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 $\delta_x \ : \ \mathcal{F} \to \mathbb{R}, \ f \mapsto f(x)$ bounded

A Very Short Introduction to Kernels

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

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

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

Talk Outline

- 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]

Talk Outline

- 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

Talk Outline

- 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
- Related problem: independence [Gretton et al., 2008]

Function Showing Difference in Distributions (1)

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

  [Fortet and Mourier, 1953]

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

Function Showing Difference in Distributions (1)

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

  [Fortet and Mourier, 1953]

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

- Classical results: $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ 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]

Function Showing Difference in Distributions (1)

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

  [Fortet and Mourier, 1953]

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

- Classical results: $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ 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]
- $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F =$the unit ball in a characteristic RKHS $\mathcal{F}$ [Fukumizu et al., 2008, Sriperumbudur et al., 2008]

Function Showing Difference in Distributions (1)

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

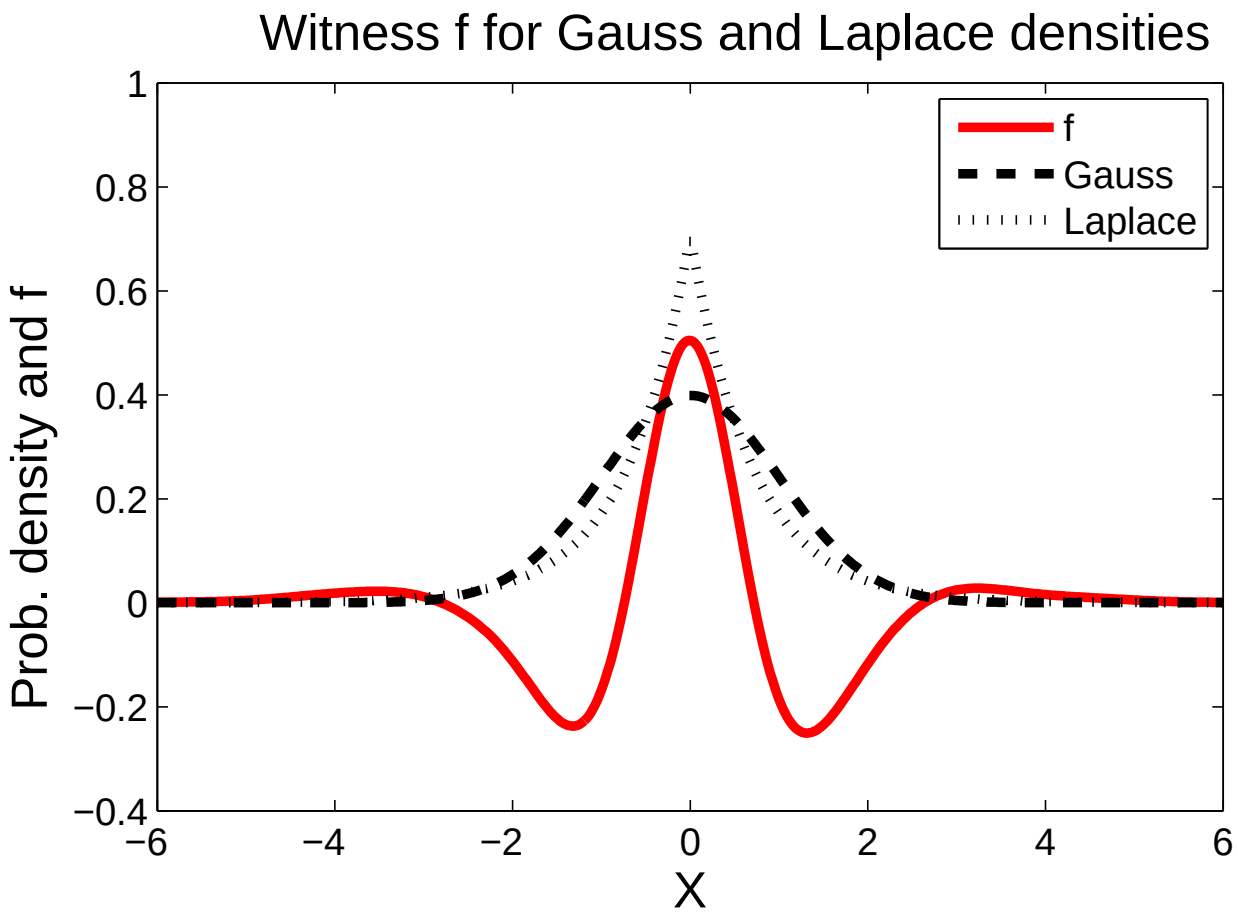
  [Fortet and Mourier, 1953]

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

- Classical results: $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ 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]
- $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ 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, ...

Function Showing Difference in Distributions (2)

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



Function Showing Difference in Distributions (3)

- The (kernel) MMD:

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

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

Function Showing Difference in Distributions (3)

- The (kernel) MMD:

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

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

using

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

Function Showing Difference in Distributions (3)

- The (kernel) MMD:

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

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

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

using

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

Function Showing Difference in Distributions (3)

- The (kernel) MMD:

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

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

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

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

using

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

Function Showing Difference in Distributions (3)

- The (kernel) MMD:

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

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

Function Showing Difference in Distributions (3)

- The (kernel) MMD:

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

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

- Reminder:

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

Characteristic kernels

Characteristic Kernels (1)

- For what kernels is MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$)

Characteristic Kernels (1)

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

Characteristic Kernels (1)

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

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

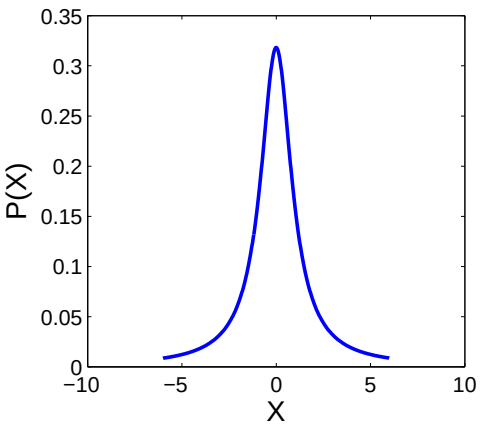
  - Λ finite non-negative Borel measure
- Fourier representation of MMD:

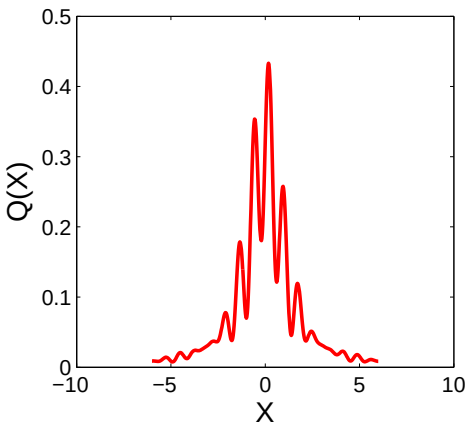
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \left[\left(\bar{\phi}_{\mathbf{P}} - \bar{\phi}_{\mathbf{Q}} \right) \Lambda \right]^\vee \right\|_{\mathcal{F}}$$

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

Characteristic Kernels (2)

