Lecture 2: Mappings of Probabilities to RKHS and Applications

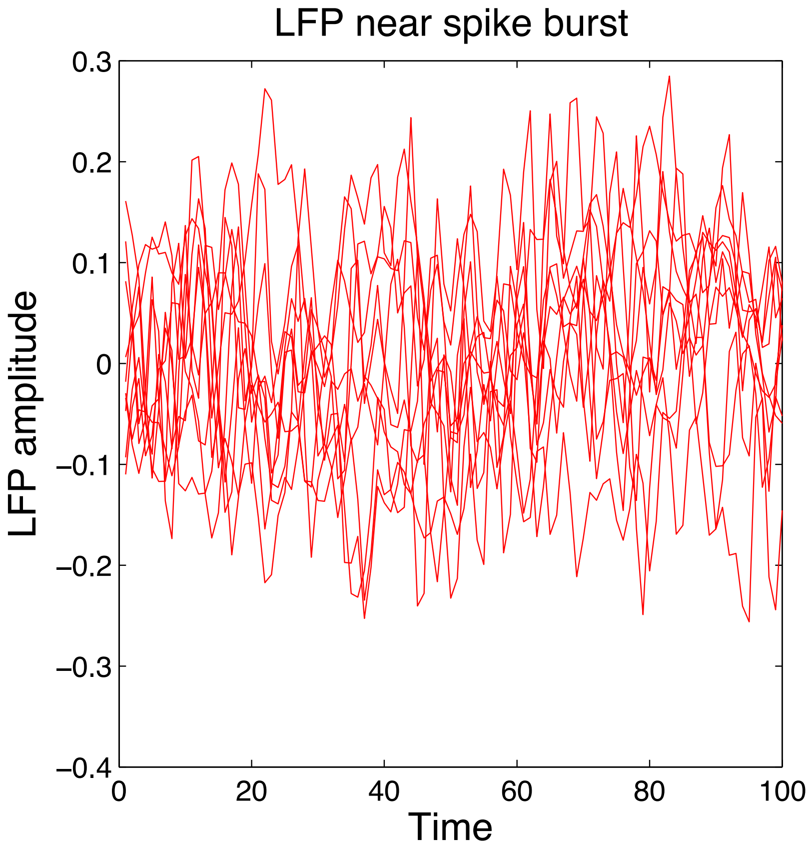
Lille, 2014

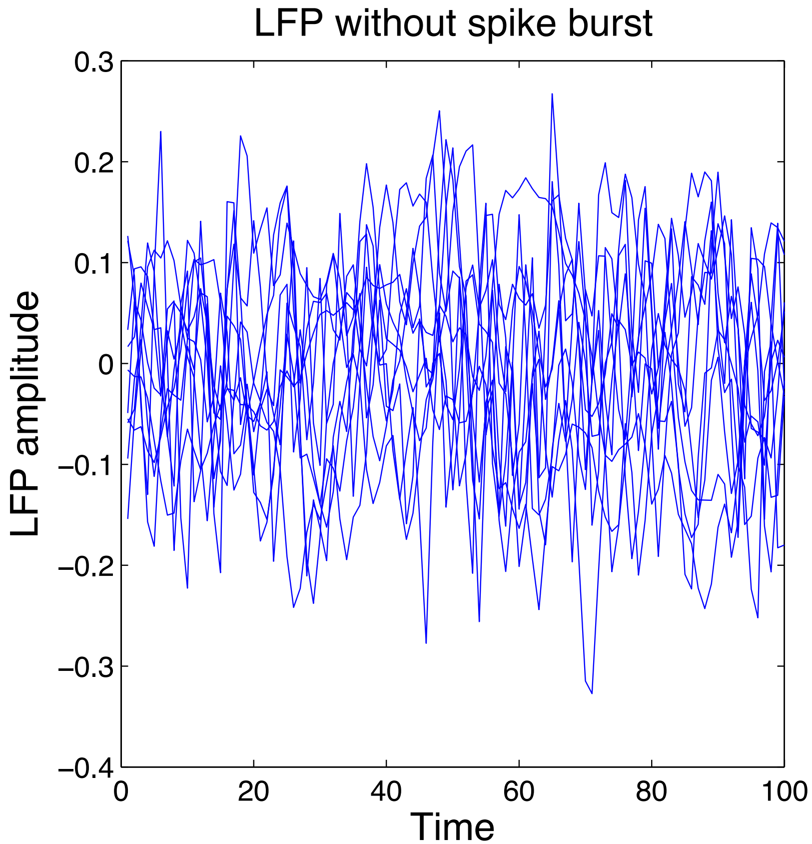
Arthur Gretton

Gatsby Unit, CSML, UCL

Detecting differences in brain signals

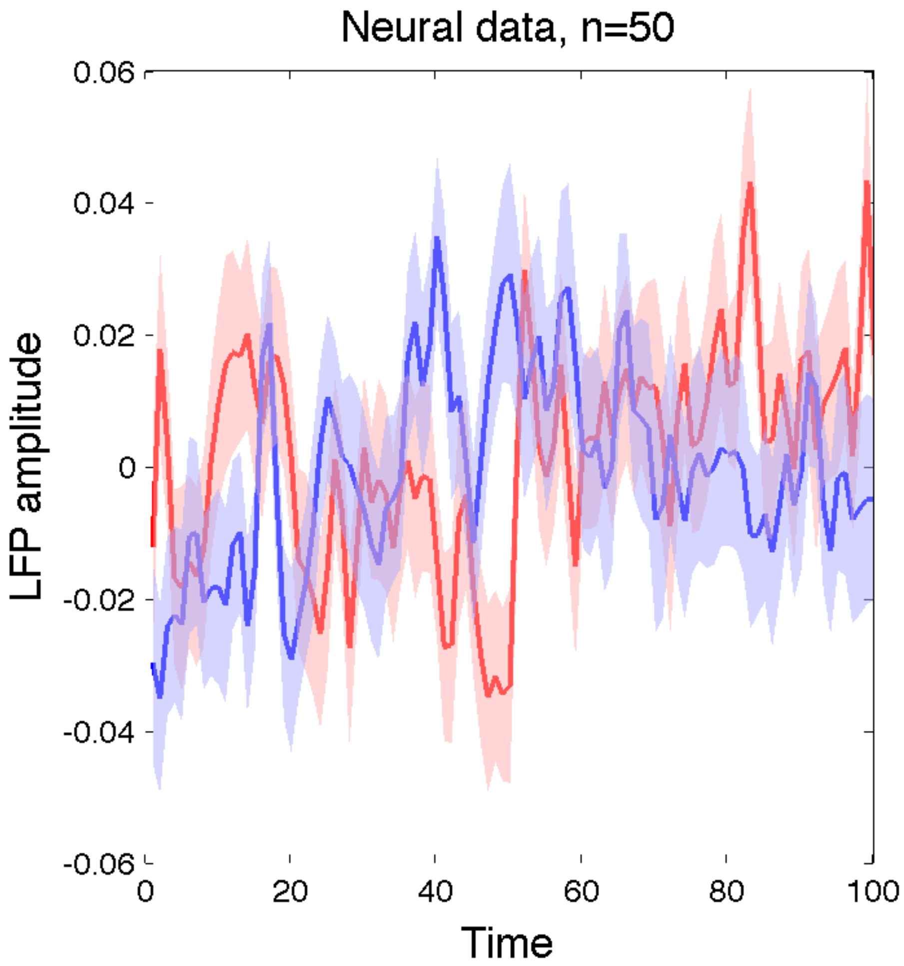
The problem: Do local field potential (LFP) signals change when measured near a spike burst?





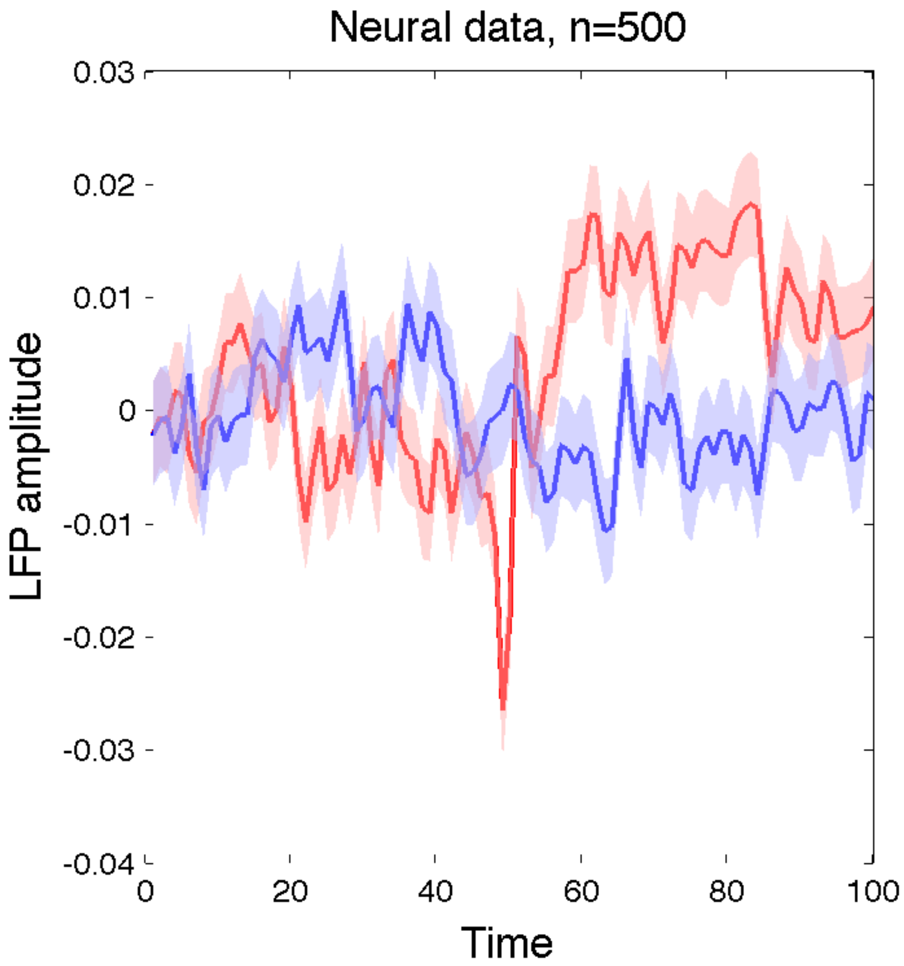
Detecting differences in brain signals

The problem: Do local field potential (LFP) signals change when measured near a spike burst?



Detecting differences in brain signals

The problem: Do local field potential (LFP) signals change when measured near a spike burst?



Detecting statistical dependence

- How do you detect dependence…
- ….in a discrete domain? [Read and Cressie, 1988]

Detecting statistical dependence

- How do you detect dependence. . .
- . . .in a discrete domain? [Read and Cressie, 1988]

Detecting statistical dependence

- How do you detect dependence...
- ...in a discrete domain? [Read and Cressie, 1988]

P(A,T)	On time	Late
Alarm	0.27	0.03
No alarm	0.07	0.63

Detecting statistical dependence

- How do you detect dependence. . .
- . . .in a discrete domain? [Read and Cressie, 1988]

P(A,T)	On time	Late
Alarm	0.10	0.20
No alarm	0.24	0.46

Detecting statistical dependence

- How do you detect dependence...
- ...in a discrete domain? [Read and Cressie, 1988]

X_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

X_2: No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

...

$\overset{?}{\Longleftrightarrow}$

Y_1: Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat et concerne l'aide financiére qu'on a annoncée pour les agriculteurs. La plupart des agriculteurs n'ont encore rien reu de cet argent.

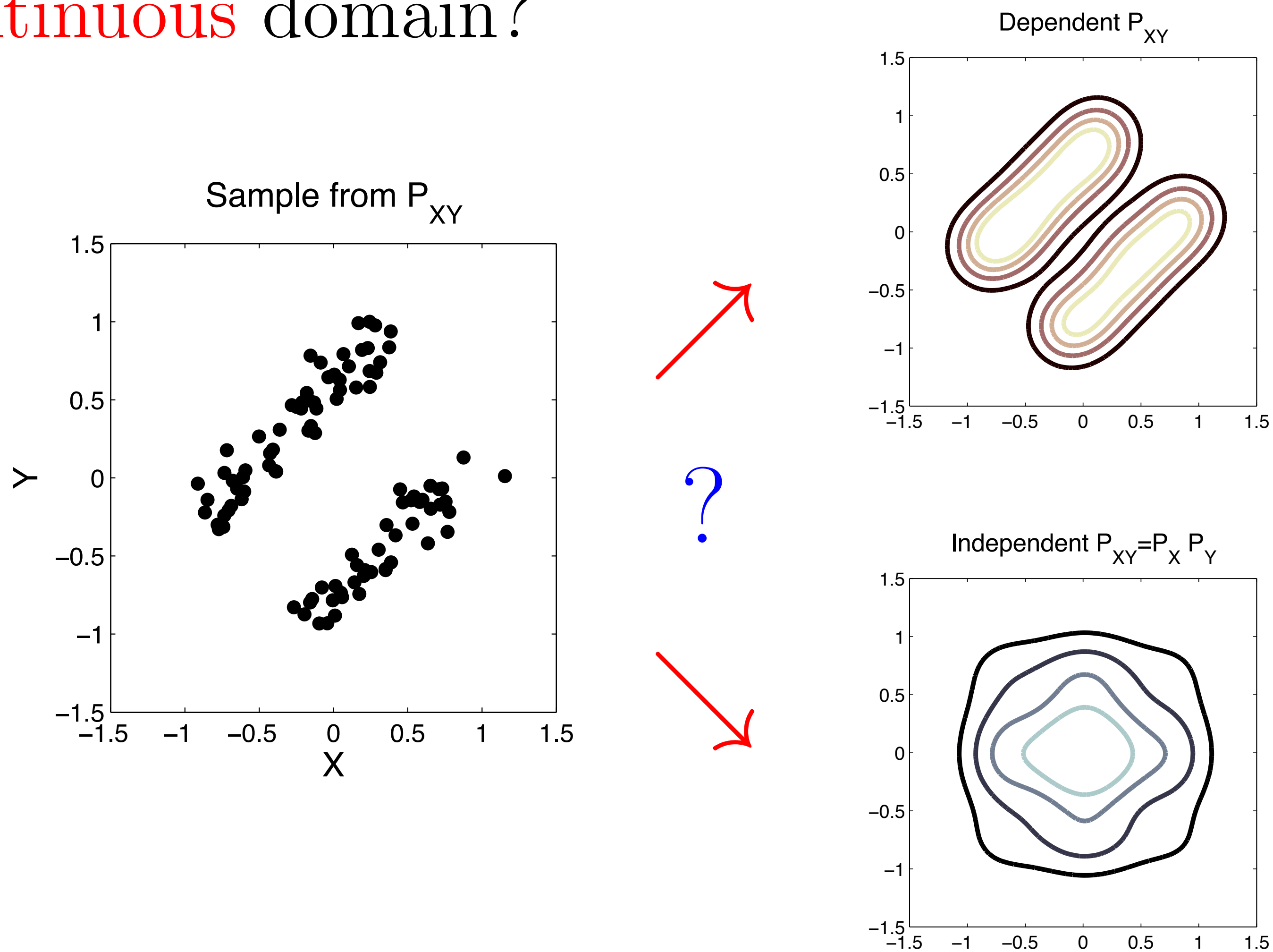
Y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

...

