# Hilbert Space Representations of Probability Distributions

*Arthur Gretton*

joint work with Karsten Borgwardt, Kenji Fukumizu, Malte Rasch, Bernhard Schölkopf, Alex Smola, Le Song, Choon Hui Teo

Max Planck Institute for Biological Cybernetics,
Tübingen, Germany

# Overview



- The two sample problem: are samples $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$ generated from the same distribution?
- Kernel independence testing: given a sample of $m$ pairs $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, are the random variables $\mathsf{x}$ and $\mathsf{y}$ independent?

# Kernels, feature maps

- Hilbert space of functions $f \in \mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$
- RKHS: evaluation operator $\delta_x \ : \ x \to \mathbb{R}$ continuous

# A very short introduction to kernels



- Hilbert space of functions $f \in \mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$
- RKHS: evaluation operator $\delta_x \ : \ x \longrightarrow \mathbb{R}$ continuous
- Riesz: Unique representer of evaluation $k(x, \cdot) \in \mathcal{F}$:

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{F}}$$

  - $k(x, \cdot)$ feature map
  - $k \ : \ \mathcal{X} \mapsto \mathbb{R}$ is kernel function

# A very short introduction to kernels



- Hilbert space of functions $f \in \mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$
- RKHS: evaluation operator $\delta_x : x \to \mathbb{R}$ continuous
- Riesz: Unique representer of evaluation $k(x, \cdot) \in \mathcal{F}$:

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{F}}$$

  - $k(x, \cdot)$ feature map
  - $k : \mathcal{X} \mapsto \mathbb{R}$ is kernel function
- Inner product between two feature maps:

$$\langle k(x_1, \cdot), k(x_2, \cdot) \rangle_{\mathcal{F}} = k(x_1, x_2)$$

# A Kernel Method for the Two Sample Problem

# The two-sample problem



- Given:
  - $m$ samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ drawn i.i.d. from **P**
  - samples $\boldsymbol{y}$ drawn from **Q**
- Determine: Are **P** and **Q** different?

# The two-sample problem

- Given:
  - $m$ samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ drawn i.i.d. from **P**
  - samples $\boldsymbol{y}$ drawn from **Q**
- Determine: Are **P** and **Q** different?
- Applications:
  - Microarray data aggregation
  - Speaker/author identification
  - Schema matching

# The two-sample problem



- Given:
  - $m$ samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ drawn i.i.d. from $\mathsf{P}$
  - samples $\boldsymbol{y}$ drawn from $\mathsf{Q}$
- Determine: Are $\mathsf{P}$ and $\mathsf{Q}$ different?
- Applications:
  - Microarray data aggregation
  - Speaker/author identification
  - Schema matching
- Where is our test useful?
  - High dimensionality
  - Low sample size
  - Structured data (strings and graphs): currently the only method

# Overview (two-sample problem)



- How to detect $\mathbf{P} \neq \mathbf{Q}$?
  - Distance between means in space of features
  - Function revealing differences in distributions
  - Same thing: the MMD [Gretton et al., 2007, Borgwardt et al., 2006]

# Overview (two-sample problem)

- How to detect $\mathbf{P} \neq \mathbf{Q}$?
  - Distance between means in space of features
  - Function revealing differences in distributions
  - Same thing: the MMD [Gretton et al., 2007, Borgwardt et al., 2006]
- Hypothesis test using MMD
  - Asymptotic distribution of MMD
  - Large deviation bounds

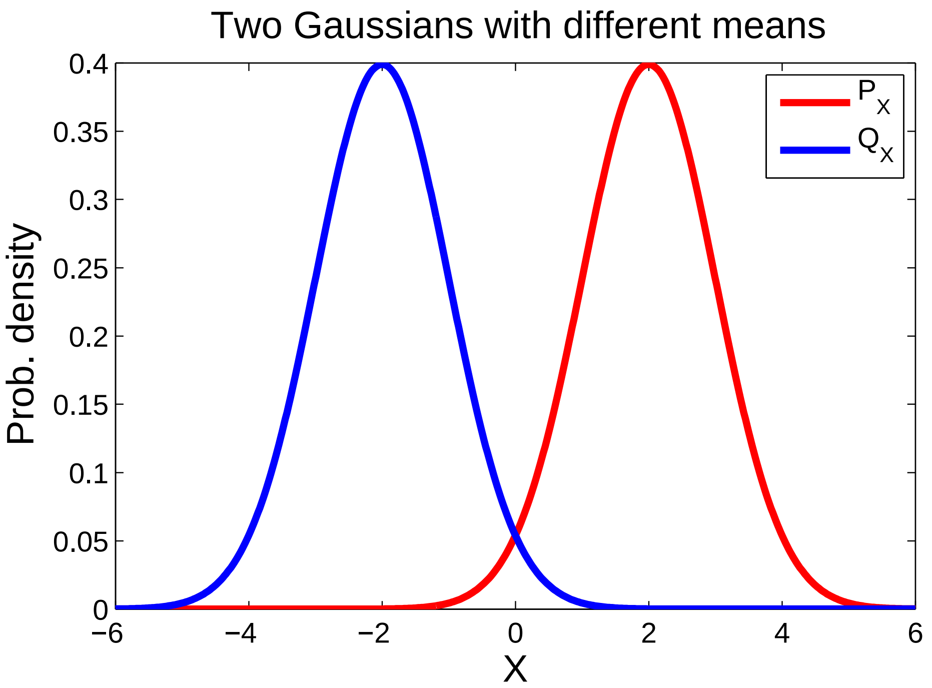
# Overview (two-sample problem)

- How to detect $\mathbf{P} \neq \mathbf{Q}$?
  - Distance between means in space of features
  - Function revealing differences in distributions
  - Same thing: the MMD [Gretton et al., 2007, Borgwardt et al., 2006]
- Hypothesis test using MMD
  - Asymptotic distribution of MMD
  - Large deviation bounds
- Experiments

# Mean discrepancy (1)



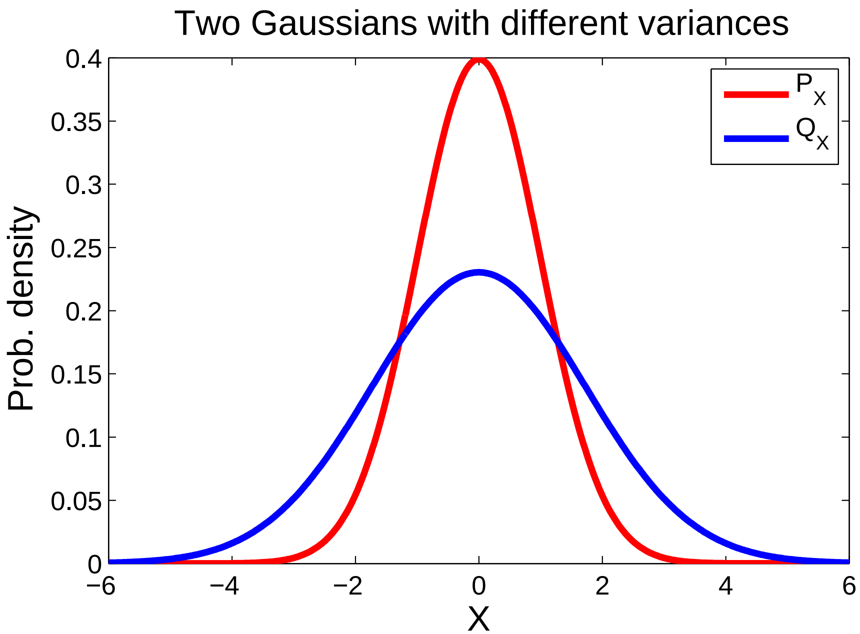
- Simple example: 2 Gaussians with different means
- Answer: t-test



# Mean discrepancy (2)

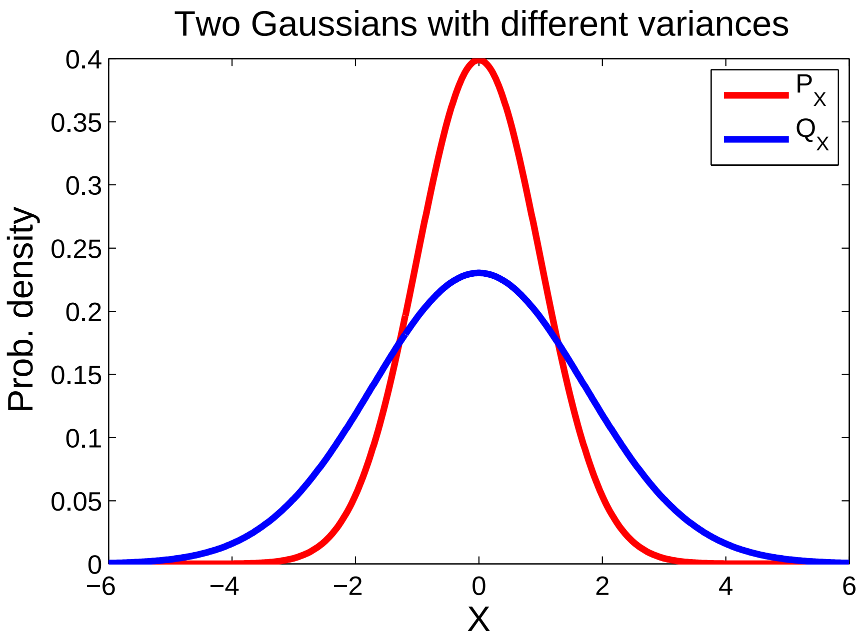


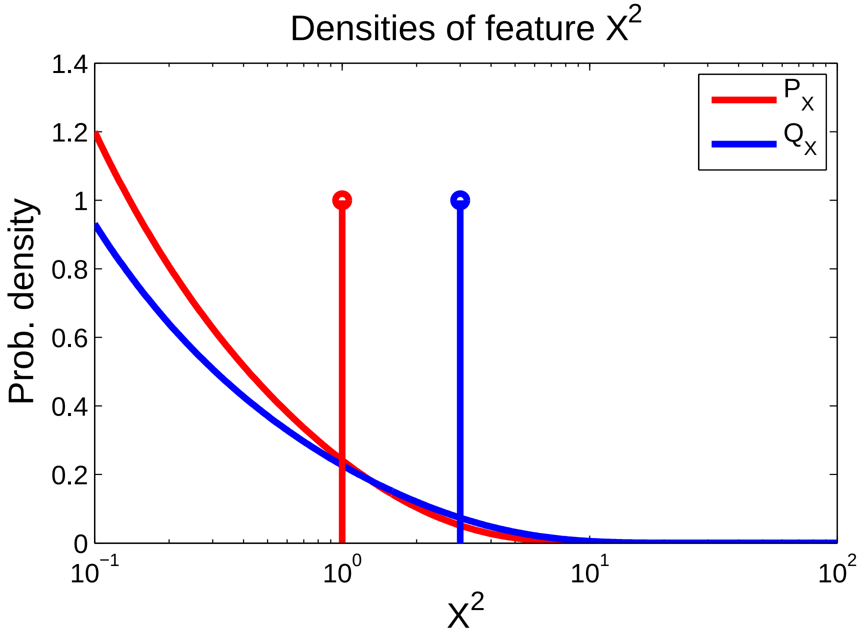
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $x^2$