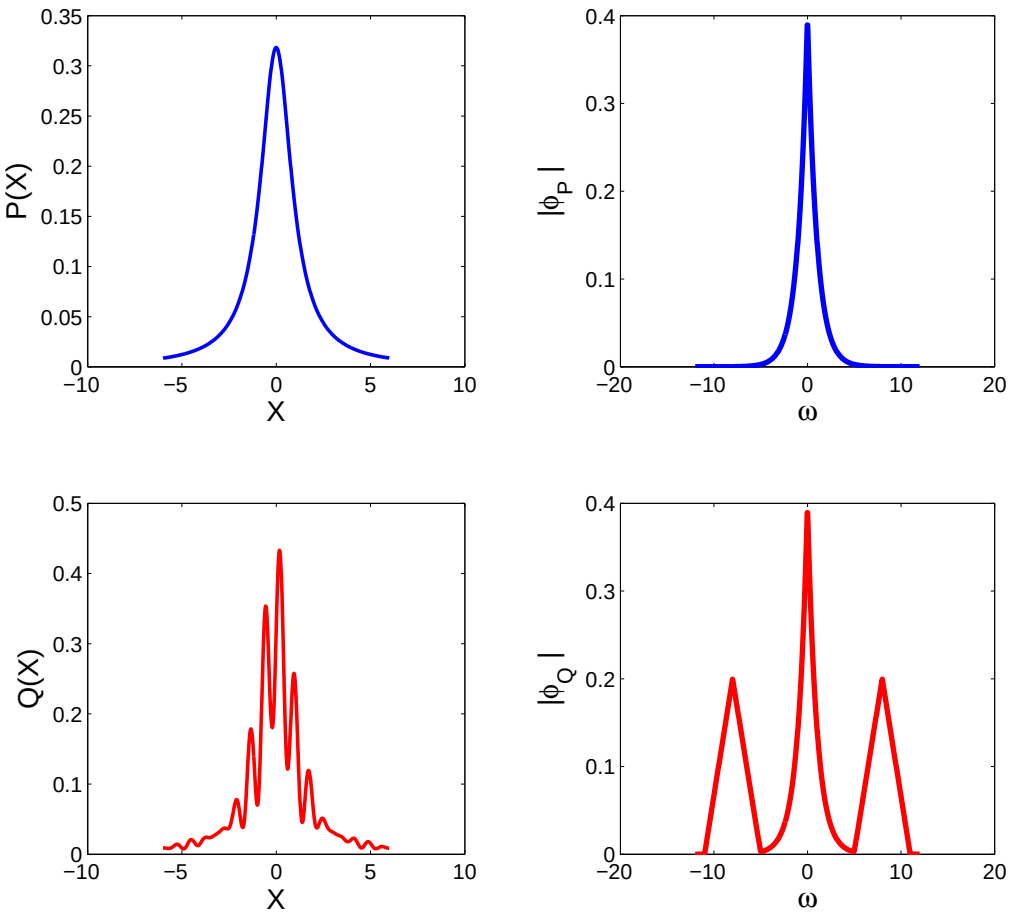
- Example: **P** differs from **Q** at (roughly) one frequency





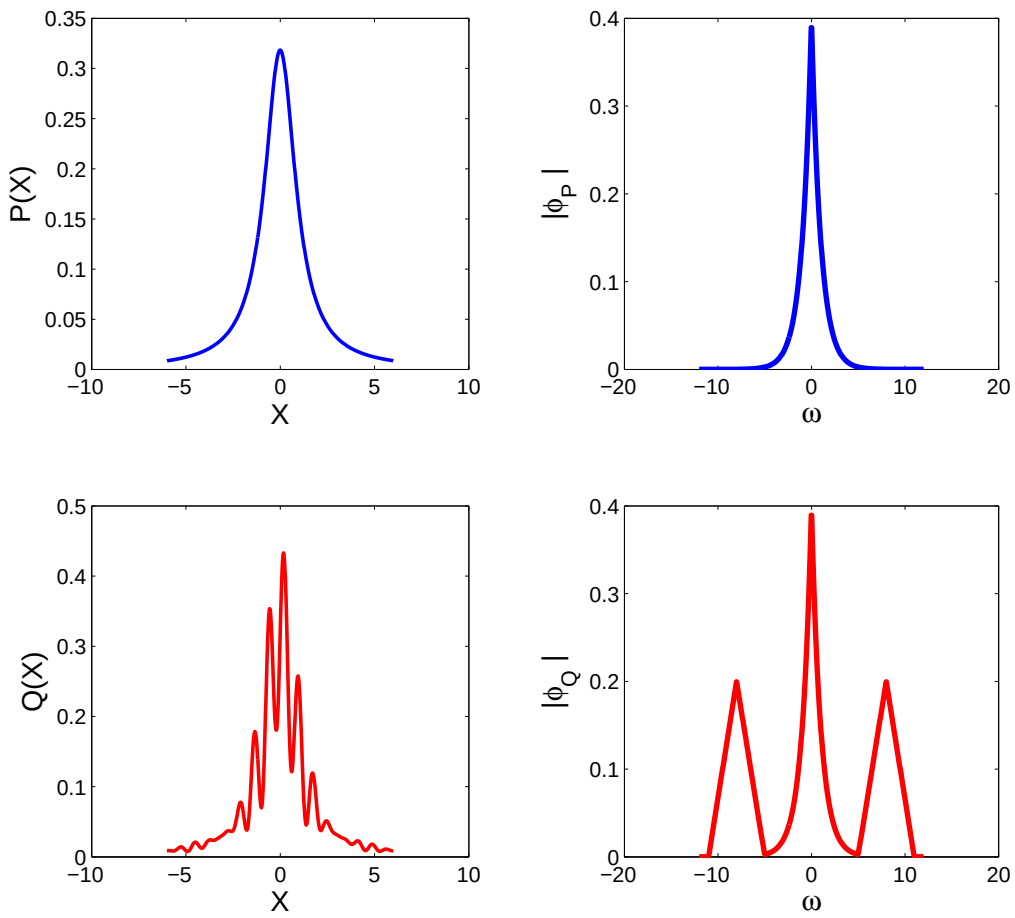
Characteristic Kernels (2)

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

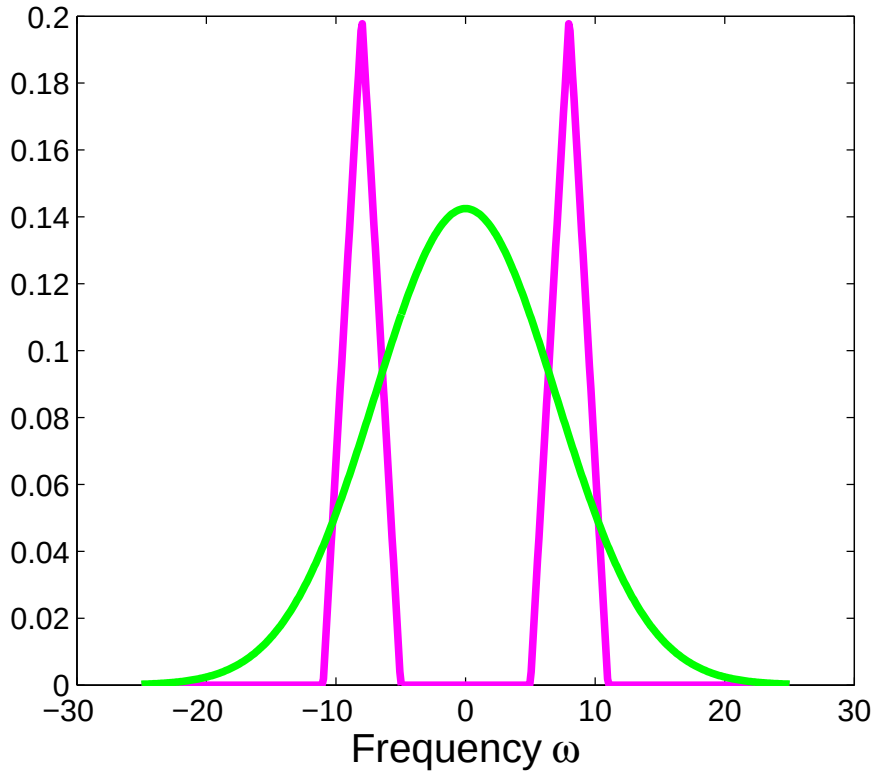


Characteristic Kernels (2)

