

Painless embeddings of distributions: the function space view

Part 1 - introduction to embeddings

Kenji Fukumizu, Arthur Gretton, Alex Smola

MPI for Biological Cybernetics

Helsinki, July 5 2008



A Very Short Introduction to Kernels



- Hilbert space of functions \mathcal{F} from \mathcal{X} to \mathbb{R}
- **RKHS**: evaluation operator $\delta_x : \mathcal{F} \rightarrow \mathbb{R}, f \mapsto f(x)$ **bounded**



A Very Short Introduction to Kernels



MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

- Hilbert space of functions \mathcal{F} from \mathcal{X} to \mathbb{R}
- **RKHS**: evaluation operator $\delta_x : \mathcal{F} \rightarrow \mathbb{R}, f \mapsto f(x)$ **bounded**
- **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**
- $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ **kernel function**
- **Why useful**: Convergence in **norm** of \mathcal{F} implies **pointwise** convergence



A Very Short Introduction to Kernels



MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

- Hilbert space of functions \mathcal{F} from \mathcal{X} to \mathbb{R}
- **RKHS**: evaluation operator $\delta_x : \mathcal{F} \rightarrow \mathbb{R}, f \mapsto f(x)$ **bounded**
- **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**
- $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ **kernel function**
- **Why useful**: Convergence in **norm** of \mathcal{F} implies **pointwise** convergence
- **Inner product between two feature maps**:

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



A Very Short Introduction to Kernels



MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

- Hilbert space of functions \mathcal{F} from \mathcal{X} to \mathbb{R}
- **RKHS**: evaluation operator $\delta_x : \mathcal{F} \rightarrow \mathbb{R}, f \mapsto f(x)$ **bounded**
- **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**
- $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ **kernel function**
- **Why useful**: Convergence in **norm** of \mathcal{F} implies **pointwise** convergence
- **Inner product between two feature maps**:

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

- **Moore-Aronszajn**: **Every RKHS** has a **UNIQUE** positive definite kernel
 - ...and vice versa

Introduction to distribution embeddings



Embeddings and Metrics



- How to define a **metric** on the space of **probability measures**
 - Function revealing differences in distributions
 - Distance between means in space of features (**RKHS**)
 - **Same thing: the MMD** [Gretton et al., 2007, Borgwardt et al., 2006]



- How to define a **metric** on the space of **probability measures**
 - Function revealing differences in distributions
 - Distance between means in space of features (**RKHS**)
 - **Same thing: the MMD** [Gretton et al., 2007, Borgwardt et al., 2006]
- For which feature spaces are mappings **unique**?
 - **Characteristic RKHSs** [Fukumizu et al., 2008, Sriperumbudur et al., 2008]
 - Easy to check for **translation invariant kernels**



- How to define a **metric** on the space of **probability measures**
 - Function revealing differences in distributions
 - Distance between means in space of features (**RKHS**)
 - **Same thing: the MMD** [Gretton et al., 2007, Borgwardt et al., 2006]
- For which feature spaces are mappings **unique**?
 - **Characteristic RKHSs** [Fukumizu et al., 2008, Sriperumbudur et al., 2008]
 - Easy to check for **translation invariant kernels**
- Is difference in distributions **significant**? [Gretton et al., 2007]



- How to define a **metric** on the space of **probability measures**
 - Function revealing differences in distributions
 - Distance between means in space of features (**RKHS**)
 - **Same thing: the MMD** [Gretton et al., 2007, Borgwardt et al., 2006]
- For which feature spaces are mappings **unique**?
 - **Characteristic RKHSs** [Fukumizu et al., 2008, Sriperumbudur et al., 2008]
 - Easy to check for **translation invariant kernels**
- Is difference in distributions **significant**? [Gretton et al., 2007]
- Related problem: **independence** [Gretton et al., 2008]



- Idea: **avoid density estimation** when comparing distributions **P** and **Q**

[Fortet and Mourier, 1953]

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



- Idea: **avoid density estimation** when comparing distributions \mathbf{P} and \mathbf{Q}

[Fortet and Mourier, 1953]

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

- **Classical results:** $\text{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when
 - $F =$ bounded continuous [Dudley, 2002]
 - $F =$ bounded variation 1 (Kolmogorov metric) [Müller, 1997]
 - $F =$ bounded Lipschitz (Earth mover's distances) [Dudley, 2002]



- Idea: **avoid density estimation** when comparing distributions \mathbf{P} and \mathbf{Q}

[Fortet and Mourier, 1953]

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

- **Classical results:** $\text{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when
 - $F =$ bounded continuous [Dudley, 2002]
 - $F =$ bounded variation 1 (Kolmogorov metric) [Müller, 1997]
 - $F =$ bounded Lipschitz (Earth mover's distances) [Dudley, 2002]
- $\text{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F =$ the unit ball in a **characteristic** RKHS \mathcal{F} [Fukumizu et al., 2008, Sriperumbudur et al., 2008]



- Idea: **avoid density estimation** when comparing distributions \mathbf{P} and \mathbf{Q}

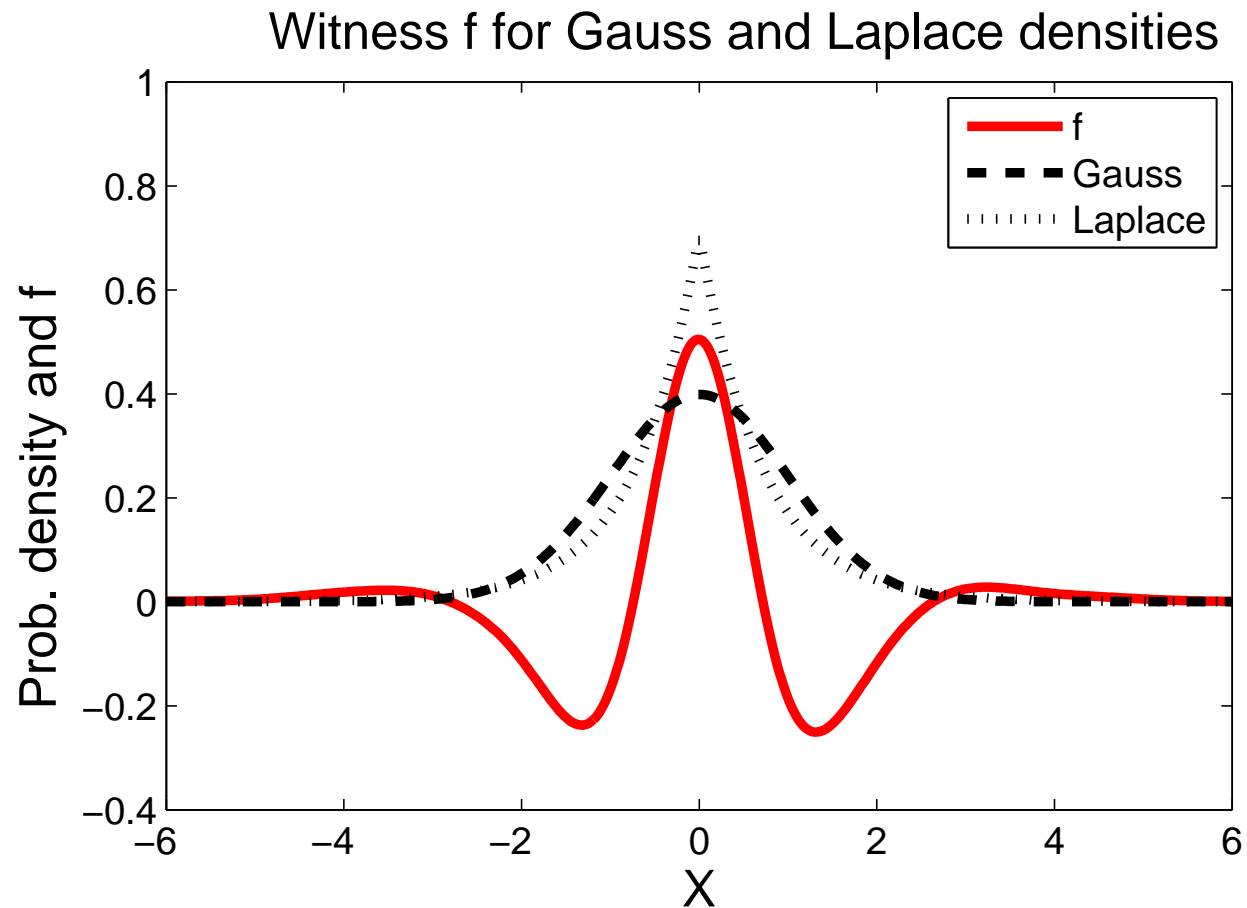
[Fortet and Mourier, 1953]

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

- **Classical results:** $\text{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when
 - $F =$ bounded continuous [Dudley, 2002]
 - $F =$ bounded variation 1 (Kolmogorov metric) [Müller, 1997]
 - $F =$ bounded Lipschitz (Earth mover's distances) [Dudley, 2002]
- $\text{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F =$ the unit ball in a **characteristic** RKHS \mathcal{F} [Fukumizu et al., 2008, Sriperumbudur et al., 2008]
 - **Examples:** Gaussian, Laplace, ...



- Gauss **P** vs Laplace **Q**





- The (kernel) MMD:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}; F)$$

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



- The (kernel) MMD:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}; F)$$

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

using

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



- The (kernel) MMD:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}; F)$$

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

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

using

$$\mathbf{E}_{\mathbf{P}}(f(x)) = \mathbf{E}_{\mathbf{P}}[\langle \phi(x), f \rangle_{\mathcal{F}}]$$

$$=: \langle \mu_x, f \rangle_{\mathcal{F}}$$



- The (kernel) MMD:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}; F)$$

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

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

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

using

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



- The (kernel) MMD:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}; F)$$

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

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

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

$$= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}}$$

$$= \mathbf{E}_{\mathbf{P}, \mathbf{P}'} k(x, x') + \mathbf{E}_{\mathbf{Q}, \mathbf{Q}'} k(y, y') - 2\mathbf{E}_{\mathbf{P}, \mathbf{Q}} k(x, y)$$

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



- The (kernel) MMD:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}; F)$$

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

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

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

$$= \langle \mu_x - \mu_y, \mu_x - \mu_y \rangle_{\mathcal{F}}$$

$$= \mathbf{E}_{\mathbf{P}, \mathbf{P}'} k(x, x') + \mathbf{E}_{\mathbf{Q}, \mathbf{Q}'} k(y, y') - 2\mathbf{E}_{\mathbf{P}, \mathbf{Q}} k(x, y)$$

- Note:

$$\mu_x := \int k(\cdot, x) d\mathbf{P}(x)$$

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



Characteristic Kernels (1)



- For what kernels is MMD a **metric** (MMD = 0 iff **P** = **Q**)



Characteristic Kernels (1)