Are the French text extracts translations of the English ones?

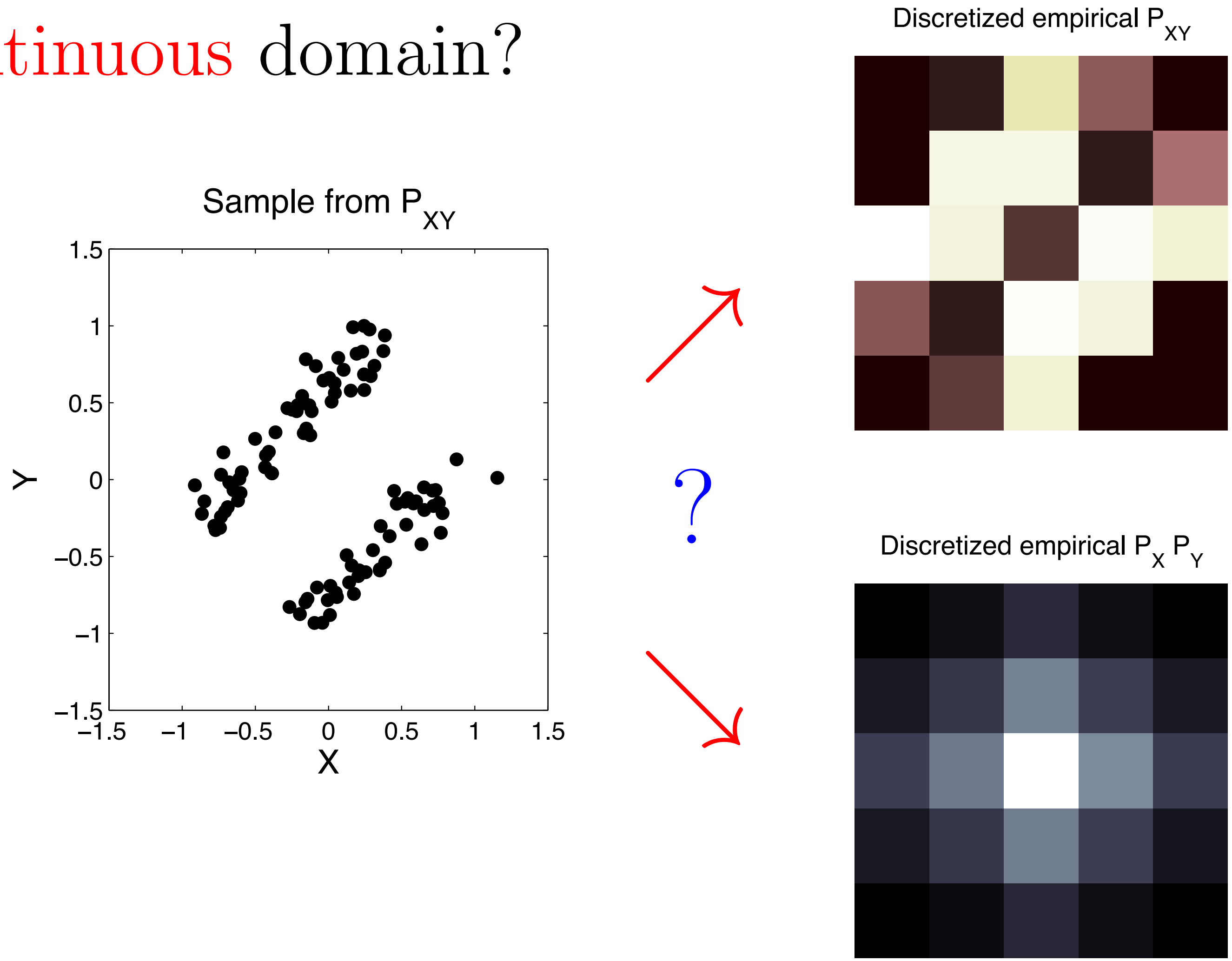
Detecting statistical dependence, continuous domains

- How do you detect dependence…
- …in a continuous domain?



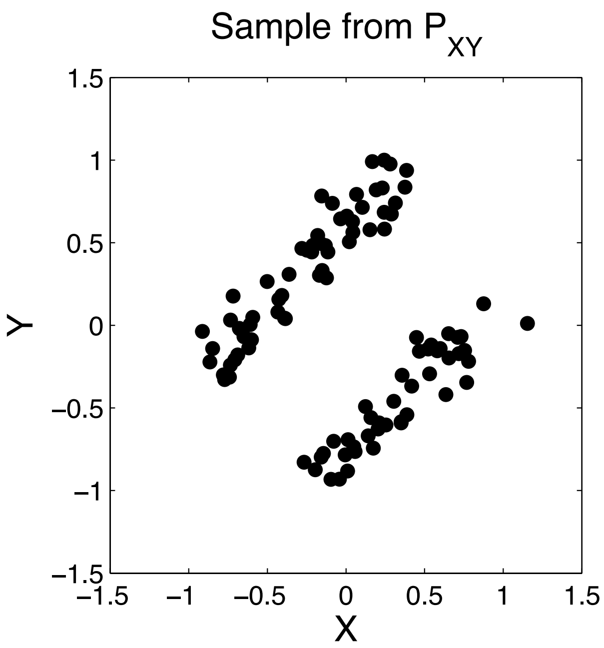
Detecting statistical dependence, continuous domains

- How do you detect dependence...
- ...in a continuous domain?

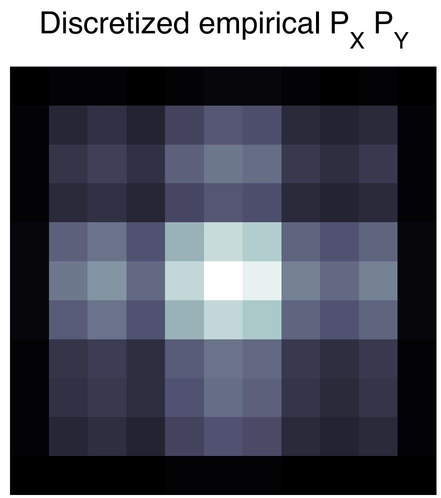


Detecting statistical dependence, continuous domains

- How do you detect dependence. . .
- . . .in a continuous domain?



?



Detecting statistical dependence, continuous domains

- How do you detect dependence...
- ...in a continuous domain?
- Problem: fails even in "low" dimensions! [NIPS07a, ALT08]
  - X and Y in $\mathbb{R}^4$, statistic=Power divergence, samples= 1024, cases where dependence detected=0/500
- Too few points per bin

Detecting statistical dependence, continuous domains

- How do you detect dependence...
- ...in a continuous domain?
- Problem: fails even in "low" dimensions! [NIPS07a, ALT08]
  - X and Y in $\mathbb{R}^4$, statistic=Power divergence, samples$=1024$, cases where dependence detected$=0/500$
- Too few points per bin

Can we represent and compare distributions in high dimensions?

Further motivating questions

- Compare distributions with high dimension/ low sample size/ “complex” structure
  - Microarray data (aggregation problem)
  - Neuroscience: naturalistic stimulus, complex response
  - Images and text on web (kernels on structured data)

Outline

- Kernel metric on the space of probability measures
  - Function revealing differences in distributions
  - Distance between means in space of features (RKHS)
- Characteristic kernels: feature space mappings of probabilities unique

Outline

- Kernel metric on the space of probability measures
  - Function revealing differences in distributions
  - Distance between means in space of features (RKHS)
- Characteristic kernels: feature space mappings of probabilities unique
- Dependence detection
  - Covariance and Correlation in feature space

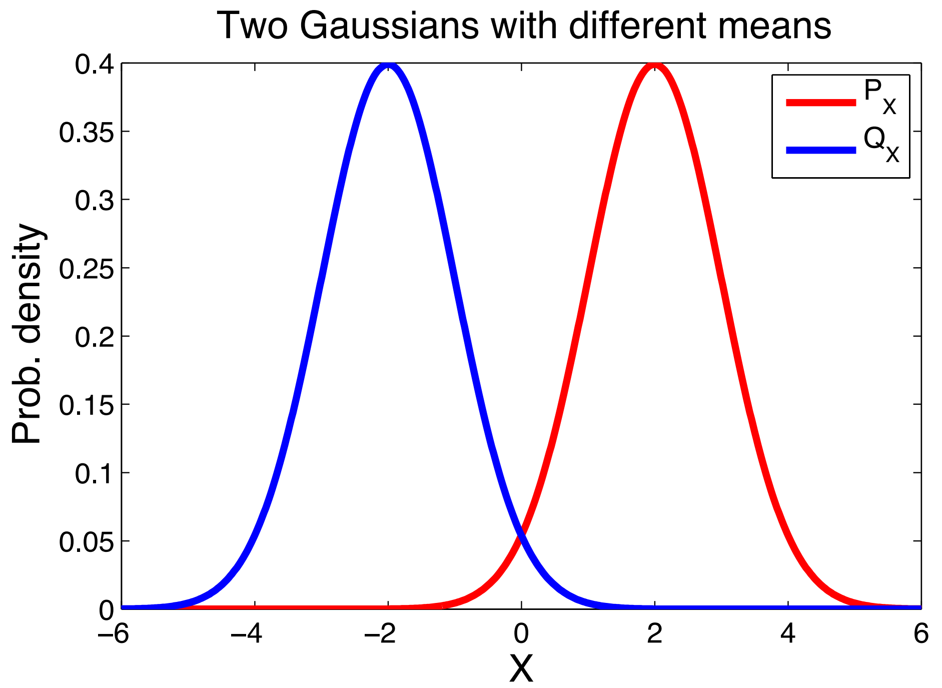
Outline

- Kernel metric on the space of probability measures
  - Function revealing differences in distributions
  - Distance between means in space of features (RKHS)
- Characteristic kernels: feature space mappings of probabilities unique
- Dependence detection
  - Covariance and Correlation in feature space
- Advanced topics
  - Testing for big data
  - Bayesian inference without models
  - Three way interactions
  - Energy distance/distance covariance is special case of RKHS distances