# Mean discrepancy (2)

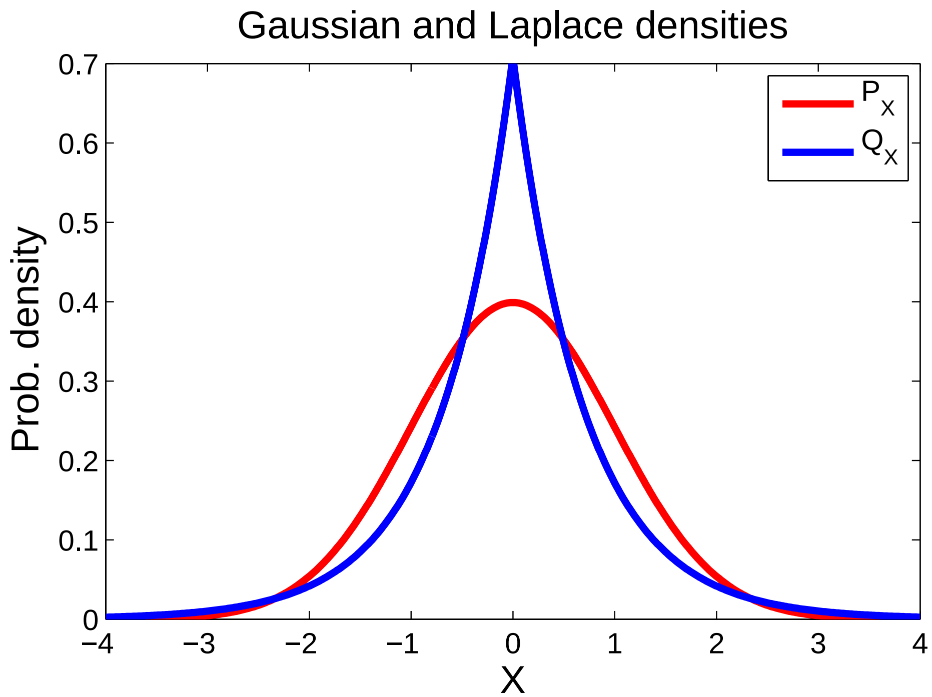
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $x^2$



# Mean discrepancy (3)

- Gaussian and Laplace distributions
- Same mean *and* same variance
- Difference in means using higher order features



# Function revealing difference in distributions (1)



- Idea: avoid density estimation when testing $\mathbf{P} \neq \mathbf{Q}$

  [Fortet and Mourier, 1953]

$$D(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[ \mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right].$$

# Function revealing difference in distributions (1)

- Idea: avoid density estimation when testing $\mathbf{P} \neq \mathbf{Q}$

  [Fortet and Mourier, 1953]

$$D(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[ \mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right].$$

- $D(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when $F$ =bounded continuous functions [Dudley, 2002]

# Function revealing difference in distributions (1)



- Idea: avoid density estimation when testing $\mathbf{P} \neq \mathbf{Q}$

[Fortet and Mourier, 1953]

$$D(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[ \mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right].$$

- $D(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when $F =$bounded continuous functions [Dudley, 2002]

- $D(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F =$the unit ball in a universal RKHS $\mathcal{F}$ [via Steinwart, 2001]

# Function revealing difference in distributions (1)



- Idea: avoid density estimation when testing $\mathbf{P} \neq \mathbf{Q}$

  [Fortet and Mourier, 1953]

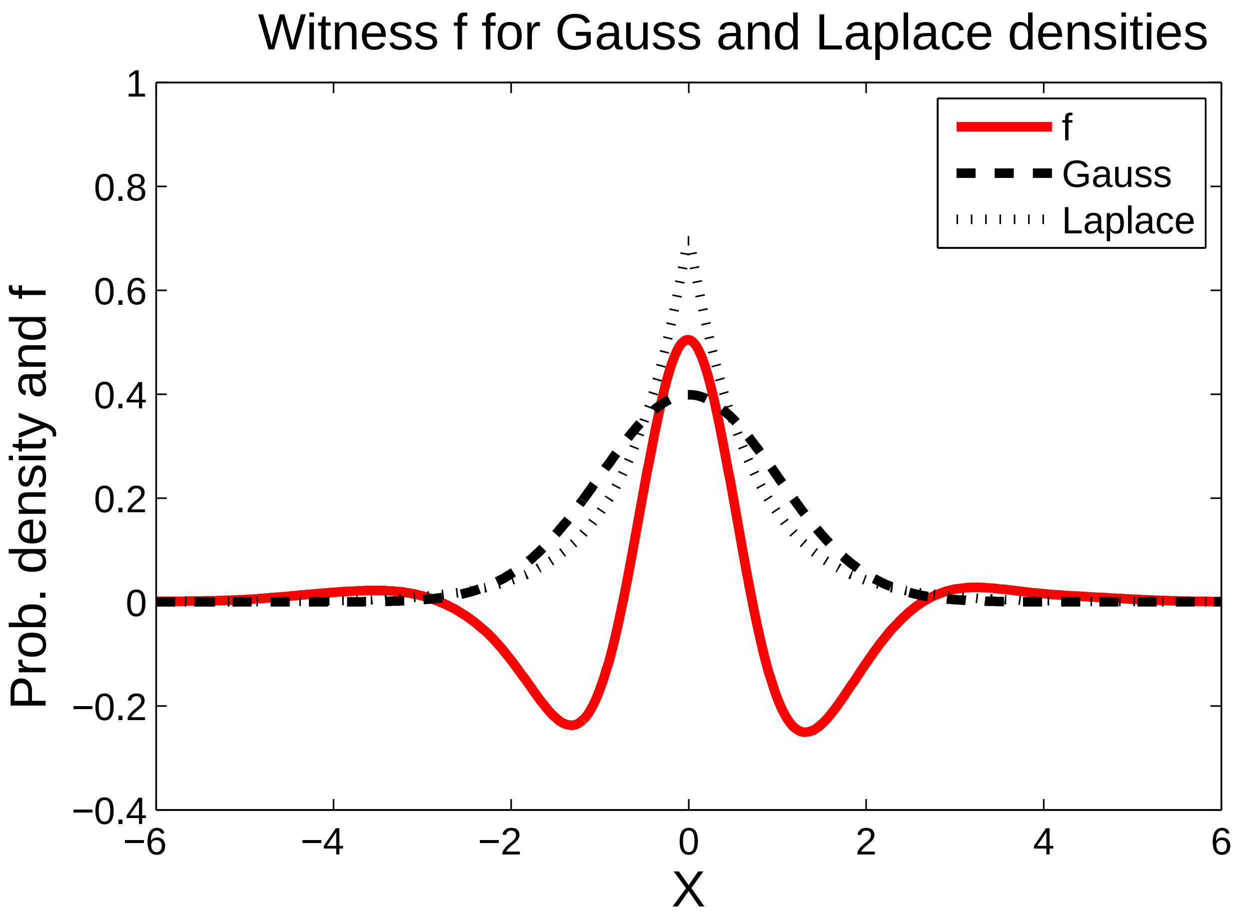
$$D(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[ \mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right].$$

- $D(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when $F =$bounded continuous functions [Dudley, 2002]

- $D(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F =$the unit ball in a universal RKHS $\mathcal{F}$ [via Steinwart, 2001]
  - Examples: Gaussian, Laplace [see also Fukumizu et al., 2004]

# Function revealing difference in distributions (2)

- Gauss vs Laplace revisited



# Function revealing difference in distributions (3)



- The (kernel) MMD:

$$\mathrm{MMD}(\mathsf{P}, \mathsf{Q}; F)$$

$$= \left( \sup_{f \in F} \left[ \mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y}) \right] \right)^2$$

# Function revealing difference in distributions (3)

- The (kernel) MMD:

$$\mathrm{MMD}(\mathsf{P}, \mathsf{Q}; F)$$

$$= \left( \sup_{f \in F} \left[ \mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y}) \right] \right)^2$$

using

$$\begin{aligned} \mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) &= \mathbf{E}_{\mathsf{P}} \left[ \langle \phi(\mathsf{x}), f \rangle_{\mathcal{F}} \right] \\ &=: \langle \mu_x, f \rangle_{\mathcal{F}} \end{aligned}$$

# Function revealing difference in distributions (3)

MAX-PLANCK-GESELLSCHAFT    BIOLOGISCHE KYBERNETIK

- The (kernel) MMD:

$$
\begin{aligned}
&\mathrm{MMD}(\mathsf{P}, \mathsf{Q}; F) \\
&= \left( \sup_{f \in F} \left[ \mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y}) \right] \right)^2 \\
&= \left( \sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2
\end{aligned}
$$

using

$$
\begin{aligned}
\mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) &= \mathbf{E}_{\mathsf{P}} \left[ \langle \phi(\mathsf{x}), f \rangle_{\mathcal{F}} \right] \\
&=: \langle \mu_x, f \rangle_{\mathcal{F}}
\end{aligned}
$$

# Function revealing difference in distributions (3)



- The (kernel) MMD:

$$\mathrm{MMD}(\mathsf{P}, \mathsf{Q}; F)$$

$$= \left( \sup_{f \in F} \left[ \mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y}) \right] \right)^2$$

$$= \left( \sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2$$

$$= \|\mu_x - \mu_y\|_{\mathcal{F}}^2$$

using

$$\|\mu\|_{\mathcal{F}} = \sup_{f \in F} \langle f, \mu \rangle_{\mathcal{F}}$$

# Function revealing difference in distributions (3)

- The (kernel) MMD:

$$
\begin{aligned}
&\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) \\
&= \left( \sup_{f \in F} \left[ \mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right] \right)^2 \\
&= \left( \sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2 \\
&= \| \mu_x - \mu_y \|_{\mathcal{F}}^2 \\
&= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}} \\
&= \mathbf{E}_{\mathbf{P},\mathbf{P}'} k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_{\mathbf{Q},\mathbf{Q}'} k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})
\end{aligned}
$$

- $\mathsf{x}'$ is a R.V. independent of $\mathsf{x}$ with distribution $\mathbf{P}$
- $\mathsf{y}'$ is a R.V. independent of $\mathsf{y}$ with distribution $\mathbf{Q}$.

# Function revealing difference in distributions (3)



- The (kernel) MMD:

$$
\begin{aligned}
&\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) \\
&= \left( \sup_{f \in F} \left[ \mathbf{E}_{\mathbf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} f(\mathsf{y}) \right] \right)^2 \\
&= \left( \sup_{f \in F} \langle f, \mu_x - \mu_y \rangle_{\mathcal{F}} \right)^2 \\
&= \|\mu_x - \mu_y\|^2_{\mathcal{F}} \\
&= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}} \\
&= \mathbf{E}_{\mathbf{P},\mathbf{P}'} k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_{\mathbf{Q},\mathbf{Q}'} k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})
\end{aligned}
$$

- $\mathsf{x}'$ is a R.V. independent of $\mathsf{x}$ with distribution $\mathbf{P}$
- $\mathsf{y}'$ is a R.V. independent of $\mathsf{y}$ with distribution $\mathbf{Q}$.