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

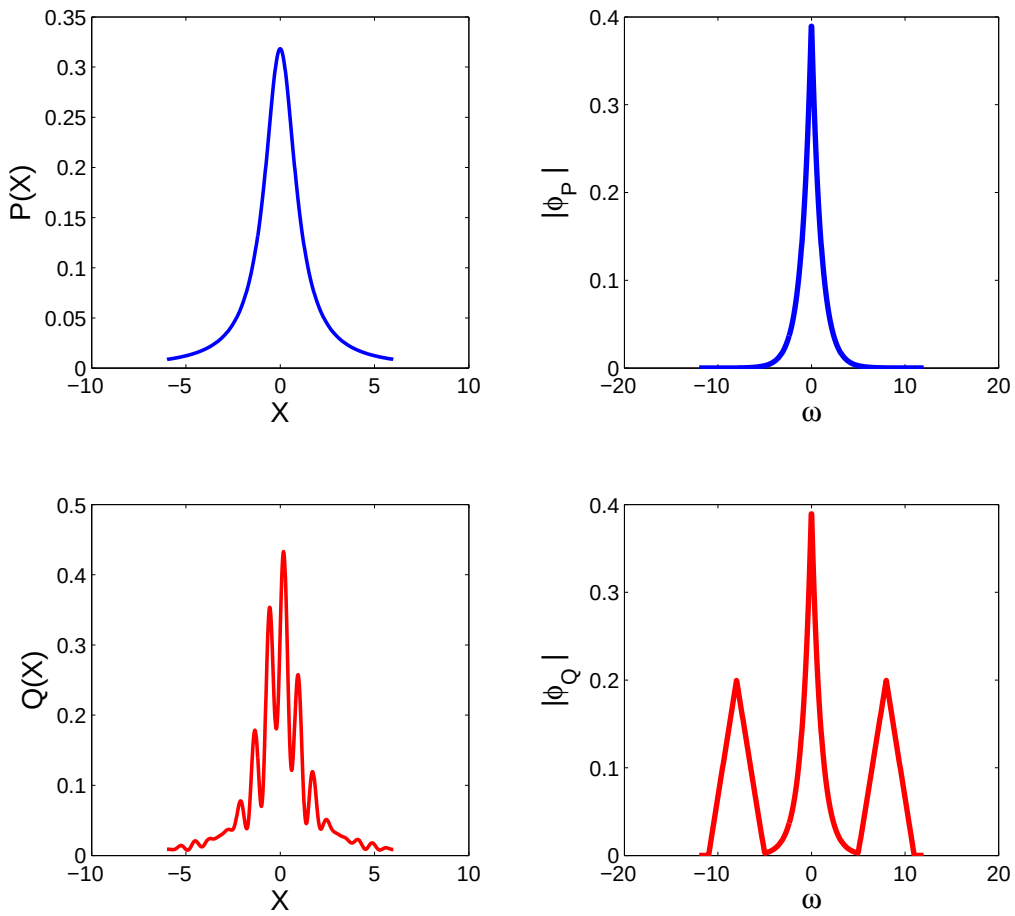


- **Gaussian** kernel
- Difference in distribution spectra

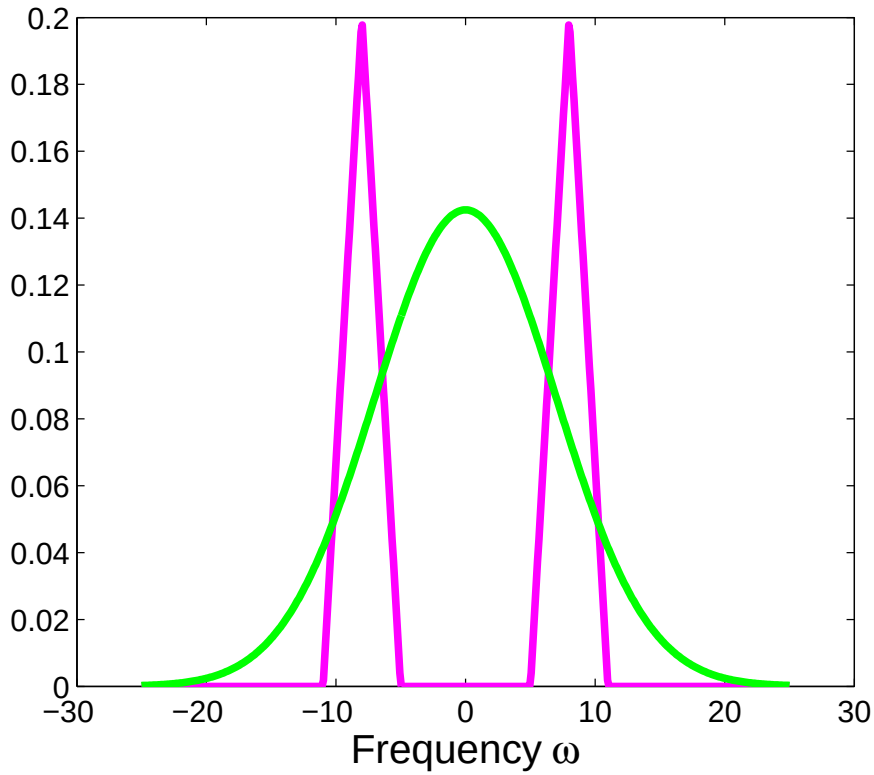


Characteristic Kernels (2)

