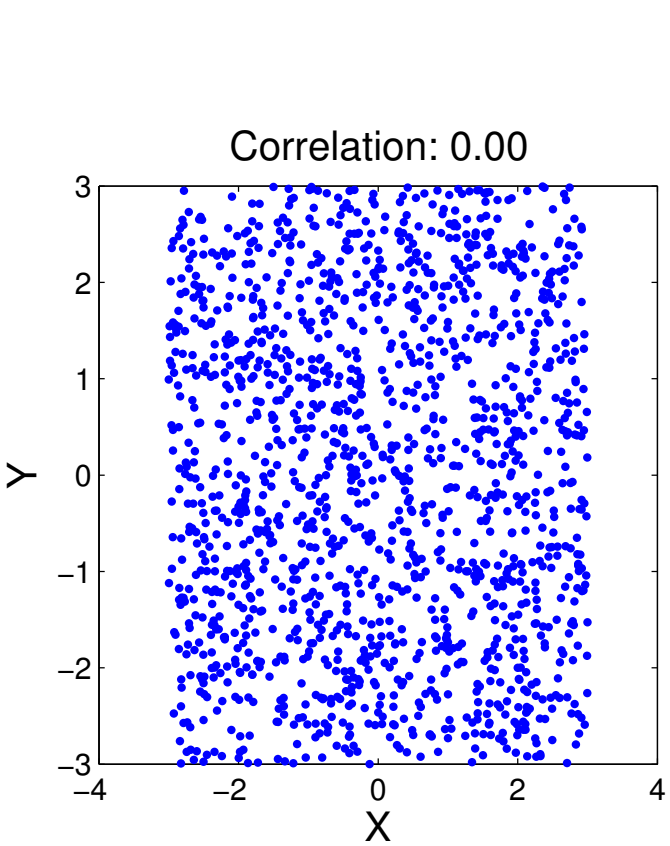


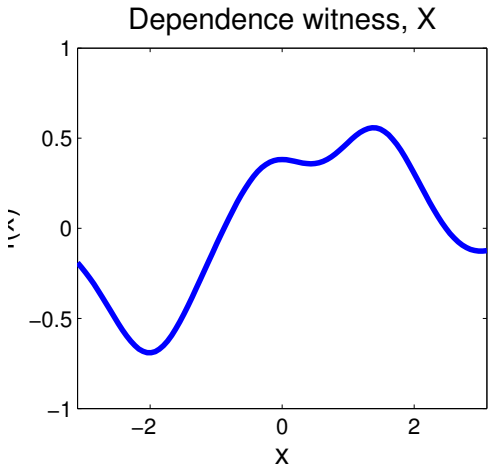


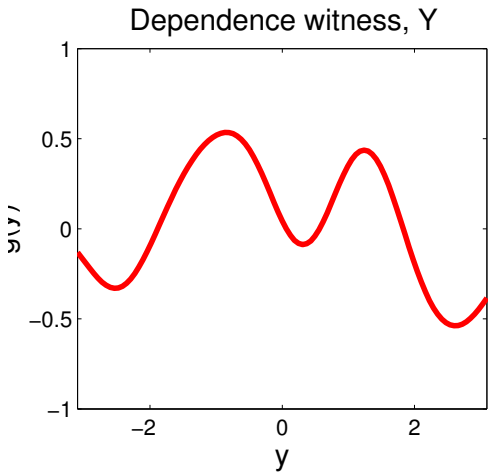
Hard-to-detect dependence

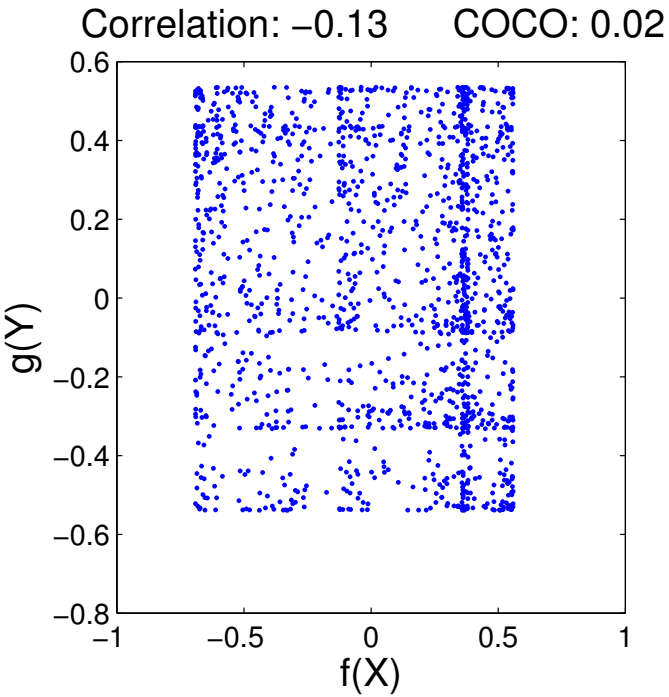
COCO vs frequency of perturbation from independence.

Case of $\omega =??$







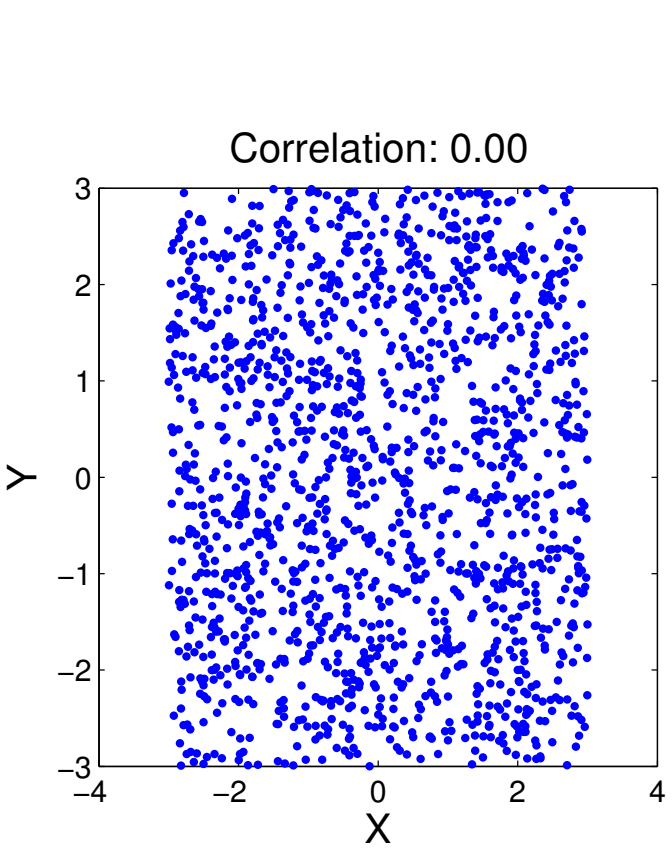


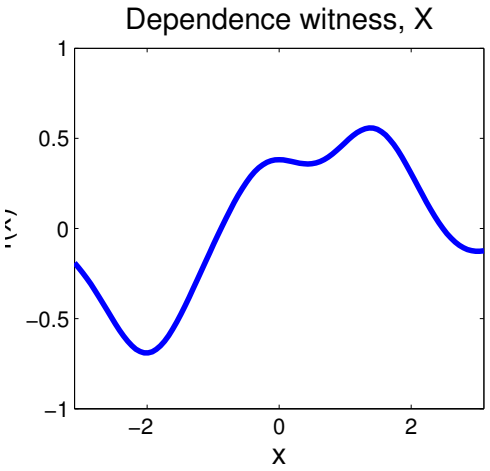
Hard-to-detect dependence

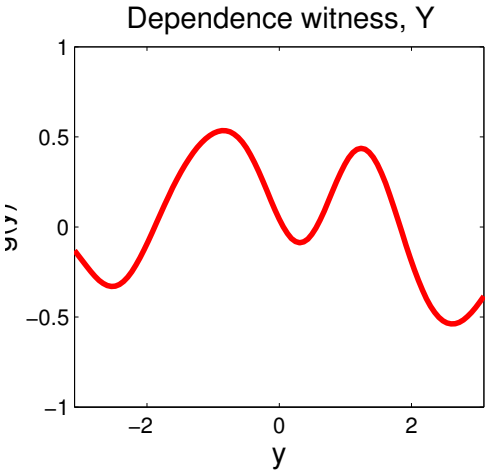
COCO vs frequency of perturbation from independence.

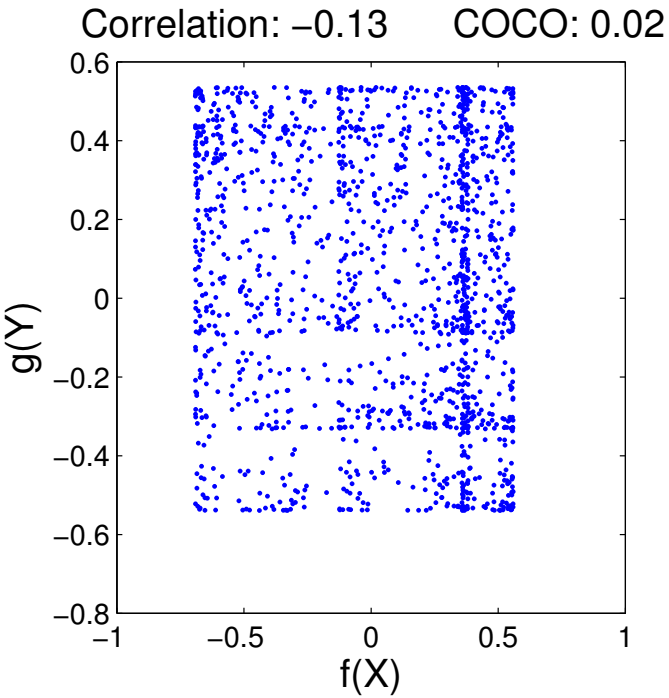
Case of uniform noise!

This bias will decrease with increasing sample size.









Hard-to-detect dependence

COCO vs frequency of perturbation from independence.