- Kernel between measures [Hein and Bousquet, 2005]

$$\mathfrak{K}(\mathbf{P}, \mathbf{Q}) = \mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})$$

# Statistical test using MMD (1)



- Two hypotheses:
  - $H_0$: null hypothesis ($\mathbf{P} = \mathbf{Q}$)
  - $H_1$: alternative hypothesis ($\mathbf{P} \neq \mathbf{Q}$)

# Statistical test using MMD (1)



- Two hypotheses:
  - $H_0$: null hypothesis ($\mathbf{P} = \mathbf{Q}$)
  - $H_1$: alternative hypothesis ($\mathbf{P} \neq \mathbf{Q}$)
- Observe samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ from $\mathbf{P}$ and $\boldsymbol{y}$ from $\mathbf{Q}$
- If empirical $\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F)$ is
  - "far from zero": reject $H_0$
  - "close to zero": accept $H_0$

# Statistical test using MMD (1)



- Two hypotheses:
  - $H_0$: null hypothesis ($\mathbf{P} = \mathbf{Q}$)
  - $H_1$: alternative hypothesis ($\mathbf{P} \neq \mathbf{Q}$)
- Observe samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ from $\mathbf{P}$ and $\boldsymbol{y}$ from $\mathbf{Q}$
- If empirical $\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F)$ is
  - "far from zero": reject $H_0$
  - "close to zero": accept $H_0$
- How good is a test?
  - Type I error: We reject $H_0$ although it is true
  - Type II error: We accept $H_0$ although it is false

# Statistical test using MMD (1)



- Two hypotheses:
  - $H_0$: null hypothesis ($\mathbf{P} = \mathbf{Q}$)
  - $H_1$: alternative hypothesis ($\mathbf{P} \neq \mathbf{Q}$)
- Observe samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ from $\mathbf{P}$ and $\boldsymbol{y}$ from $\mathbf{Q}$
- If empirical $\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F)$ is
  - "far from zero": reject $H_0$
  - "close to zero": accept $H_0$
- How good is a test?
  - Type I error: We reject $H_0$ although it is true
  - Type II error: We accept $H_0$ although it is false
- Good test has a low type II error for user-defined Type I error

# Statistical test using MMD (2)



- "far from zero" vs "close to zero" - threshold?

# Statistical test using MMD (2)

- “far from zero” vs “close to zero” - threshold?
- One answer: asymptotic distribution of $\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F)$

# Statistical test using MMD (2)



- "far from zero" vs "close to zero" - threshold?
- One answer: asymptotic distribution of $\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F)$
- An unbiased empirical estimate (quadratic cost):

$$\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F) = \frac{1}{m(m-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{h((x_i, y_i), (x_j, y_j))}$$

# Statistical test using MMD (2)



- "far from zero" vs "close to zero" - threshold?
- One answer: asymptotic distribution of $\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F)$
- An unbiased empirical estimate (quadratic cost):

$$\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F) = \frac{1}{m(m-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{h((x_i, y_i), (x_j, y_j))}$$

- When $\mathbf{P} \neq \mathbf{Q}$, asymptotically normal [Hoeffding, 1948, Serfling, 1980]

# Statistical test using MMD (2)

- "far from zero" vs "close to zero" - threshold?
- One answer: asymptotic distribution of $\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F)$
- An unbiased empirical estimate (quadratic cost):

$$\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F) = \frac{1}{m(m-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{h((x_i, y_i),(x_j, y_j))}$$

- When $\mathbf{P} \neq \mathbf{Q}$, asymptotically normal [Hoeffding, 1948, Serfling, 1980]

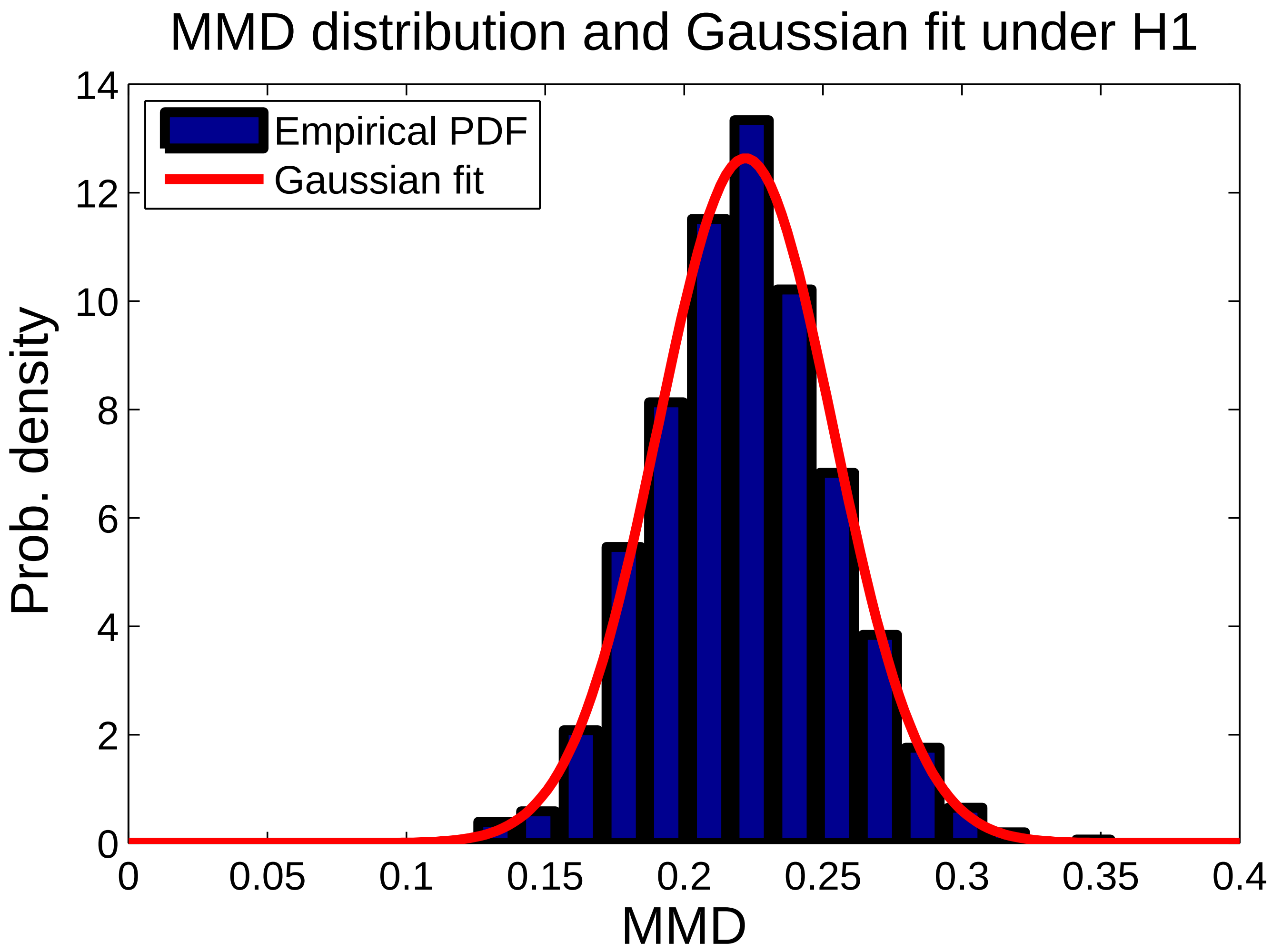
- Expression for the variance: $z_i := (x_i, y_i)$

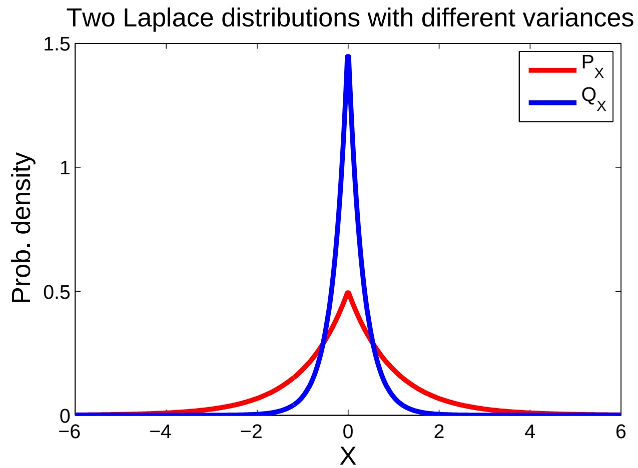
$$\sigma_u^2 = \frac{2^2}{m} \left( \mathbf{E}_{\mathsf{z}} \left[ (\mathbf{E}_{\mathsf{z}'} h(\mathsf{z}, \mathsf{z}'))^2 \right] - \left[ \mathbf{E}_{\mathsf{z}, \mathsf{z}'} (h(\mathsf{z}, \mathsf{z}')) \right]^2 \right) + O(m^{-2})$$

# Statistical test using MMD (3)



- Example: laplace distributions with different variance



# Statistical test using MMD (4)



- When $\mathsf{P} = \mathsf{Q}$, U-statistic degenerate: $\mathbf{E}_{\mathsf{z}'}[h(\mathsf{z}, \mathsf{z}')] = 0$ [Anderson et al., 1994]
- Distribution is

$$m\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F) \sim \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2\right]$$

- where
  - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
  - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathsf{P}_x(x) = \lambda_i \psi_i(x')$

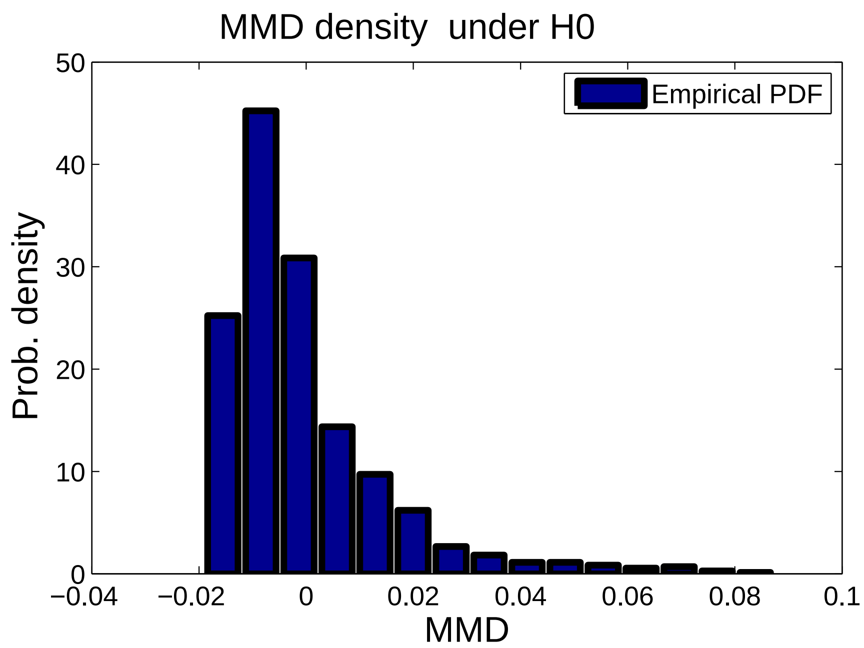
# Statistical test using MMD (4)



- When $\mathbf{P} = \mathbf{Q}$, U-statistic degenerate: $\mathbf{E}_{z'}[h(z, z')] = 0$ [Anderson et al., 1994]
- Distribution is

$$m\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F) \sim \sum_{l=1}^{\infty} \lambda_l \left[ z_l^2 - 2 \right]$$

- where
  - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
  - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathbf{P}_x(x) = \lambda_i \psi_i(x')$



# Statistical test using MMD (5)

- Given $\mathbf{P} = \mathbf{Q}$, want threshold $T$ such that $\mathbf{P}(\mathrm{MMD} > T) \leq 0.05$