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

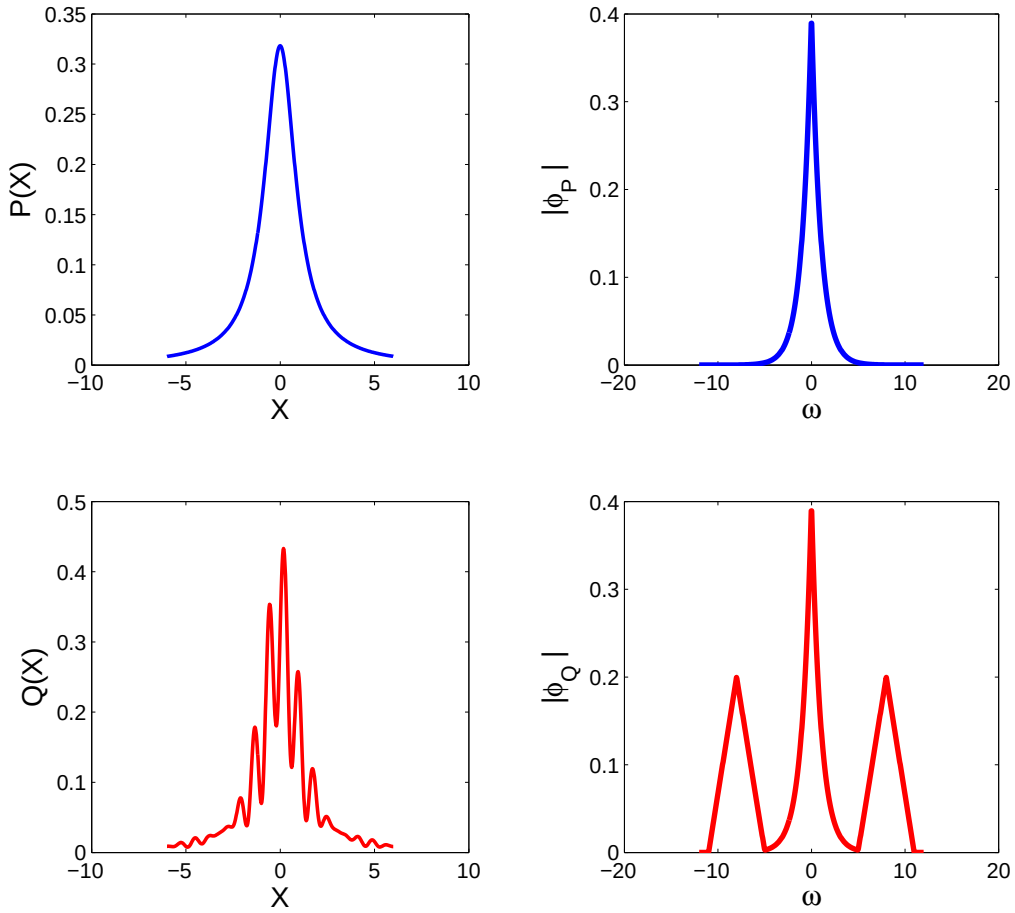


- **Gaussian** kernel characteristic
- Difference in distribution spectra

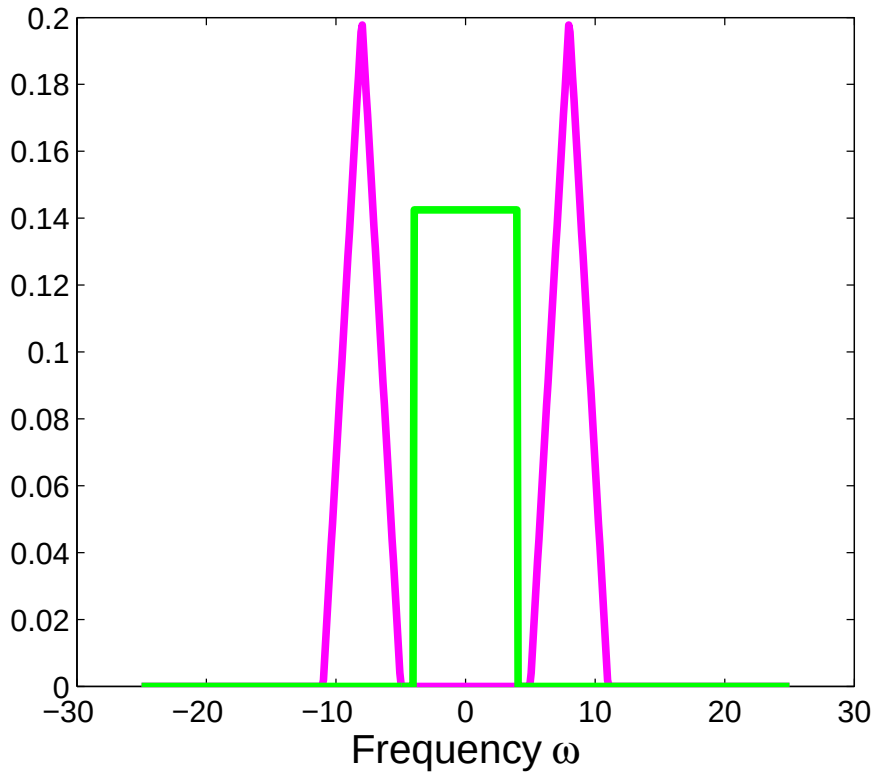


Characteristic Kernels (3)

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

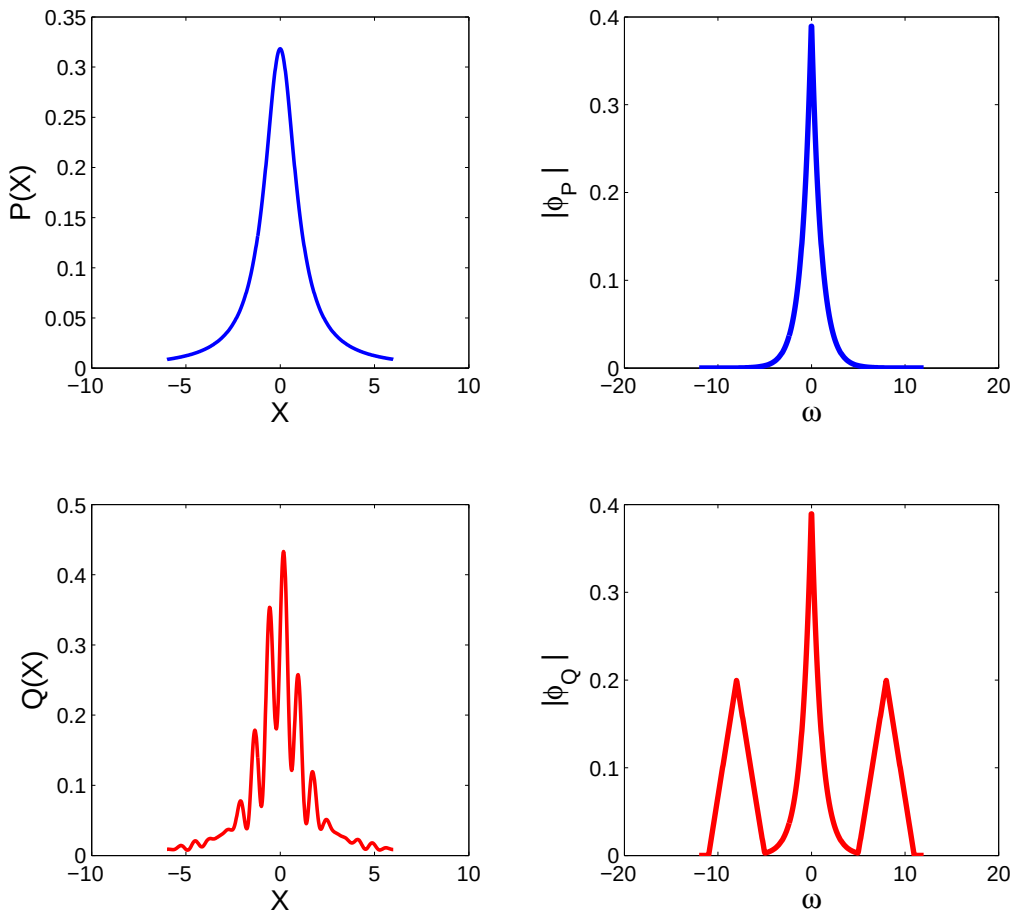


- **Sinc** kernel NOT characteristic
- Difference in distribution spectra

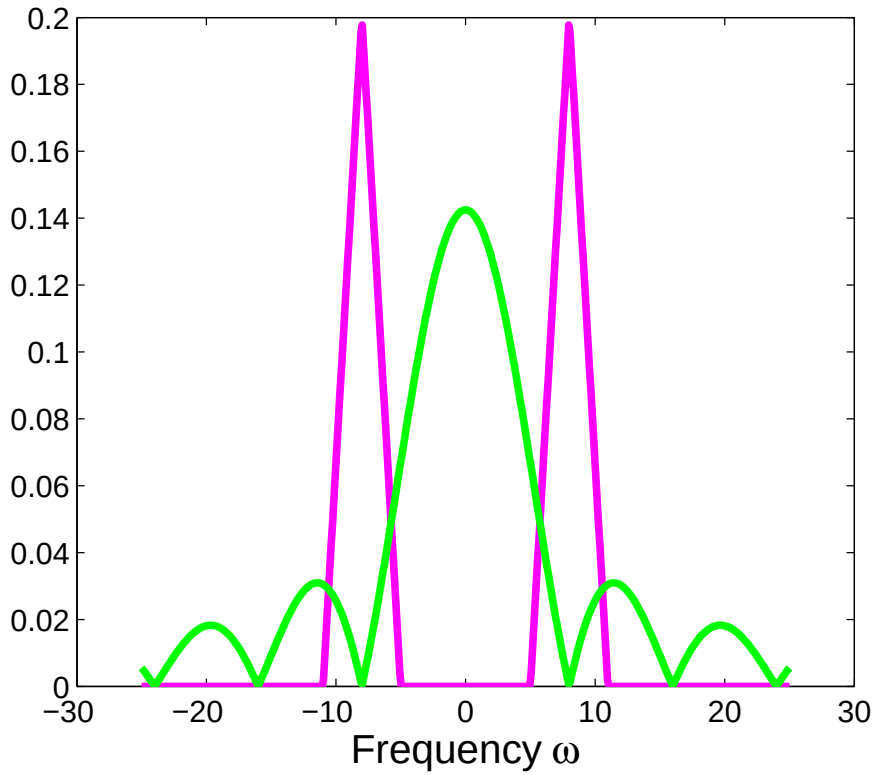


Characteristic Kernels (4)