- For what kernels is MMD a **metric** (MMD = 0 iff **P** = **Q**)
- **Translation invariant** kernels: $k(x, y) = k(x - y)$



Characteristic Kernels (1)



- For what kernels is MMD a **metric** (MMD = 0 iff **P** = **Q**)
- **Translation invariant** kernels: $k(x, y) = k(x - y)$
- **Bochner's theorem**:

$$k(x) = \int_{\mathbb{R}^d} e^{-ix^\top \omega} d\Lambda(\omega)$$

- Λ finite non-negative Borel measure



Characteristic Kernels (1)



- For what kernels is MMD a **metric** (MMD = 0 iff **P** = **Q**)
- **Translation invariant** kernels: $k(x, y) = k(x - y)$

- **Bochner's theorem:**

$$k(x) = \int_{\mathbb{R}^d} e^{-ix^\top \omega} d\Lambda(\omega)$$

- Λ finite non-negative Borel measure

- **Fourier representation of MMD:**

$$\text{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| [(\bar{\phi}_{\mathbf{P}} - \bar{\phi}_{\mathbf{Q}}) \Lambda]^\vee \right\|_{\mathcal{F}}$$

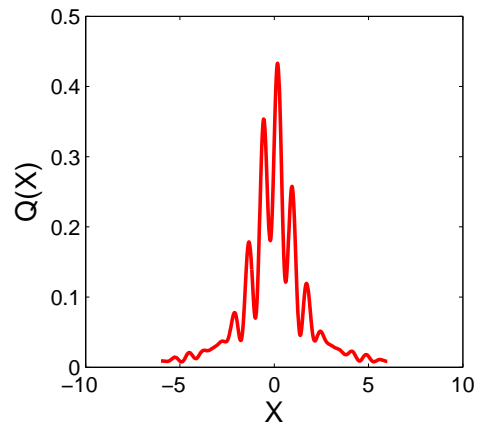
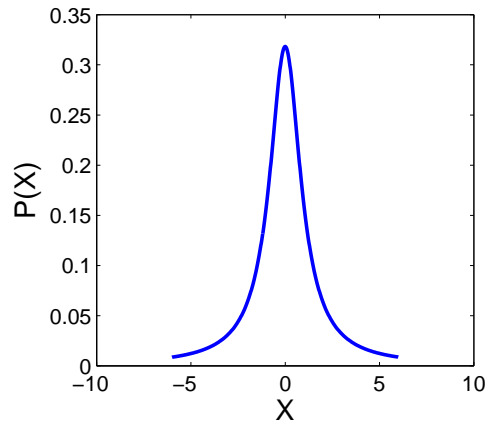
- $\phi_{\mathbf{P}}$ characteristic function of **P**
- f^\wedge is Fourier transform, f^\vee is inverse Fourier transform



Characteristic Kernels (2)

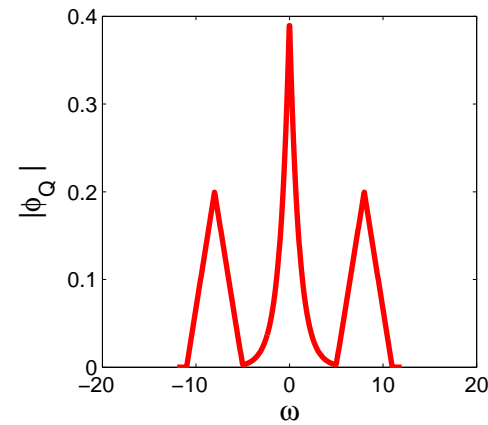
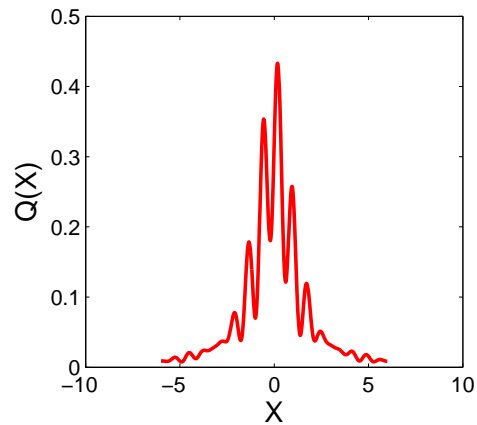
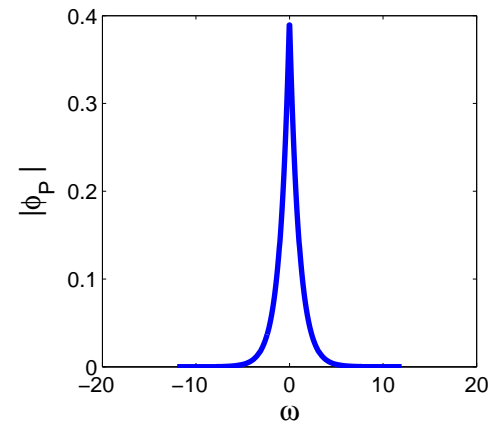
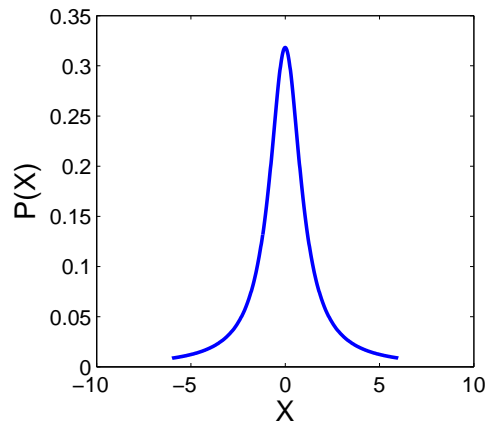


- Example: **P** differs from **Q** at (roughly) one frequency



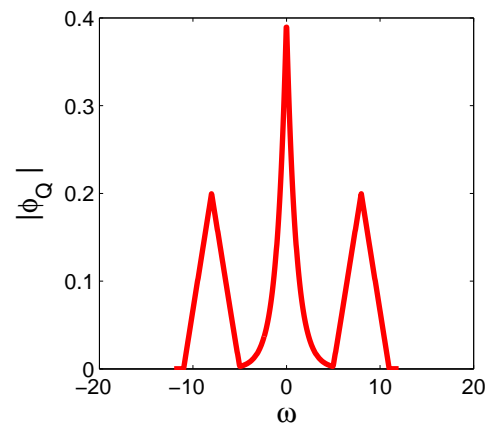
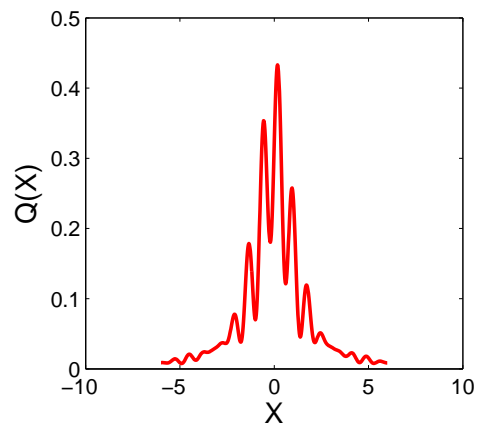
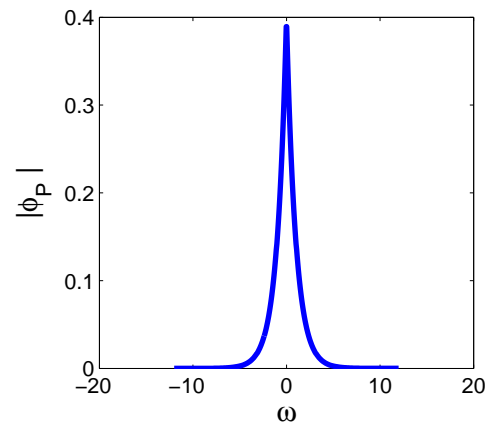
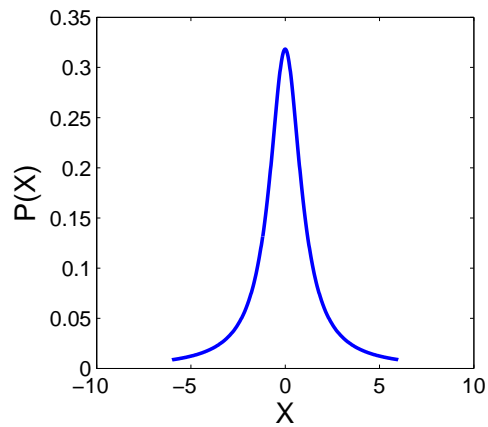
Characteristic Kernels (2)

- Example: **P** differs from **Q** at (roughly) one frequency

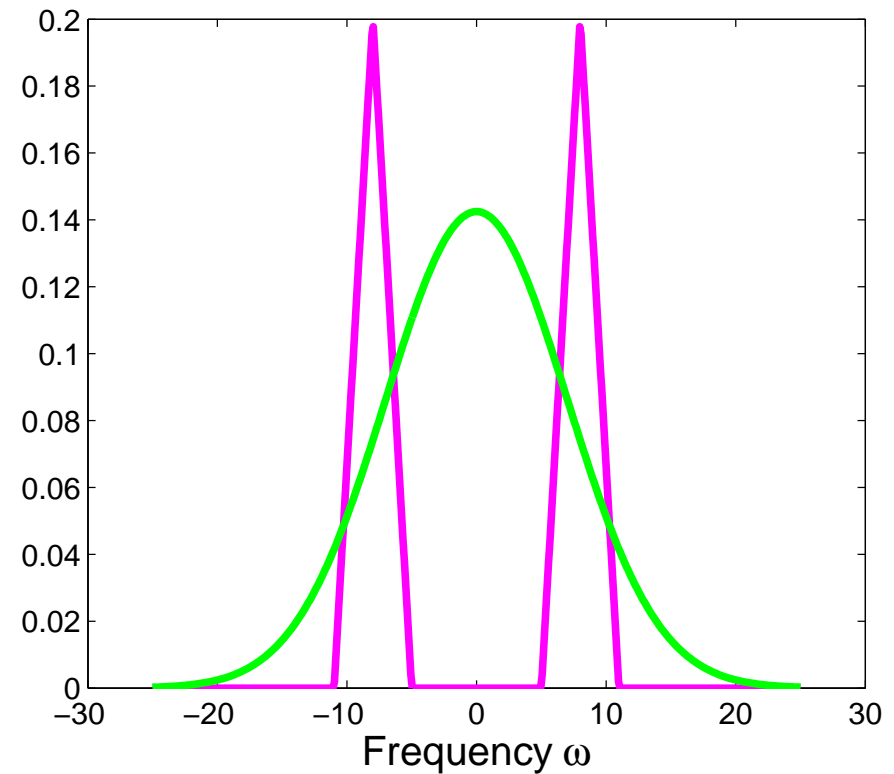


Characteristic Kernels (2)