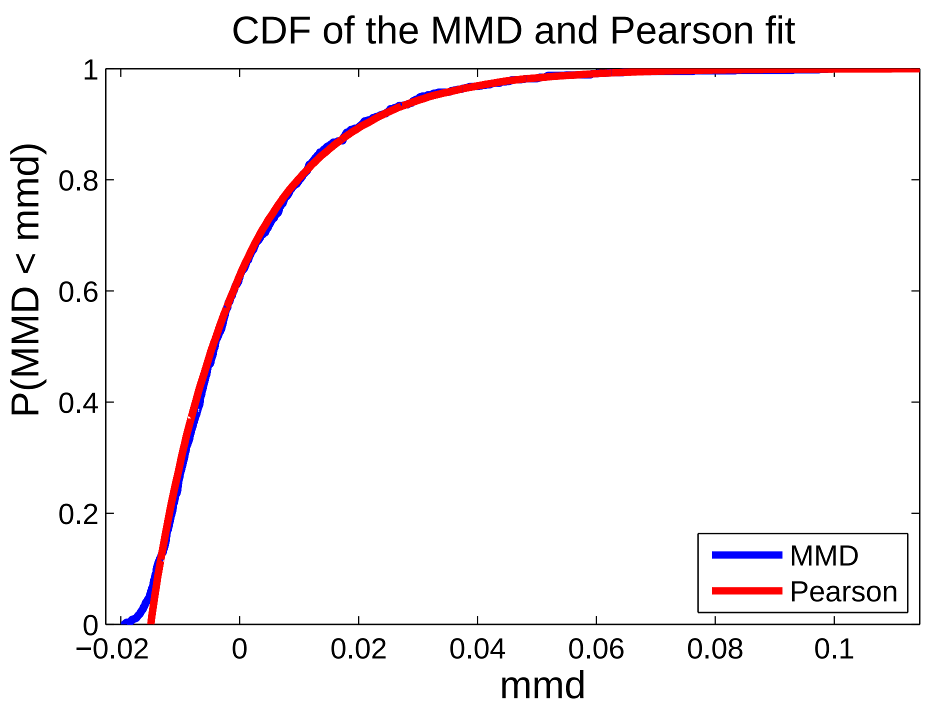
# Statistical test using MMD (5)



- Given $\mathbf{P} = \mathbf{Q}$, want threshold $T$ such that $\mathbf{P}(\mathrm{MMD} > T) \leq 0.05$
- Bootstrap for empirical CDF [Arcones and Giné, 1992]
- Pearson curves by matching first four moments [Johnson et al., 1994]
- Large deviation bounds [Hoeffding, 1963, McDiarmid, 1969]
- Other...

# Statistical test using MMD (5)

- Given $\mathbf{P} = \mathbf{Q}$, want threshold $T$ such that $\mathbf{P}(\mathrm{MMD} > T) \leq 0.05$
- Bootstrap for empirical CDF [Arcones and Giné, 1992]
- Pearson curves by matching first four moments [Johnson et al., 1994]
- Large deviation bounds [Hoeffding, 1963, McDiarmid, 1969]
- Other...



# Experiments

- Small sample size: Pearson more accurate than bootstrap
- Large sample size: bootstrap faster

# Experiments



- Small sample size: Pearson more accurate than bootstrap
- Large sample size: bootstrap faster
- Cancer subtype ($m = 25$, 2118 dimensions):
  - For Pearson, Type I 3.5%, Type II 0%
  - For bootstrap, Type I 0.9%, Type II 0%

# Experiments

- Small sample size: Pearson more accurate than bootstrap
- Large sample size: bootstrap faster
- Cancer subtype ($m = 25$, 2118 dimensions):
  - For Pearson, Type I 3.5%, Type II 0%
  - For bootstrap, Type I 0.9%, Type II 0%
- Neural spikes ($m = 1000$, 100 dimensions):
  - For Pearson, Type I 4.8%, Type II 3.4%
  - For bootstrap, Type I 5.4%, Type II 3.3%

# Experiments

- Small sample size: Pearson more accurate than bootstrap
- Large sample size: bootstrap faster
- Cancer subtype ($m = 25$, 2118 dimensions):
  - For Pearson, Type I 3.5%, Type II 0%
  - For bootstrap, Type I 0.9%, Type II 0%
- Neural spikes ($m = 1000$, 100 dimensions):
  - For Pearson, Type I 4.8%, Type II 3.4%
  - For bootstrap, Type I 5.4%, Type II 3.3%
- Further experiments: comparison with t-test, Friedman-Rafsky tests [Friedman and Rafsky, 1979], Biau-Györfi test [Biau and Gyorfi, 2005], and Hall-Tajvidi test [Hall and Tajvidi, 2002].

# Conclusions (two-sample problem)

- The MMD: distance between means in feature spaces
- When feature spaces universal RKHSs, MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$
- Statistical test of whether $\mathbf{P} \neq \mathbf{Q}$ using asymptotic distribution:
  - Pearson approximation for low sample size
  - Bootstrap for large sample size
- Useful in high dimensions and for structured data

# Dependence Detection with Kernels

# Kernel dependence measures



- Independence testing
  - Given: $m$ samples $\boldsymbol{z} := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ from $\mathbf{P}$
  - Determine: Does $\mathbf{P} = \mathbf{P}_{\mathsf{x}}\mathbf{P}_{\mathsf{y}}$?

# Kernel dependence measures



- Independence testing
  - Given: $m$ samples $\boldsymbol{z} := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ from $\mathbf{P}$
  - Determine: Does $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$?
- Kernel dependence measures
  - Zero only at independence
  - Take into account high order moments
  - Make "sensible" assumptions about smoothness

# Kernel dependence measures



- Independence testing
  - Given: $m$ samples $\boldsymbol{z} := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ from $\mathbf{P}$
  - Determine: Does $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$?
- Kernel dependence measures
  - Zero only at independence
  - Take into account high order moments
  - Make "sensible" assumptions about smoothness
- Covariance operators in spaces of features
  - Spectral norm (COCO) [Gretton et al., 2005c,d]
  - Hilbert-Schmidt norm (HSIC) [Gretton et al., 2005a]

# Function revealing dependence (1)

- Idea: avoid density estimation when testing $\mathsf{P} = \mathsf{P}_x \mathsf{P}_y$ [Rényi, 1959]

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

# Function revealing dependence (1)

- Idea: avoid density estimation when testing $\mathsf{P} = \mathsf{P}_x \mathsf{P}_y$ [Rényi, 1959]

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

- $\mathrm{COCO}(\mathsf{P}; F, G) = 0$ iff $\mathsf{x}, \mathsf{y}$ independent, when $F$ and $G$ are respective unit balls in universal RKHSs $\mathcal{F}$ and $\mathcal{G}$ [via Steinwart, 2001]
  - Examples: Gaussian, Laplace [see also Bach and Jordan, 2002]

# Function revealing dependence (1)



- Idea: avoid density estimation when testing $\mathsf{P} = \mathsf{P}_x \mathsf{P}_y$ [Rényi, 1959]

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

- $\mathrm{COCO}(\mathsf{P}; F, G) = 0$ iff $\mathsf{x}, \mathsf{y}$ independent, when $F$ and $G$ are respective unit balls in universal RKHSs $\mathcal{F}$ and $\mathcal{G}$ [via Steinwart, 2001]
  - Examples: Gaussian, Laplace [see also Bach and Jordan, 2002]

## In geometric terms:

- Covariance operator: $C_{xy} : \mathcal{G} \rightarrow \mathcal{F}$ such that

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

# Function revealing dependence (1)



- Idea: avoid density estimation when testing $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$ [Rényi, 1959]

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

- $\mathrm{COCO}(\mathbf{P}; F, G) = 0$ iff $\mathsf{x}, \mathsf{y}$ independent, when $F$ and $G$ are respective unit balls in universal RKHSs $\mathcal{F}$ and $\mathcal{G}$ [via Steinwart, 2001]
  - Examples: Gaussian, Laplace [see also Bach and Jordan, 2002]

## In geometric terms:

- Covariance operator: $C_{xy} : \mathcal{G} \to \mathcal{F}$ such that

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

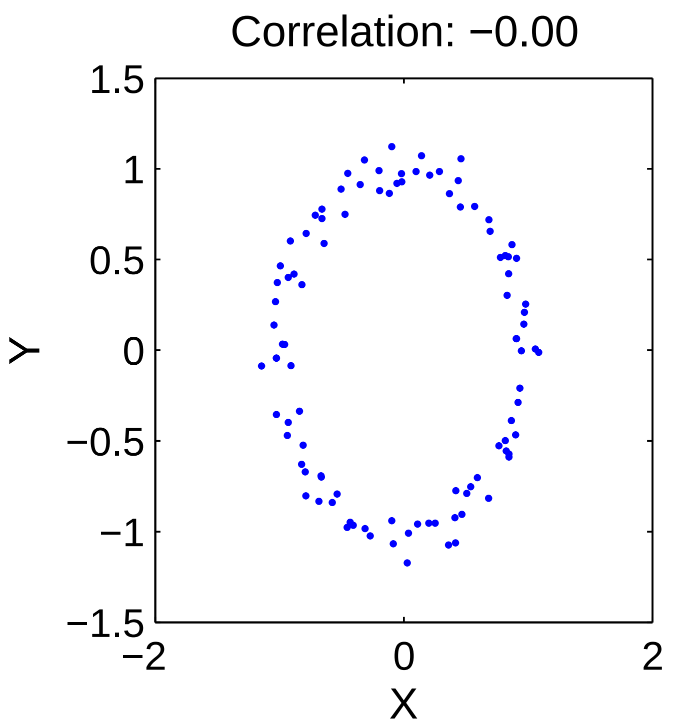
- COCO is the spectral norm of $C_{xy}$ [Gretton et al., 2005c,d]:

$$\mathrm{COCO}(\mathbf{P}; F, G) := \|C_{xy}\|_{\mathrm{S}}$$

# Function revealing dependence (2)

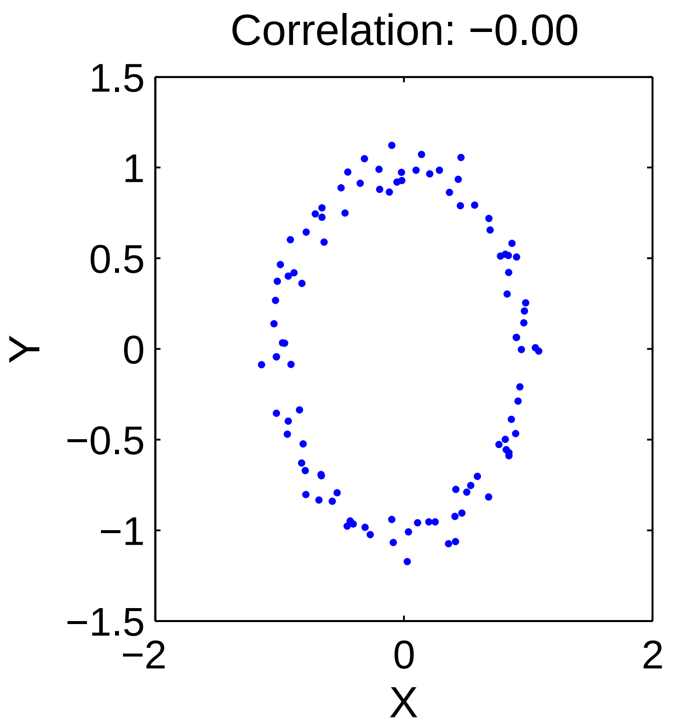


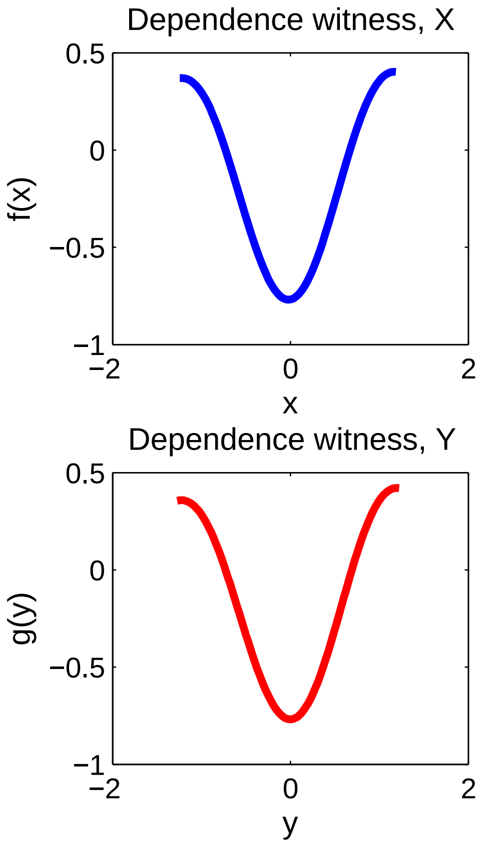
- Ring-shaped density, correlation approx. zero [example from Fukumizu, Bach, and Gretton, 2005]