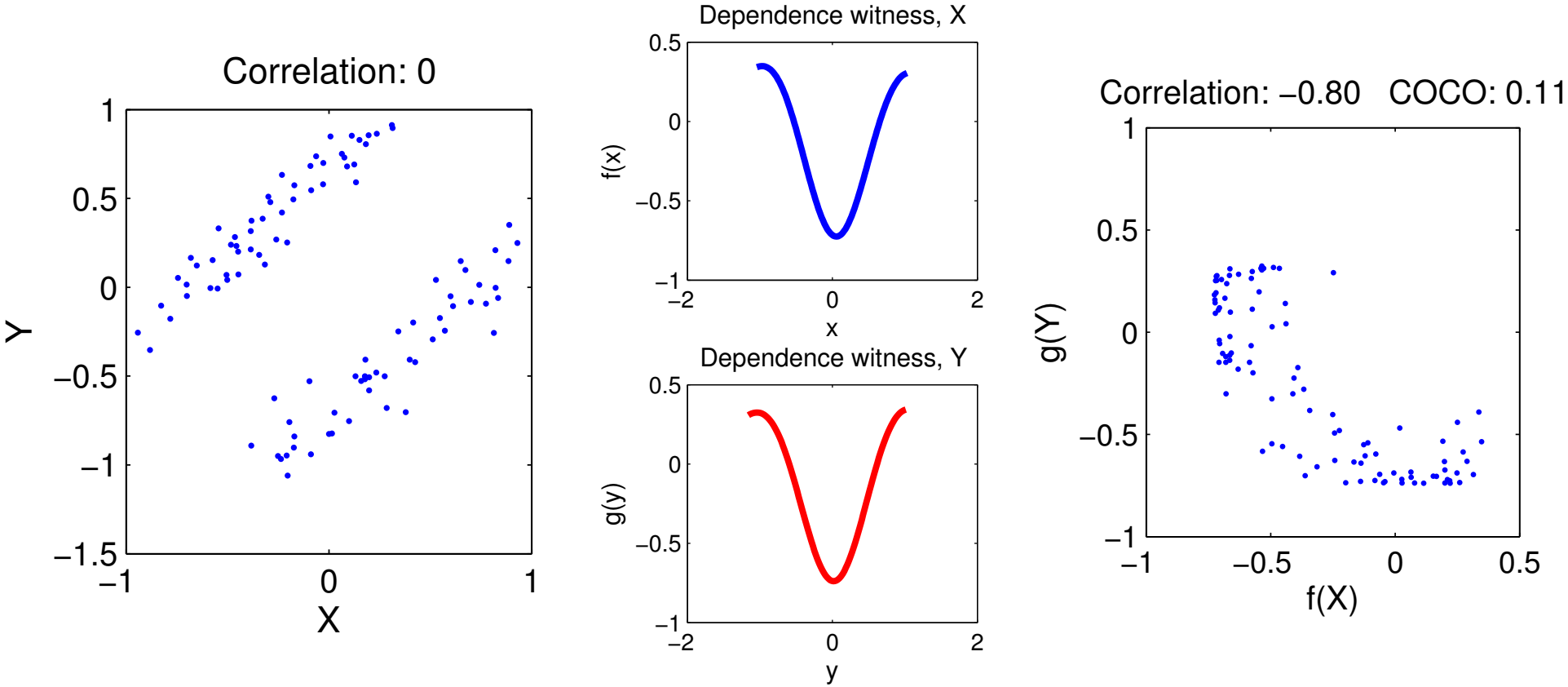
- As dependence is encoded at higher frequencies, the smooth mappings f, g achieve lower linear covariance.
- Even for independent variables, COCO will **not** be zero at finite sample sizes, since some mild linear dependence will be induced by f, g (bias)
- This bias will decrease with increasing sample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

More functions revealing dependence

- Can we do better than COCO?

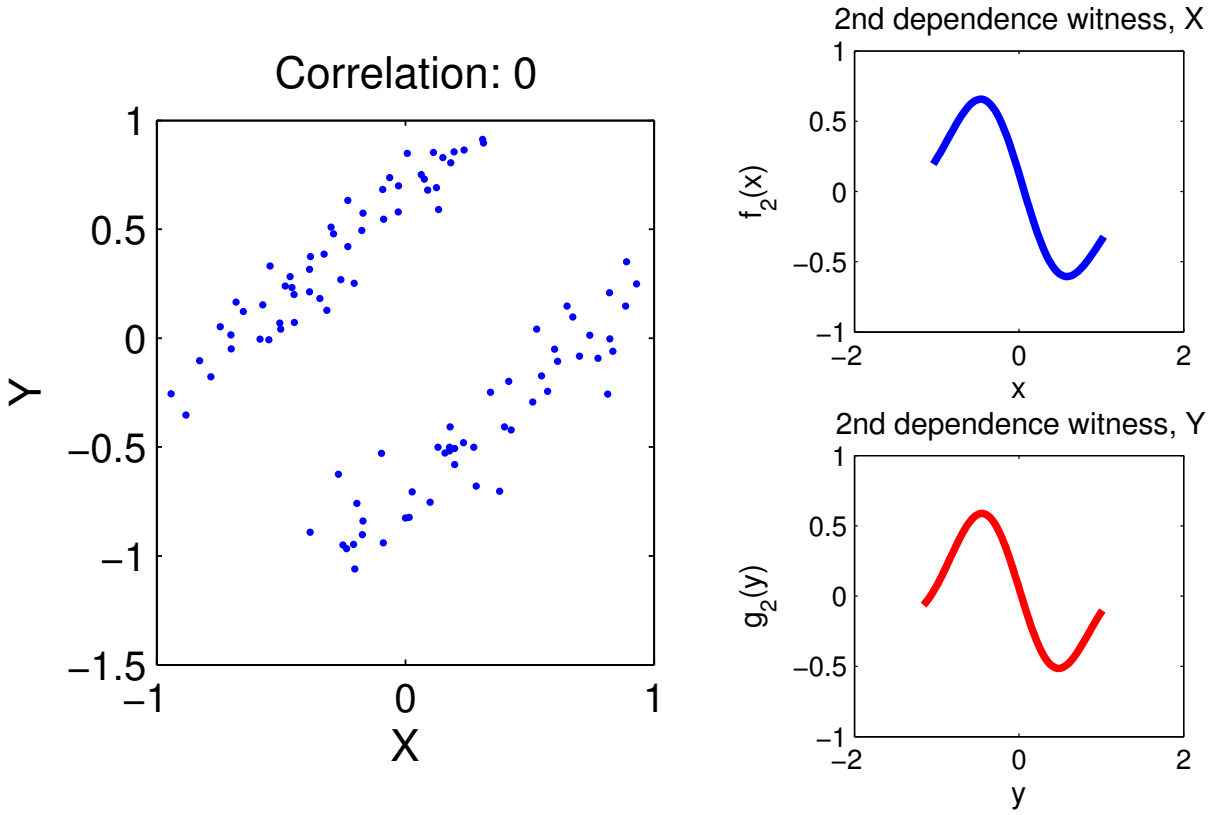
More functions revealing dependence

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



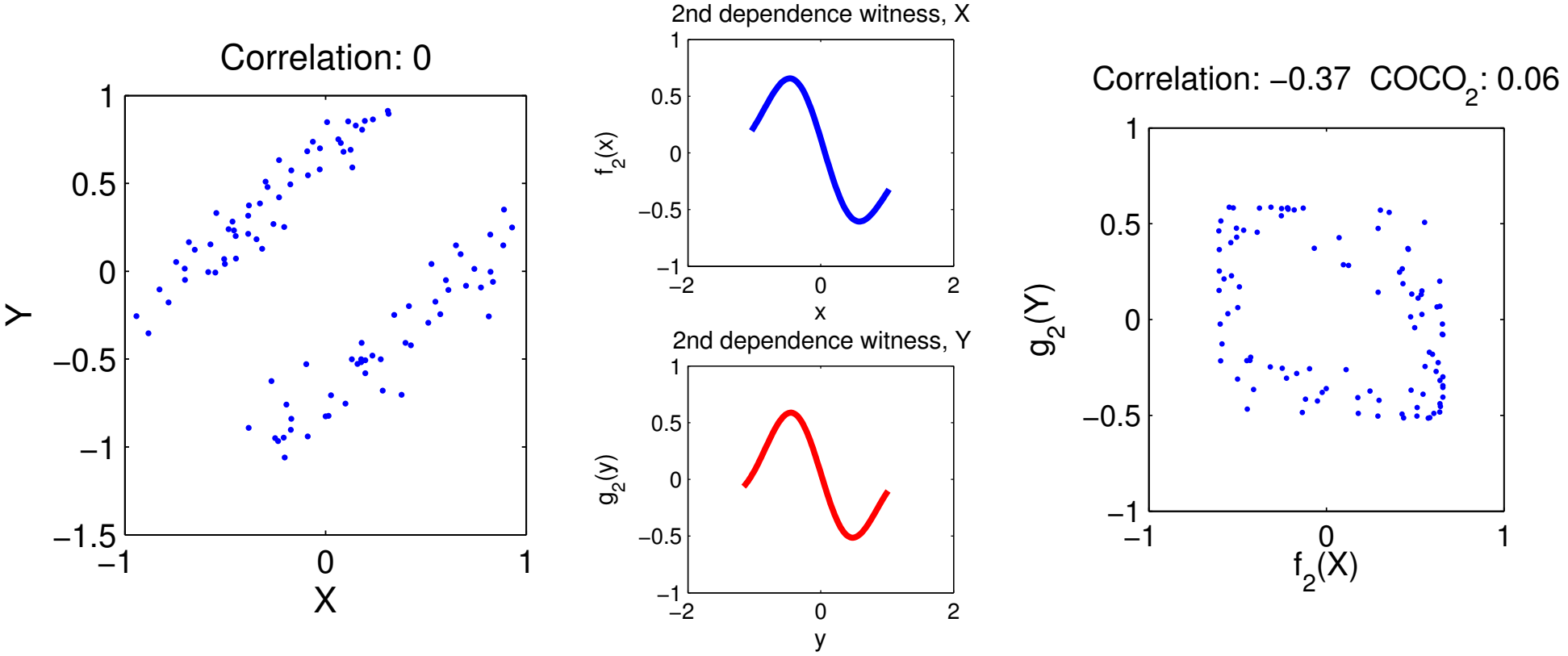
More functions revealing dependence

- 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 $\mathcal{F}$:

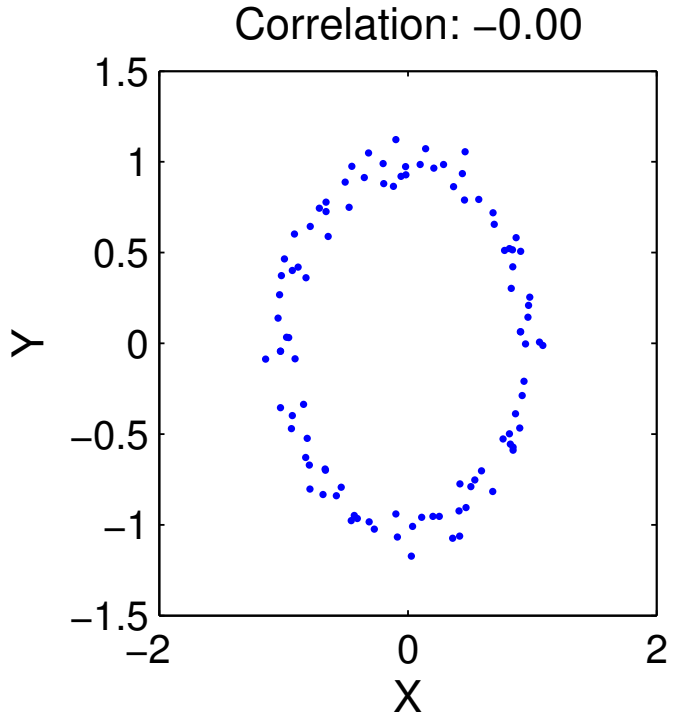
$$f^* = \arg\min_{f \in \mathcal{F}} \left(\mathbf{E}_{XY} \left(Y - \langle f, \phi(X) \rangle_{\mathcal{F}} \right)^2 + \lambda \|f\|_{\mathcal{F}}^2 \right),$$

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 $\mathcal{F}$:

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

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



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., 2012]

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

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., 2012]

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

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]

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

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

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

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

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

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

$$k_\rho(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').$$

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

$$k_\rho(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').$$

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

$$k_\rho(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').$$

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

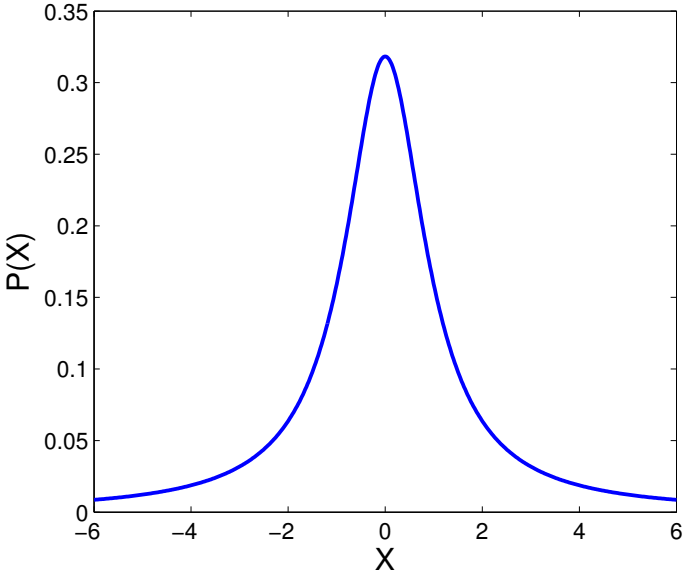
Call k_ρ a distance induced kernel

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 $0 < q \leq 2$.

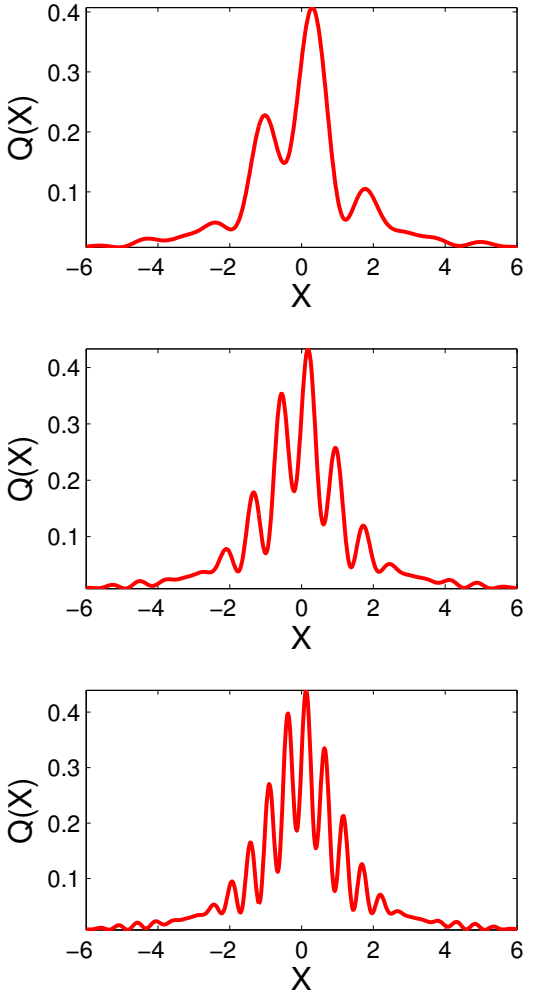
Energy distance is MMD with a distance induced kernel
Distance covariance is HSIC with distance induced kernels

Two-sample testing benchmark

Two-sample testing example in 1-D:

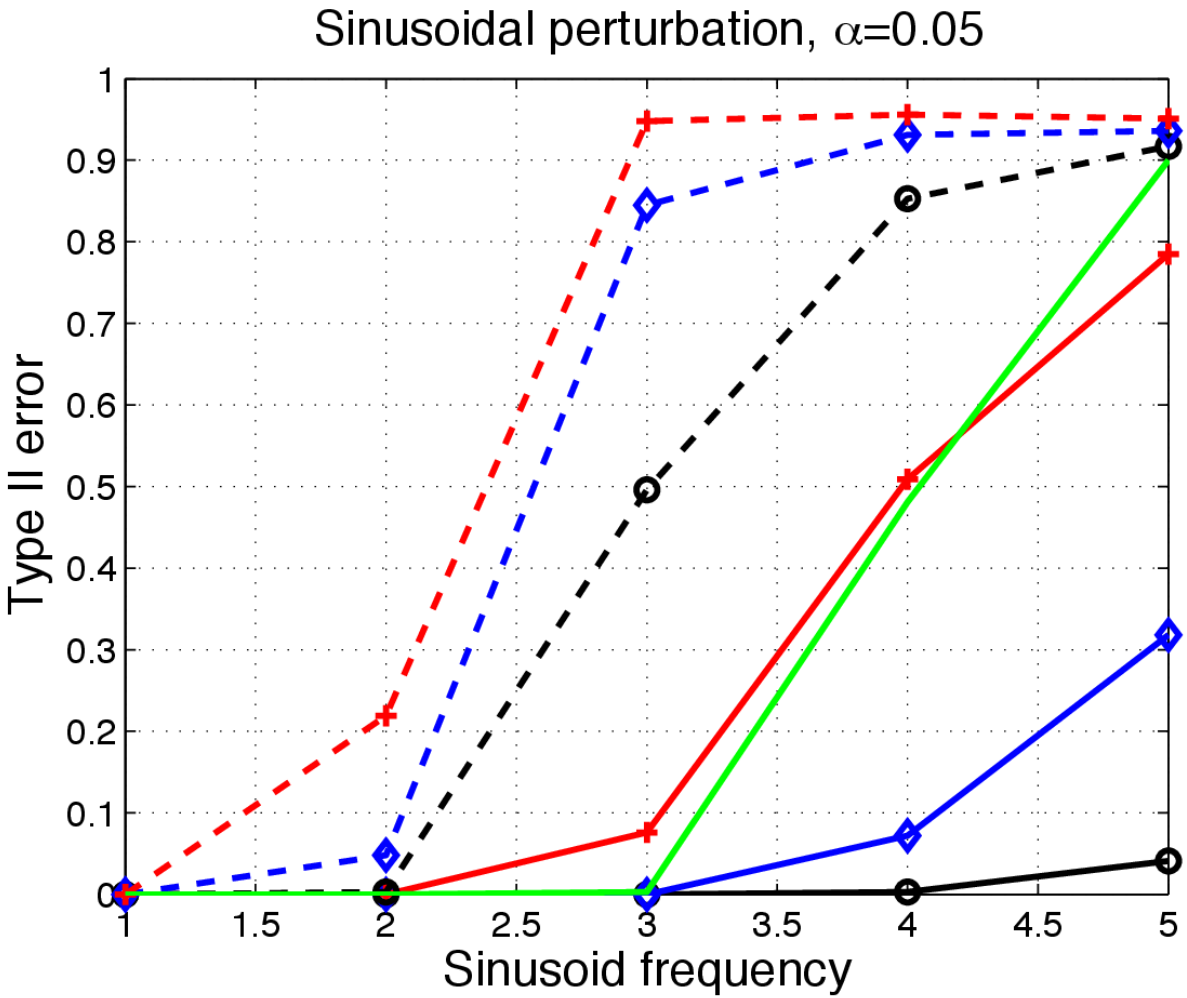


VS



Two-sample test, MMD with distance kernel

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

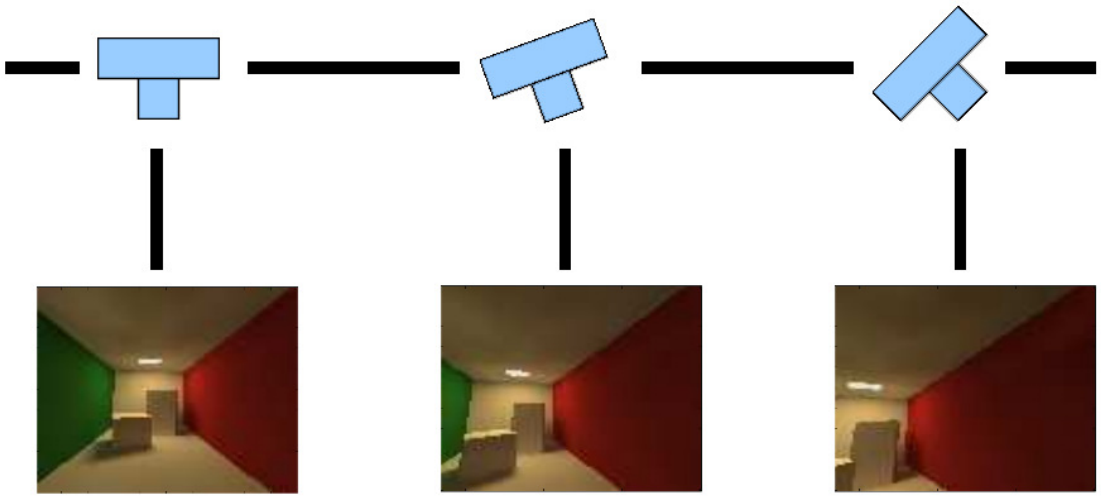


Key:

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

Nonparametric Bayesian inference using distribution embeddings

Motivating Example: Bayesian inference without a model



- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0,1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.