- Example: **P** differs from **Q** at (roughly) one frequency

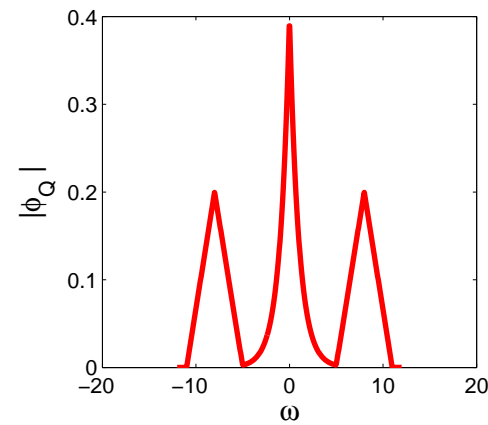
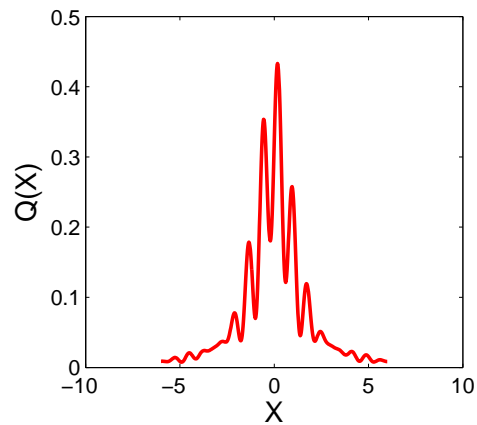
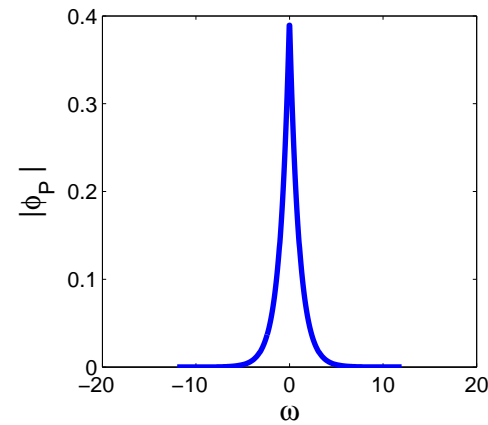
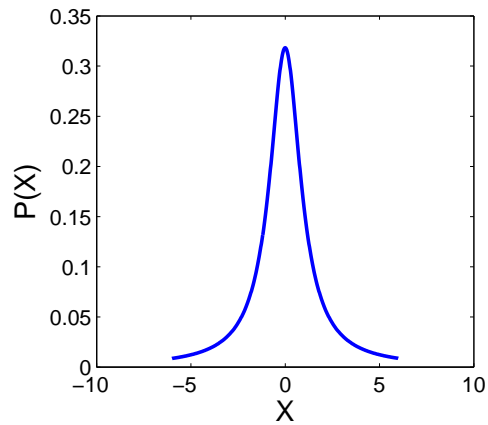


- Gaussian kernel
- Difference in distribution spectra

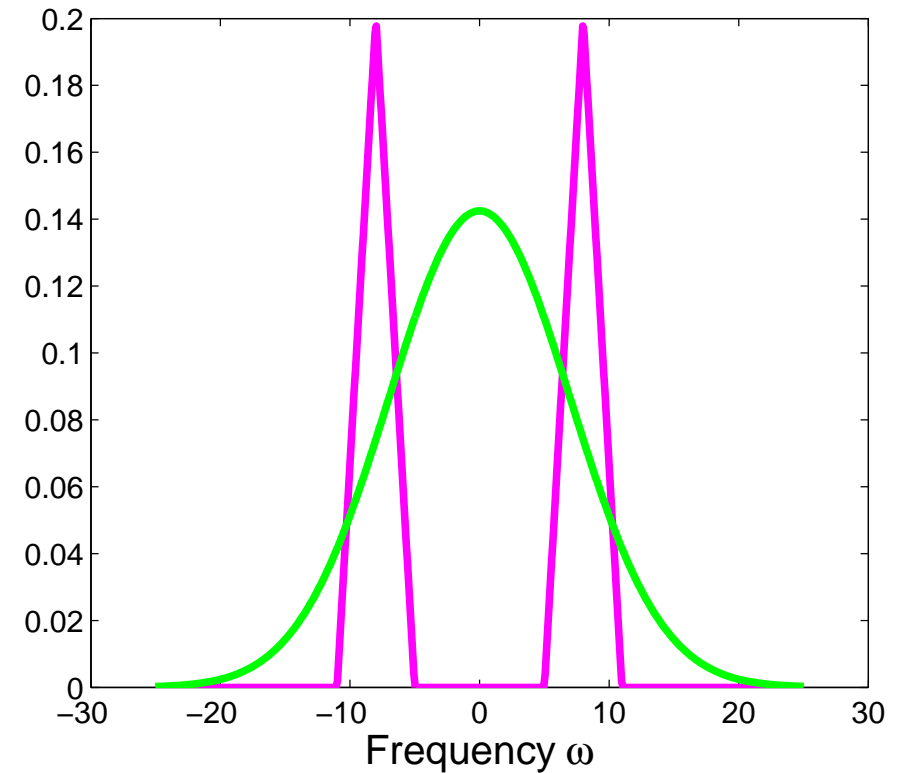


Characteristic Kernels (2)

- Example: **P** differs from **Q** at (roughly) one frequency

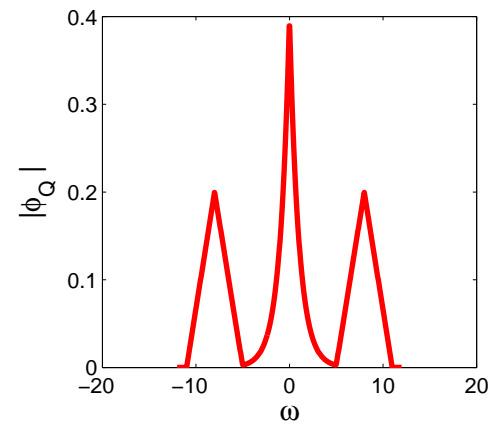
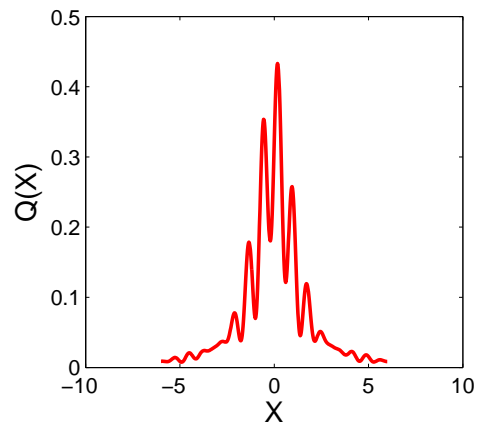
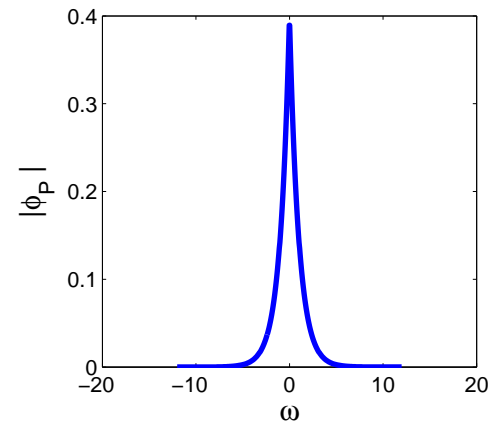
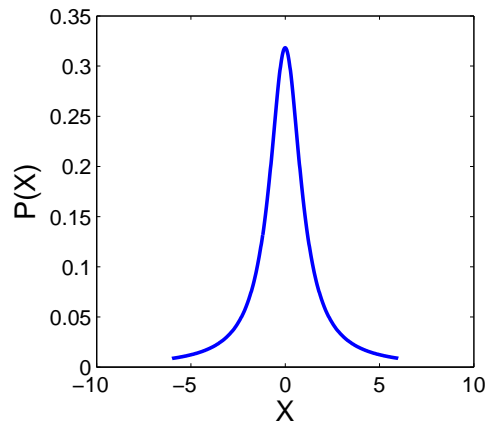


- Gaussian kernel characteristic
- Difference in distribution spectra

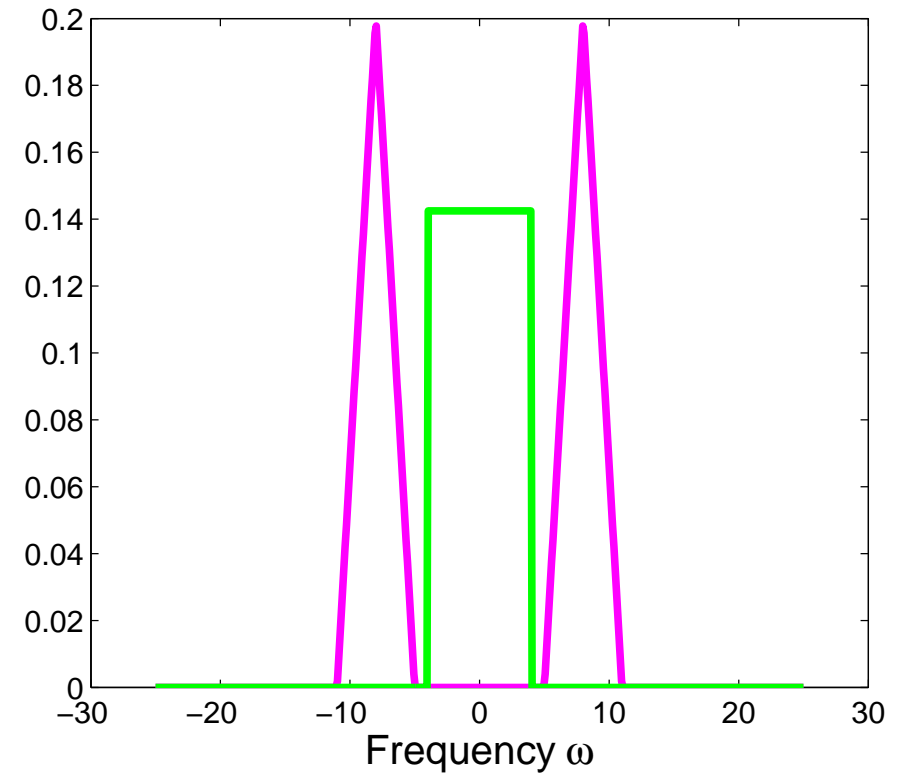


Characteristic Kernels (3)