Kernel distance between distributions

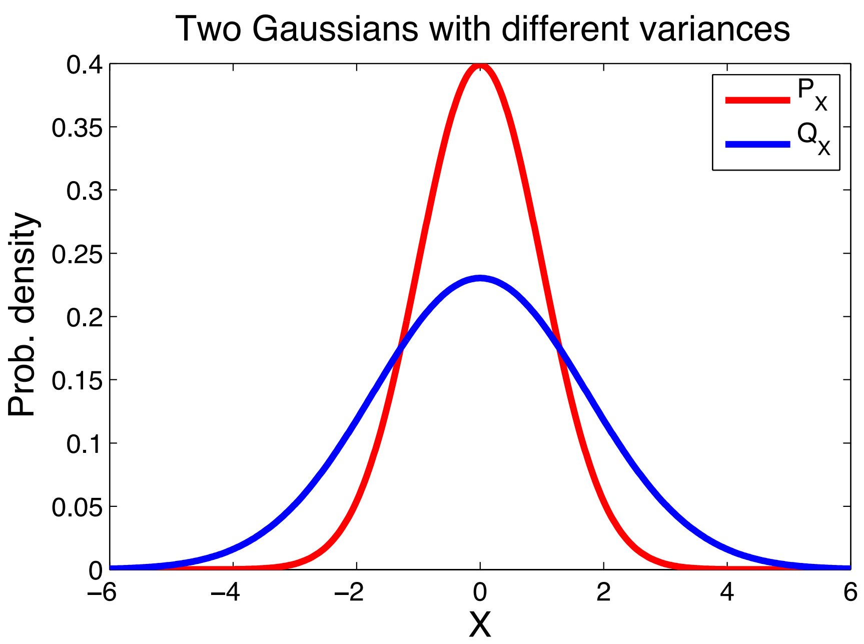
Feature mean difference

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



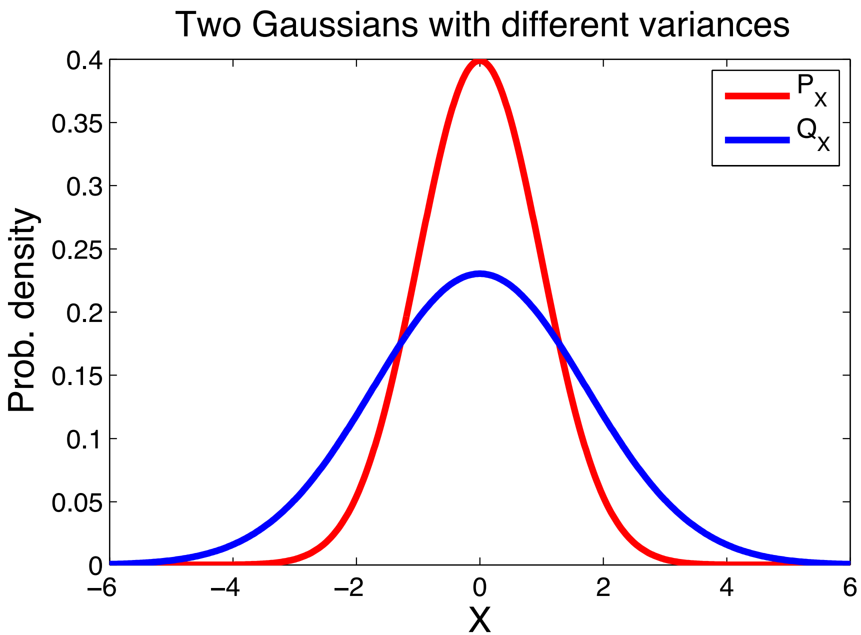
Feature mean difference

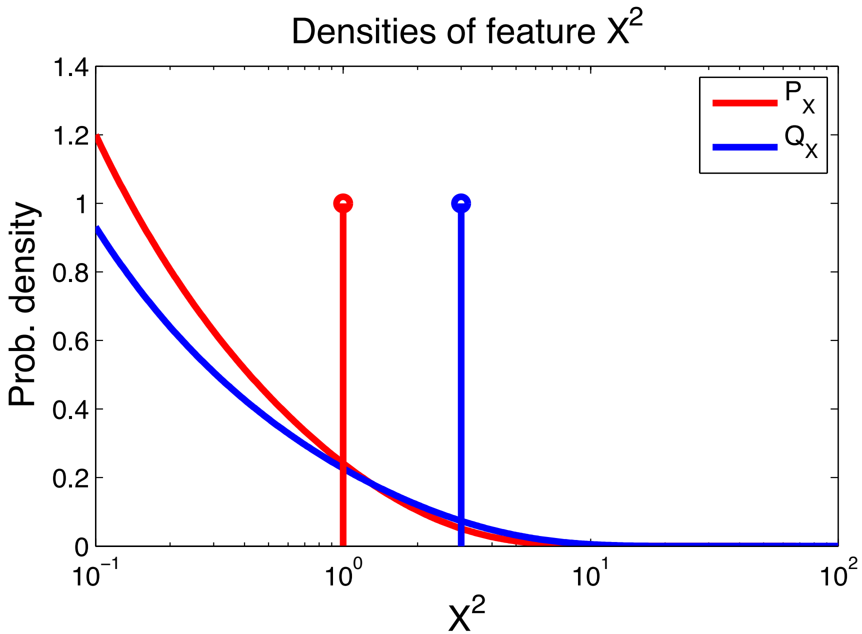
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

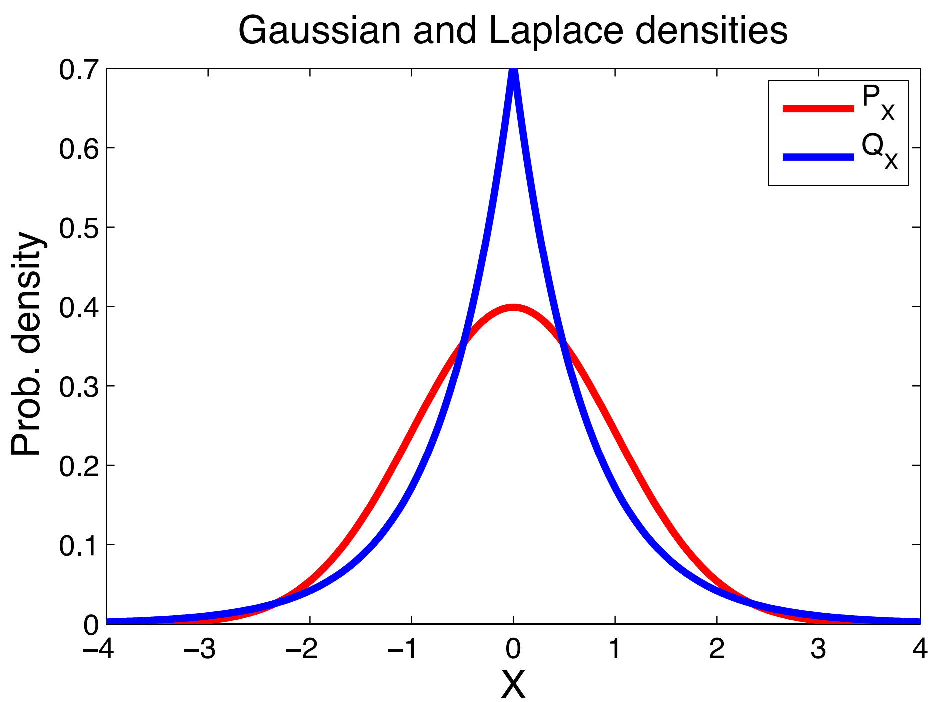
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$





Feature mean difference

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



Reminder: feature maps and the RKHS

- Feature map of $x \in \mathbb{R}^2$, written φ_x

$$\varphi^{(p)}(x) = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad \varphi^{(g)}(x) = \left[\ldots \sqrt{\lambda_i} e_i(x) \ldots\right] \in \ell_2$$

Reminder: feature maps and the RKHS

- Feature map of $x \in \mathbb{R}^2$, written φ_x

$$\varphi^{(p)}(x) = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad \varphi^{(g)}(x) = \left[\ldots \sqrt{\lambda_i} e_i(x) \ldots \right] \in \ell_2$$

- Inner product between feature maps:

$$\left\langle \varphi^{(p)}(x), \varphi^{(p)}(y) \right\rangle_{\mathcal{F}} = \langle x, y \rangle^2 \qquad \left\langle \varphi^{(g)}(x), \varphi^{(g)}(y) \right\rangle_{\mathcal{F}} = \exp\left(-\lambda \left\| x - y \right\|^2 \right)$$

Reminder: feature maps and the RKHS

- Feature map of $x \in \mathbb{R}^2$, written φ_x

$$\varphi^{(p)}(x) = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad \varphi^{(g)}(x) = \left[\ldots \sqrt{\lambda_i} e_i(x) \ldots \right] \in \ell_2$$

- Inner product between feature maps:

$$\left\langle \varphi^{(p)}(x), \varphi^{(p)}(y) \right\rangle_{\mathcal{F}} = \langle x, y \rangle^2 \qquad \left\langle \varphi^{(g)}(x), \varphi^{(g)}(y) \right\rangle_{\mathcal{F}} = \exp\left(-\lambda \|x - y\|^2 \right)$$

- In general,

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

for positive definite $k(x, y)$

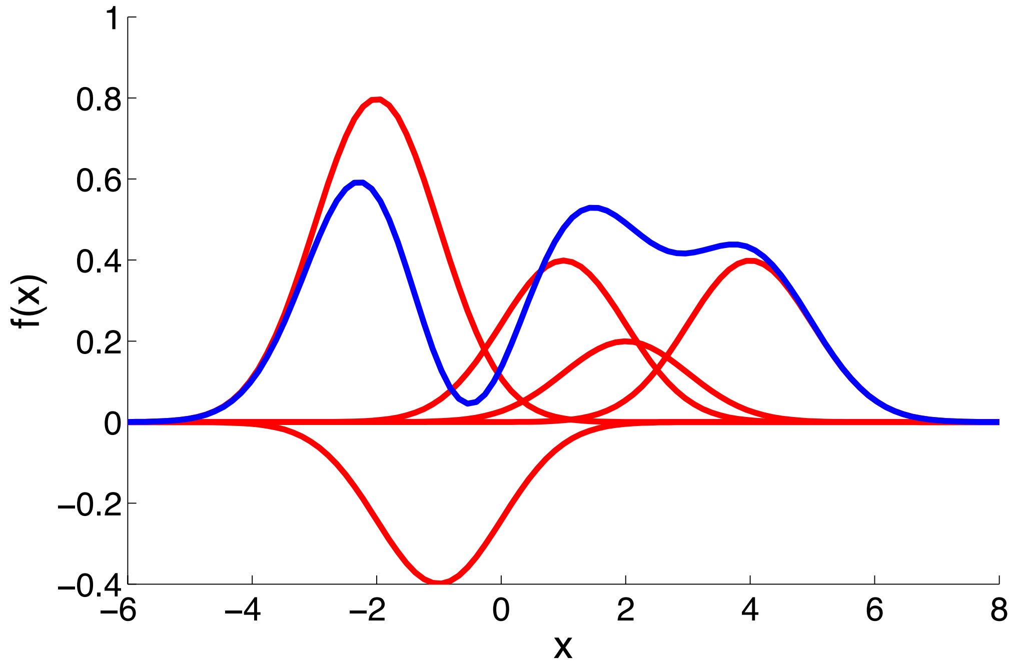
$$\boxed{\text{Kernels are inner products of feature maps}}$$

Reminder: feature view of RKHS functions

- Reproducing property:

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m} \alpha_i \langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{F}} = \langle f(\cdot), \varphi(x) \rangle_{\mathcal{F}}$$

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i \varphi(x_i)$$



The mean in feature space

For finite dimensional feature spaces, we can define expectations in terms of inner products.

$$\phi(x) = k(\cdot, x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \qquad f(\cdot) = \begin{bmatrix} a \\ b \end{bmatrix}$$

Then

$$f(x) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} = \langle f(\cdot), \phi(x) \rangle_{\mathcal{F}} .$$

The mean in feature space

For finite dimensional feature spaces, we can define expectations in terms of inner products.

$$\phi(x) = k(\cdot, x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \qquad f(\cdot) = \begin{bmatrix} a \\ b \end{bmatrix}$$

Then

$$f(x) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} = \langle f(\cdot), \phi(x) \rangle_{\mathcal{F}} .$$

Consider random variable $\mathsf{x} \sim \mathbf{P}$

$$\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) = \mathbf{E}_{\mathbf{P}} \left(\begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathsf{x} \\ \mathsf{x}^2 \end{bmatrix} \right) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathbf{E}_{\mathbf{P}} \mathsf{x} \\ \mathbf{E}_{\mathbf{P}} (\mathsf{x}^2) \end{bmatrix} =: \begin{bmatrix} a \\ b \end{bmatrix}^\top \mu_{\mathbf{P}} .$$

Does this reasoning translate to infinite dimensions?

Does the feature space mean exist?

Does there exist an element $\mu_{\mathsf{P}} \in \mathcal{F}$ such that

$$\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) = \langle f(\cdot), \mu_{\mathsf{P}}(\cdot) \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

Does the feature space mean exist?

Does there exist an element $\mu_{\mathsf{P}} \in \mathcal{F}$ such that

$$\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) = \langle f(\cdot), \mu_{\mathsf{P}}(\cdot) \rangle_{\mathcal{F}} \qquad \forall f \in \mathcal{F}$$

Recall the concept of bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when $\forall f \in \mathcal{F}$,

$$|Af| \leq \lambda_A \|f\|_{\mathcal{F}}.$$

Riesz representation theorem: In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,

$$Af = \langle f(\cdot), g_A(\cdot) \rangle_{\mathcal{F}}$$

Does the feature space mean exist?

Existence of mean embedding: If $\mathbf{E}_{\mathsf{P}}\sqrt{k(\mathsf{x},\mathsf{x})} < \infty$ then $\mu_{\mathsf{P}} \in \mathcal{F}$.

Proof:

The linear operator $T_{\mathsf{P}}f := \mathbf{E}_{\mathsf{P}}f(\mathsf{x})$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$|T_{\mathsf{P}}f| \leq |\mathbf{E}_{\mathsf{P}}f(\mathsf{x})| \leq \mathbf{E}_{\mathsf{P}}|f(\mathsf{x})| = \mathbf{E}_{\mathsf{P}}|\langle f(\cdot), \phi(\mathsf{x})\rangle_{\mathcal{F}}| \leq \mathbf{E}_{\mathsf{P}}\left(\sqrt{k(\mathsf{x},\mathsf{x})}\,\|f\|_{\mathcal{F}}\right).$$

Hence by Riesz (with $\lambda_{T_{\mathsf{P}}} = \mathbf{E}_{\mathsf{P}}\sqrt{k(\mathsf{x},\mathsf{x})}$), $\exists \mu_{\mathsf{P}} \in \mathcal{F}$ such that

$$T_{\mathsf{P}}f = \langle f(\cdot), \mu_{\mathsf{P}}(\cdot)\rangle_{\mathcal{F}}.$$

$\mu_{\mathbf{P}}$ is feature map of probability

Embedding of **P** to feature space

- Mean embedding $\mu_{\mathbf{P}} \in \mathcal{F}$

$$\langle \mu_{\mathbf{P}}(\cdot), f(\cdot)\rangle_{\mathcal{F}} = E_{\mathbf{P}} f(\mathsf{x}).$$

- What does prob. feature map look like?

$$\mu_{\mathbf{P}}(x) = \langle \mu_{\mathbf{P}}(\cdot), \varphi(x)\rangle_{\mathcal{F}}$$

$$= \langle \mu_{\mathbf{P}}(\cdot), k(\cdot, x)\rangle_{\mathcal{F}} = E_{\mathbf{P}} k(\mathsf{x}, x).$$

Expectation of kernel!