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

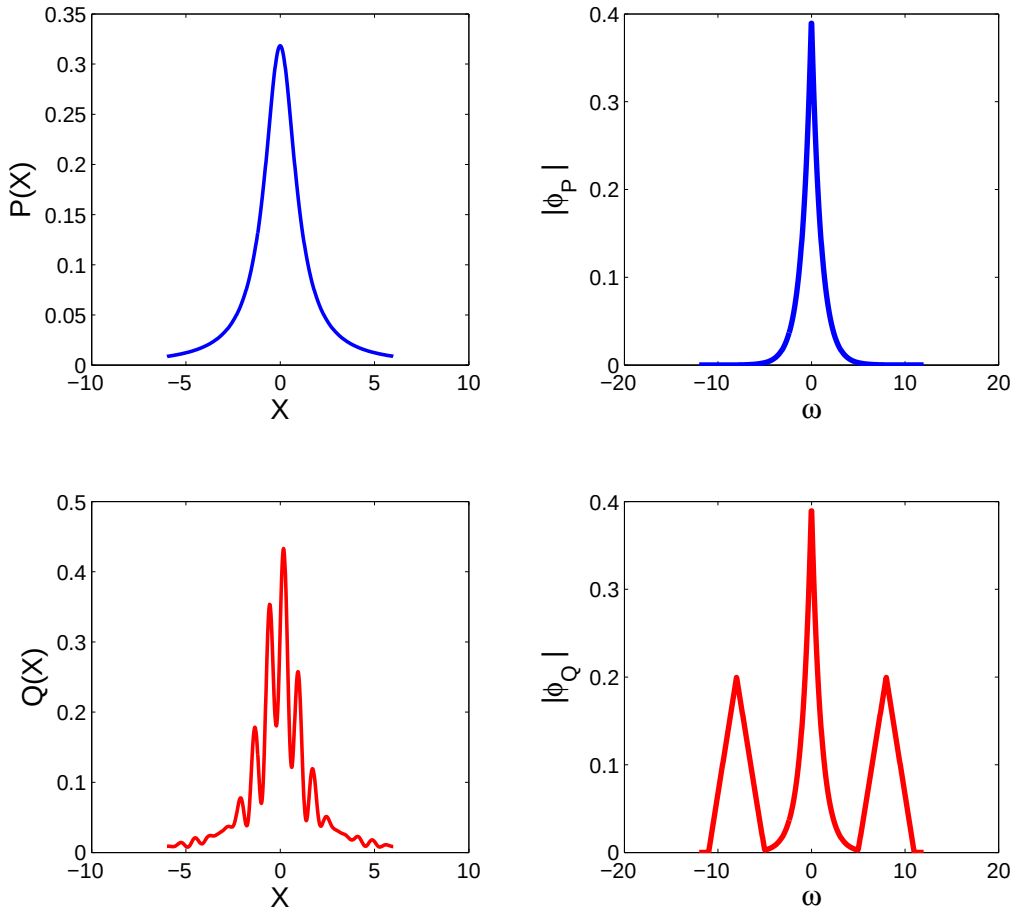


- **B-Spline** kernel ???
- Difference in distribution spectra

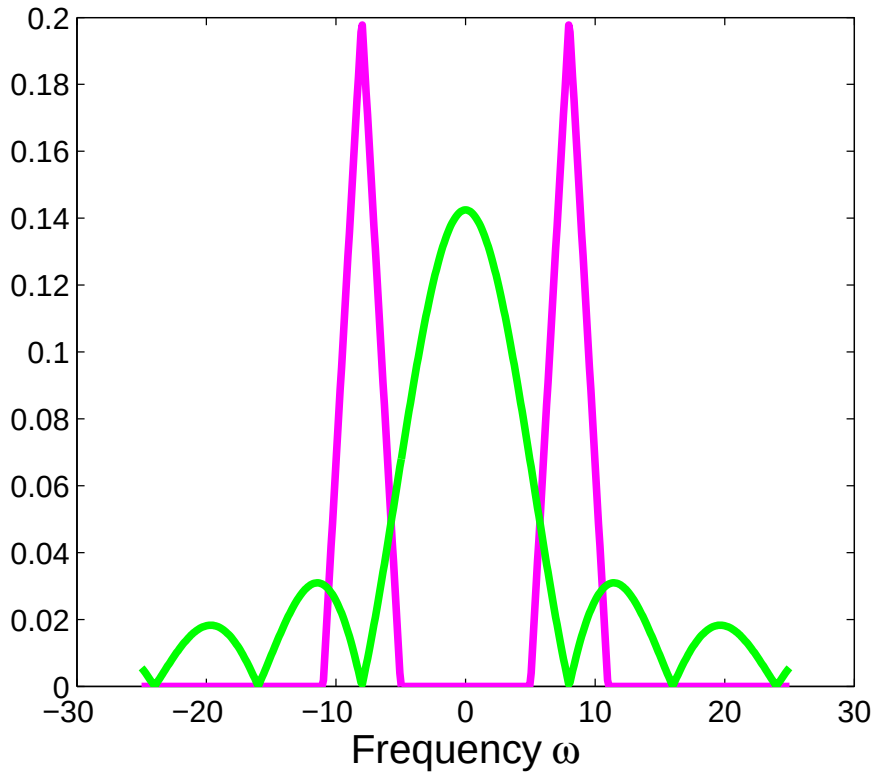


Characteristic Kernels (4)

- 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 $\mathrm{supp}(\Lambda) = \mathbb{R}^d$

Characteristic Kernels (5)

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

Characteristic Kernels (5)

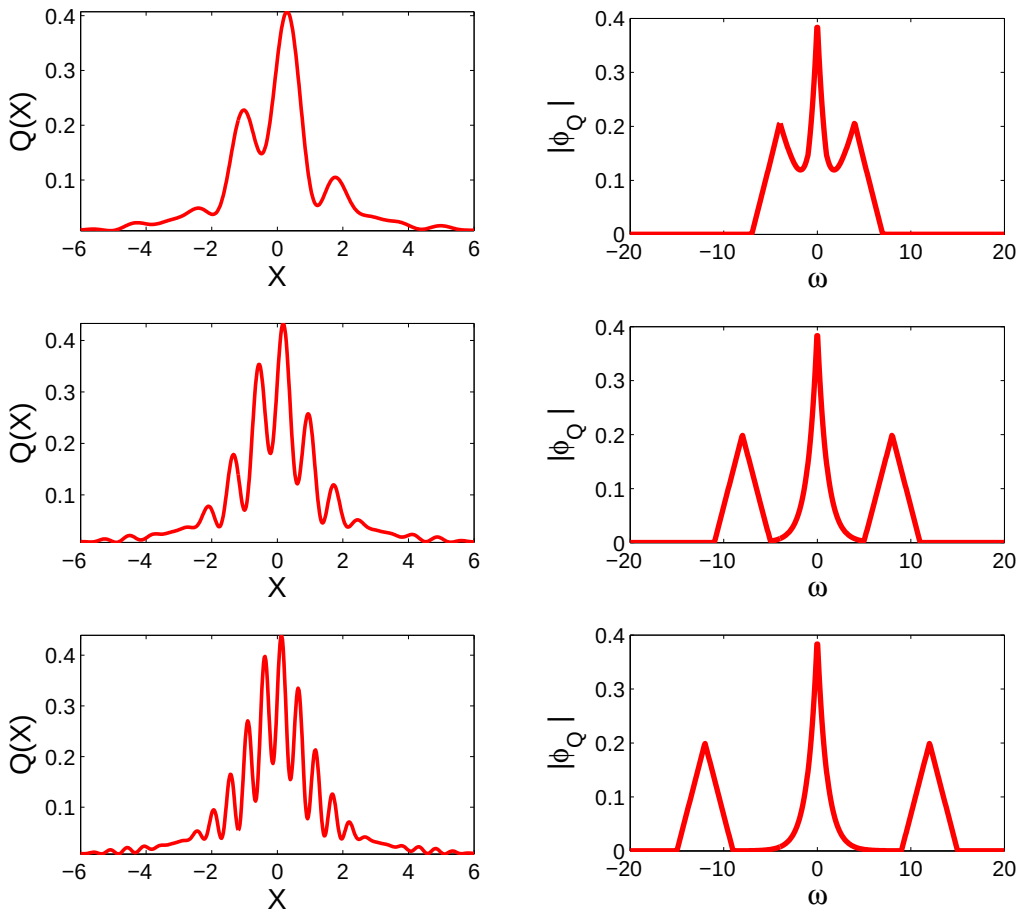
- Main theorem: k characteristic if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$
- Corollary: if k continuous with compact support, then k characteristic
- Another example: universal kernels on compact domains, includes NON-translation invariant [Steinwart, 2001]