# Function revealing dependence (2)

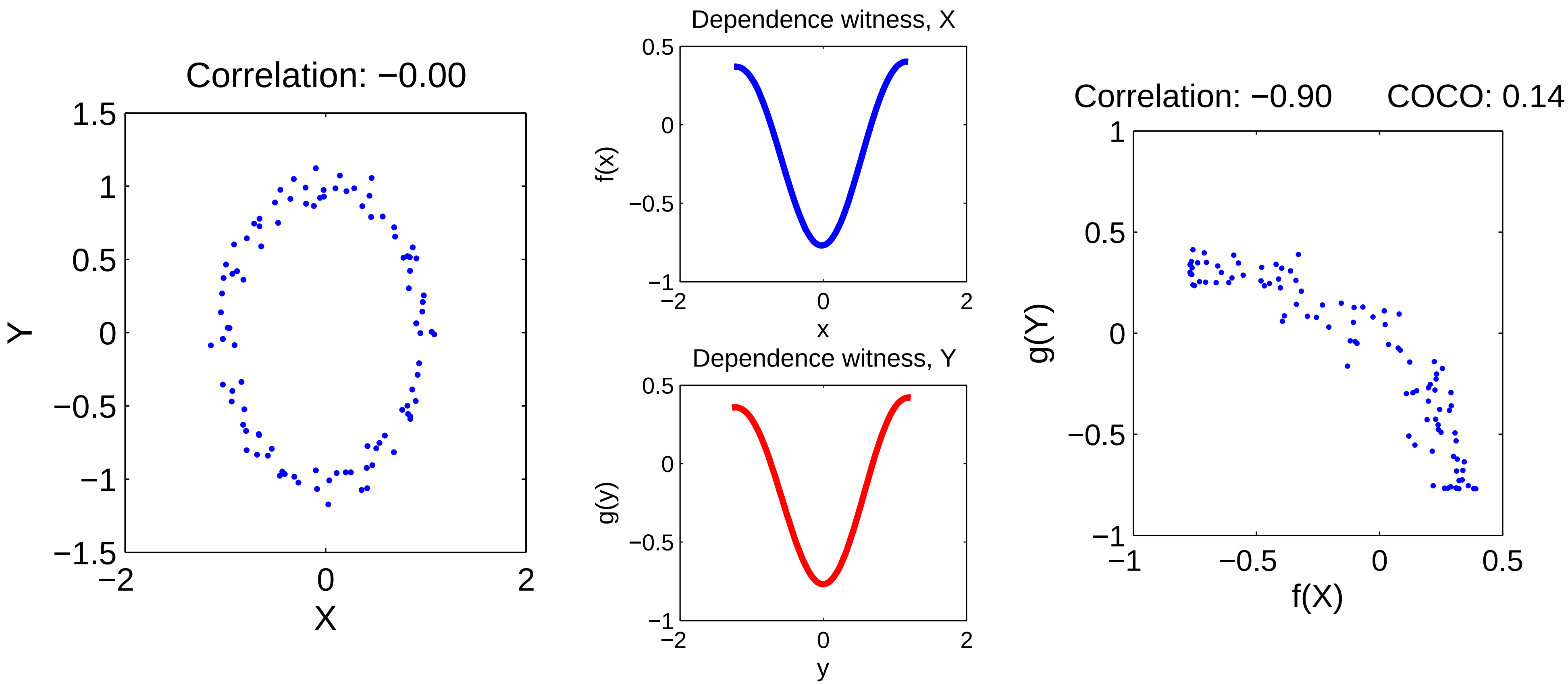
- Ring-shaped density, correlation approx. zero [example from Fukumizu, Bach, and Gretton, 2005]



# Function revealing dependence (2)

- Ring-shaped density, correlation approx. zero [example from Fukumizu, Bach, and Gretton, 2005]



# Function revealing dependence (3)



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

$$\begin{bmatrix} \mathbf{0} & \frac{1}{m}\widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \frac{1}{m}\widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{\mathbf{K}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{L}} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}.$$

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

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

and $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{F}}, \quad l(y_i, y_j) = \langle \psi(y_i), \psi(y_j) \rangle_{\mathcal{G}}$

# Function revealing dependence (3)





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

$$\begin{bmatrix} \mathbf{0} & \frac{1}{m}\widetilde{\mathbf{K}}\widetilde{\mathbf{L}} \\ \frac{1}{m}\widetilde{\mathbf{L}}\widetilde{\mathbf{K}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{\mathbf{K}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{L}} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}.$$

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

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

and $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{F}}, \quad l(y_i, y_j) = \langle \psi(y_i), \psi(y_j) \rangle_{\mathcal{G}}$

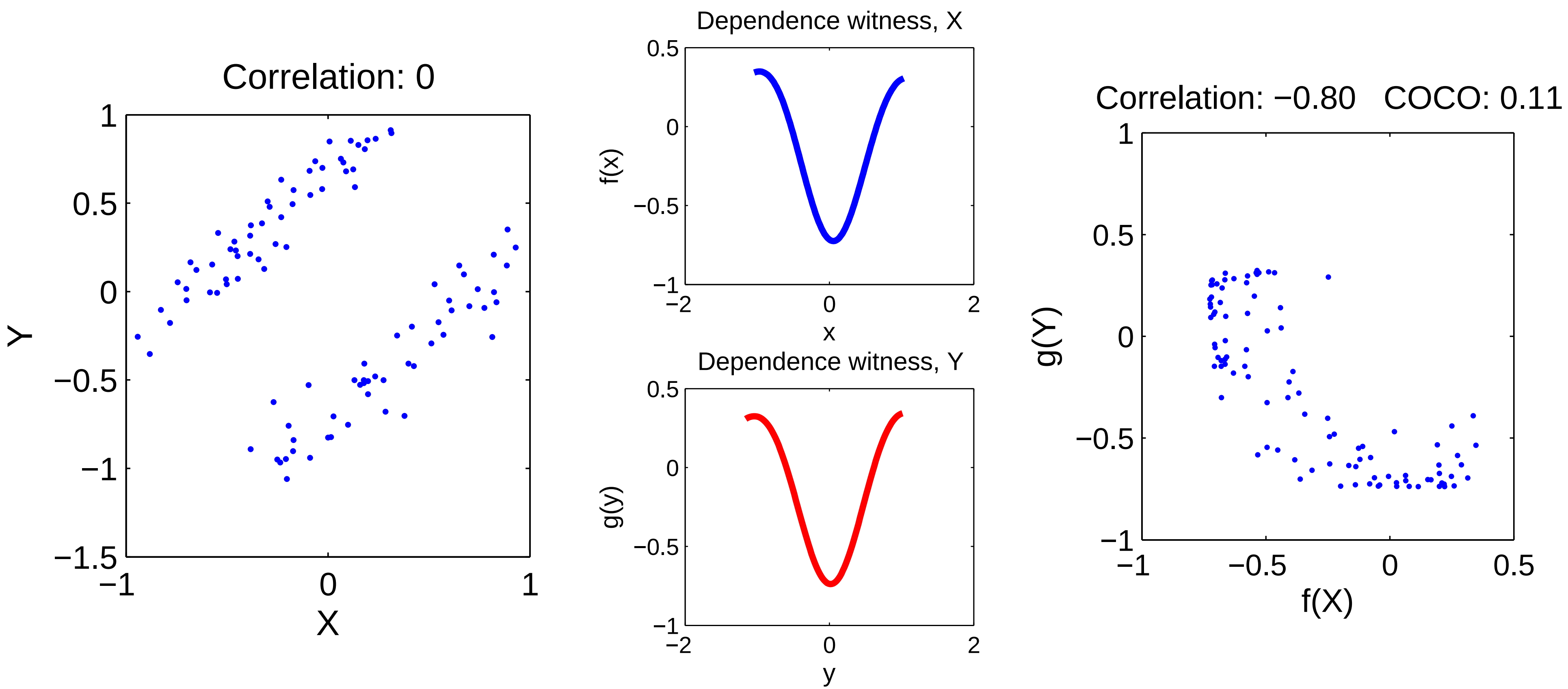
- Witness function for $x$:

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

# Function revealing dependence (4)



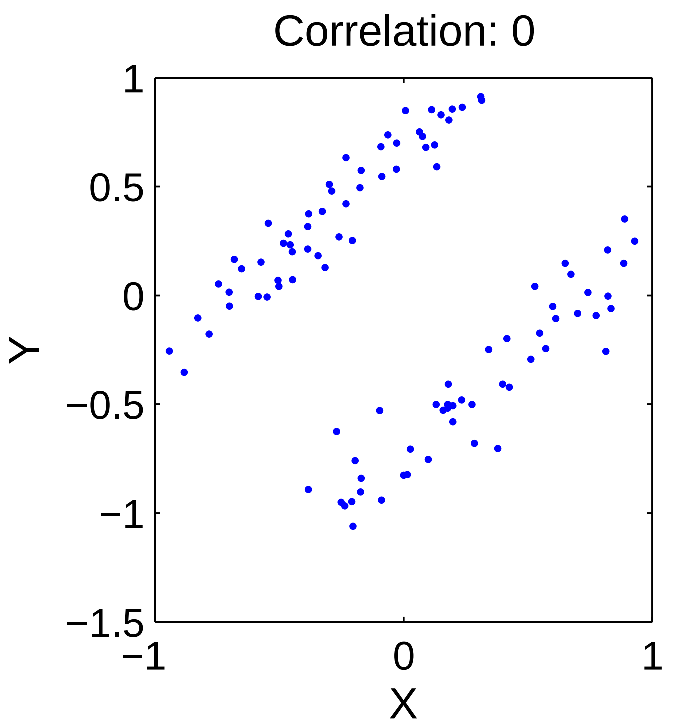
- Can we do better?
- A second example with zero correlation