- Empirical estimate:

$$\hat{\mu}_{\mathbf{P}}(x) = \frac{1}{m}\sum_{i=1}^{m} k(x_i, x) \qquad x_i \sim \mathbf{P}$$

$\mu_{\mathbf{P}}$ is feature map of probability

Embedding of $\mathbf{P}$ to feature space

- Mean embedding $\mu_{\mathbf{P}} \in \mathcal{F}$

$$\langle \mu_{\mathbf{P}}(\cdot), f(\cdot) \rangle_{\mathcal{F}} = E_{\mathbf{P}} f(\mathsf{x}).$$

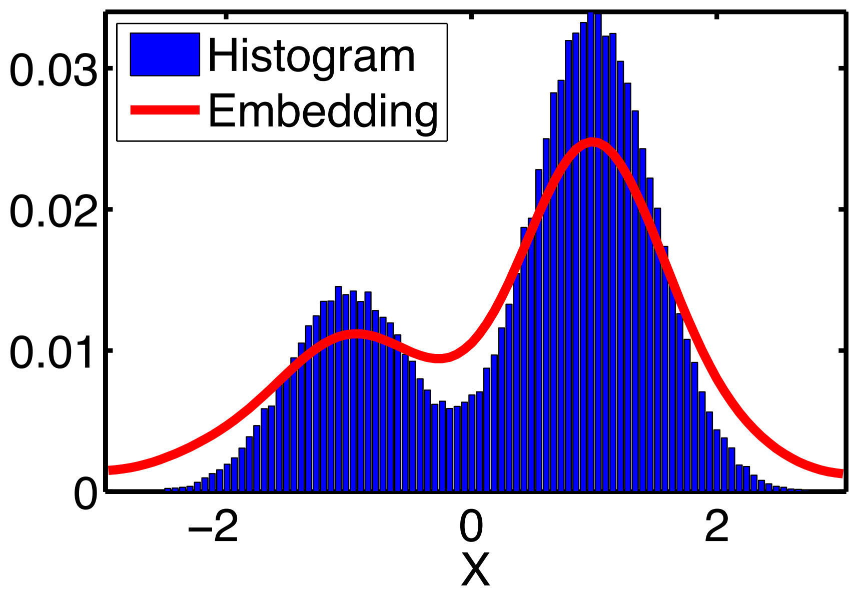
- What does prob. feature map look like?

$$\mu_{\mathbf{P}}(x) = \langle \mu_{\mathbf{P}}(\cdot), \varphi(x) \rangle_{\mathcal{F}}$$
$$= \langle \mu_{\mathbf{P}}(\cdot), k(\cdot, x) \rangle_{\mathcal{F}} = E_{\mathbf{P}} k(\mathsf{x}, x).$$

Expectation of kernel!

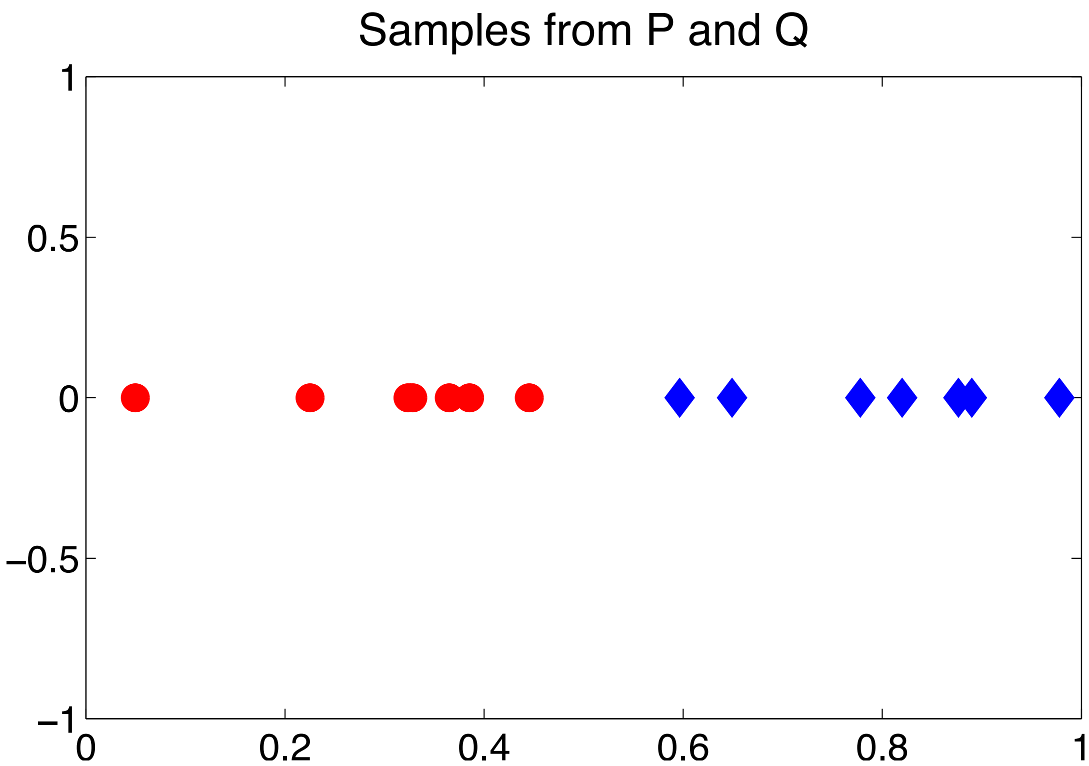
- Empirical estimate:

$$\hat{\mu}_{\mathbf{P}}(x) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, x) \qquad x_i \sim \mathbf{P}$$



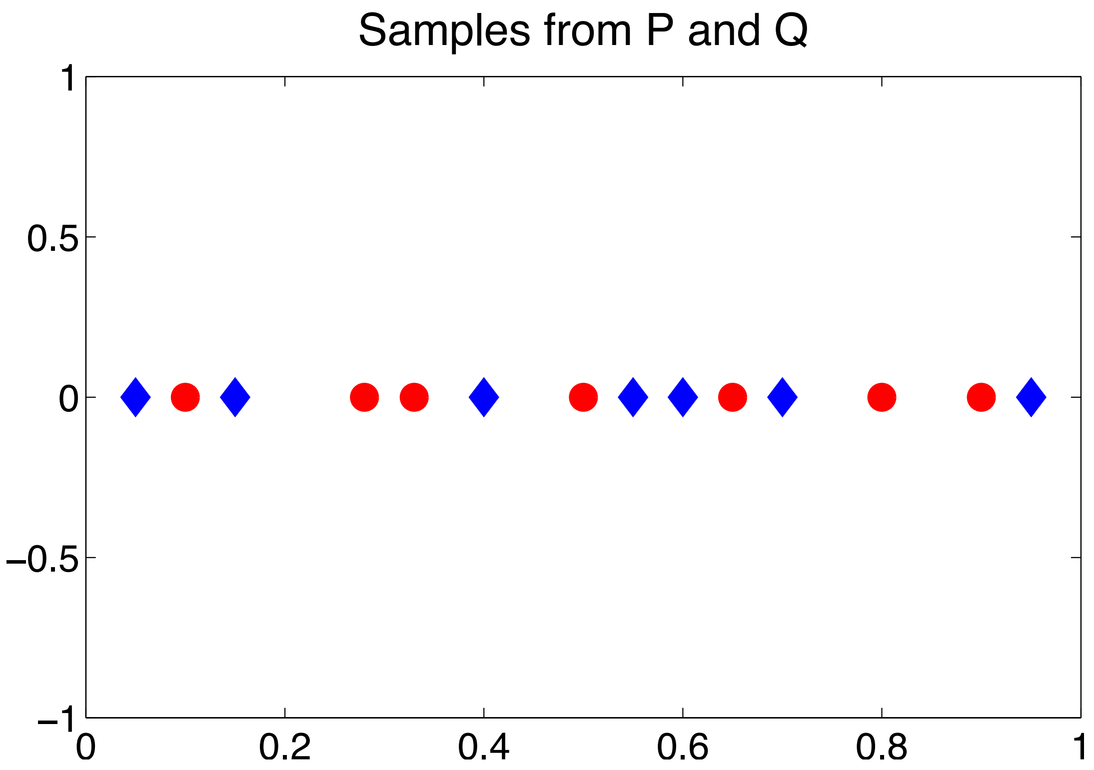
Function Showing Difference in Distributions

- Are **P** and **Q** different?



Function Showing Difference in Distributions

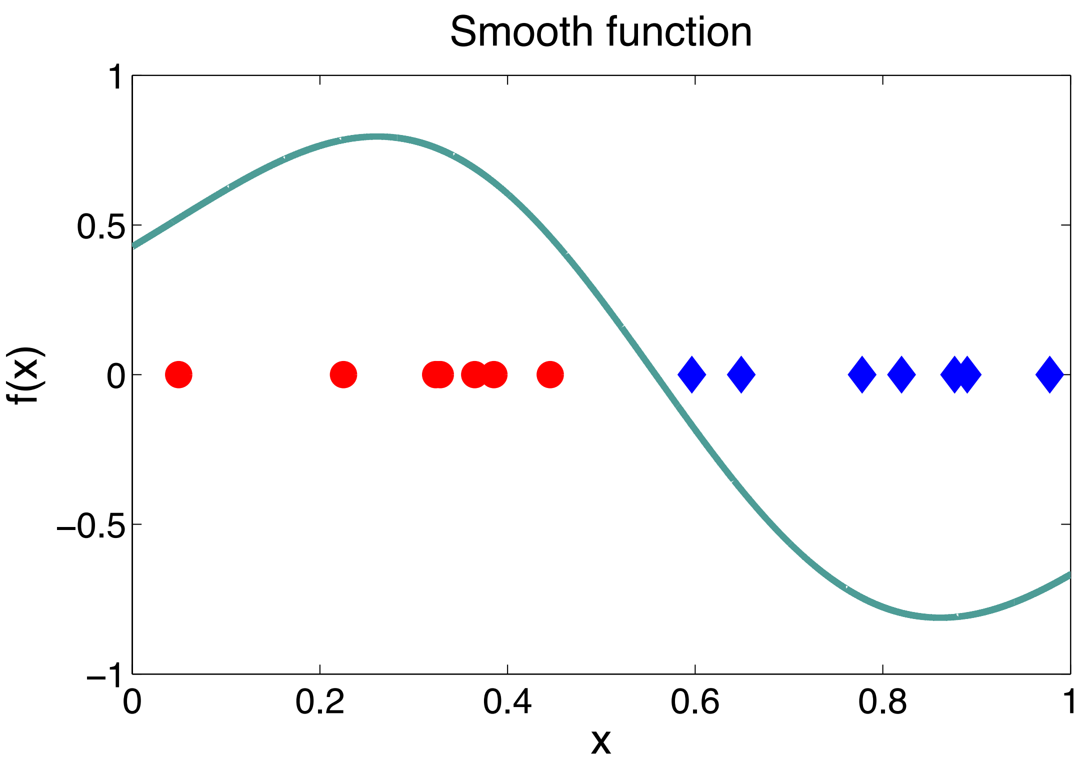
- Are **P** and **Q** different?



Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

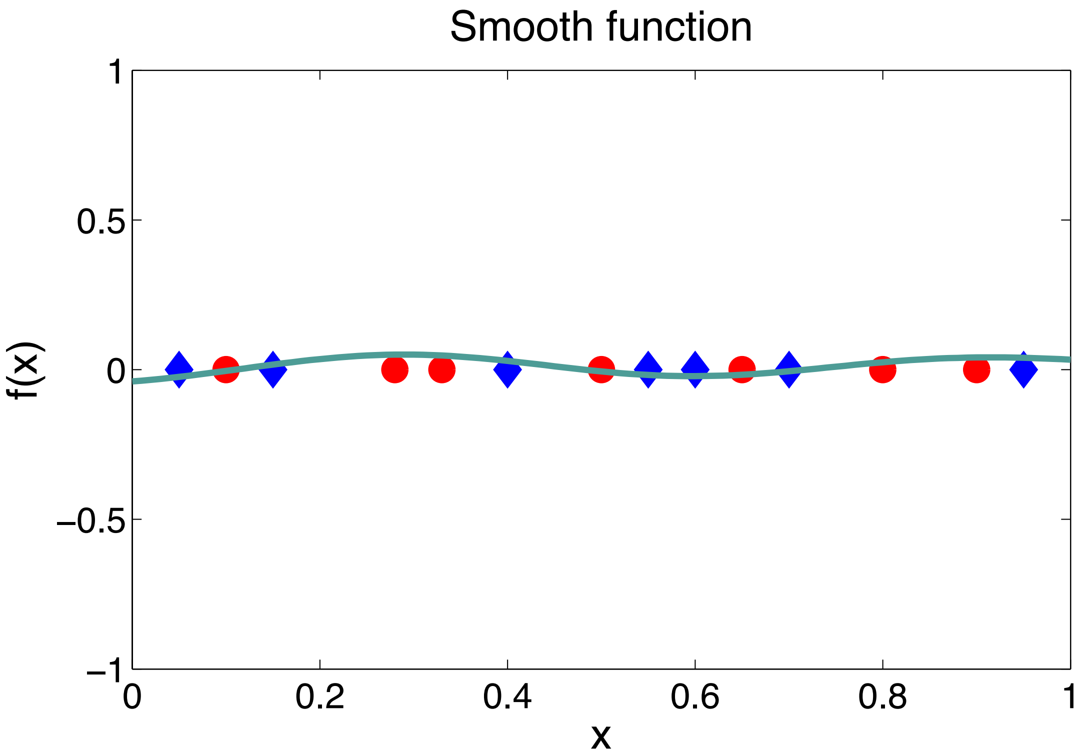
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathrm{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathrm{y}) \right].$$



Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

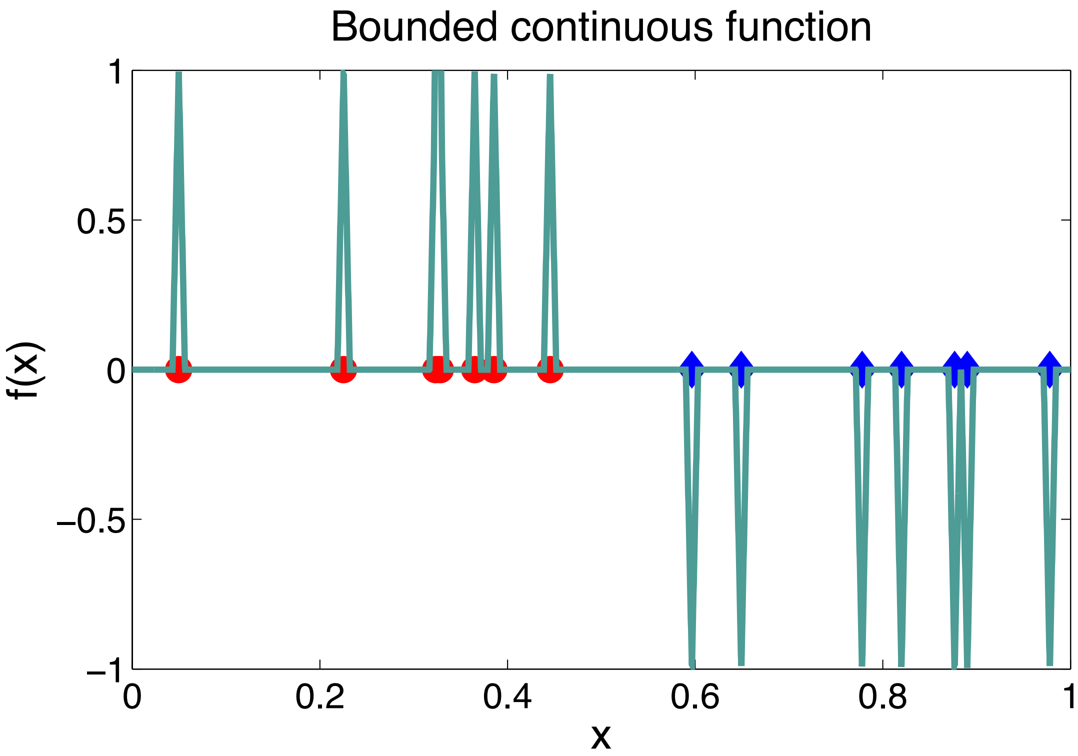
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathrm{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathrm{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

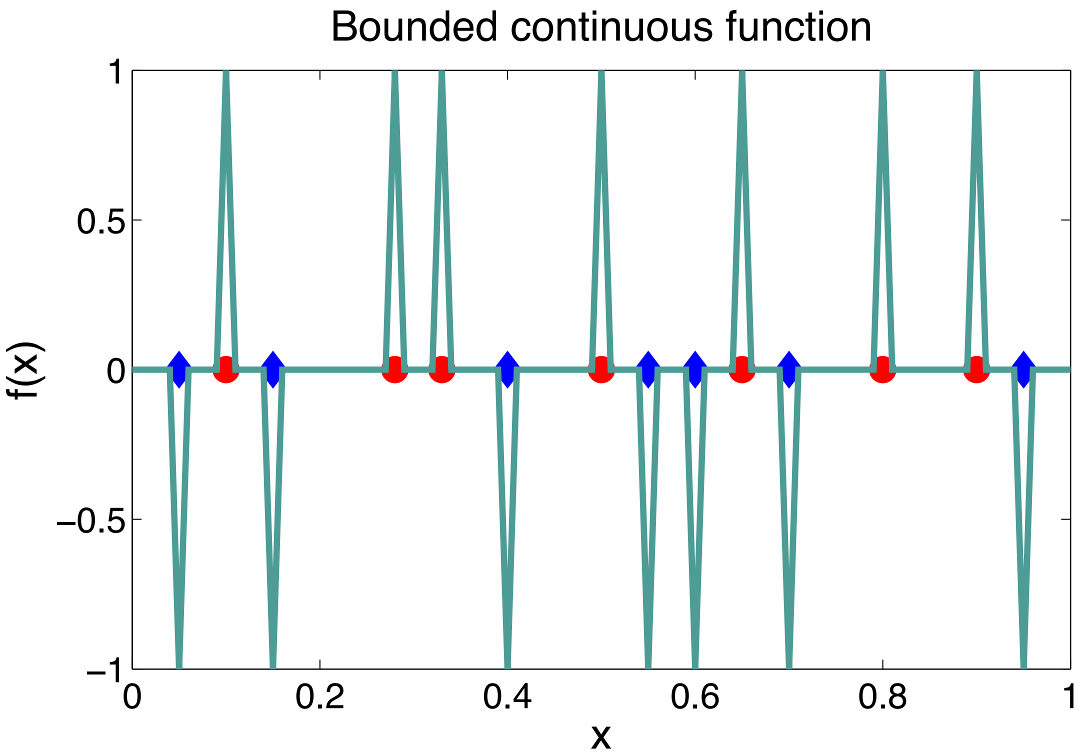
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

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

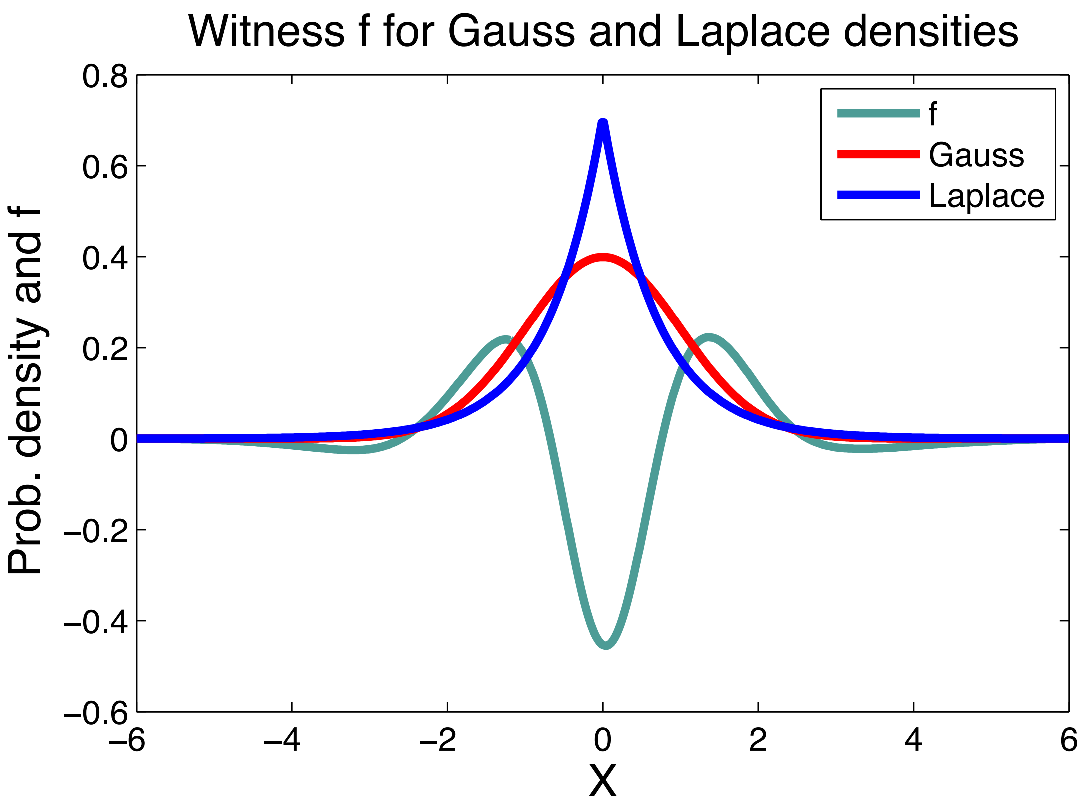


Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

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

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



Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for **P** vs **Q**

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

- 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

- Maximum mean discrepancy: smooth function for $\mathbf{P}$ vs $\mathbf{Q}$

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

- 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 $\mathcal{F}$ [ISMB06, NIPS06a, NIPS07b, NIPS08a, JMLR10]

Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for P vs Q

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

- Classical results: $\mathrm{MMD}(\mathsf{P},\mathsf{Q};F) = 0$ iff $\mathsf{P} = \mathsf{Q}$, when
  - $F =$bounded continuous [Dudley, 2002]
  - $F =$ bounded variation 1 (Kolmogorov metric) [Müller, 1997]
  - $F =$ bounded Lipschitz (Earth mover's distances) [Dudley, 2002]
- $\mathrm{MMD}(\mathsf{P},\mathsf{Q};F) = 0$ iff $\mathsf{P} = \mathsf{Q}$ when $F =$the unit ball in a characteristic RKHS $\mathcal{F}$ [ISMB06, NIPS06a, NIPS07b, NIPS08a, JMLR10]

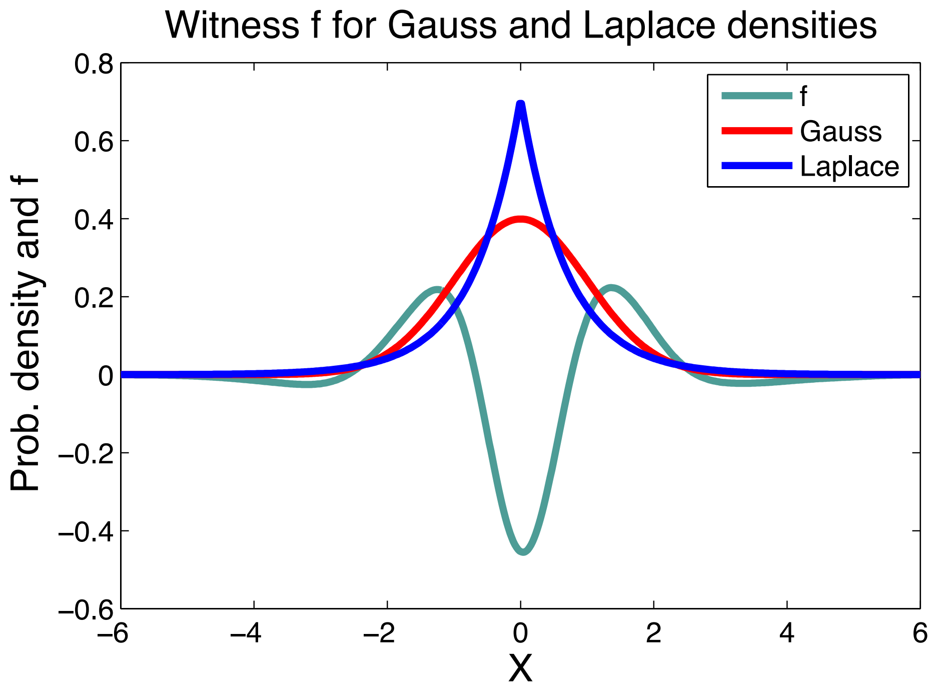
$\boxed{\text{How do smooth functions relate to feature maps?}}$

Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

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

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



Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

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

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

use

$$\mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) \quad =: \quad \langle \mu_{\mathsf{P}}, f \rangle_{\mathcal{F}}$$

Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

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

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

$$= \left(\sup_{f \in F} \langle f, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \right)^2$$

use

$$\mathbf{E}_{\mathbf{P}}(f(\mathsf{x})) =: \langle \mu_{\mathbf{P}}, f \rangle_{\mathcal{F}}$$

Function view vs feature mean view

- The (kernel) MMD: [ISMB06, NIPS06a]

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

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

$$= \left(\sup_{f\in F}\langle f, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\rangle_{\mathcal{F}}\right)^2$$

$$= \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2$$

use

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

$\boxed{\text{Function view and feature view equivalent}}$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathsf{P}$ and $\{y_i\}_{i=1}^m \sim \mathsf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i}^m \left[k(x_i, x_j) + k(y_i, y_j)\right] - \sum_{i=1}^m \sum_{j=1}^m \left[k(y_i, x_j) + k(x_i, y_j)\right]$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) + k(y_i, y_j)\right] - \sum_{i=1}^m \sum_{j=1}^m \left[k(y_i, x_j) + k(x_i, y_j)\right]$$

- Proof:

$$\begin{aligned} \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\ &= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) + k(y_i, y_j)\right]$$
$$- \sum_{i=1}^m \sum_{j=1}^m \left[k(y_i, x_j) + k(x_i, y_j)\right]$$

- Proof:

$$\begin{aligned} \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\ &= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i}^m \left[k(x_i, x_j) + k(y_i, y_j)\right] \\ - \sum_{i=1}^m \sum_{j=1}^m \left[k(y_i, x_j) + k(x_i, y_j)\right]$$

- Proof:

$$\begin{aligned} \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\ &= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \\ &= \langle \mathbf{E}_{\mathbf{P}} \varphi(\mathsf{x}), \mathbf{E}_{\mathbf{P}} \varphi(\mathsf{x}) \rangle + \ldots \end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m \left[k(x_i, x_j) + k(y_i, y_j)\right] - \sum_{i=1}^m \sum_{j=1}^m \left[k(y_i, x_j) + k(x_i, y_j)\right]$$

- Proof:

$$\begin{aligned}
\|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\
&= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \\
&= \langle \mathbf{E}_{\mathbf{P}} \varphi(\mathsf{x}), \mathbf{E}_{\mathbf{P}} \varphi(\mathsf{x}) \rangle + \ldots \\
&= \mathbf{E}_{\mathbf{P}} \langle \varphi(\mathsf{x}), \varphi(\mathsf{x}') \rangle + \ldots
\end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathbf{P}$ and $\{y_i\}_{i=1}^m \sim \mathbf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i}^m \left[k(x_i, x_j) + k(y_i, y_j)\right]$$
$$- \sum_{i=1}^m \sum_{j=1}^m \left[k(y_i, x_j) + k(x_i, y_j)\right]$$

- Proof:

$$\begin{aligned}
\|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\
&= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \\
&= \langle \mathbf{E}_{\mathbf{P}} \varphi(\mathsf{x}), \mathbf{E}_{\mathbf{P}} \varphi(\mathsf{x}) \rangle + \dots \\
&= \mathbf{E}_{\mathbf{P}} \left\langle \varphi(\mathsf{x}), \varphi(\mathsf{x}') \right\rangle + \dots \\
&= \mathbf{E}_{\mathbf{P}} k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_{\mathbf{Q}} k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y})
\end{aligned}$$