Challenges:

- No parametric model of camera dynamics (only samples)
- No parametric model of map from camera angle to image (only samples)
- Want to do filtering: Bayesian inference

ABC: an approach to Bayesian inference without a model

Bayes rule:

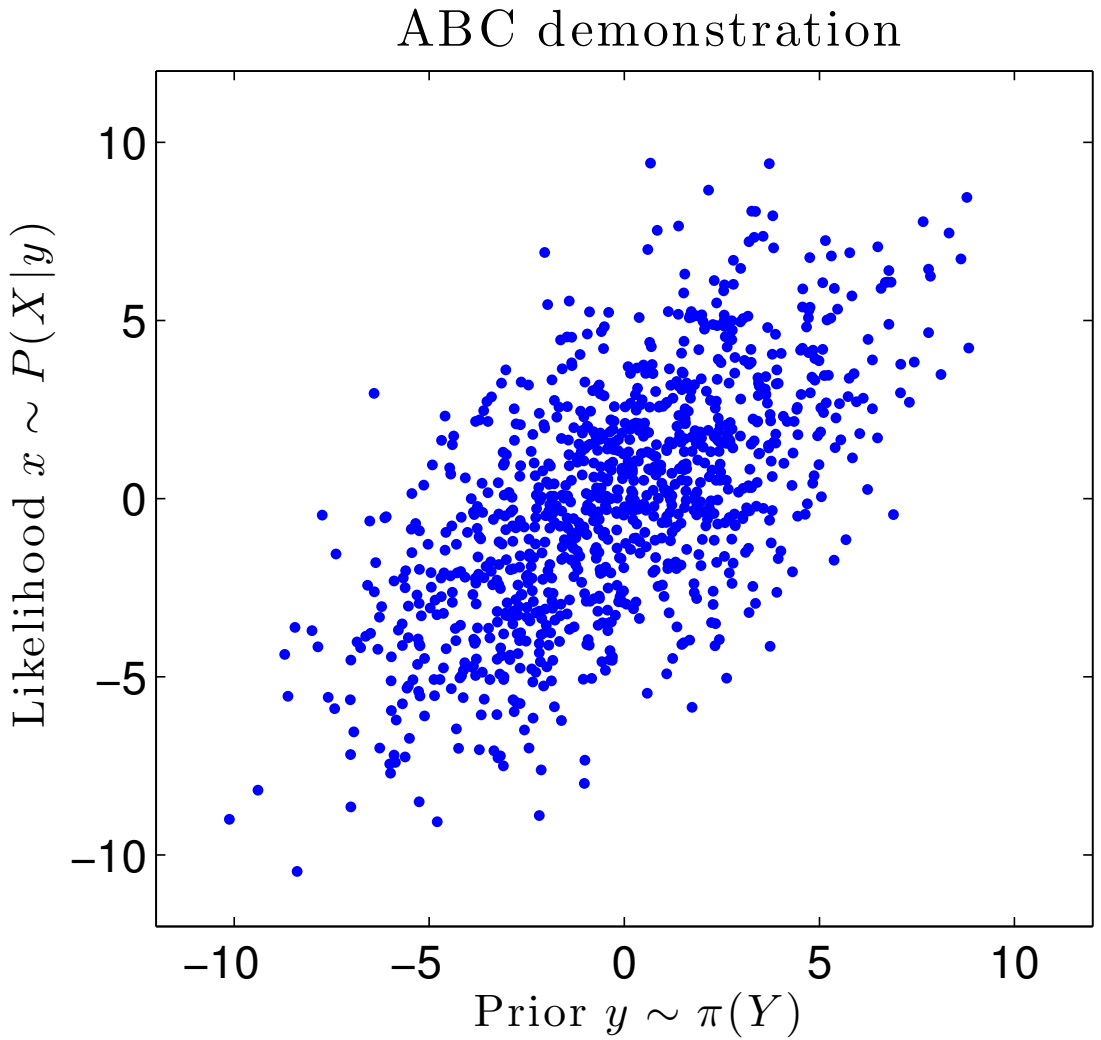
$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- $\pi(y)$ is prior

One approach: Approximate Bayesian Computation (ABC)

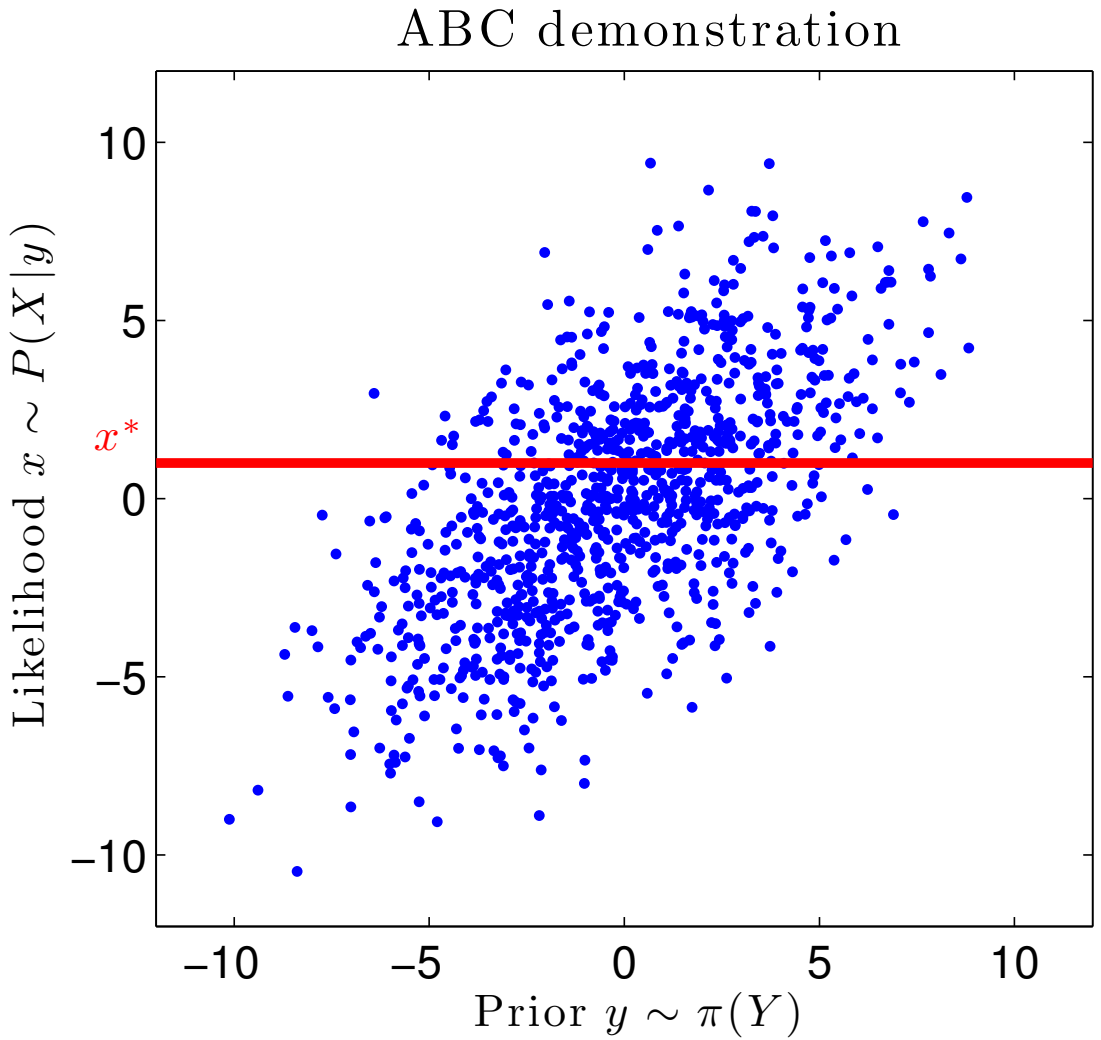
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