- Example: **P** differs from **Q** at (roughly) one frequency

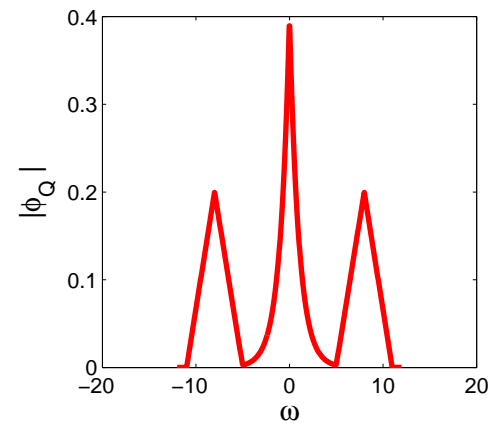
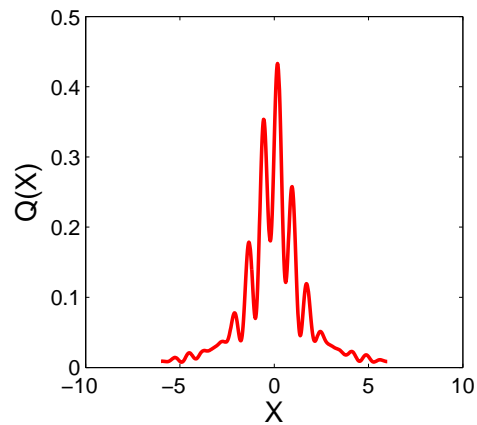
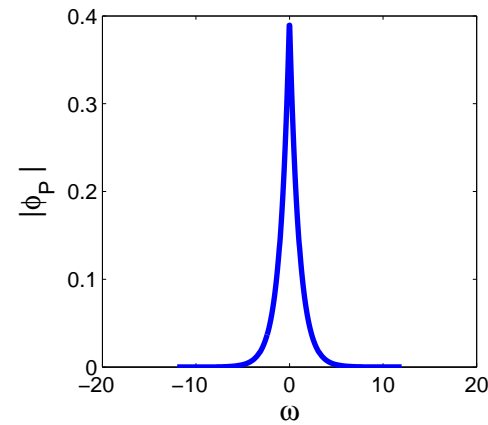
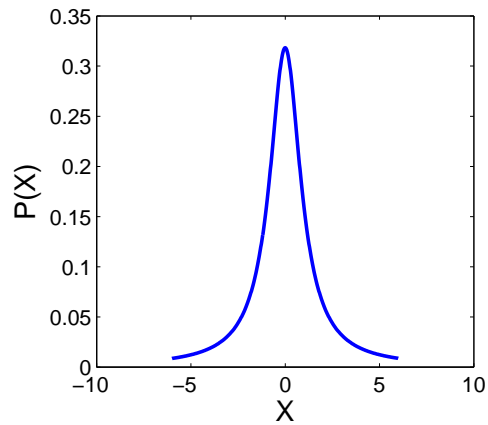


- Sinc kernel **NOT** characteristic
- Difference in distribution spectra



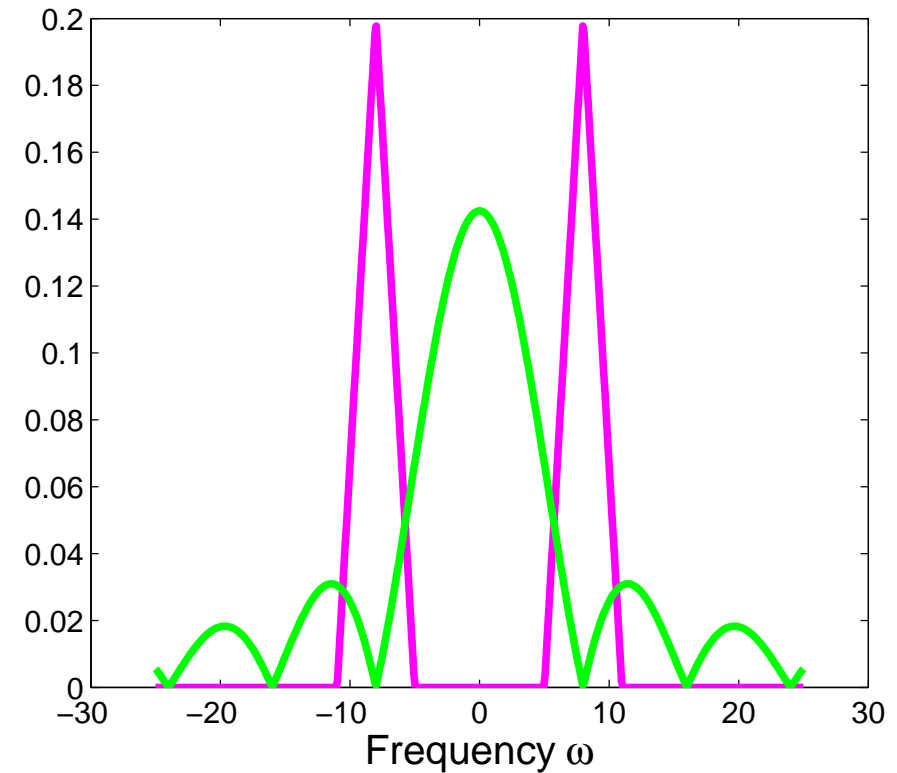
Characteristic Kernels (4)

- Example: **P** differs from **Q** at (roughly) one frequency



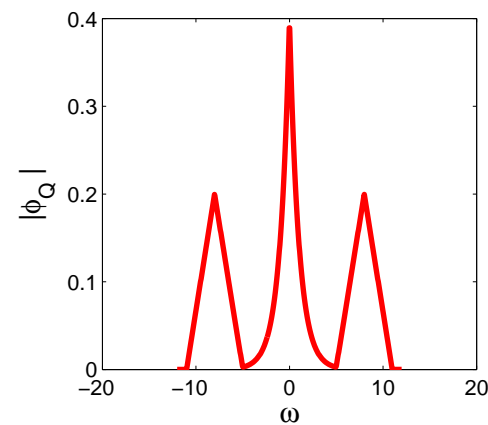
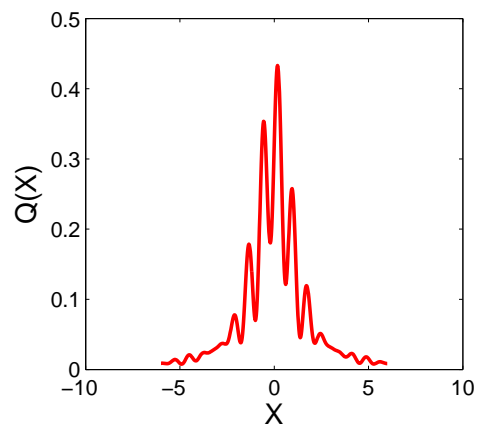
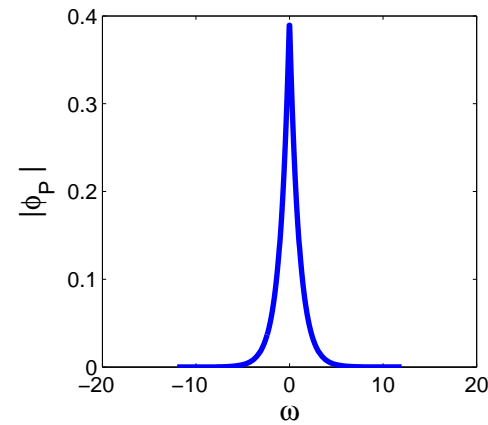
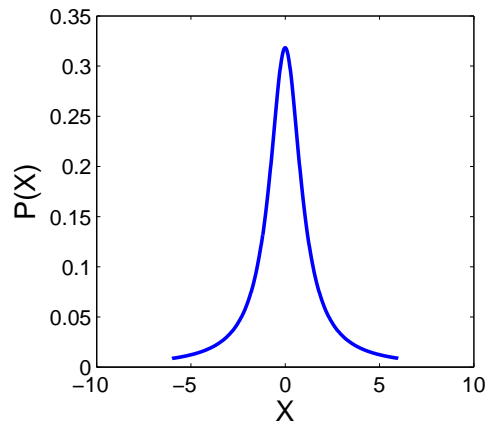
- **B-Spline** kernel ???

- Difference in distribution spectra

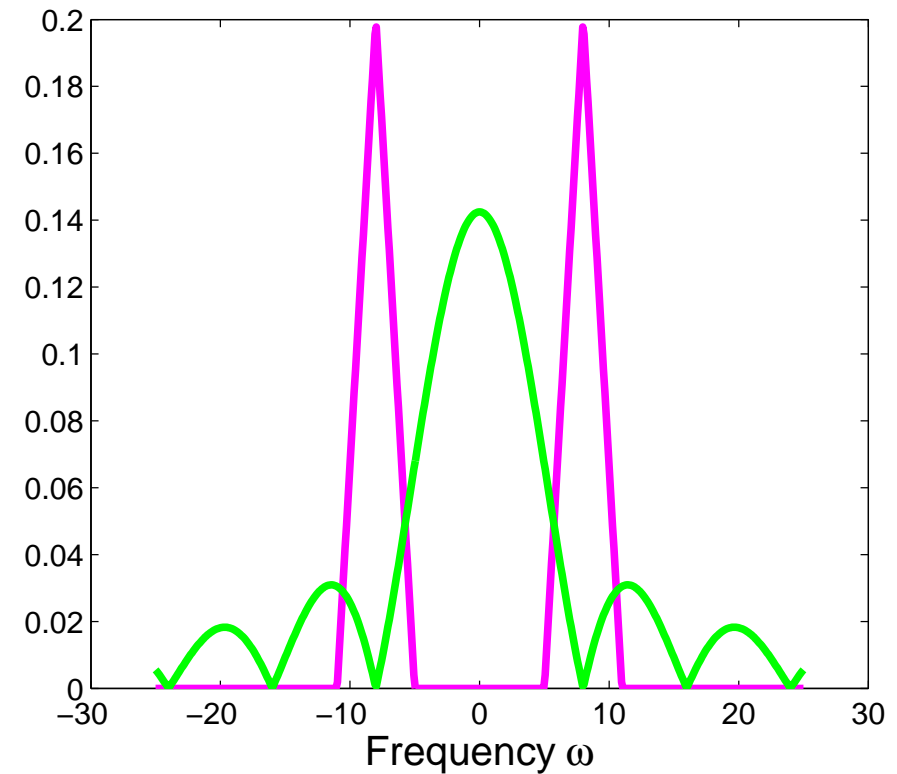


Characteristic Kernels (4)

- Example: **P** differs from **Q** at (roughly) one frequency



- **B-Spline** kernel characteristic
- Difference in distribution spectra





Characteristic Kernels (5)



- **Main theorem:** k characteristic if and only if
 $\text{supp}(\Lambda) = \mathbb{R}^d$



Characteristic Kernels (5)