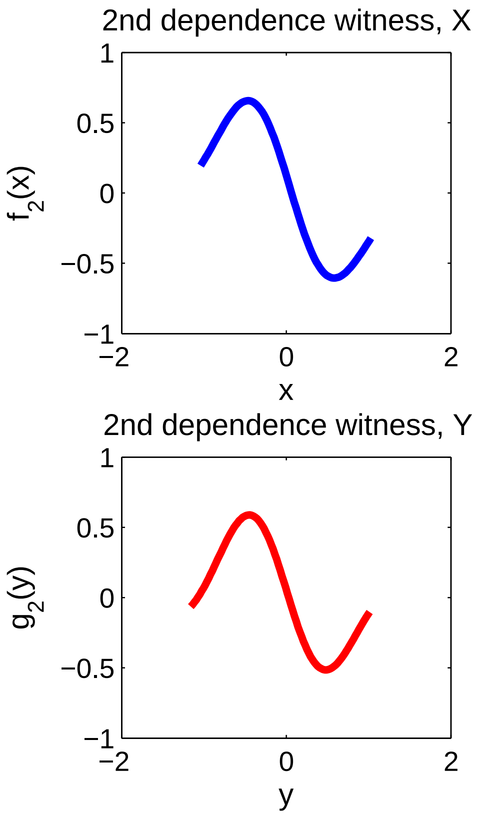


# Function revealing dependence (4)



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

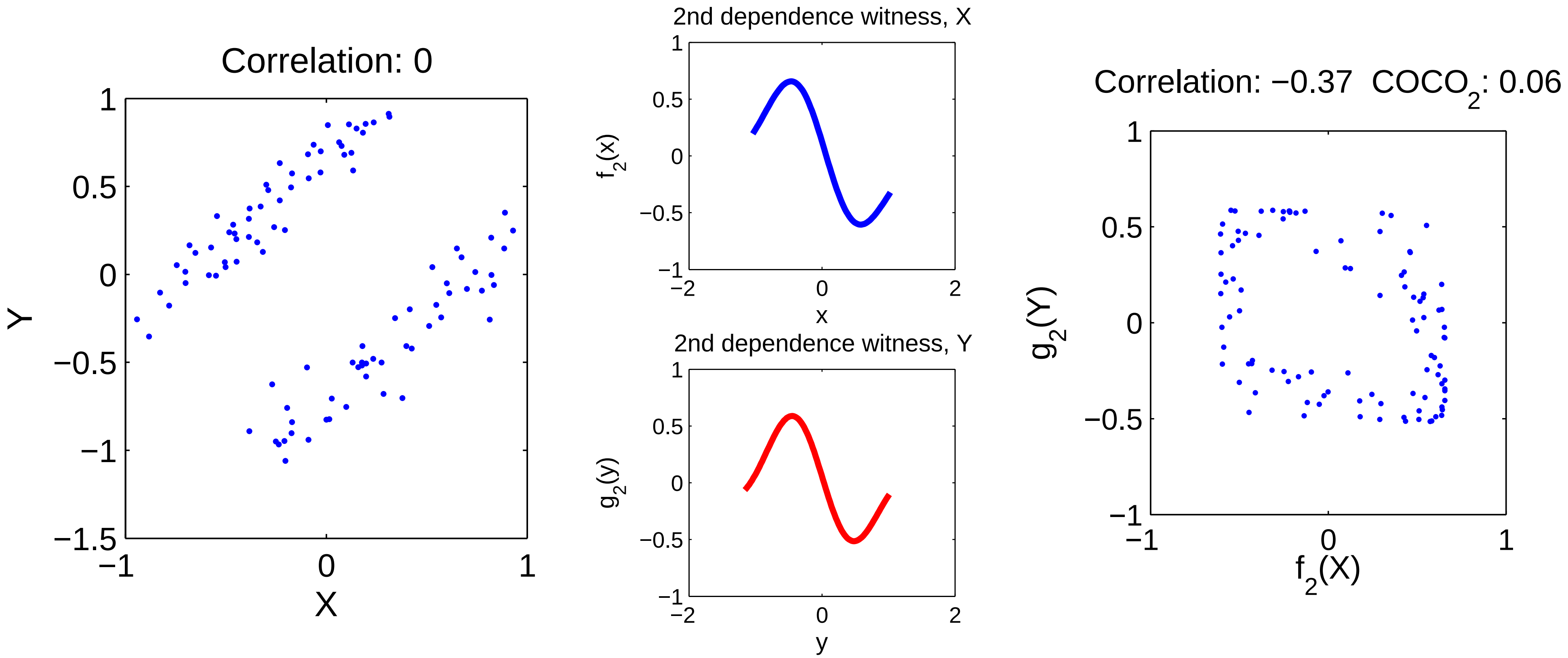


# Function revealing dependence (4)



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



# Hilbert-Schmidt Independence Criterion



- Given $\gamma_i := \mathrm{COCO}_i(\boldsymbol{z}; F, G)$, define Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005b]:

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

# Hilbert-Schmidt Independence Criterion



- Given $\gamma_i := \mathrm{COCO}_i(\boldsymbol{z}; F, G)$, define Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005b]:

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

- In limit of infinite samples:

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

- $\mathsf{x}'$ an independent copy of $\mathsf{x}$, $\mathsf{y}'$ a copy of $\mathsf{y}$

# Link between HSIC and MMD (1)



- Define the product space $\mathcal{F} \times \mathcal{G}$ with kernel

$$\left\langle \Phi(x,y), \Phi(x',y') \right\rangle = \mathfrak{K}((x,y),(x',y')) = k(x,x')l(y,y')$$

# Link between HSIC and MMD (1)



- Define the product space $\mathcal{F} \times \mathcal{G}$ with kernel

$$\left\langle \Phi(x,y), \Phi(x',y') \right\rangle = \mathfrak{K}((x,y),(x',y')) = k(x,x')l(y,y')$$

- Define the mean elements

$$\langle \mu_{xy}, \Phi(x,y) \rangle := \mathbf{E}_{x',y'} \left\langle \Phi(x',y'), \Phi(x,y) \right\rangle = \mathbf{E}_{x',y'} k(x,x')l(y,y')$$

and

$$\langle \mu_{x \perp\!\!\!\perp y}, \Phi(x,y) \rangle := \mathbf{E}_{x',y''} \left\langle \Phi(x',y''), \Phi(x,y) \right\rangle = \mathbf{E}_{x'} k(x,x') \mathbf{E}_{y'} l(y,y')$$

# Link between HSIC and MMD (1)



- Define the product space $\mathcal{F} \times \mathcal{G}$ with kernel

$$\left\langle \Phi(x,y), \Phi(x',y') \right\rangle = \mathfrak{K}((x,y),(x',y')) = k(x,x')l(y,y')$$

- Define the mean elements

$$\langle \mu_{xy}, \Phi(x,y) \rangle := \mathbf{E}_{x',y'} \left\langle \Phi(x',y'), \Phi(x,y) \right\rangle = \mathbf{E}_{x',y'} k(x,x')l(y,y')$$

and

$$\langle \mu_{x \perp\!\!\!\perp y}, \Phi(x,y) \rangle := \mathbf{E}_{x',y''} \left\langle \Phi(x',y''), \Phi(x,y) \right\rangle = \mathbf{E}_{x'} k(x,x') \mathbf{E}_{y'} l(y,y')$$

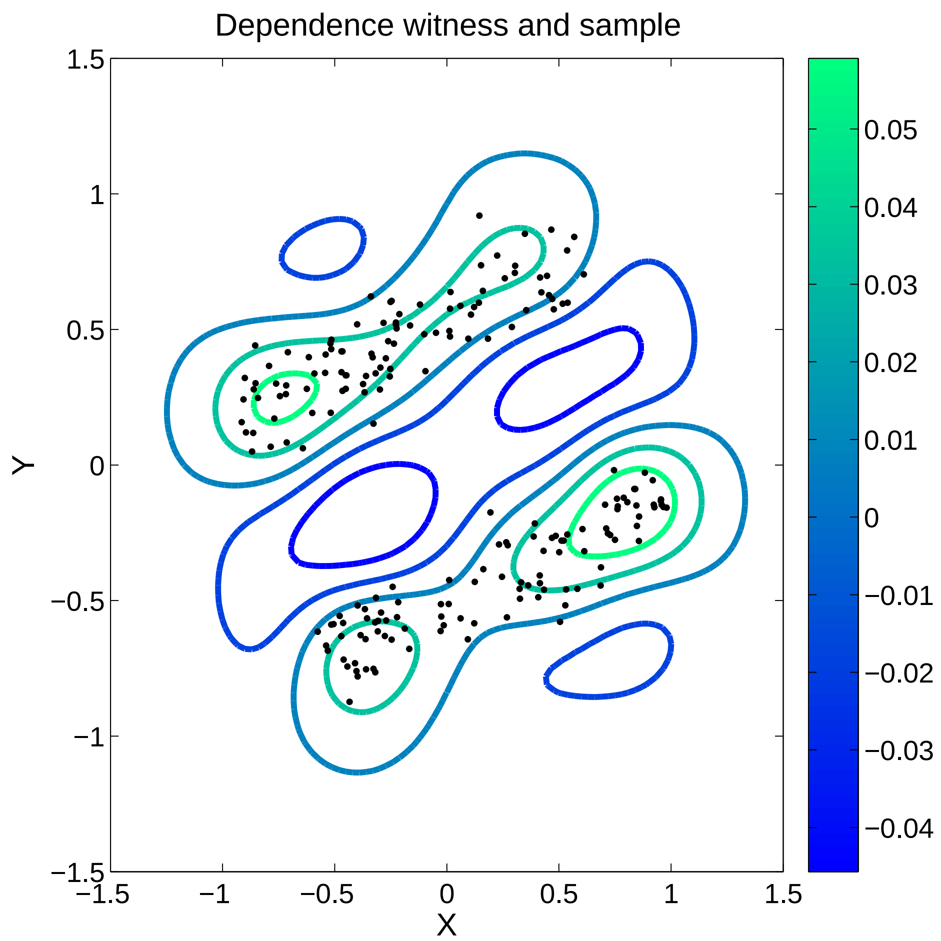
- The MMD between these two mean elements is

$$\begin{aligned} \mathrm{MMD}(\mathbf{P}, \mathbf{P}_x \mathbf{P}_y, F \times G) &= \| \mu_{xy} - \mu_{x \perp\!\!\!\perp y} \|^2_{\mathcal{F} \times \mathcal{G}} \\ &= \langle \mu_{xy} - \mu_{x \perp\!\!\!\perp y}, \mu_{xy} - \mu_{x \perp\!\!\!\perp y} \rangle \\ &= \mathrm{HSIC}(\mathbf{P}, F, G) \end{aligned}$$