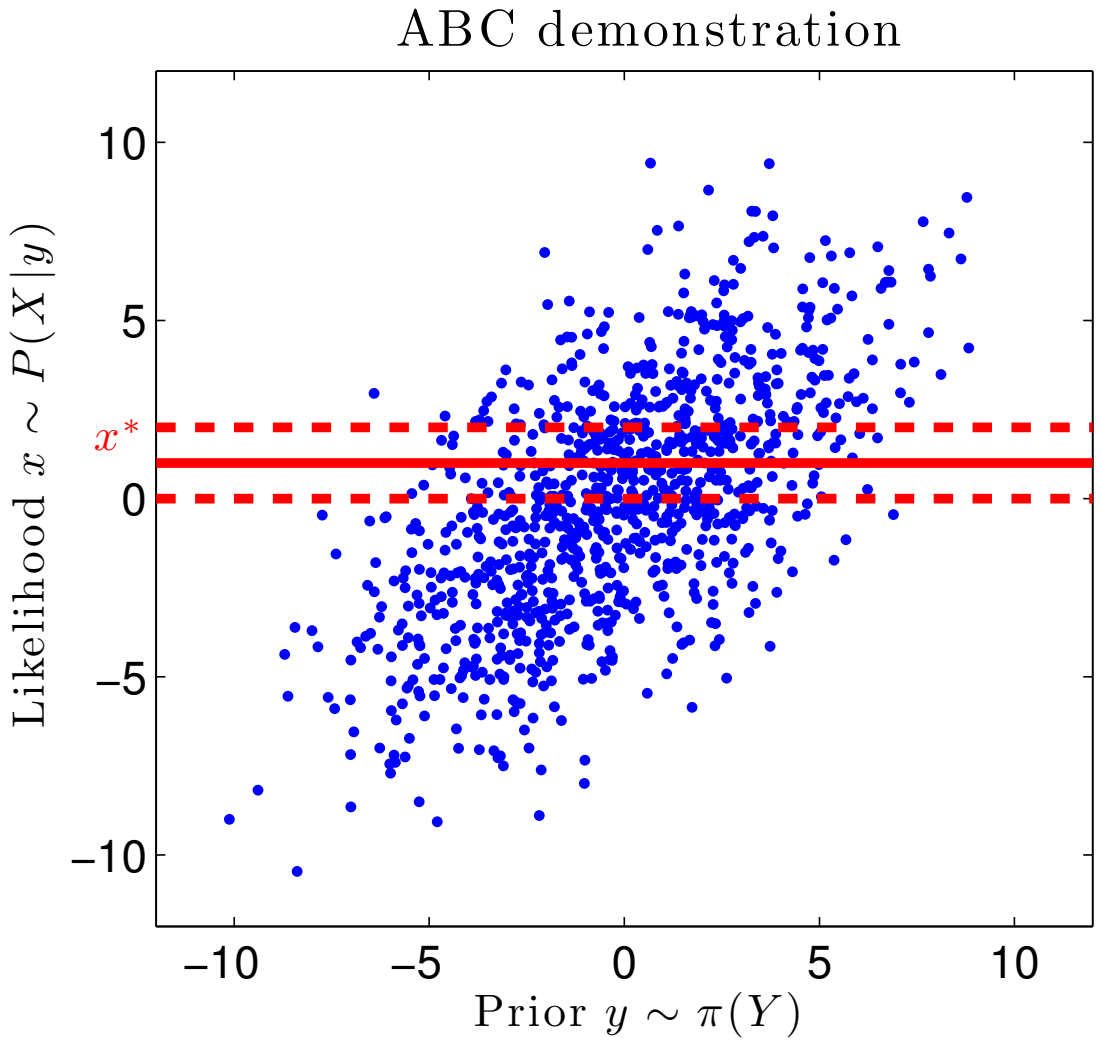
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



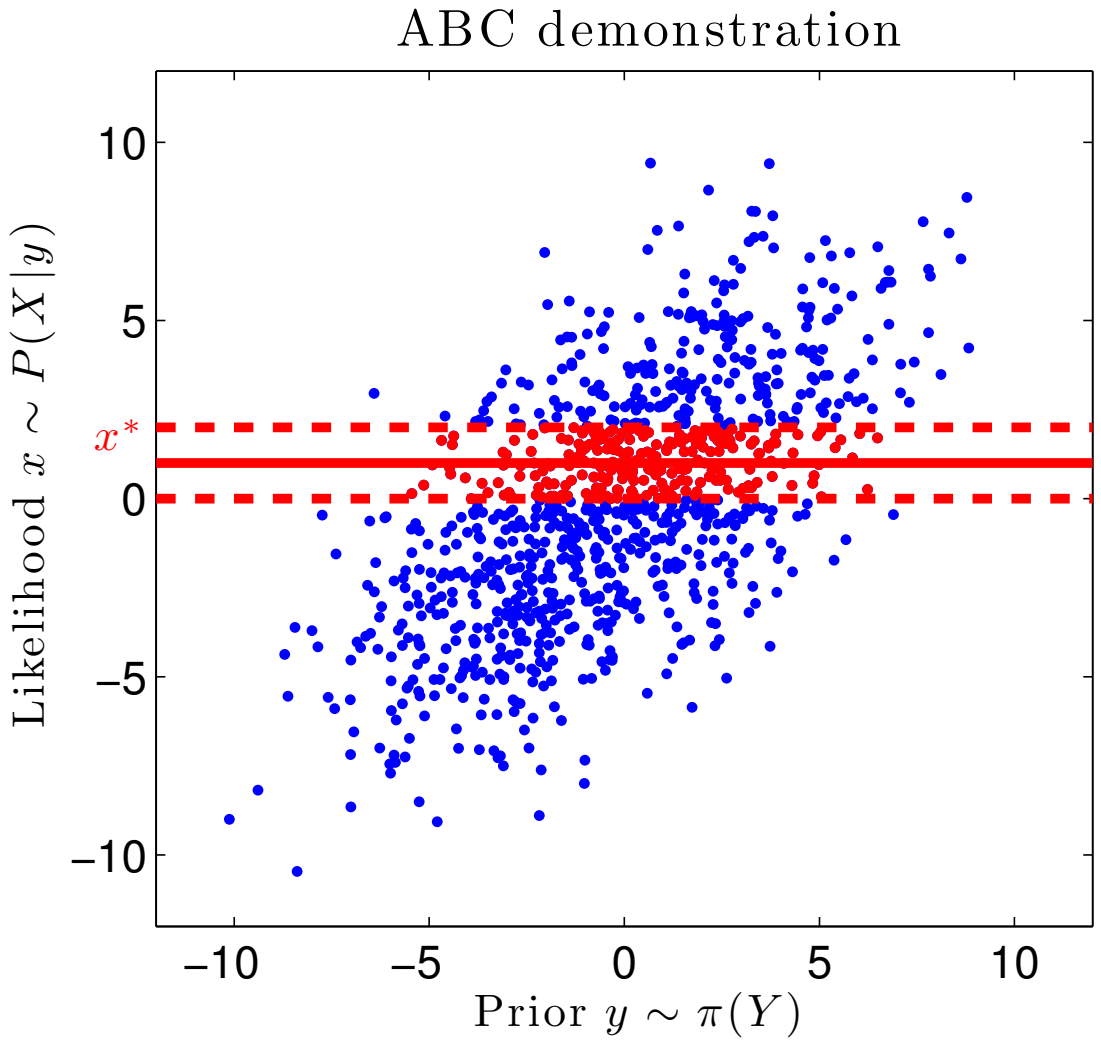
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



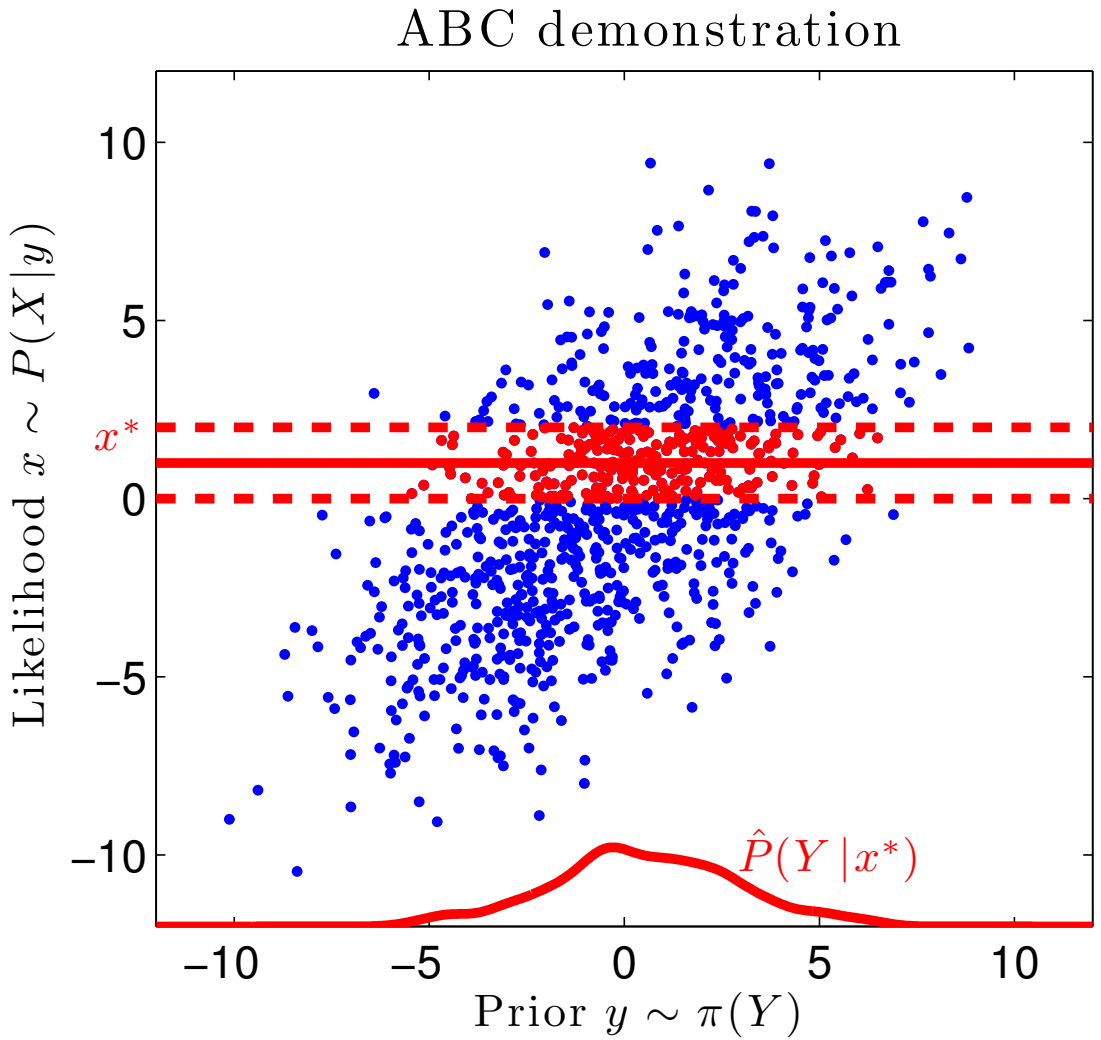
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