- **Main theorem:** k characteristic **if and only if**
 $\text{supp}(\Lambda) = \mathbb{R}^d$
- **Corollary:** if k **continuous with compact support**, then k characteristic



Characteristic Kernels (5)



- **Main theorem:** k characteristic **if and only if**
 $\text{supp}(\Lambda) = \mathbb{R}^d$
- Corollary: if k **continuous with compact support**, then k characteristic
- Another example: **universal kernels** on compact domains, **includes NON-translation invariant** [Steinwart, 2001]

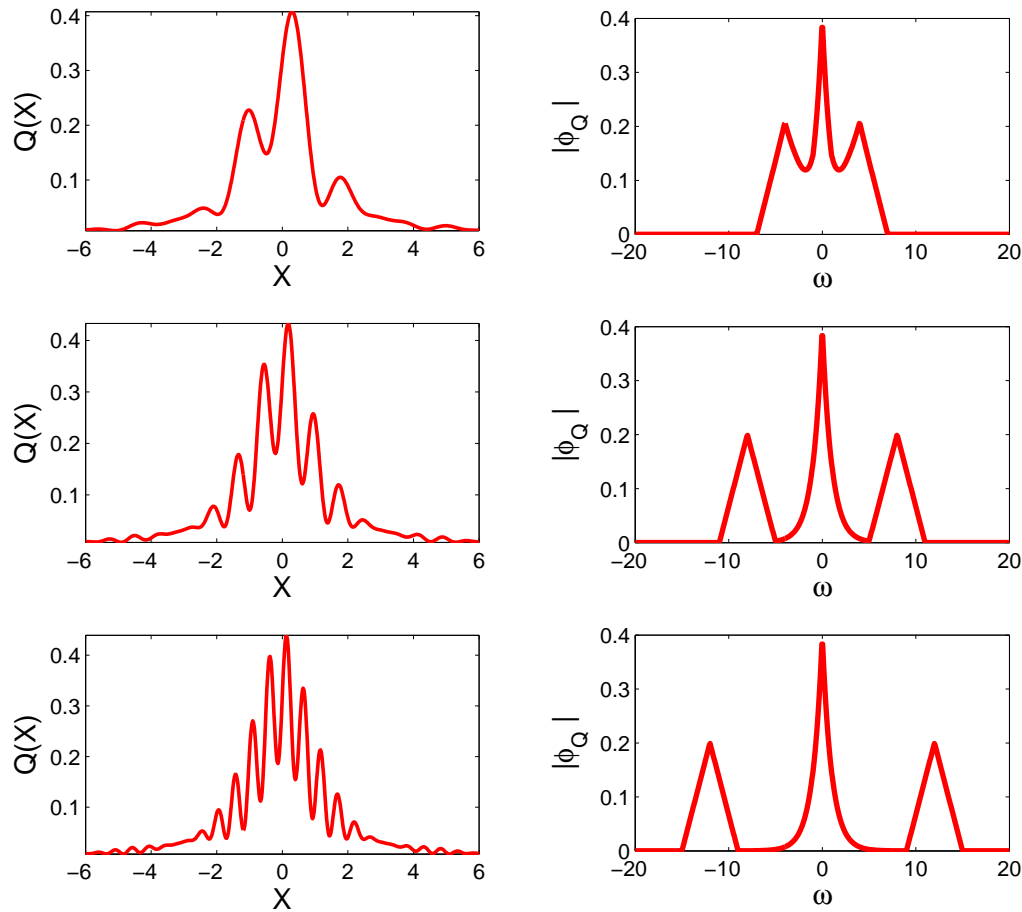


Characteristic Kernels (5)

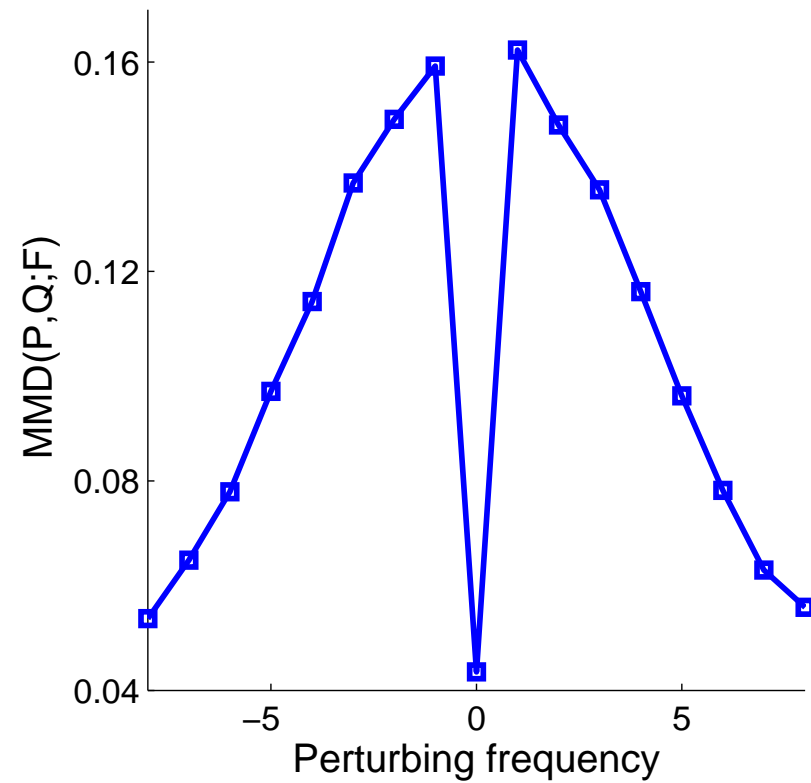


- **Main theorem:** k characteristic **if and only if**
 $\text{supp}(\Lambda) = \mathbb{R}^d$
- Corollary: if k **continuous with compact support**, then k characteristic
- Another example: **universal kernels** on compact domains, **includes NON-translation invariant** [Steinwart, 2001]
- Similar reasoning wherever extensions of **Bochner's theorem** exist:
 - Locally compact Abelian groups (**periodic domains**)
 - Compact, non-Abelian groups (**orthogonal matrices**)
 - The semigroup \mathbb{R}_n^+ (**histograms**)

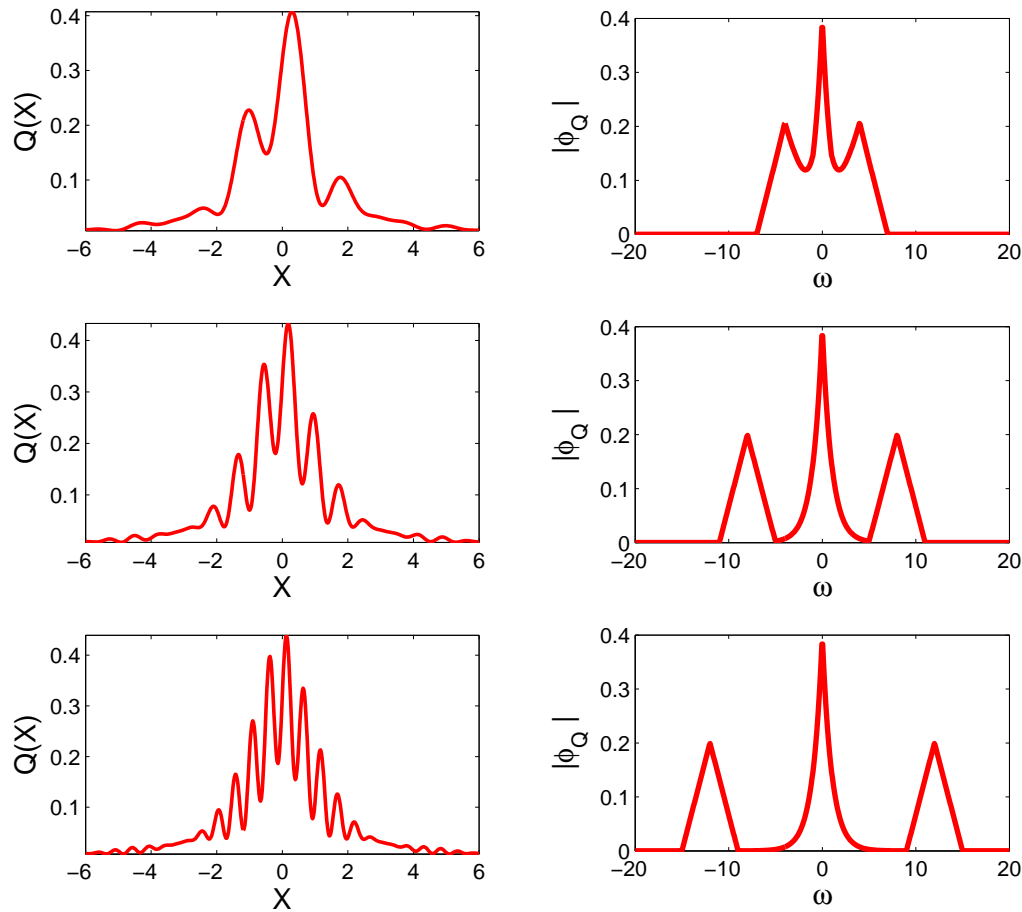
- **Gaussian** kernel example



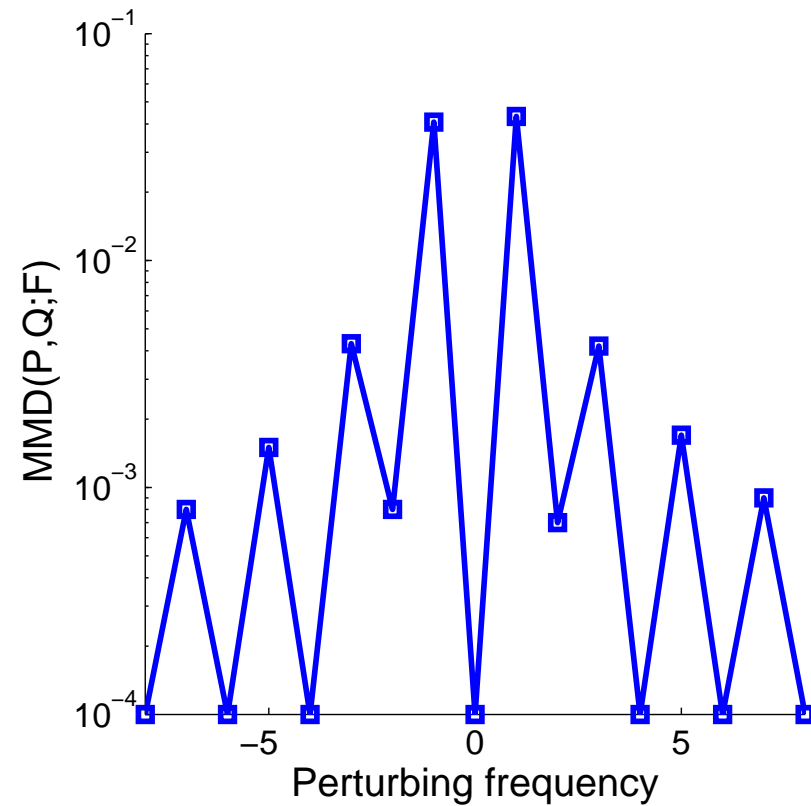
- MMD vs frequency of perturbation to **P**