Empirical estimate of MMD

- An unbiased empirical estimate: for $\{x_i\}_{i=1}^m \sim \mathsf{P}$ and $\{y_i\}_{i=1}^m \sim \mathsf{Q}$,

$$\widehat{MMD}^2 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i}^m \left[k(x_i, x_j) + k(y_i, y_j)\right]$$
$$- \sum_{i=1}^m \sum_{j=1}^m \left[k(y_i, x_j) + k(x_i, y_j)\right]$$

- Proof:

$$\begin{aligned}
\|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|_{\mathcal{F}}^2 &= \langle \mu_{\mathsf{P}} - \mu_{\mathsf{Q}}, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} \\
&= \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle \\
&= \langle \mathbf{E}_{\mathsf{P}} \varphi(\mathsf{x}), \mathbf{E}_{\mathsf{P}} \varphi(\mathsf{x}) \rangle + \ldots \\
&= \mathbf{E}_{\mathsf{P}} \left\langle \varphi(\mathsf{x}), \varphi(\mathsf{x}') \right\rangle + \ldots \\
&= \mathbf{E}_{\mathsf{P}} k(\mathsf{x}, \mathsf{x}') + \mathbf{E}_{\mathsf{Q}} k(\mathsf{y}, \mathsf{y}') - 2\mathbf{E}_{\mathsf{P},\mathsf{Q}} k(\mathsf{x}, \mathsf{y})
\end{aligned}$$

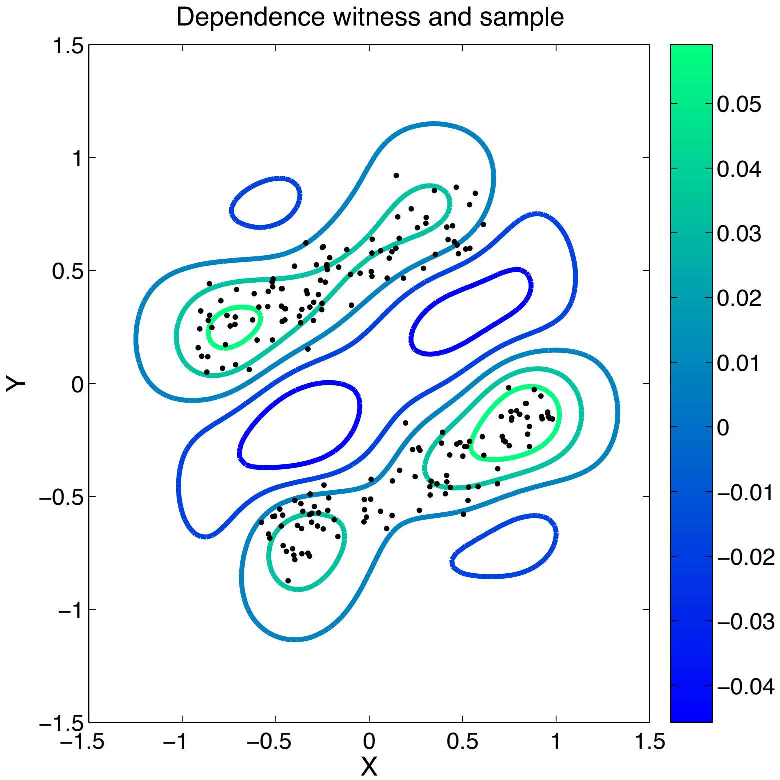
Then $\widehat{\mathbf{E}} k(\mathsf{x}, \mathsf{x}') = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i}^m k(x_i, x_j)$

MMD for independence

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

$$\left(\sup_f \left[\mathbf{E}_{\mathsf{P}_{XY}} f - \mathbf{E}_{\mathsf{P}_X \mathsf{P}_Y} f\right]\right)^2 = \sup_{\|f\| \le 1} \left\langle f, \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\rangle^2_{\mathcal{F}\times\mathcal{G}}$$

$$= \left\| \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\|^2_{\mathcal{F}\times\mathcal{G}} := HSIC(\mathsf{P}_{XY}, \mathsf{P}_X \mathsf{P}_Y)$$

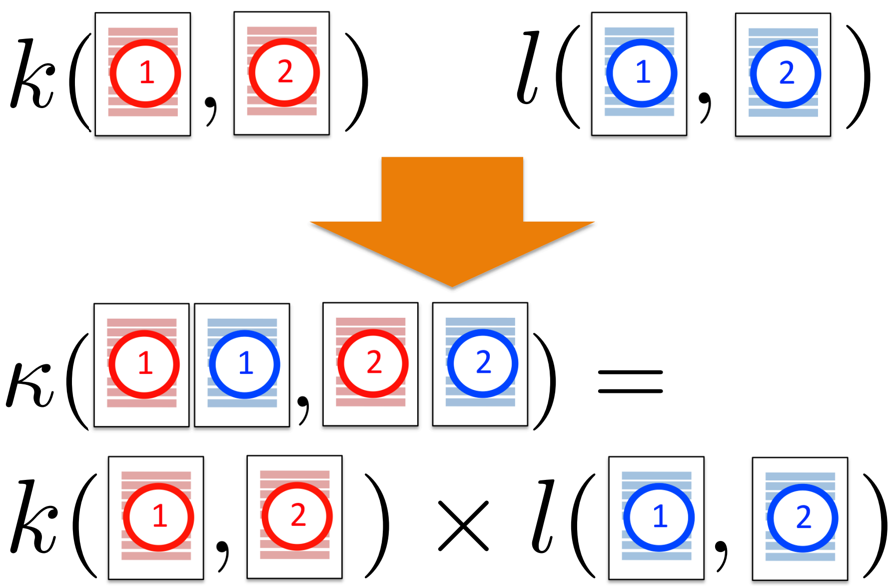


MMD for independence

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

$$\left(\sup_{f}\left[\mathbf{E}_{\mathsf{P}_{XY}} f - \mathbf{E}_{\mathsf{P}_X \mathsf{P}_Y} f\right]\right)^2 = \sup_{\|f\| \leq 1} \left\langle f, \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\rangle^2_{\mathcal{F}\times\mathcal{G}}$$

$$= \left\| \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\|^2_{\mathcal{F}\times\mathcal{G}} := HSIC(\mathsf{P}_{XY}, \mathsf{P}_X \mathsf{P}_Y)$$



Characteristic kernels (Version 1: Via Universality)

Characteristic Kernels (via universality)

Characteristic: MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

Characteristic Kernels (via universality)

Characteristic: MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

Classical result: $\mathbf{P} = \mathbf{Q}$ if and only if $\mathbb{E}_{\mathbf{P}}(f(\mathsf{x})) = \mathbb{E}_{\mathbf{Q}}(f(\mathsf{y}))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ [Dudley, 2002]

Characteristic Kernels (via universality)

Characteristic: MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

Classical result: $\mathbf{P} = \mathbf{Q}$ if and only if $\mathbb{E}_{\mathbf{P}}(f(\mathsf{x})) = \mathbb{E}_{\mathbf{Q}}(f(\mathsf{y}))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ [Dudley, 2002]

Universal RKHS: $k(x, x')$ continuous, $\mathcal{X}$ compact, and $\mathcal{F}$ dense in $C(\mathcal{X})$ with respect to L_∞ [Steinwart, 2001]

Characteristic Kernels (via universality)

Characteristic: MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

Classical result: $\mathbf{P} = \mathbf{Q}$ if and only if $\mathbb{E}_{\mathbf{P}}(f(\mathsf{x})) = \mathbb{E}_{\mathbf{Q}}(f(\mathsf{y}))$ for all $f \in C(\mathcal{X})$, the space of bounded continuous functions on $\mathcal{X}$ [Dudley, 2002]

Universal RKHS: $k(x, x')$ continuous, $\mathcal{X}$ compact, and $\mathcal{F}$ dense in $C(\mathcal{X})$ with respect to L_∞ [Steinwart, 2001]

If $\mathcal{F}$ universal, then $\text{MMD}\{\mathbf{P}, \mathbf{Q}; F\} = 0$ iff $\mathbf{P} = \mathbf{Q}$

Characteristic Kernels (via universality)

Proof:

First, it is clear that $\mathbf{P} = \mathbf{Q}$ implies $\text{MMD}\{\mathbf{P}, \mathbf{Q}; F\}$ is zero.

Converse: by the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$ $\exists g \in \mathcal{F}$

$$\|f - g\|_\infty \leq \epsilon.$$

Characteristic Kernels (via universality)

Proof:

First, it is clear that $\mathsf{P} = \mathsf{Q}$ implies $\mathrm{MMD}\left\{\mathsf{P}, \mathsf{Q}; F\right\}$ is zero.

Converse: by the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$ $\exists g \in \mathcal{F}$

$$\|f - g\|_\infty \leq \epsilon.$$

We next make the expansion

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq |\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| + |\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y})| + |\mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})|.$$

The first and third terms satisfy

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| \leq \mathbf{E}_{\mathsf{P}} |f(\mathsf{x}) - g(\mathsf{x})| \leq \epsilon.$$

Characteristic Kernels (via universality)

Proof (continued):

Next, write

$$\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) = \langle g(\cdot), \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} = 0,$$

since MMD $\{\mathsf{P}, \mathsf{Q}; F\} = 0$ implies $\mu_{\mathsf{P}} = \mu_{\mathsf{Q}}$. Hence

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq 2\epsilon$$

for all $f \in C(\mathcal{X})$ and $\epsilon > 0$, which implies $\mathsf{P} = \mathsf{Q}$.

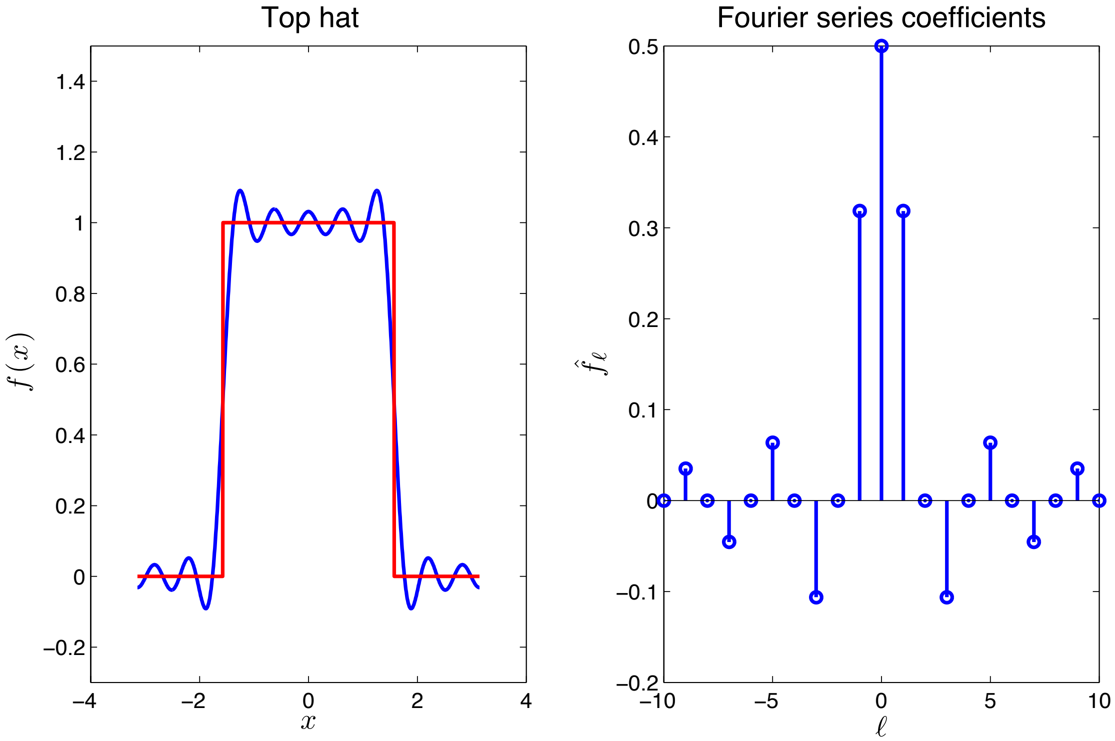
Characteristic kernels (Version 2: Via Fourier)

Characteristic Kernels (via Fourier)

Reminder: Fourier series

- Function $[-\pi, \pi]$ with periodic boundary.