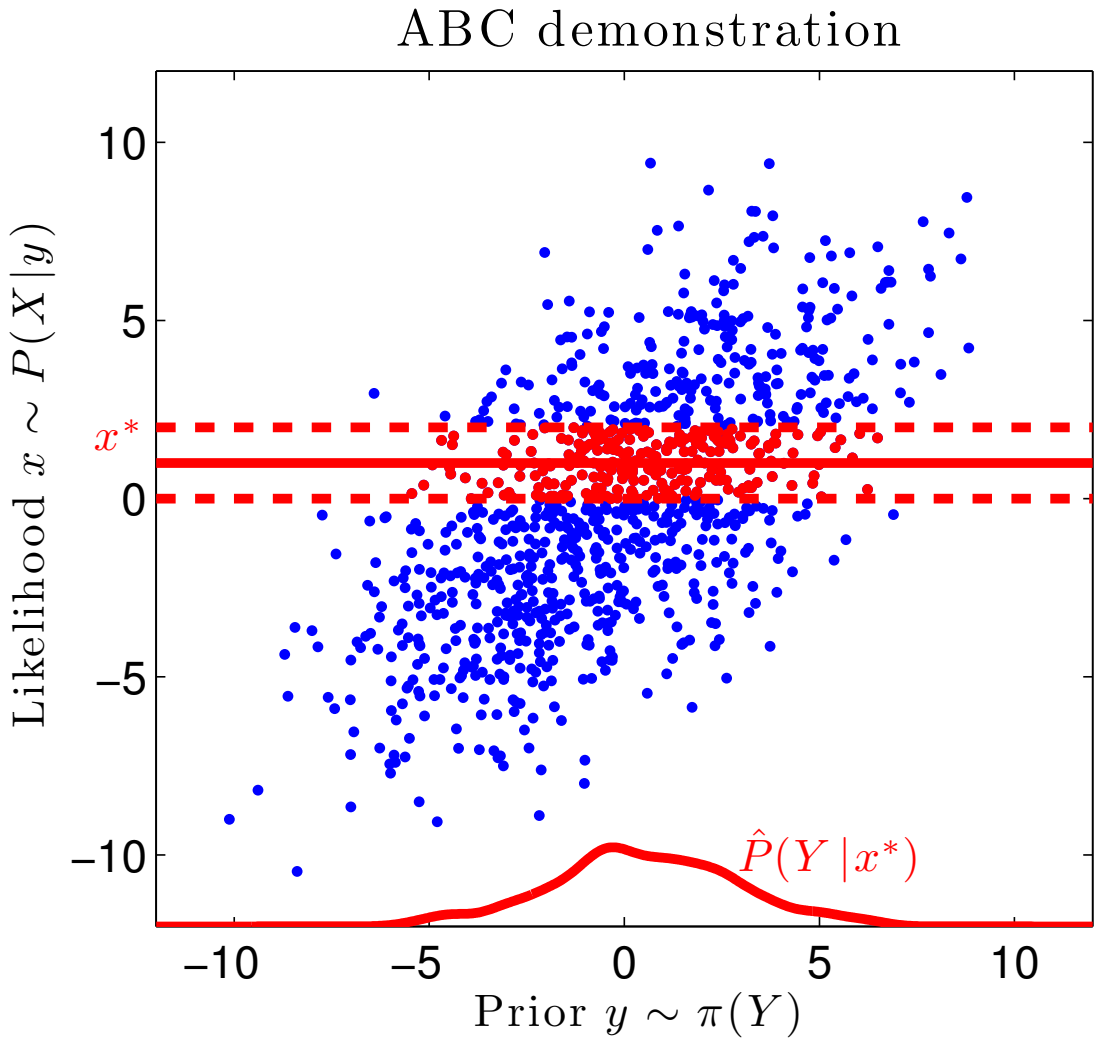
ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



ABC: an approach to Bayesian inference without a model

Approximate Bayesian Computation (ABC):



Needed: distance measure D, tolerance parameter τ.

ABC: an approach to Bayesian inference without a model

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- $\pi(y)$ is prior

ABC generates a sample from $p(Y|x^*)$ as follows:

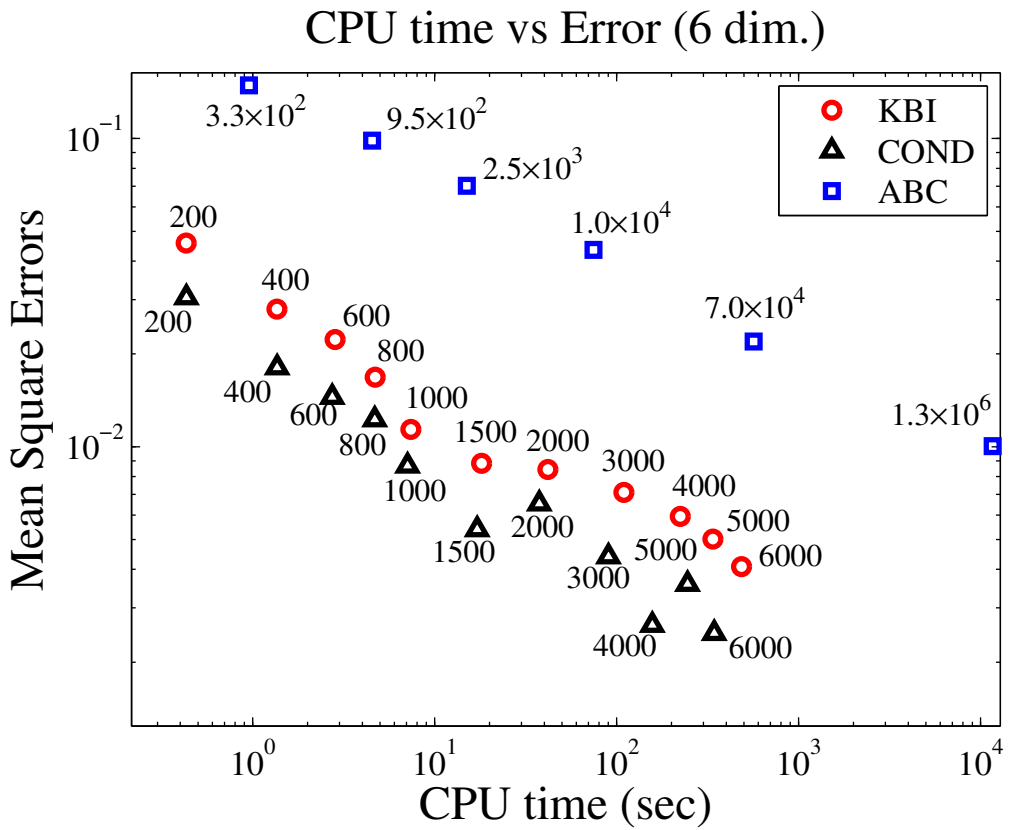
1. generate a sample y_t from the prior π,
2. generate a sample x_t from $\mathbf{P}(X|y_t)$,
3. if $D(x^*, x_t) < \tau$, accept $y = y_t$; otherwise reject,
4. go to (i).

In step (3), D is a distance measure, and τ is a tolerance parameter.

Motivating example 2: simple Gaussian case

- $p(x, y)$ is $\mathcal{N}((0, \mathbf{1}_d^T)^T, V)$ with V a randomly generated covariance

Posterior mean on x: ABC vs kernel approach



Bayes again

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- π is prior

How would this look with kernel embeddings?

Bayes again

Bayes rule:

$$\mathbf{P}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\int \mathbf{P}(x|y)\pi(y)dy}$$

- $\mathbf{P}(x|y)$ is likelihood
- π is prior

How would this look with kernel embeddings?

Define RKHS $\mathcal{G}$ on $\mathcal{Y}$ with feature map ψ_y and kernel $l(y, \cdot)$

We need a conditional mean embedding: for all $g \in \mathcal{G}$,

$$\mathbf{E}_{Y|x^*} g(Y) = \langle g, \mu_{\mathbf{P}(y|x^*)} \rangle_{\mathcal{G}}$$

This will be obtained by RKHS-valued ridge regression

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \dots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \dots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Solve

$$\breve{A} = \arg\min_{A \in \mathbb{R}^{d' \times d}} \left(\|Y - AX\|^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right),$$

where

$$\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^\top A) = \sum_{i=1}^{\min\{d,d'\}} \gamma_{A,i}^2$$

Ridge regression and the conditional feature mean

Ridge regression from $\mathcal{X} := \mathbb{R}^d$ to a finite *vector* output $\mathcal{Y} := \mathbb{R}^{d'}$ (these could be d' nonlinear features of y):

Define training data

$$X = \begin{bmatrix} x_1 & \dots & x_m \end{bmatrix} \in \mathbb{R}^{d \times m} \qquad Y = \begin{bmatrix} y_1 & \dots & y_m \end{bmatrix} \in \mathbb{R}^{d' \times m}$$

Solve

$$\breve{A} = \arg\min_{A \in \mathbb{R}^{d' \times d}} \left(\|Y - AX\|^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right),$$

where

$$\|A\|_{\mathrm{HS}}^2 = \mathrm{tr}(A^\top A) = \sum_{i=1}^{\min\{d,d'\}} \gamma_{A,i}^2$$

Solution: $\breve{A} = C_{YX} \left(C_{XX} + m\lambda I \right)^{-1}$

Ridge regression and the conditional feature mean

Prediction at new point x:

$$\begin{aligned} y^* &= \breve{A}x \\ &= C_{YX}\left(C_{XX} + m\lambda I\right)^{-1} x \\ &= \sum_{i=1}^{m} \beta_i(x) y_i \end{aligned}$$

where

$$\beta_i(x) = (K + \lambda m I)^{-1} \begin{bmatrix} k(x_1, x) & \dots & k(x_m, x) \end{bmatrix}^\top$$

and

$$K := X^\top X \qquad\qquad k(x_1, x) = x_1^\top x$$

Ridge regression and the conditional feature mean

Prediction at new point x:

$$\begin{aligned} y^* &= \breve{A}x \\ &= C_{YX}\left(C_{XX} + m\lambda I\right)^{-1} x \\ &= \sum_{i=1}^{m} \beta_i(x) y_i \end{aligned}$$

where

$$\beta_i(x) = (K + \lambda m I)^{-1} \begin{bmatrix} k(x_1, x) & \dots & k(x_m, x) \end{bmatrix}^\top$$

and

$$K := X^\top X \qquad\qquad k(x_1, x) = x_1^\top x$$

What if we do everything in kernel space?