- **B-spline** kernel example



- MMD vs frequency of perturbation to **P**





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 $\mathbf{x} := \{x_1, \dots, x_m\}$ from \mathbf{P} and \mathbf{y} from \mathbf{Q}
- If **empirical** $\text{MMD}(\mathbf{x}, \mathbf{y}; F)$ 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 $\mathbf{x} := \{x_1, \dots, x_m\}$ from \mathbf{P} and \mathbf{y} from \mathbf{Q}
- If **empirical** $\text{MMD}(\mathbf{x}, \mathbf{y}; F)$ 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 $\mathbf{x} := \{x_1, \dots, x_m\}$ from \mathbf{P} and \mathbf{y} from \mathbf{Q}
- If **empirical** $\text{MMD}(\mathbf{x}, \mathbf{y}; F)$ 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 $\text{MMD}(\mathbf{x}, \mathbf{y}; F)$



Statistical Test Using MMD (2)



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

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



Statistical Test Using MMD (2)



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

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

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



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

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

- When $\mathbf{P} \neq \mathbf{Q}$, **asymptotically normal** [Hoeffding, 1948, Serfling, 1980]
- Expression for the **variance**: $z_i := (x_i, y_i)$

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



Statistical Test Using MMD (3)



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

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

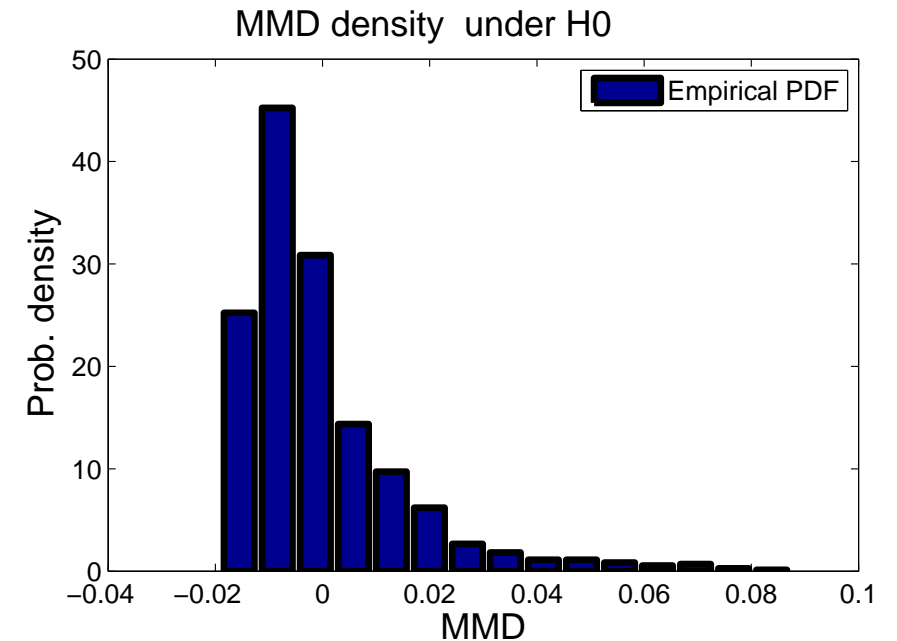
- where
 - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
 - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathbf{P}_x(x) = \lambda_i \psi_i(x')$

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

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

- where

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





Statistical Test Using MMD (4)



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



Statistical Test Using MMD (4)



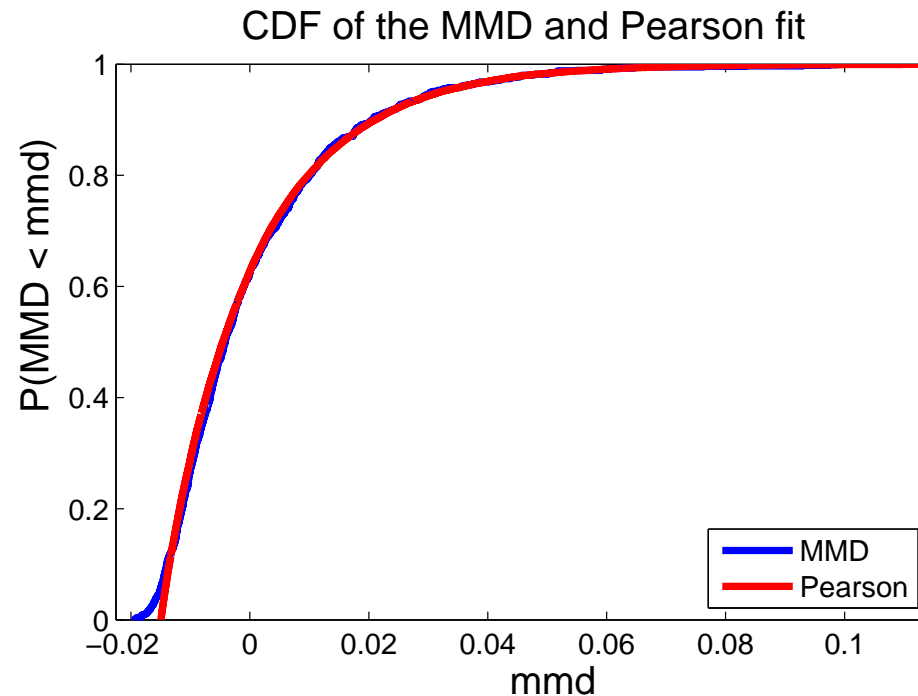
- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\text{MMD} > T) \leq 0.05$
- **Bootstrap** for empirical CDF [Arcones and Giné, 1992]
- **Pearson curves** by matching first four moments [Johnson et al., 1994]
- **Large deviation bounds** [Hoeffding, 1963, McDiarmid, 1989]
- **Other...**



Statistical Test Using MMD (4)



- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\text{MMD} > T) \leq 0.05$
- **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, 1989]
- **Other...**



Dependence Detection with Kernels



MMD for Independence (1)



- Independence
 - Determine: Does $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$?



MMD for Independence (1)



- Independence
 - Determine: Does $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$?
- MMD between mapping of \mathbf{P} and mapping of $\mathbf{P}_x \mathbf{P}_y$



MMD for Independence (1)



MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

- Independence
 - Determine: Does $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$?
- MMD between mapping of \mathbf{P} and mapping of $\mathbf{P}_x \mathbf{P}_y$
- Covariance operators in spaces of features
 - Spectral norm (COCO) [Gretton et al., 2005b,c]
 - Hilbert-Schmidt norm (HSIC) [Gretton et al., 2005a]



MMD for Independence (2)



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

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



MMD for Independence (2)



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

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

- Define the **mean elements**

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

and

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



MMD for Independence (2)



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

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

- Define the **mean elements**

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

and

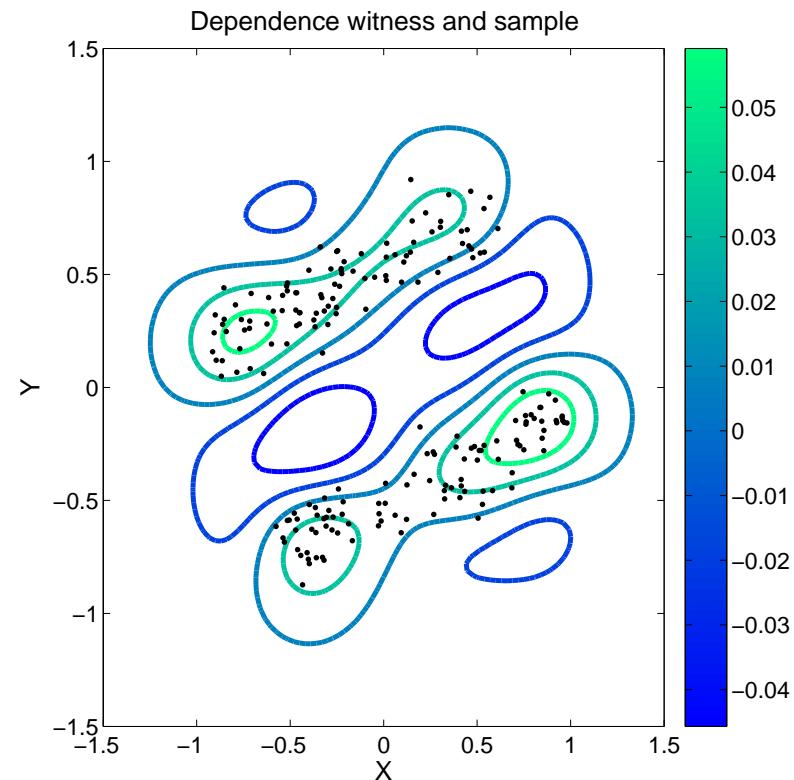
$$\langle \mu_{x \perp y}, \Phi(x, y) \rangle := \mathbf{E}_{x', y''} \langle \Phi(x', y''), \Phi(x, y) \rangle = \mathbf{E}_{x'} k(x, x') \mathbf{E}_{y'} l(y, y')$$

- The **squared distance** between these two mean elements is

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

- Witness function:

$$\sup_{\|f\| \leq 1} \langle f, \mu_{xy} - \mu_{x \perp y} \rangle_{\mathcal{F} \times \mathcal{G}} = \|\mu_{xy} - \mu_{x \perp y}\|_{\mathcal{F} \times \mathcal{G}}$$





Covariance in RKHS (1)



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

$$\text{COCO}(\mathbf{P}; F, G) := \sup_{f \in F, g \in G} (\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)])$$



Covariance in RKHS (1)



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

$$\text{COCO}(\mathbf{P}; F, G) := \sup_{f \in F, g \in G} (\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)])$$

- $\text{COCO}(\mathbf{P}; F, G) = 0$ **iff** x, y **independent**, when F and G are respective unit balls in **characteristic** RKHSs \mathcal{F}, \mathcal{G} [Fukumizu et al., 2008, Sriperumbudur et al., 2008]



Covariance in RKHS (1)



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

$$\text{COCO}(\mathbf{P}; F, G) := \sup_{f \in F, g \in G} (\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)])$$

- $\text{COCO}(\mathbf{P}; F, G) = 0$ iff x, y **independent**, when F and G are respective unit balls in **characteristic** RKHSs \mathcal{F}, \mathcal{G} [Fukumizu et al., 2008, Sriperumbudur et al., 2008]

More formally:

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

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



Covariance in RKHS (1)