Characteristic Kernels (5)

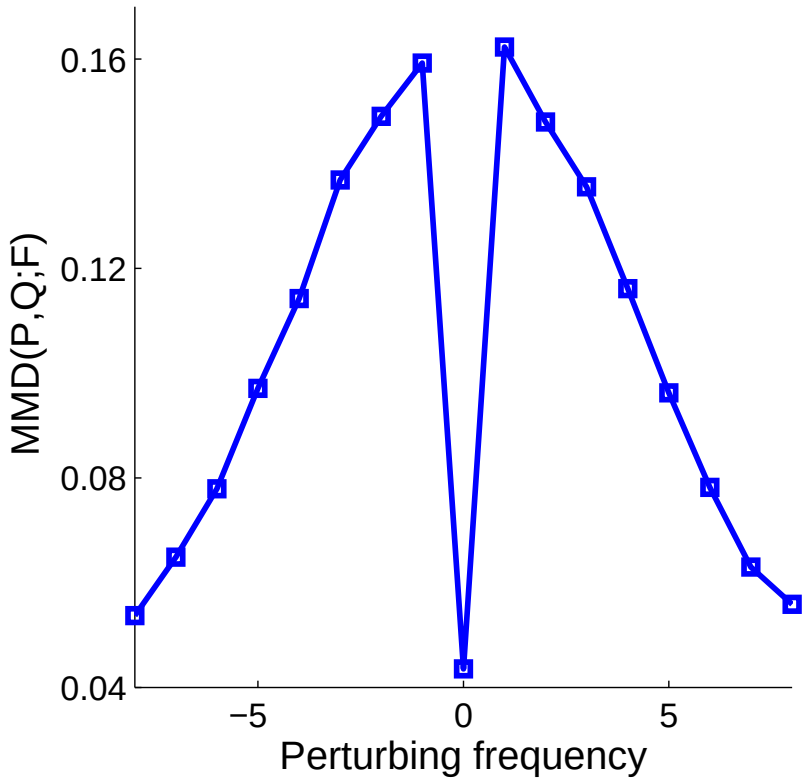
- Main theorem: k characteristic if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$
- Corollary: if k continuous with compact support, then k characteristic
- Another example: universal kernels on compact domains, includes NON-translation invariant [Steinwart, 2001]
- Similar reasoning wherever extensions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

Kernel Choice (1)

- Gaussian kernel example

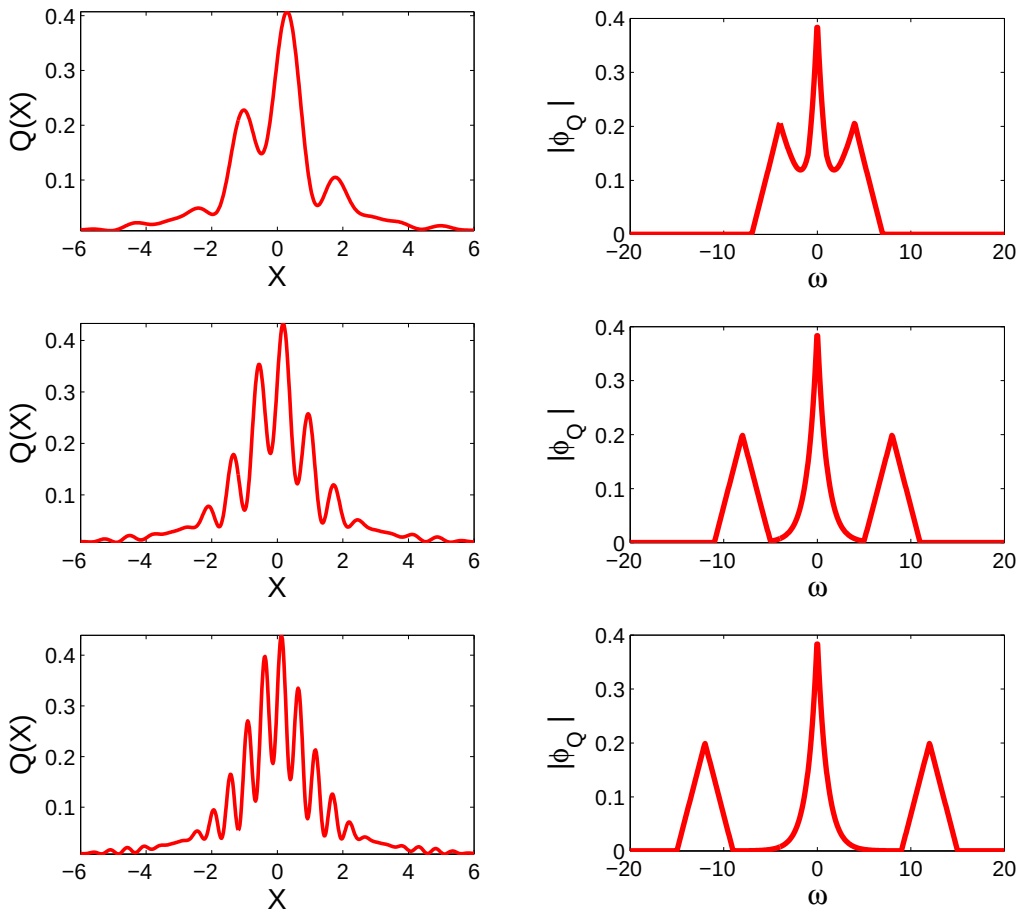


- MMD vs frequency of perturbation to **P**

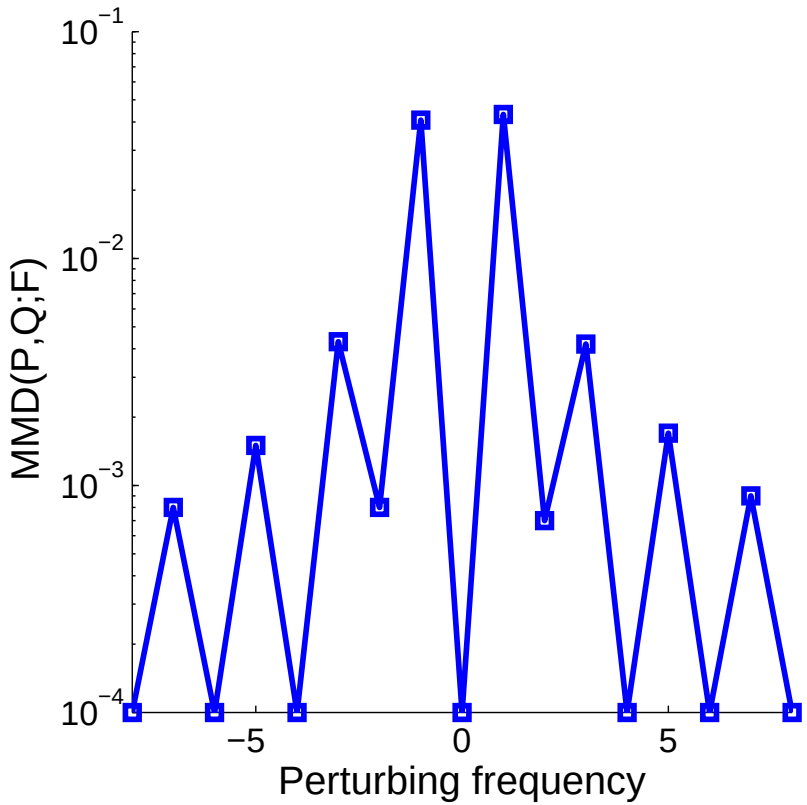


Kernel Choice (2)

- B-spline kernel example



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



Dependence detection using MMD

MMD for Independence (1)

- Independence
  - Determine: Does $\mathbf{P} = \mathbf{P}_{\mathsf{x}}\mathbf{P}_{\mathsf{y}}$?