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x) = \sum_{l=-\infty}^{\infty} \hat{f}_\ell \left(\cos(\ell x) + \imath \sin(\ell x)\right).$$

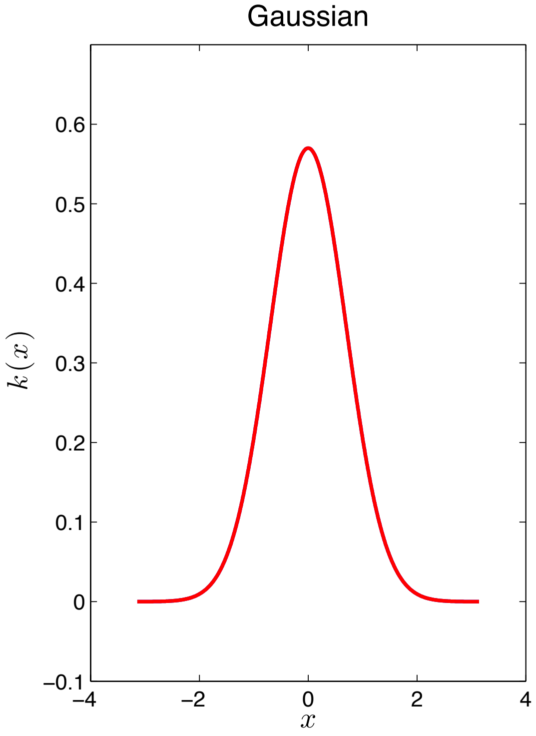


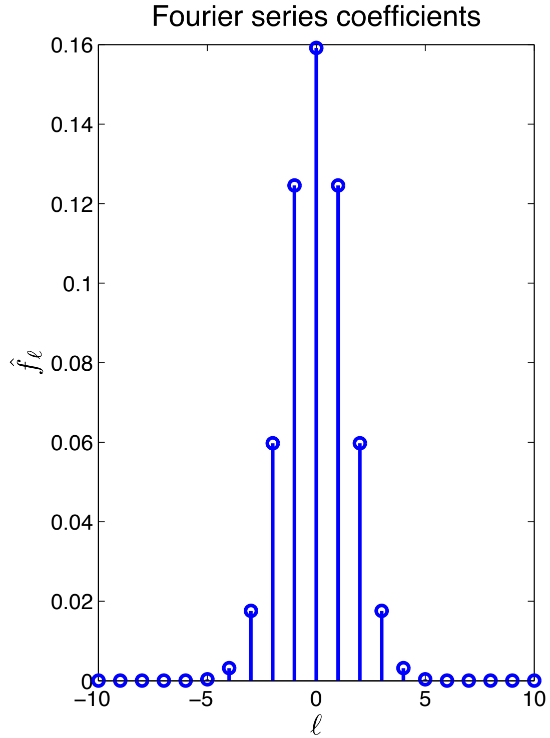
Characteristic Kernels (via Fourier)

Reminder: Fourier series of kernel

$$k(x,y) = k(x-y) = k(z), \qquad k(z) = \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath \ell z),$$

E.g. Gaussian, $\quad k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right), \qquad \hat{k}_\ell = \frac{1}{2\pi} \exp\left(\frac{-\sigma^2\ell^2}{2}\right).$





Characteristic Kernels (via Fourier)

Maximum mean embedding via Fourier series:

- Fourier series for $\mathbf{P}$ is characteristic function $\phi_{\mathbf{P}}$
- Fourier series for mean embedding is product of fourier series! (convolution theorem)

$$\mu_{\mathbf{P}}(x) = E_{\mathbf{P}} k(\mathsf{x} - x) = \int_{-\pi}^{\pi} k(x-t) d\mathbf{P}(t) \qquad \hat{\mu}_{\mathbf{P},\ell} = \hat{k}_\ell \times \phi_{\mathbf{P},\ell}$$

Characteristic Kernels (via Fourier)

Maximum mean embedding via Fourier series:

- Fourier series for $\mathbf{P}$ is characteristic function $\phi_{\mathbf{P}}$
- Fourier series for mean embedding is product of fourier series! (convolution theorem)

$$\mu_{\mathbf{P}}(x) = E_{\mathbf{P}} k(\mathsf{x} - x) = \int_{-\pi}^{\pi} k(x - t) d\mathbf{P}(t) \qquad \hat{\mu}_{\mathbf{P},\ell} = \hat{k}_\ell \times \phi_{\mathbf{P},\ell}$$

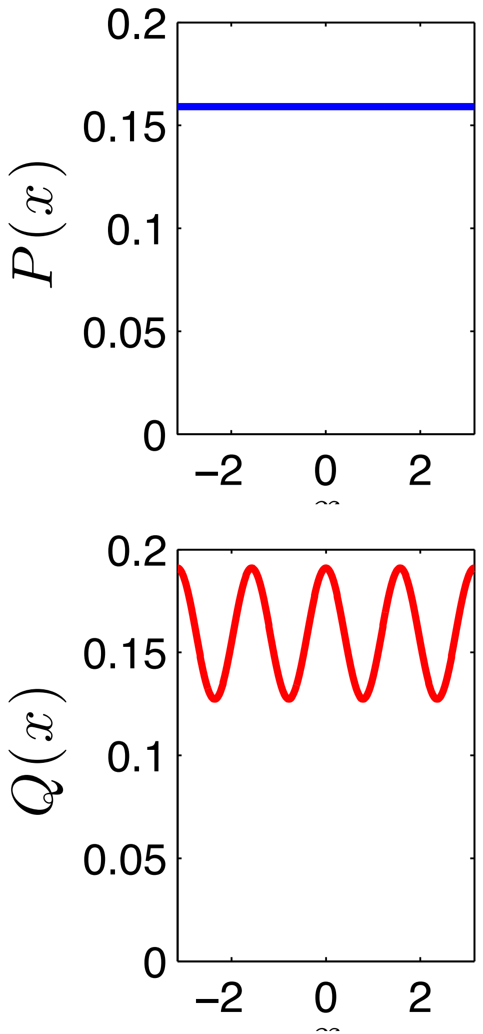
- MMD can be written in terms of Fourier series:

$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \sum_{\ell=-\infty}^{\infty} \left[\left(\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell} \right) \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

- Characteristic: MMD a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08, JMLR10]

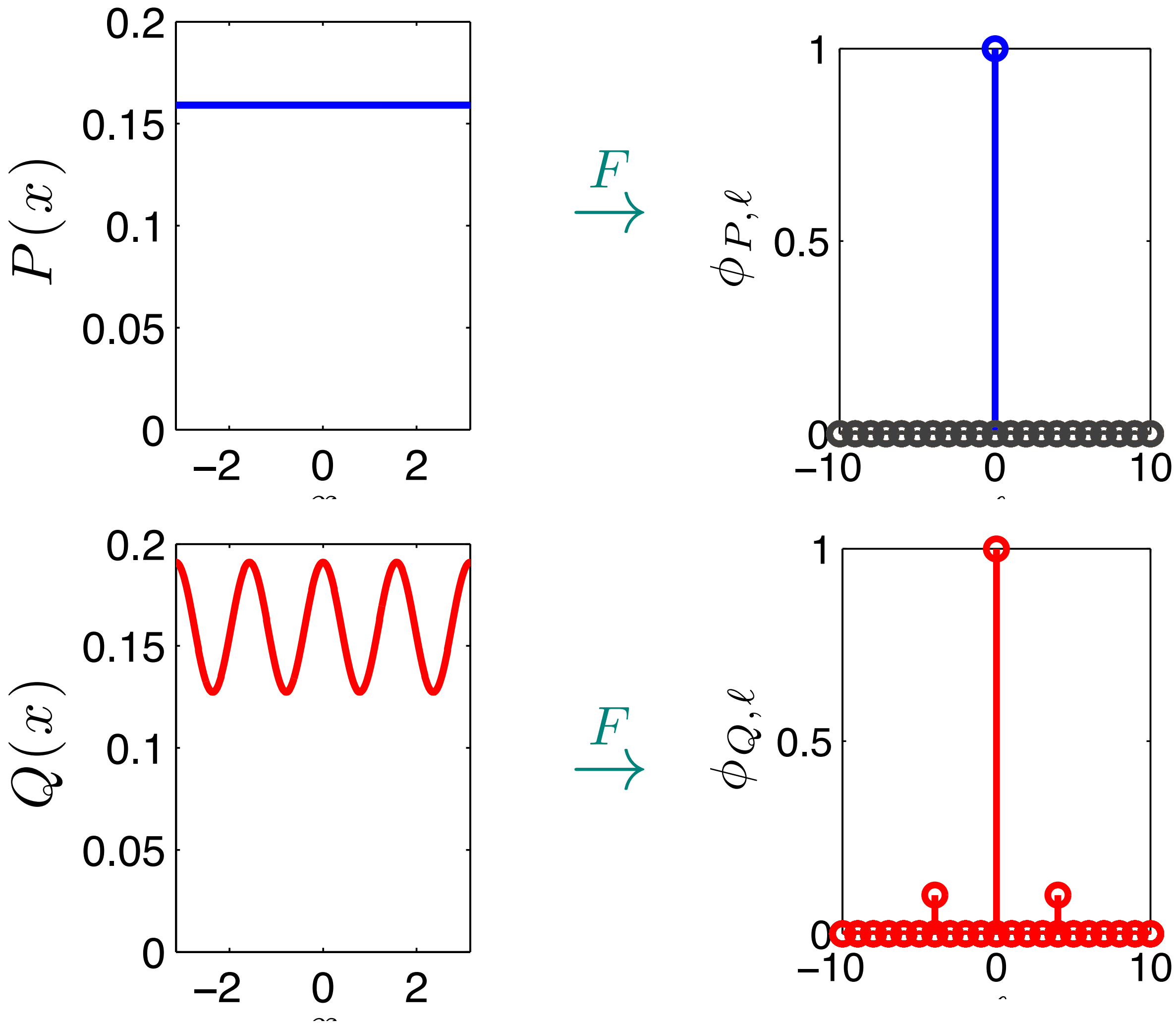
Example

- Example: **P** differs from **Q** at one frequency



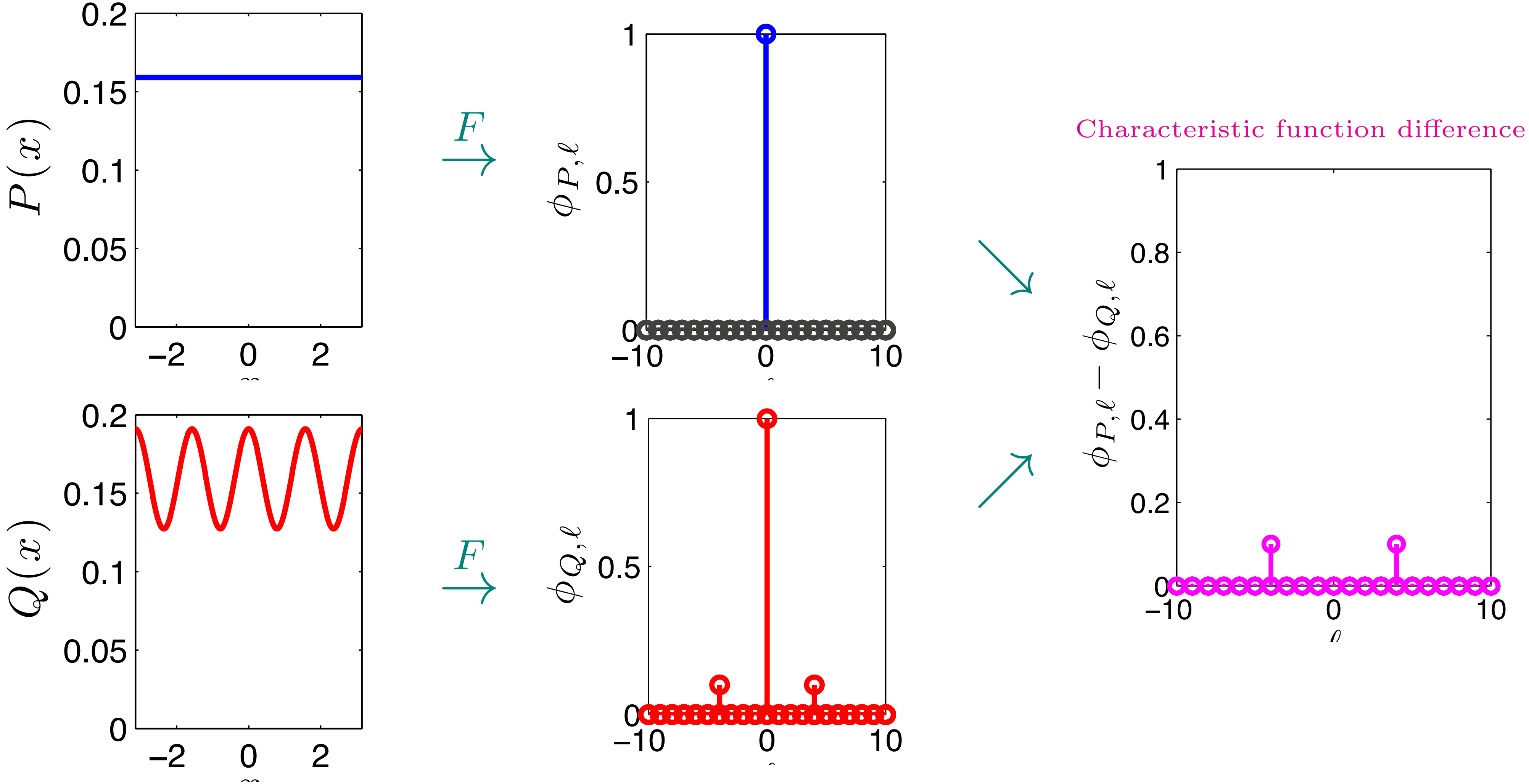
Characteristic Kernels (2)

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



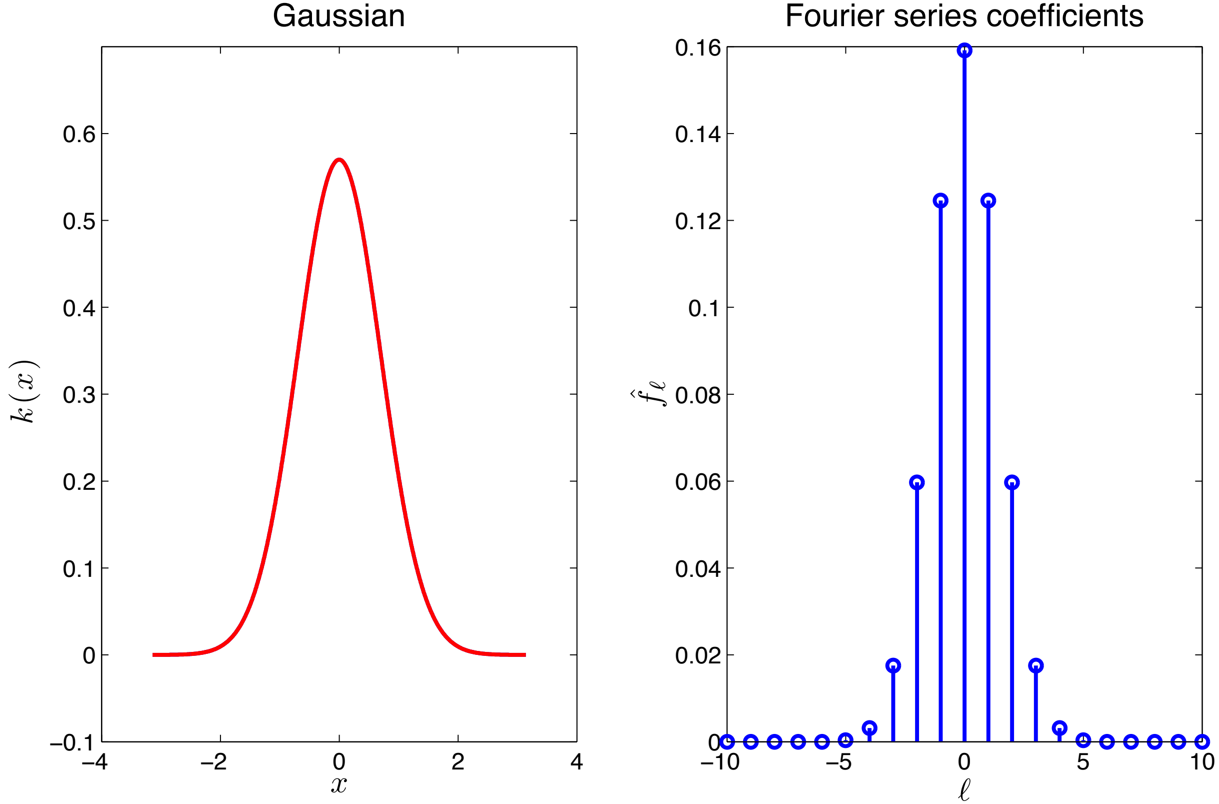
Characteristic Kernels (2)