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

$$\text{COCO}(\mathbf{P}; F, G) := \sup_{f \in F, g \in G} (\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)])$$

- $\text{COCO}(\mathbf{P}; F, G) = 0$ iff x, y **independent**, when F and G are respective unit balls in **characteristic** RKHSs \mathcal{F}, \mathcal{G} [Fukumizu et al., 2008, Sriperumbudur et al., 2008]

More formally:

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

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

- COCO is the **spectral norm** of Σ_{xy} [Gretton et al., 2005b,c]:

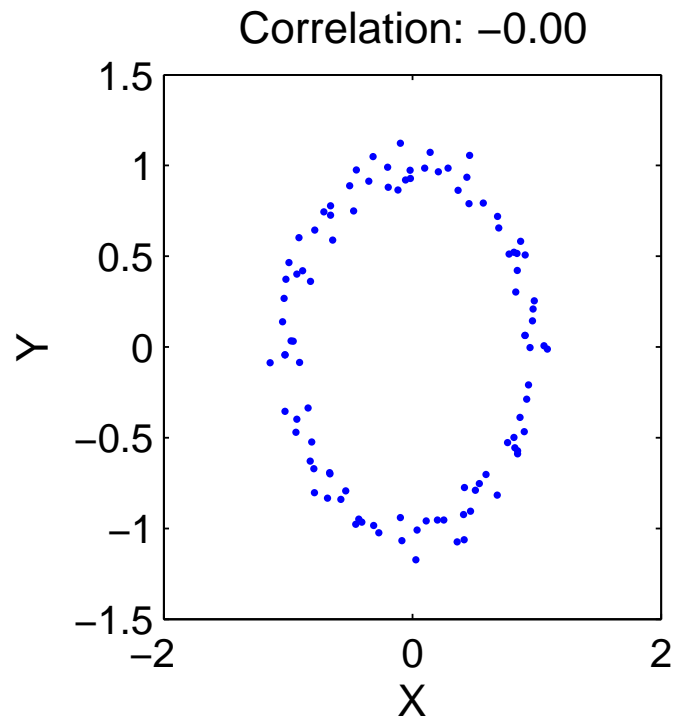
$$\text{COCO}(\mathbf{P}; F, G) := \|\Sigma_{xy}\|_S$$



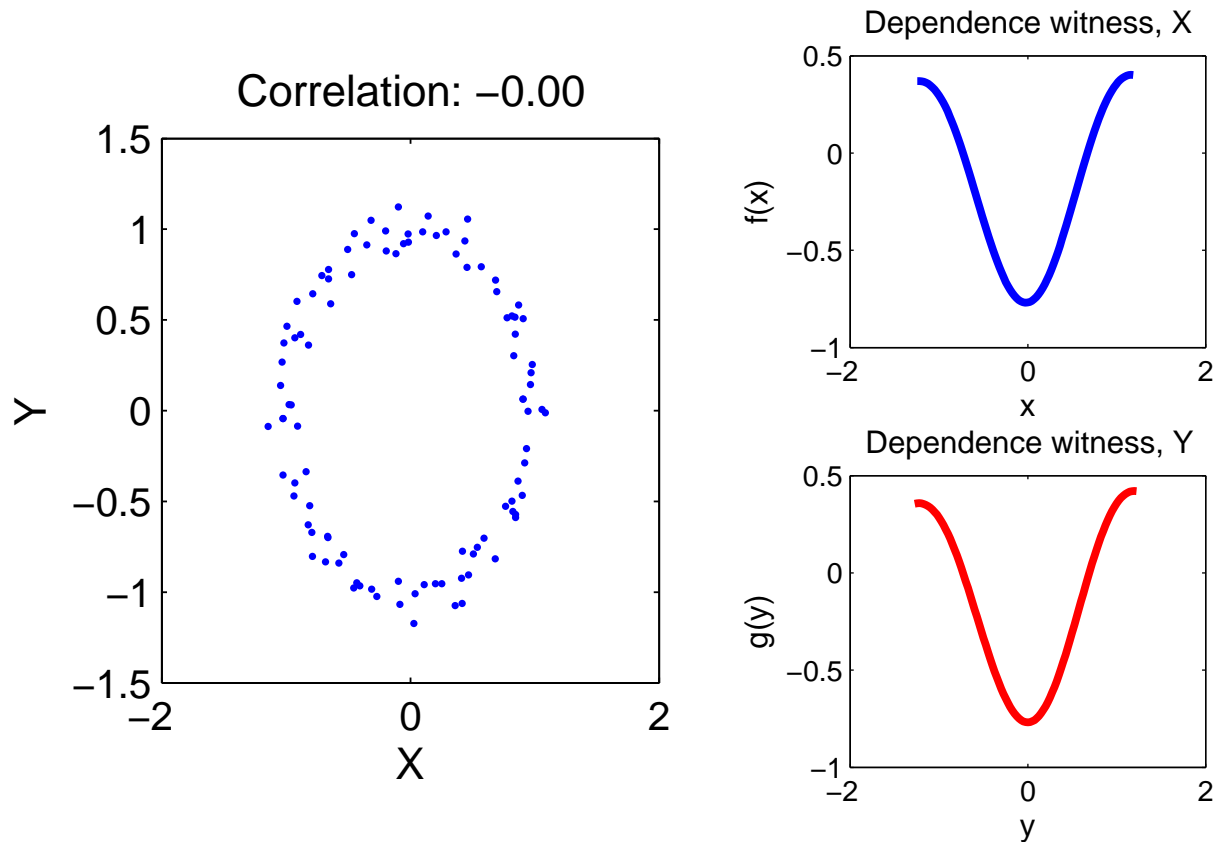
Covariance in RKHS (2)



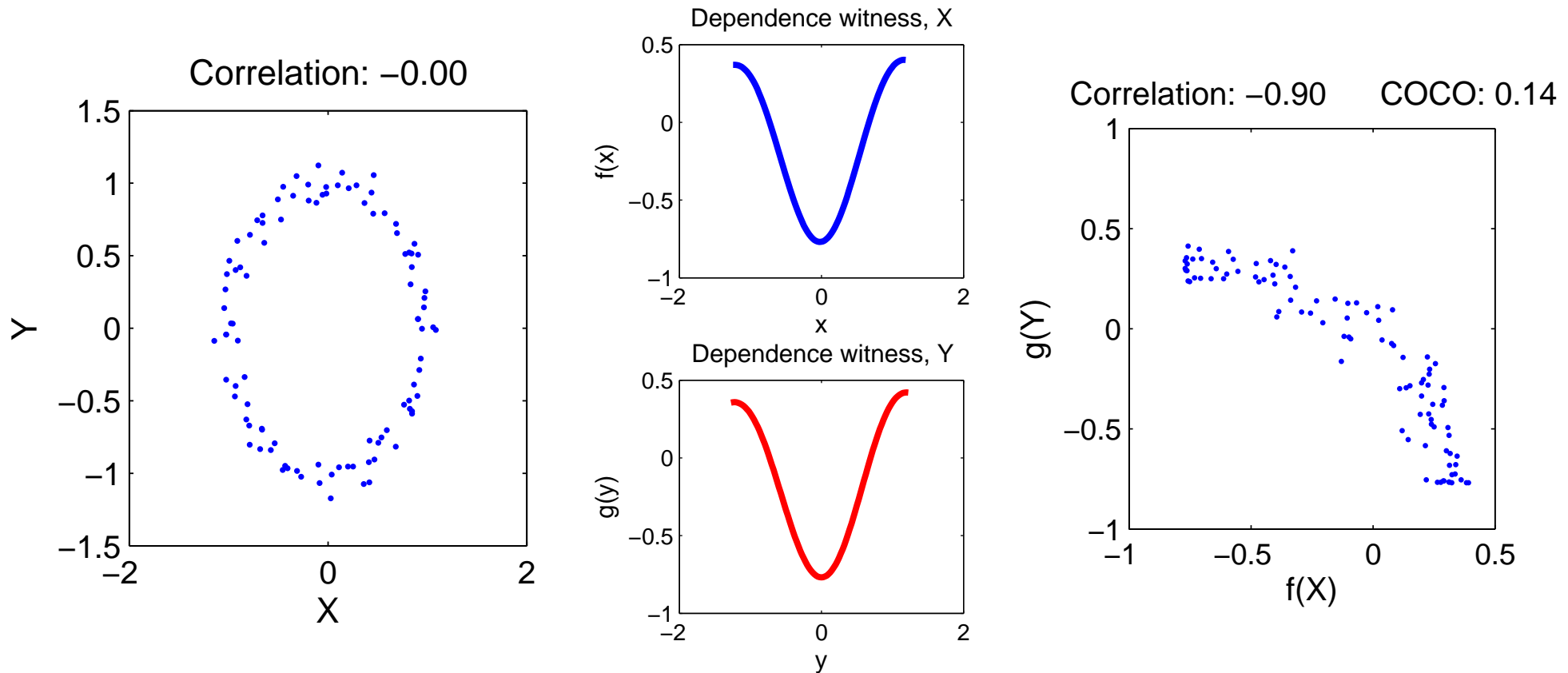
- Ring-shaped density, **correlation approx. zero** [example from Fukumizu, Bach, and Gretton, 2005]