Ridge regression and the conditional feature mean

Recall our setup:

- Given training *pairs:*

$$(x_i, y_i) \sim \mathbf{P}_{XY}$$

- $\mathcal{F}$ on $\mathcal{X}$ with feature map φ_x and kernel $k(x, \cdot)$
- $\mathcal{G}$ on $\mathcal{Y}$ with feature map ψ_y and kernel $l(y, \cdot)$

We define the covariance between feature maps:

$$C_{XX} = \mathbf{E}_X \ (\varphi_X \otimes \varphi_X) \qquad C_{XY} = \mathbf{E}_{XY} \ (\varphi_X \otimes \psi_Y)$$

and matrices of feature mapped training data

$$X = \begin{bmatrix} \varphi_{x_1} & \dots & \varphi_{x_m} \end{bmatrix} \quad Y := \begin{bmatrix} \psi_{y_1} & \dots & \psi_{y_m} \end{bmatrix}$$

Ridge regression and the conditional feature mean

Objective: [Weston et al. (2003), Micchelli and Pontil (2005), Caponnetto and De Vito (2007), Grunewalder et al. (2012, 2013)]

$$\breve{A} = \arg \min_{A \in \mathrm{HS}(\mathcal{F},\mathcal{G})} \left(\mathbf{E}_{XY} \left\| Y - AX \right\|_{\mathcal{G}}^2 + \lambda \|A\|_{\mathrm{HS}}^2 \right), \qquad \|A\|_{\mathrm{HS}}^2 = \sum_{i=1}^{\infty} \gamma_{A,i}^2$$

Solution same as vector case:

$$\breve{A} = C_{YX} \left(C_{XX} + m\lambda I \right)^{-1},$$

Prediction at new x using kernels:

$$\begin{aligned} \breve{A}\varphi_x &= \begin{bmatrix} \psi_{y_1} & \dots & \psi_{y_m} \end{bmatrix} (K + \lambda m I)^{-1} \begin{bmatrix} k(x_1, x) & \dots & k(x_m, x) \end{bmatrix} \\ &= \sum_{i=1}^{m} \beta_i(x) \psi_{y_i} \end{aligned}$$

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

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|_{\mathcal{G}}^2$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := A\varphi_x$$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|_{\mathcal{G}}^2$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := A\varphi_x$$

We need A to have the property

$$\begin{aligned} \mathbf{E}_{Y|x} g(Y) &\approx \langle g, \mu_{Y|x} \rangle_{\mathcal{G}} \\ &= \langle g, A\varphi_x \rangle_{\mathcal{G}} \\ &= \langle A^* g, \varphi_x \rangle_{\mathcal{F}} = (A^* g)(x) \end{aligned}$$

Ridge regression and the conditional feature mean

How is loss $\|Y - AX\|_{\mathcal{G}}^2$ relevant to conditional expectation of some $\mathbf{E}_{Y|x} g(Y)$? Define: [Song et al. (2009), Grunewalder et al. (2013)]

$$\mu_{Y|x} := A\varphi_x$$

We need A to have the property

$$\begin{aligned}\mathbf{E}_{Y|x} g(Y) &\approx \langle g, \mu_{Y|x} \rangle_{\mathcal{G}} \\ &= \langle g, A\varphi_x \rangle_{\mathcal{G}} \\ &= \langle A^* g, \varphi_x \rangle_{\mathcal{F}} = (A^* g)(x)\end{aligned}$$

Natural risk function for conditional mean

$$\mathcal{L}(A, \mathbf{P}_{XY}) := \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\underbrace{\left(\mathbf{E}_{Y|X} g(Y)\right)}_{\text{Target}}(X) - \underbrace{(A^* g)}_{\text{Estimator}}(X) \right]^2,$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - \left(A^* g \right)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - \left(A^* g \right)(X) \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[\langle g, \psi_Y \rangle_{\mathcal{G}} - \langle A^* g, \varphi_X \rangle_{\mathcal{F}} \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[\langle g, \psi_Y \rangle_{\mathcal{G}} - \langle g, A\varphi_X \rangle_{\mathcal{G}} \right]^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - \left(A^* g \right)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - \left(A^* g \right)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2 \\
&\leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2
\end{aligned}$$

Ridge regression and the conditional feature mean

The squared loss risk provides an upper bound on the natural risk.

$$\mathcal{L}(A, \mathbf{P}_{XY}) \leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2$$

Proof: Jensen and Cauchy Schwarz

$$\begin{aligned}
\mathcal{L}(A, \mathbf{P}_{XY}) &:= \sup_{\|g\| \leq 1} \mathbf{E}_X \left[\left(\mathbf{E}_{Y|X} g(Y) \right)(X) - (A^* g)(X) \right]^2 \\
&\leq \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \left[g(Y) - (A^* g)(X) \right]^2 \\
&= \mathbf{E}_{XY} \sup_{\|g\| \leq 1} \langle g, \psi_Y - A\varphi_X \rangle_{\mathcal{G}}^2 \\
&\leq \mathbf{E}_{XY} \left\| \psi_Y - A\varphi_X \right\|_{\mathcal{G}}^2
\end{aligned}$$

If we assume $\mathbf{E}_Y[g(Y)|X = x] \in \mathcal{F}$ then upper bound tight (next slide).

Conditions for ridge regression = conditional mean

Conditional mean obtained by ridge regression when $\mathbf{E}_Y[g(Y)|X = x] \in \mathcal{F}$

Given a function $g \in \mathcal{G}$. Assume $E_{Y|X}[g(Y)|X = \cdot] \in \mathcal{F}$. Then

$$\boxed{C_{XX} E_{Y|X}[g(Y)|X = \cdot] = C_{XY} g.}$$

Why this is useful:

$$\begin{aligned} E_{Y|X}[g(Y)|X = x] &= \langle E_{Y|X}[g(Y)|X = \cdot], \varphi_x \rangle_{\mathcal{F}} \\ &= \langle C_{XX}^{-1} C_{XY} g, \varphi_x \rangle_{\mathcal{F}} \\ &= \langle g, \underbrace{C_{YX} C_{XX}^{-1}}_{\text{regression}} \varphi_x \rangle_{\mathcal{G}} \end{aligned}$$

Conditions for ridge regression = conditional mean

Conditional mean obtained by ridge regression when $\mathbf{E}_Y[g(Y)|X = x] \in \mathcal{F}$

Given a function $g \in \mathcal{G}$. Assume $E_{Y|X}[g(Y)|X = \cdot] \in \mathcal{F}$. Then

$$\boxed{C_{XX} E_{Y|X}[g(Y)|X = \cdot] = C_{XY} g.}$$

Proof: [Fukumizu et al., 2004]

For all $f \in \mathcal{F}$, by definition of C_{XX},

$$\begin{aligned}
&\left\langle f, C_{XX} E_{Y|X}[g(Y)|X = \cdot] \right\rangle_{\mathcal{F}} \\
&= \operatorname{cov}\left(f, E_{Y|X}[g(Y)|X = \cdot]\right) \\
&= E_X\left(f(X)\, E_{Y|X}[g(Y)|X]\right) \\
&= E_{XY}(f(X)g(Y)) \\
&= \langle f, C_{XY} g \rangle ,
\end{aligned}$$

by definition of C_{XY}.

Kernel Bayes' law

- Prior: $Y \sim \pi(y)$
- Likelihood: $(X|y) \sim \mathbf{P}(x|y)$ with some joint $\mathbf{P}(x, y)$

Kernel Bayes' law

- Prior: $Y \sim \pi(y)$
- Likelihood: $(X|y) \sim \mathbf{P}(x|y)$ with some joint $\mathbf{P}(x, y)$
- Joint distribution: $\mathbf{Q}(x, y) = \mathbf{P}(x|y)\pi(y)$

Warning: $\mathbf{Q} \neq \mathbf{P}$, *change of measure* from $\mathbf{P}(y)$ to $\pi(y)$

- Marginal for x:

$$\mathbf{Q}(x) := \int \mathbf{P}(x|y)\pi(y)dy.$$

- Bayes' law:

$$\mathbf{Q}(y|x) = \frac{\mathbf{P}(x|y)\pi(y)}{\mathbf{Q}(x)}$$

Kernel Bayes' law

- Posterior embedding via the usual conditional update,

$$\mu_{\mathbf{Q}(y|x)} = C_{\mathbf{Q}(y,x)} C^{-1}_{\mathbf{Q}(x,x)} \phi_x .$$

Kernel Bayes' law

- Posterior embedding via the usual conditional update,

$$\mu_{\mathbf{Q}(y|x)} = C_{\mathbf{Q}(y,x)} C^{-1}_{\mathbf{Q}(x,x)} \phi_x .$$

- Given mean embedding of prior: $\mu_\pi(y)$
- Define marginal covariance:

$$C_{\mathbf{Q}(x,x)} = \int (\varphi_x \otimes \varphi_x) \; \mathbf{P}(x|y)\pi(y)dx = C_{(xx)y} C^{-1}_{yy} \mu_{\pi(y)}$$

- Define cross-covariance:

$$C_{\mathbf{Q}(y,x)} = \int (\phi_y \otimes \varphi_x) \; \mathbf{P}(x|y)\pi(y)dx = C_{(yx)y} C^{-1}_{yy} \mu_{\pi(y)} .$$

Kernel Bayes' law: consistency result

- How to compute posterior expectation from data?
- Given samples: $\{(x_i, y_i)\}_{i=1}^n$ from $\mathbf{P}_{\mathsf{xy}}$, $\{(u_j)\}_{j=1}^n$ from prior π.
- Want to compute $\mathbb{E}[g(Y)|X = x]$ for g in $\mathcal{G}$
- For any $x \in \mathcal{X}$,

$$\left|\mathbf{g}_y^T R_{Y|X} \mathbf{k}_X(x) - \mathbb{E}[f(Y)|X = x]\right| = O_p(n^{-\frac{4}{27}}), \quad (n \to \infty),$$

where

  - $\mathbf{g}_y = (g(y_1), \ldots, g(y_n))^T \in \mathbb{R}^n$.
  - $\mathbf{k}_X(x) = (k(x_1, x), \ldots, k(x_n, x))^T \in \mathbb{R}^n$
  - $R_{Y|X}$ learned from the samples, contains the u_j

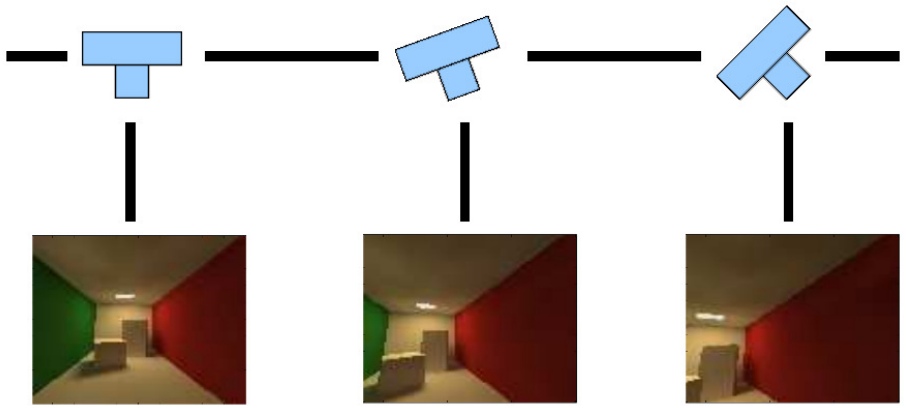
Smoothness assumptions:

- $\pi/p_Y \in \mathcal{R}(C_{YY}^{1/2})$, where p_Y p.d.f. of $\mathbf{P}_Y$,
- $E[g(Y)|X = \cdot] \in \mathcal{R}(C^2_{\mathbf{Q}(xx)})$.