# Link between HSIC and MMD (2)



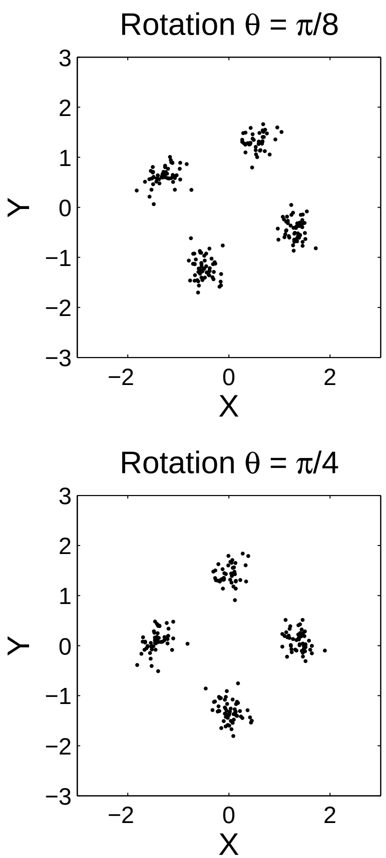
- Witness function for HSIC



# Independence test: verifying ICA and ISA

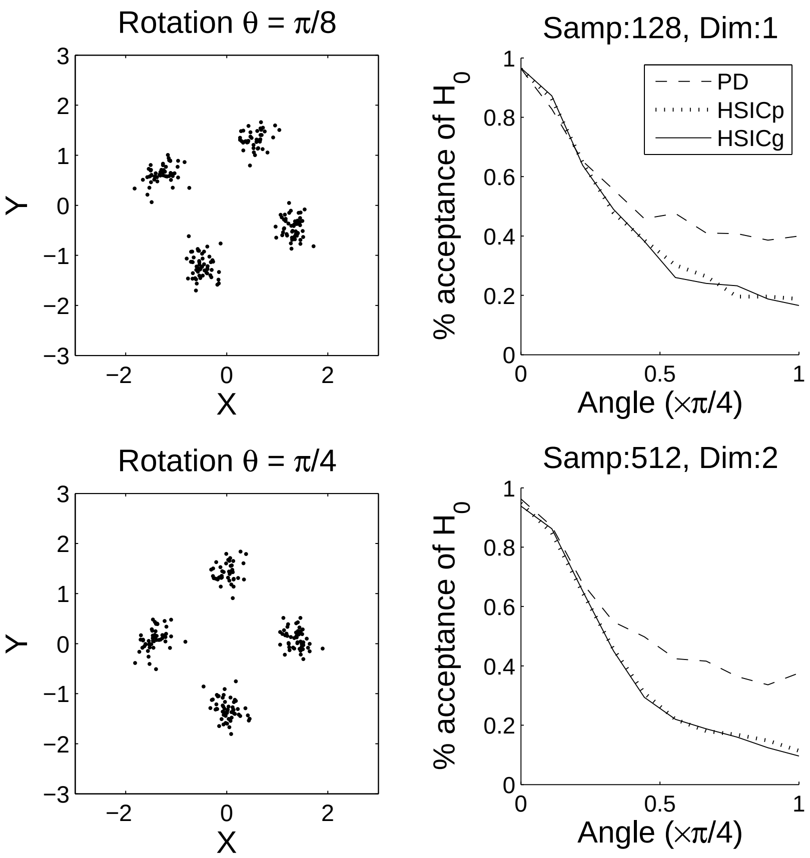


- HSICp: null distribution via sampling
- HSICg: null distribution via moment matching
- Compare with contingency table test (PD) [Read and Cressie, 1988]



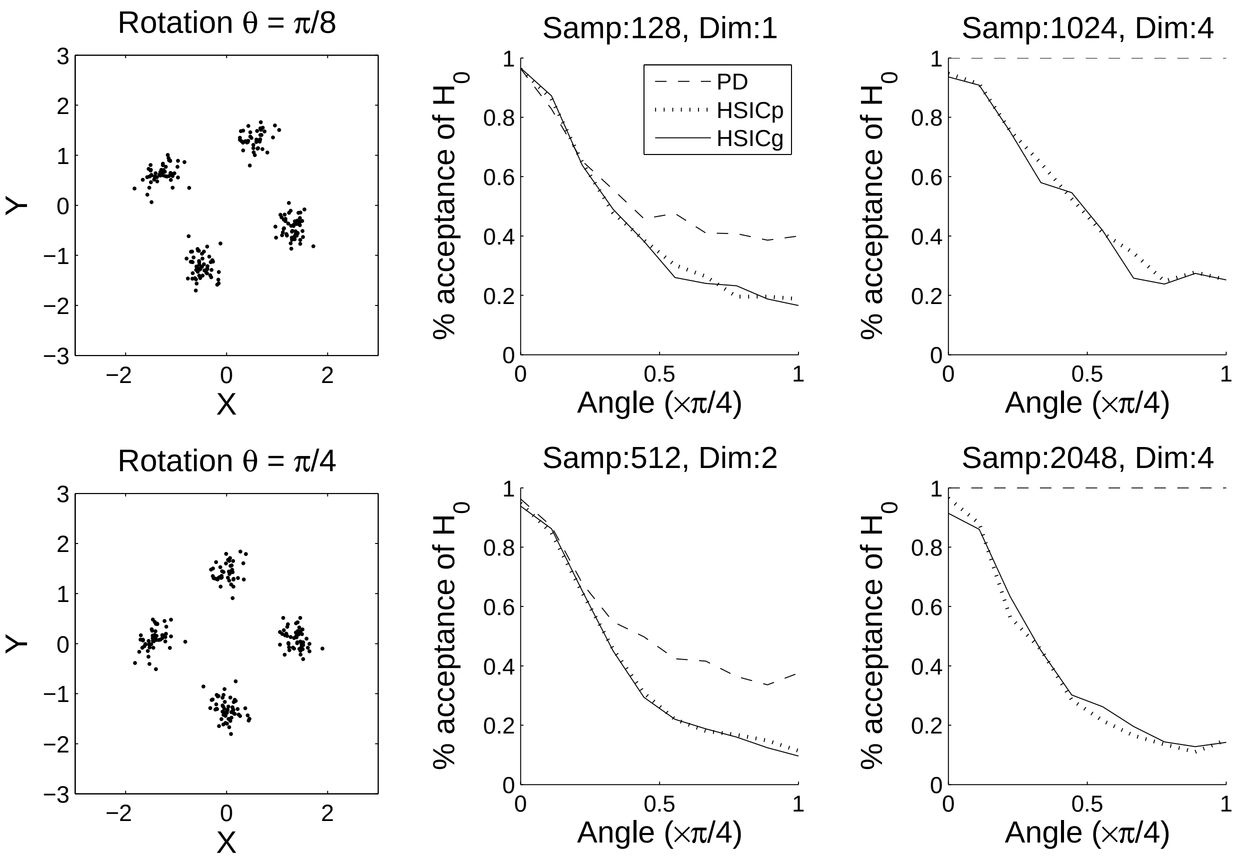
# Independence test: verifying ICA and ISA

- HSICp: null distribution via sampling
- HSICg: null distribution via moment matching
- Compare with contingency table test (PD) [Read and Cressie, 1988]



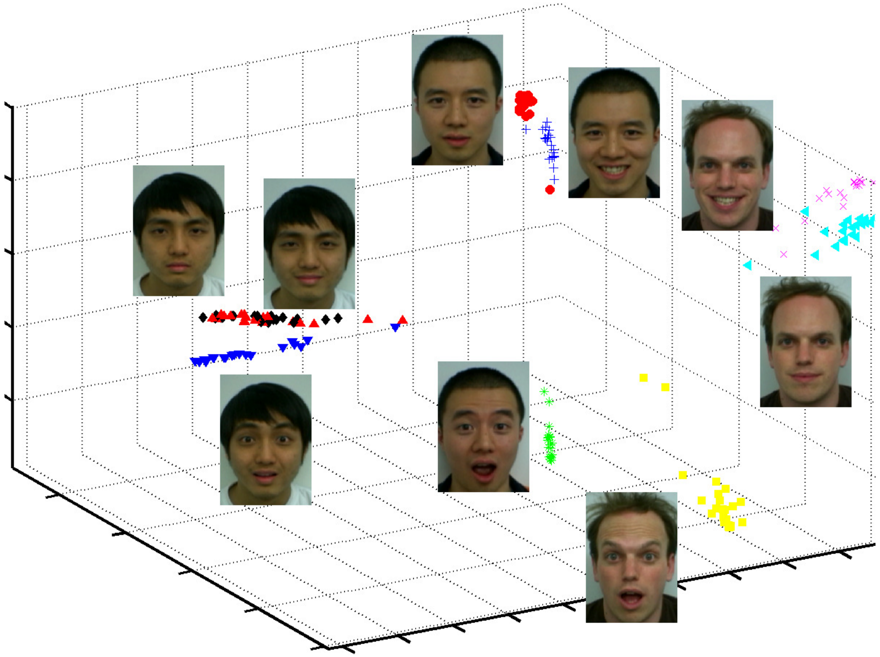
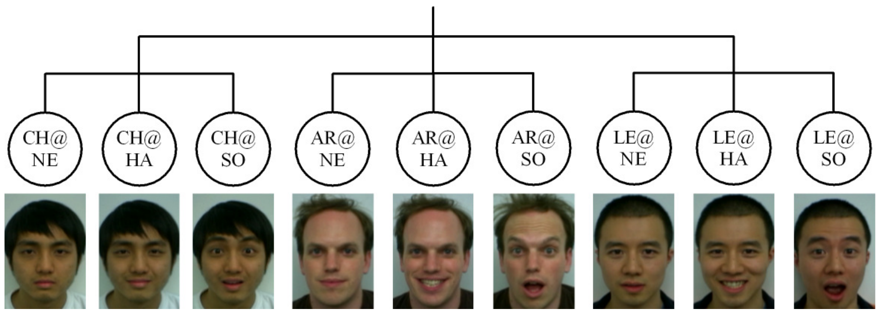
# Independence test: verifying ICA and ISA

- HSICp: null distribution via sampling
- HSICg: null distribution via moment matching
- Compare with contingency table test (PD) [Read and Cressie, 1988]



# Other applications of HSIC

- Feature selection [Song et al., 2007c,a]
- Clustering [Song et al., 2007b]

# Conclusions (dependence measures)

- COCO and HSIC: norms of covariance operator between feature spaces
- When feature spaces universal RKHSs, COCO $=$ HSIC $= 0$ iff $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$
- Statistical test possible using asymptotic distribution
- Independent component analysis
  - high accuracy
  - less sensitive to initialisation

Questions?

# References

N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50:41–54, 1994.

M. Arcones and E. Giné. On the bootstrap of $u$ and $v$ statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.

G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric $l_1$-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, UK, 2002.

R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, 70:266–285, 1953.