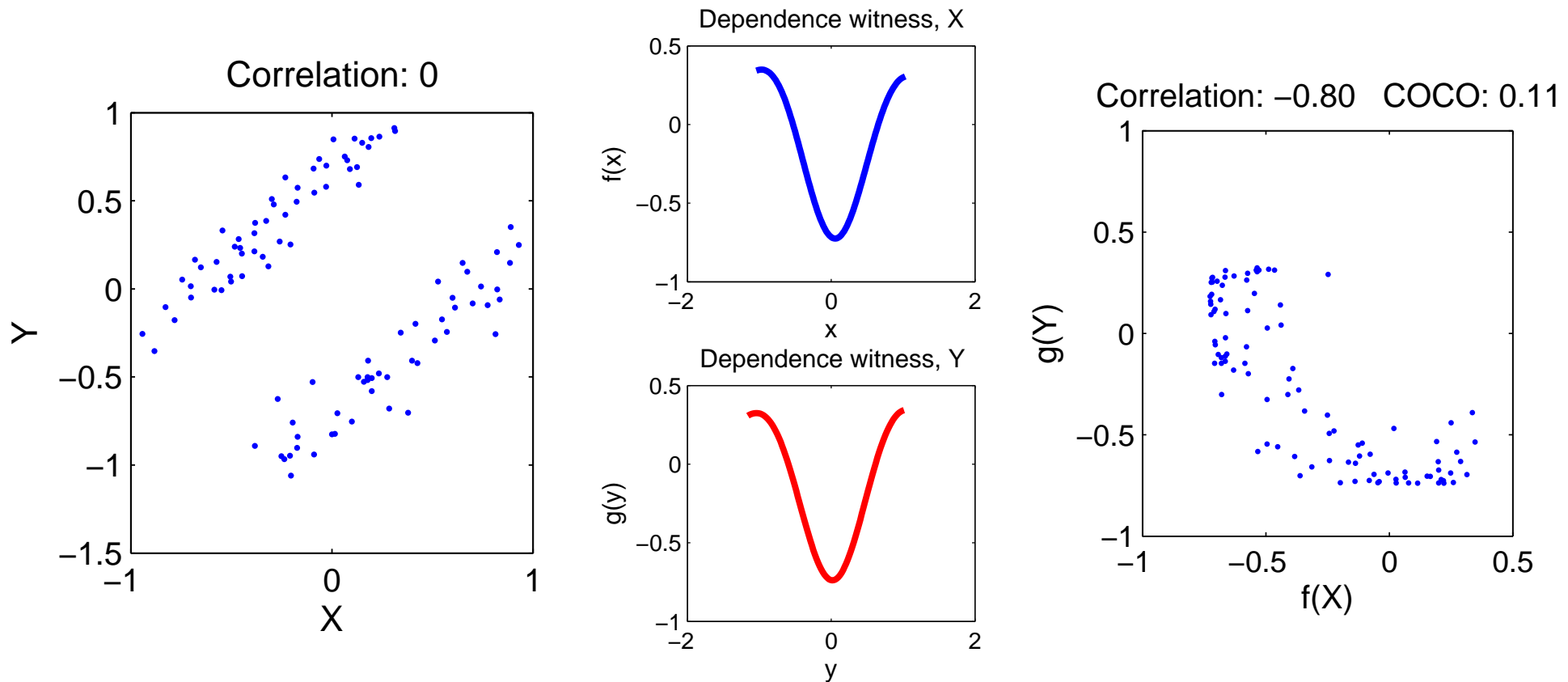
- Ring-shaped density, **correlation approx. zero** [example from Fukumizu, Bach, and Gretton, 2005]



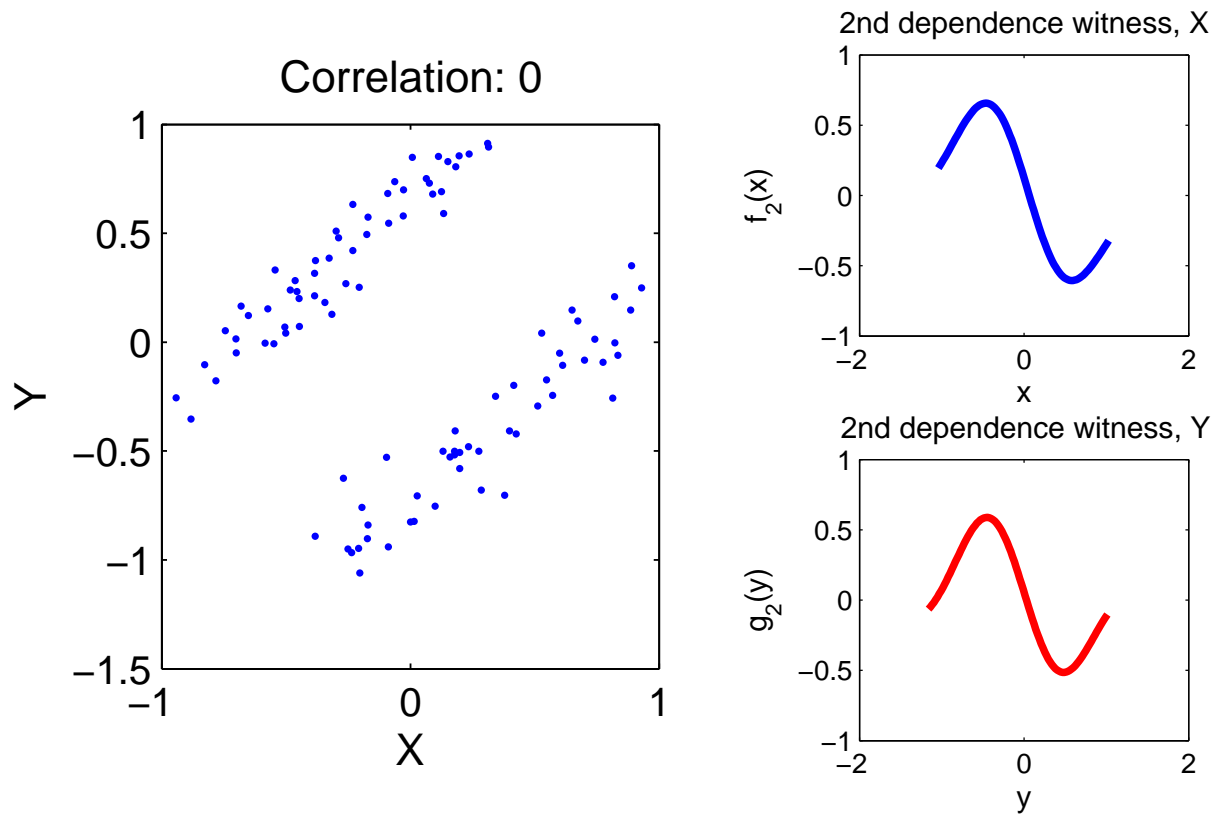
- Ring-shaped density, **correlation approx. zero** [example from Fukumizu, Bach, and Gretton, 2005]



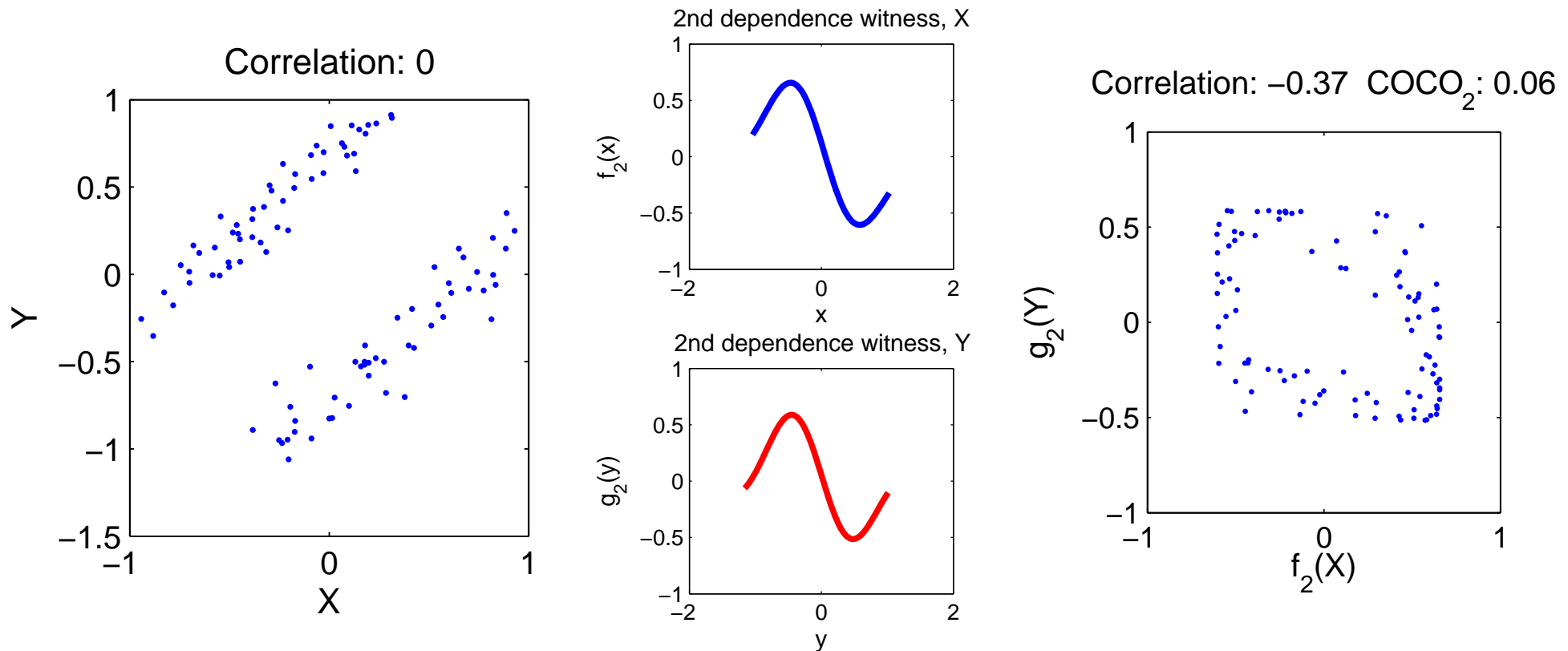
- Can we do better?
- A second example with zero correlation



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



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





Hilbert-Schmidt Independence Criterion



MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

- Given $\gamma_i := \text{COCO}_i(\mathbf{z}; F, G)$, $\mathbf{z} := \{(x_1, y_1), \dots, (x_m, y_m)\}$
- **Hilbert-Schmidt Independence Criterion (HSIC)** [Gretton et al., 2005a]

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



Hilbert-Schmidt Independence Criterion



MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

- Given $\gamma_i := \text{COCO}_i(\mathbf{z}; F, G)$, $\mathbf{z} := \{(x_1, y_1), \dots, (x_m, y_m)\}$
- **Hilbert-Schmidt Independence Criterion (HSIC)** [Gretton et al., 2005a]

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

- In limit of infinite samples:

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



Hilbert-Schmidt Independence Criterion



MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

- Given $\gamma_i := \text{COCO}_i(\mathbf{z}; F, G)$, $\mathbf{z} := \{(x_1, y_1), \dots, (x_m, y_m)\}$
- **Hilbert-Schmidt Independence Criterion (HSIC)** [Gretton et al., 2005a]

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

- In limit of infinite samples:

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

- (Biased) **empirical HSIC** a v-statistic

$$\text{HSIC}(\mathbf{z}; F, G) := \frac{1}{m^2} \text{trace}(\mathbf{KHLH})$$



- Embeddings of distributions into RKHSs
- **Injective** for **characteristic kernels**
 - Easy to check for **translation invariant kernels**:
support of spectrum is \mathbb{R}^d .
- **HSIC**: distance between embedding of **\mathbf{P}** and of **$\mathbf{P}_x \mathbf{P}_y$**
- Next part:

$$HSIC(\mathbf{z}; F, G) := \frac{1}{m^2} \text{trace}(\mathbf{KHLH})$$

Questions?



References

- N. Anderson, P. Hall, and D. Titterton. 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.
- 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 (ISMB)*, 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.
- K. Fukumizu, F. Bach, and A. Gretton. Consistency of kernel canonical correlation analysis. Technical Report 942, Institute of Statistical Mathematics, Tokyo, Japan, 2005.
- K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, 2008.
- 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, *ALT*, pages 63–77. Springer-Verlag, 2005a.
- A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129, 2005b.