Experiment: Kernel Bayes' law vs EKF

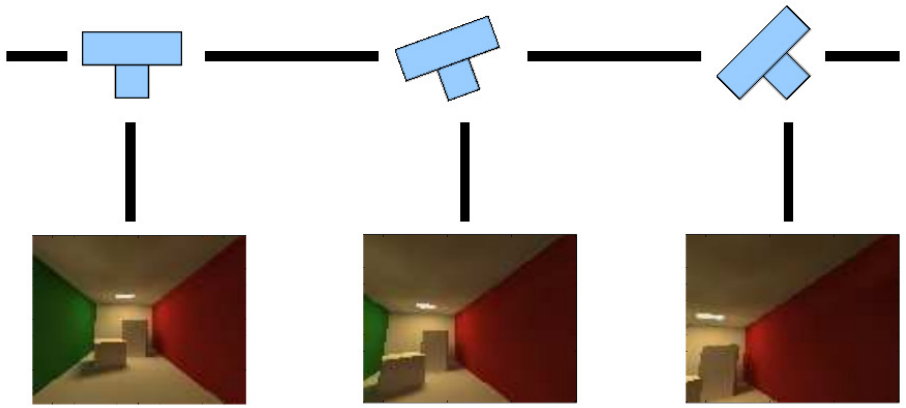
Experiment: Kernel Bayes' law vs EKF

- Compare with extended Kalman filter (EKF) on camera orientation task
- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0,1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.



Experiment: Kernel Bayes' law vs EKF

- Compare with extended Kalman filter (EKF) on camera orientation task
- 3600 downsampled frames of 20×20 RGB pixels ($Y_t \in [0,1]^{1200}$)
- 1800 training frames, remaining for test.
- Gaussian noise added to Y_t.



Average MSE and standard errors (10 runs)

	KBR (Gauss)	KBR (Tr)	Kalman (9 dim.)	Kalman (Quat.)
$\sigma^2 = 10^{-4}$	0.210 ± 0.015	0.146 ± 0.003	1.980 ± 0.083	0.557 ± 0.023
$\sigma^2 = 10^{-3}$	0.222 ± 0.009	0.210 ± 0.008	1.935 ± 0.064	0.541 ± 0.022

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

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]

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

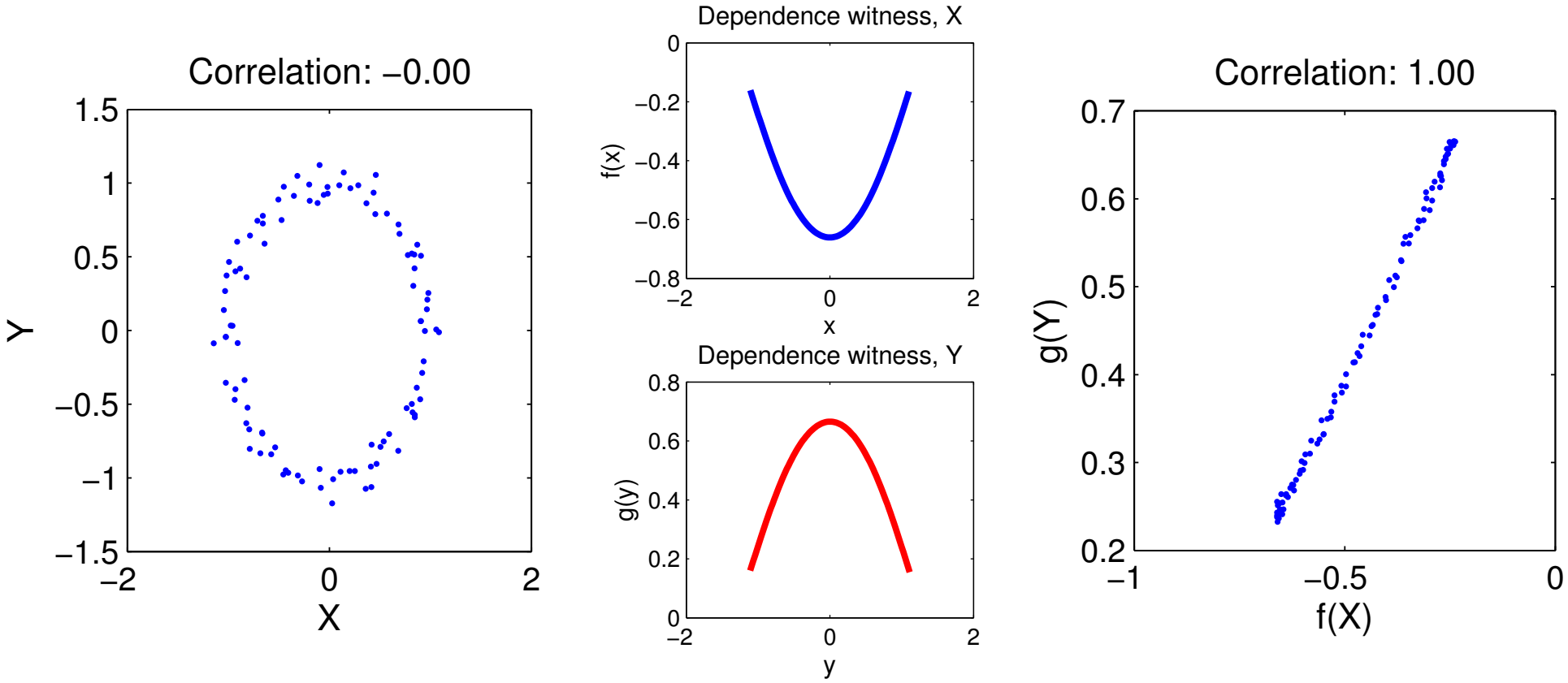
  - First canonical correlate

- Regularized empirical estimate of HS norm: [NIPS07b]

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

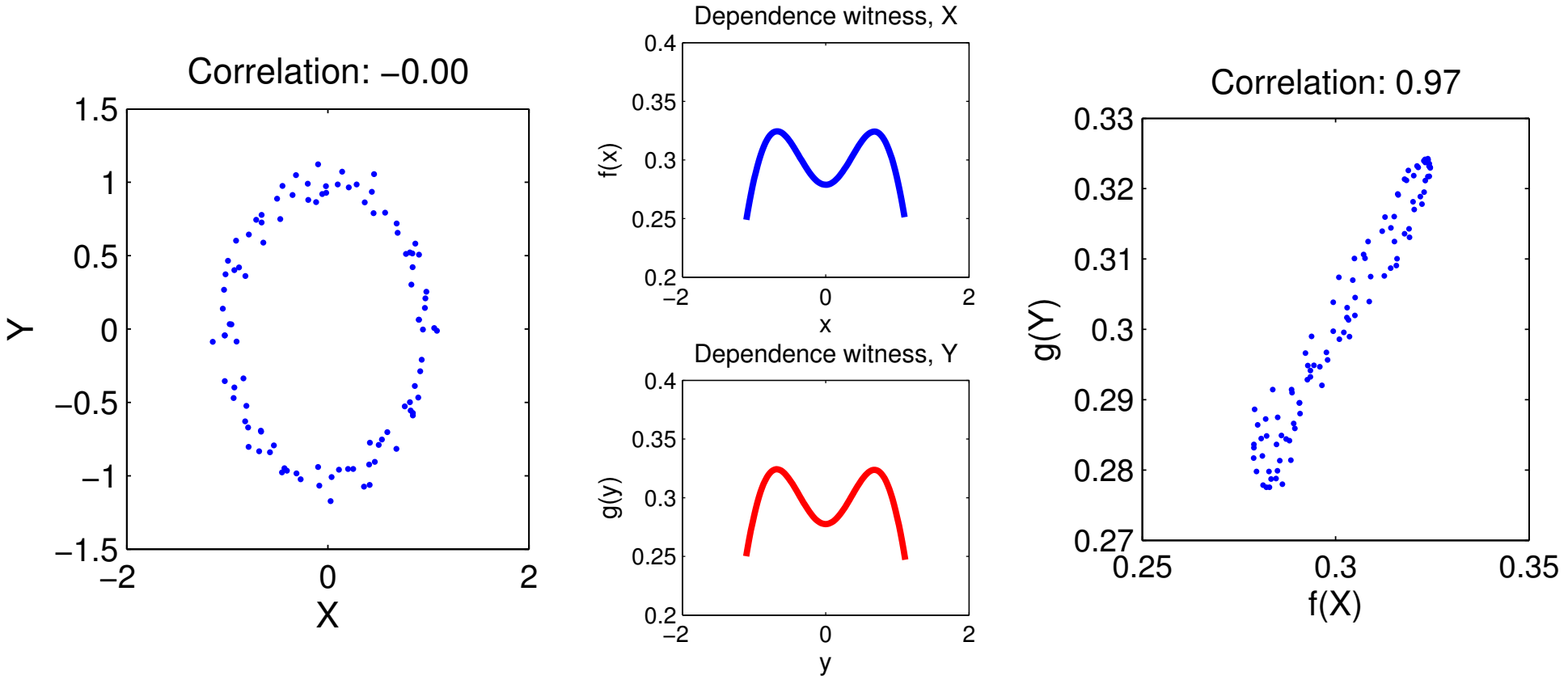
Kernel CCA: Illustration

- 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.

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, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

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.

K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14: 3753–3783, 2013.

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, 2012.

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

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. 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.

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.