MMD for Independence (1)

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

MMD for Independence (1)

- Independence
  - Determine: Does $\mathbf{P} = \mathbf{P}_{\mathsf{x}}\mathbf{P}_{\mathsf{y}}$?
- 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)

- $\mathcal{F}$ with kernel $k(x, x')$, $\mathcal{G}$ with kernel $l(y, y')$
- Define the product space $\mathcal{F} \times \mathcal{G}$ with kernel

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

MMD for Independence (2)

- $\mathcal{F}$ with kernel $k(x, x')$, $\mathcal{G}$ with kernel $l(y, y')$
- 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

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

MMD for Independence (2)

- $\mathcal{F}$ with kernel $k(x, x')$, $\mathcal{G}$ with kernel $l(y, y')$
- 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

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

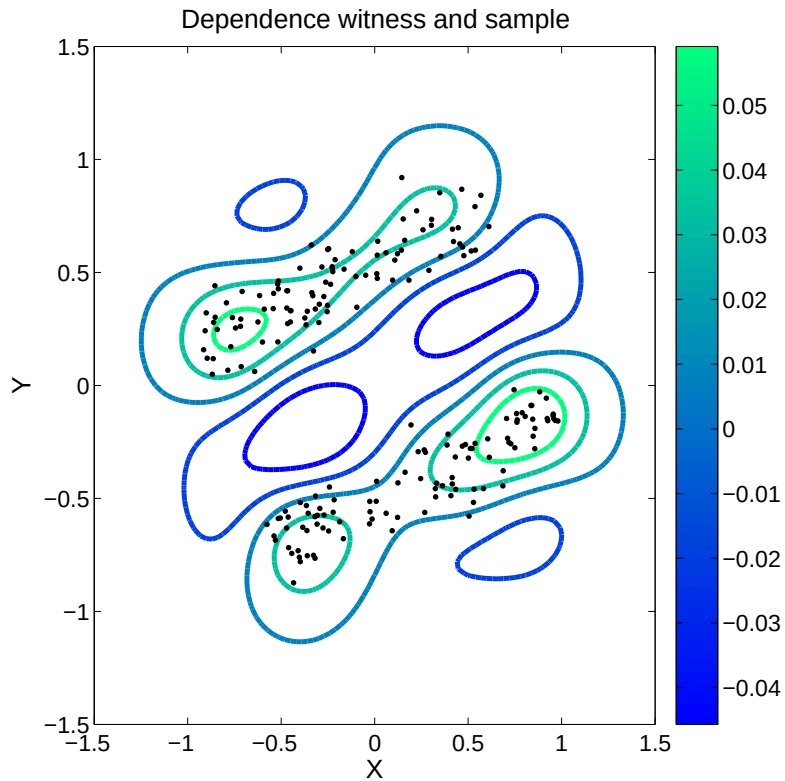
- The squared distance between these two mean elements is

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

MMD for Independence (3)

- Witness function:

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



Covariance in RKHS (1)

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

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

Covariance in RKHS (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 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]

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

- $\mathrm{COCO}(\mathbf{P}; F, G) = 0$ iff 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: $\Sigma_{xy} : \mathcal{G} \to \mathcal{F}$ such that

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

Covariance in RKHS (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 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: $\Sigma_{xy} : \mathcal{G} \to \mathcal{F}$ such that

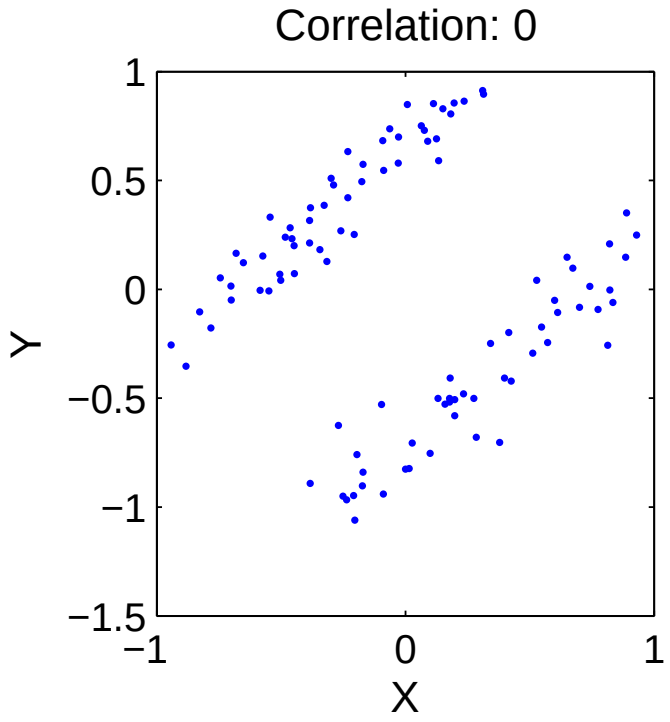
$$\langle f, \Sigma_{xy} g \rangle_{\mathcal{F}} = \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]$$

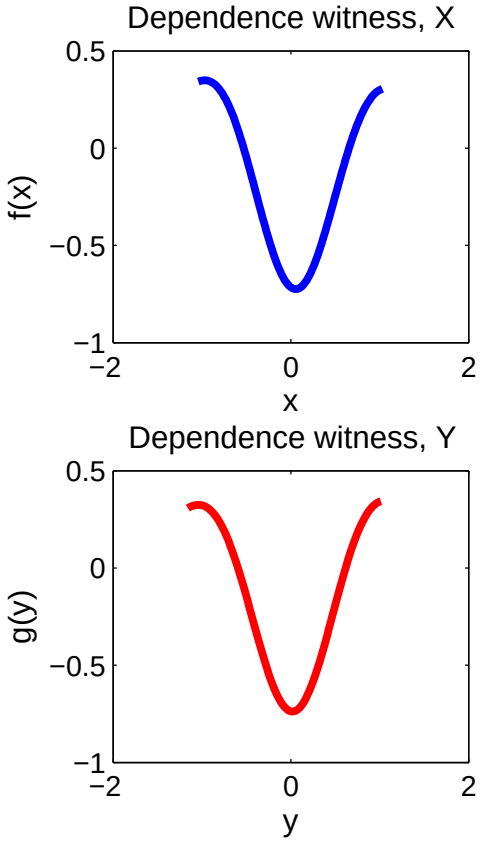
- COCO is the spectral norm of Σ_{xy} [Gretton et al., 2005b,c]:

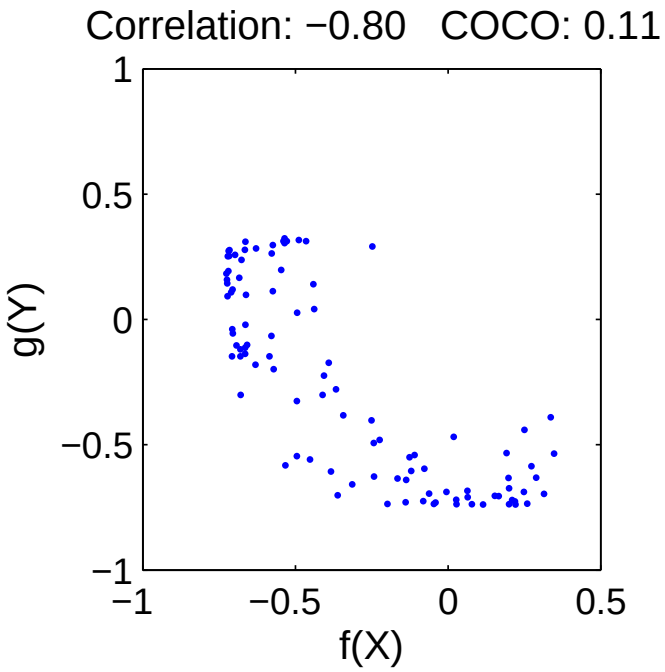
$$\mathrm{COCO}(\mathbf{P}; F, G) := \|\Sigma_{xy}\|_{\mathrm{S}}$$

Covariance in RKHS (2)

- Example: variables with zero correlation
- Can we do better?