J. Friedman and L. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, 2004.

K. Fukumizu, F. Bach, and A. Gretton. Consistency of kernel canonical correlation analysis. Technical Report

# Hard-to-detect dependence (1)



- COCO can be $\approx 0$ for dependent RVs with highly non-smooth densities

# Hard-to-detect dependence (1)

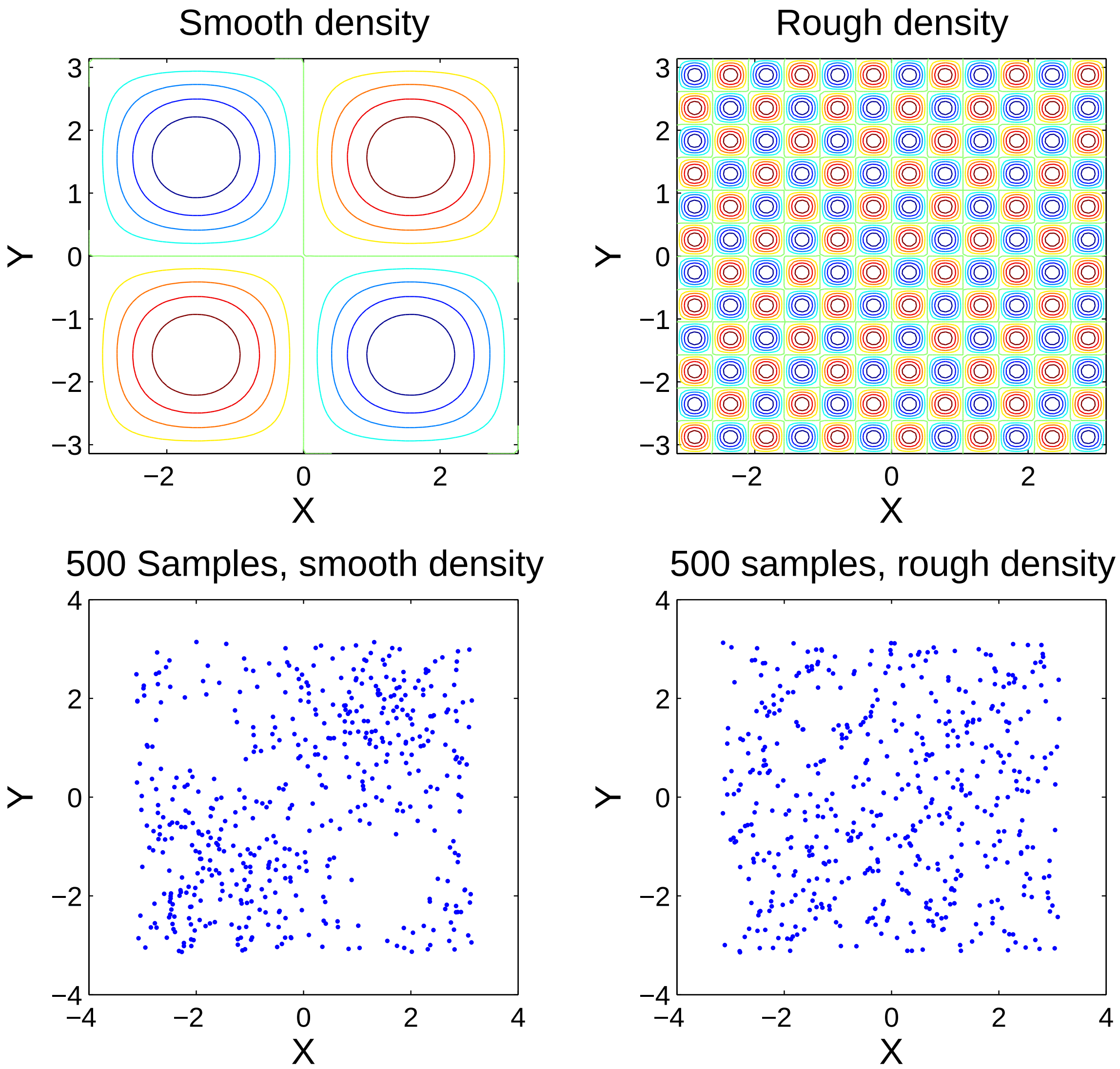
- COCO can be $\approx 0$ for dependent RVs with highly non-smooth densities
- Reason: norms in the denominator

$$\mathrm{COCO}(\mathsf{P}; F, G) := \sup_{f \in \mathcal{F},\, g \in \mathcal{G}} \frac{\mathrm{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)}{\|\mathbf{f}\|_{\mathcal{F}} \|\mathbf{g}\|_{\mathcal{G}}}$$

- RESULT: not detectable with finite sample size
- More formally: see Ingster [1989]

# Hard-to-detect dependence (2)



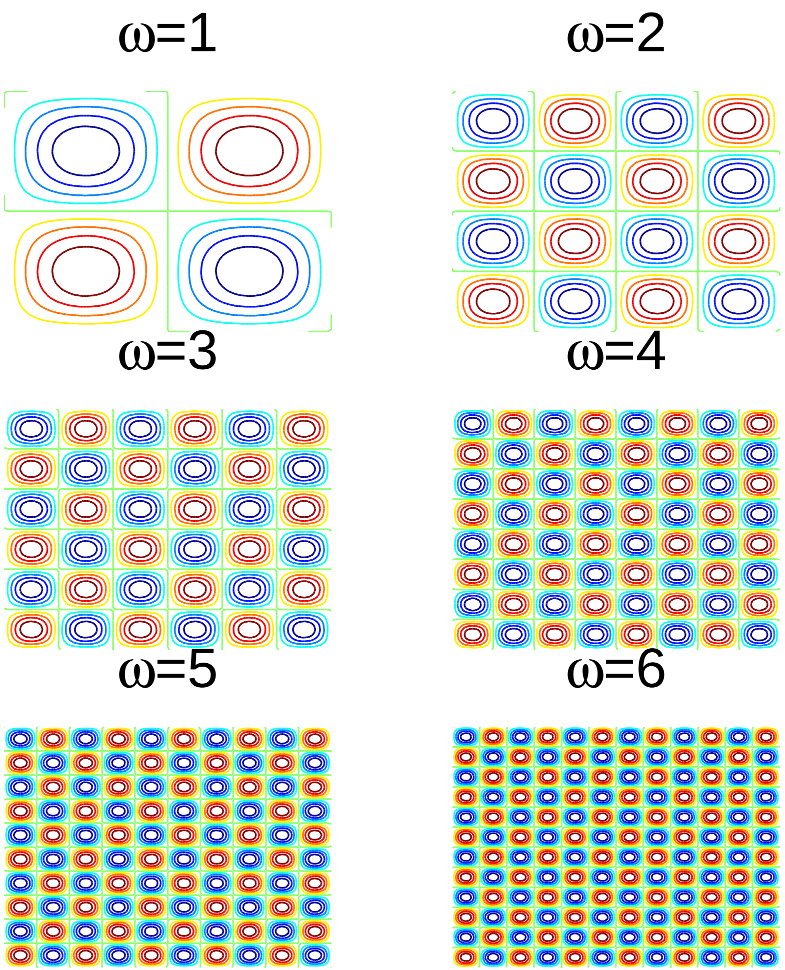


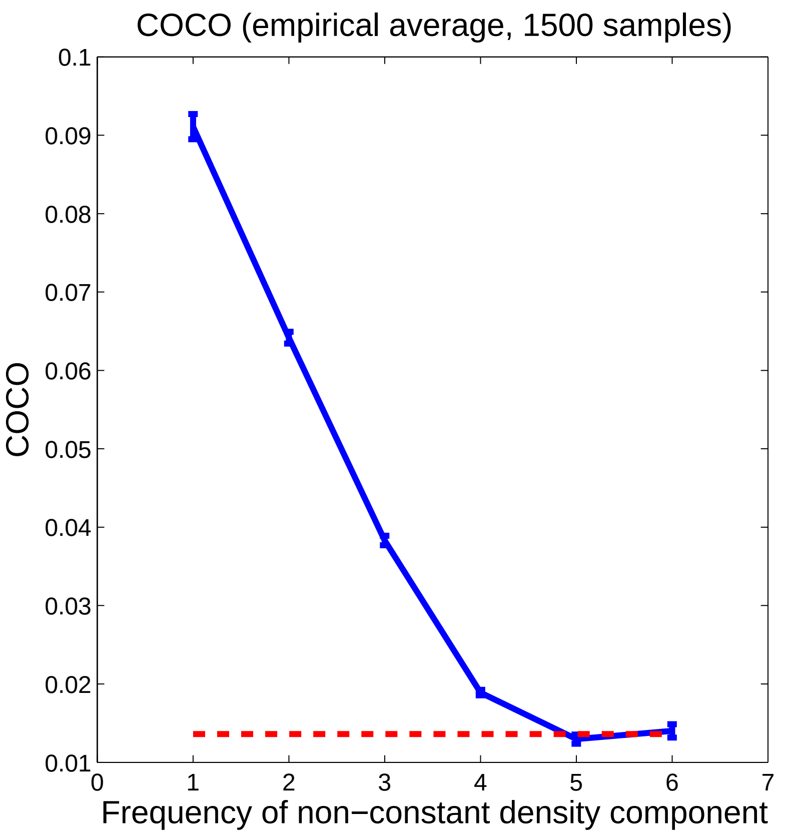

Density takes the form:

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

# Hard-to-detect dependence (3)

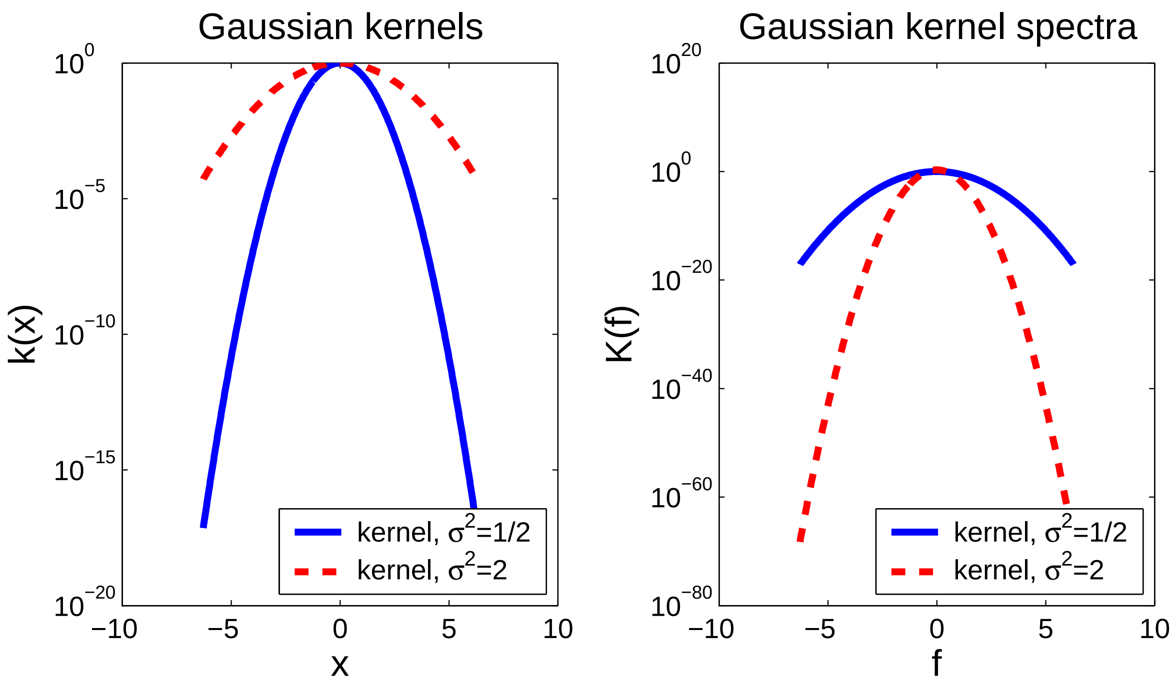
- Example: sinusoids of increasing frequency



# Choosing kernel size (1)

- The RKHS norm of $f$ is $\|f\|^2_{\mathcal{H}_\mathcal{X}} := \sum_{i=1}^{\infty} \tilde{f}_i^2 \left(\tilde{k}_i\right)^{-1}$.
- If kernel decays quickly, its spectrum decays slowly:
  - then non-smooth functions have smaller RKHS norm
- Example: spectrum of two Gaussian kernels



# Choosing kernel size (2)

- Could we just decrease kernel size?
- Yes, but only up to a point