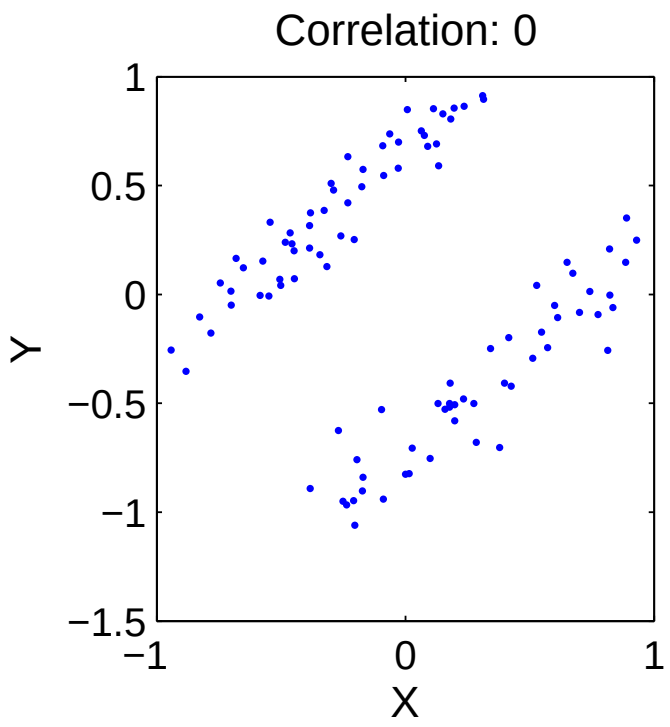


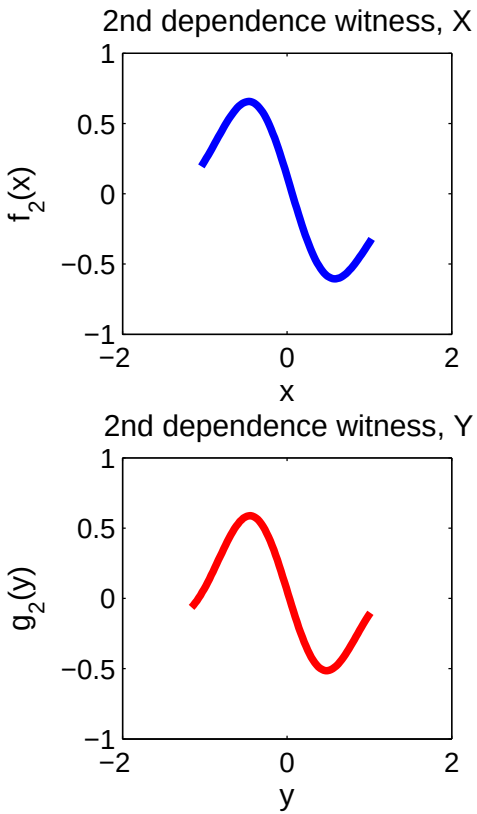




Covariance in RKHS (2)

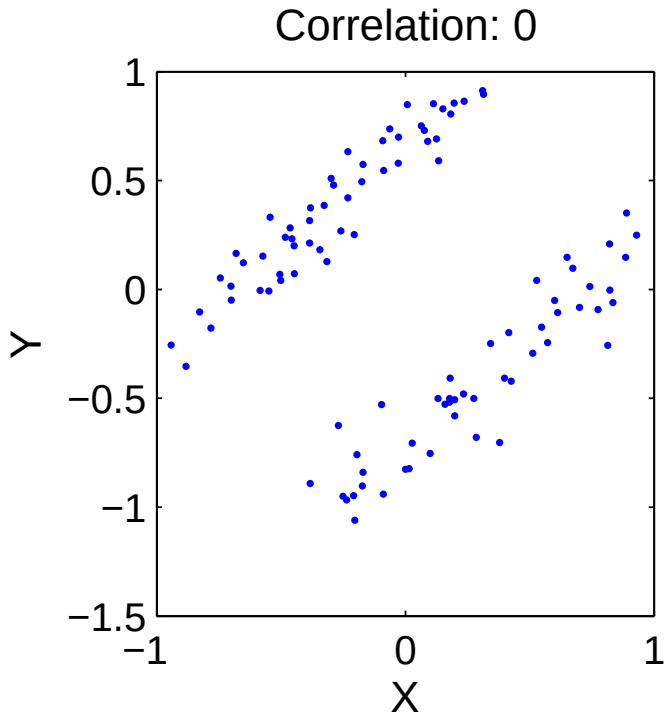
- Example: variables with zero correlation
- Can we do better?

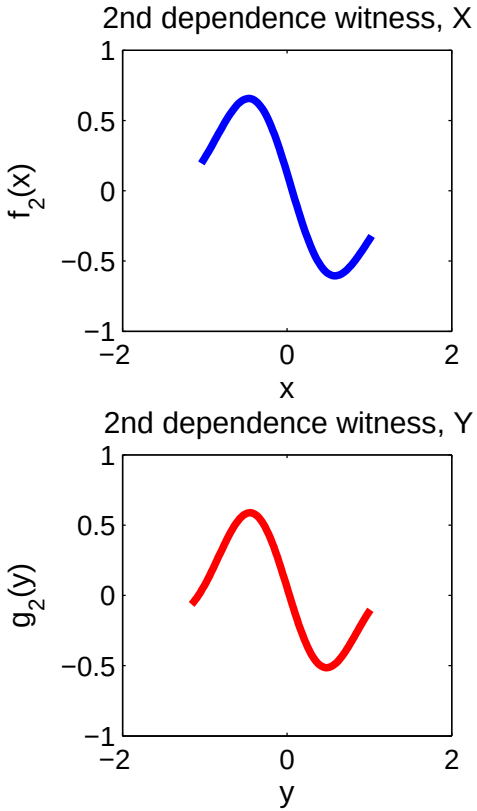


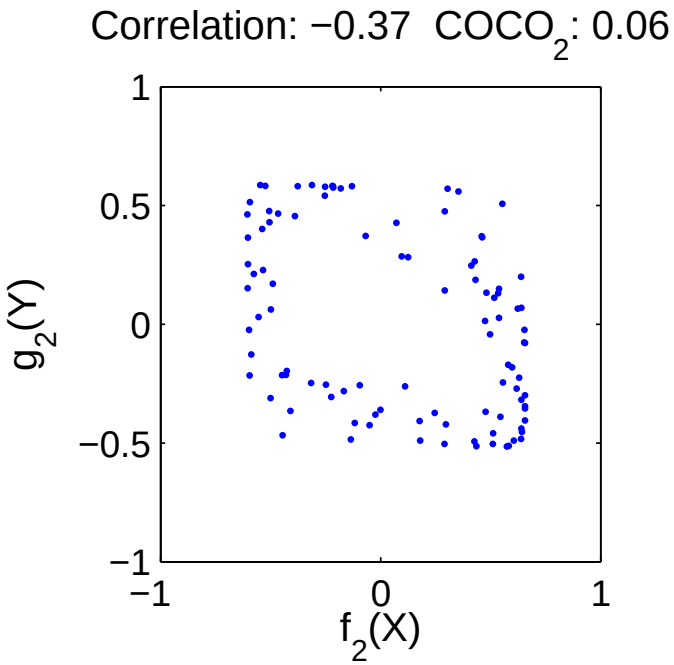


Covariance in RKHS (2)

- Example: variables with zero correlation
- Can we do better?







Hilbert-Schmidt Independence Criterion

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

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

Hilbert-Schmidt Independence Criterion

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

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

- In limit of infinite samples:

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

Hilbert-Schmidt Independence Criterion

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

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

- In limit of infinite samples:

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

- (Biased) empirical HSIC a v-statistic

$$HSIC(\boldsymbol{z}; F, G) := \frac{1}{m^2}\mathrm{trace}(\mathbf{KHLH}) \qquad H = I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$$

End of Part 1

- 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 $\mathbf{P}_{\mathsf{x}}\mathbf{P}_{\mathsf{y}}$
- Next part:

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

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.

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.