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



Example

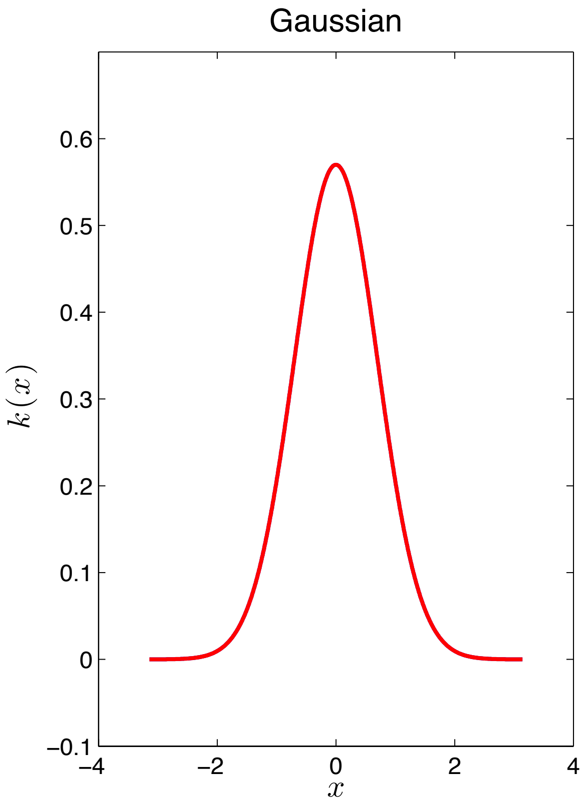
Is the Gaussian kernel characteristic?

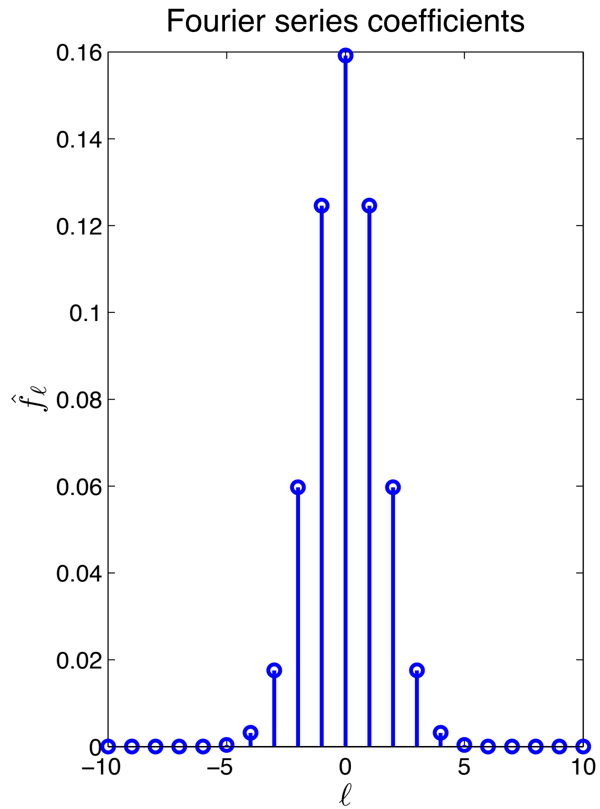


$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \sum_{\ell=-\infty}^{\infty} \left[(\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell}) \, \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

Example

Is the Gaussian kernel characteristic? YES

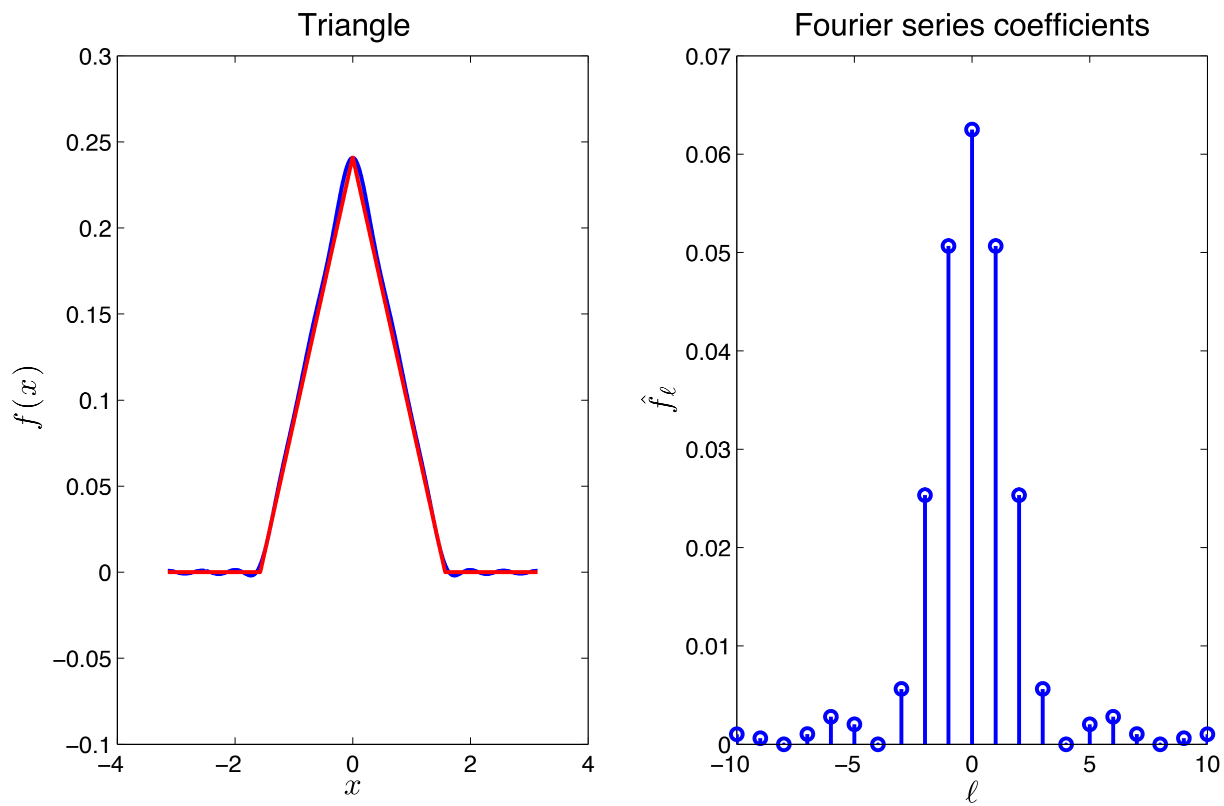




$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \sum_{\ell=-\infty}^{\infty} \left[\left(\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell}\right) \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

Example

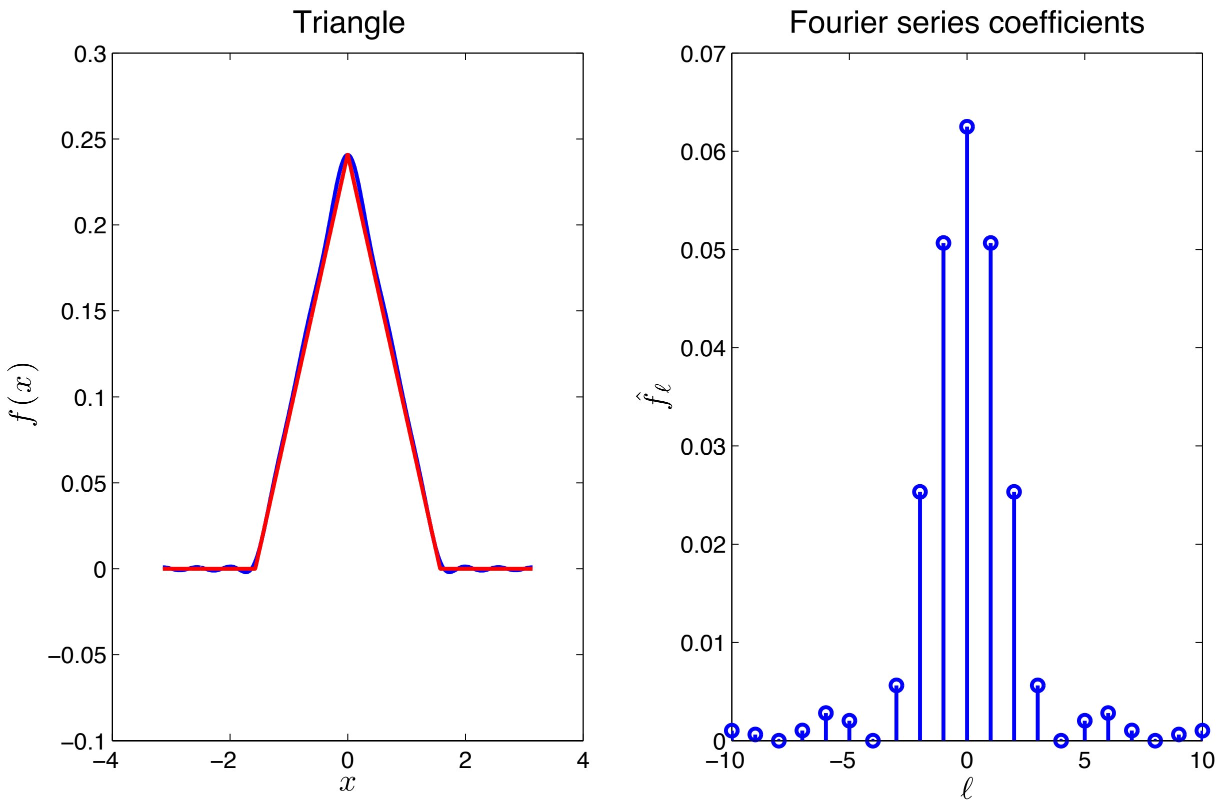
Is the triangle kernel characteristic?



$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \sum_{\ell=-\infty}^{\infty} \left[\left(\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell} \right) \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

Example

Is the triangle kernel characteristic? NO



$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \sum_{\ell=-\infty}^{\infty} \left[(\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell}) \, \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$?

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$?
- Characteristic function of $\mathbf{P}$ via Fourier transform

$$\phi_{\mathbf{P}}(\omega) = \int_{\mathbb{R}^d} e^{ix^\top \omega} d\mathbf{P}(x)$$

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$?
- Characteristic function of $\mathbf{P}$ via Fourier transform

$$\phi_{\mathbf{P}}(\omega) = \int_{\mathbb{R}^d} e^{ix^\top \omega} d\mathbf{P}(x)$$

- Translation invariant kernels: $k(x, y) = k(x - y) = k(z)$
- Bochner's theorem:

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

  - Λ finite non-negative Borel measure

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$?
- Characteristic function of $\mathbf{P}$ via Fourier transform

$$\phi_{\mathbf{P}}(\omega) = \int_{\mathbb{R}^d} e^{ix^\top \omega} d\mathbf{P}(x)$$

- Translation invariant kernels: $k(x, y) = k(x - y) = k(z)$
- Bochner's theorem:

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

  - Λ finite non-negative Borel measure

Characteristic Kernels (via Fourier)

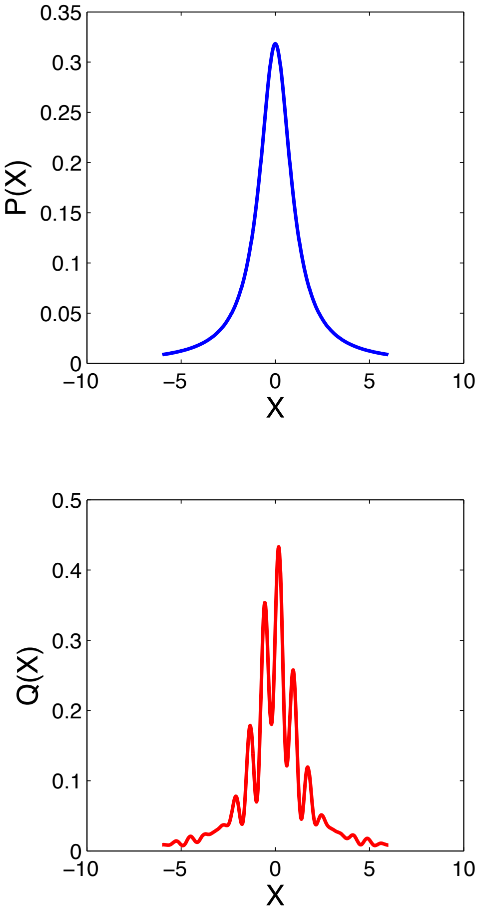
- Fourier representation of MMD:

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

  - $\phi_{\mathbf{P}}$ characteristic function of $\mathbf{P}$
  - $f^{\wedge}$ is Fourier transform, $f^{\vee}$ is inverse Fourier transform
  - $\mu_{\mathbf{P}} := \int k(\cdot, x)\, d\mathbf{P}(x)$

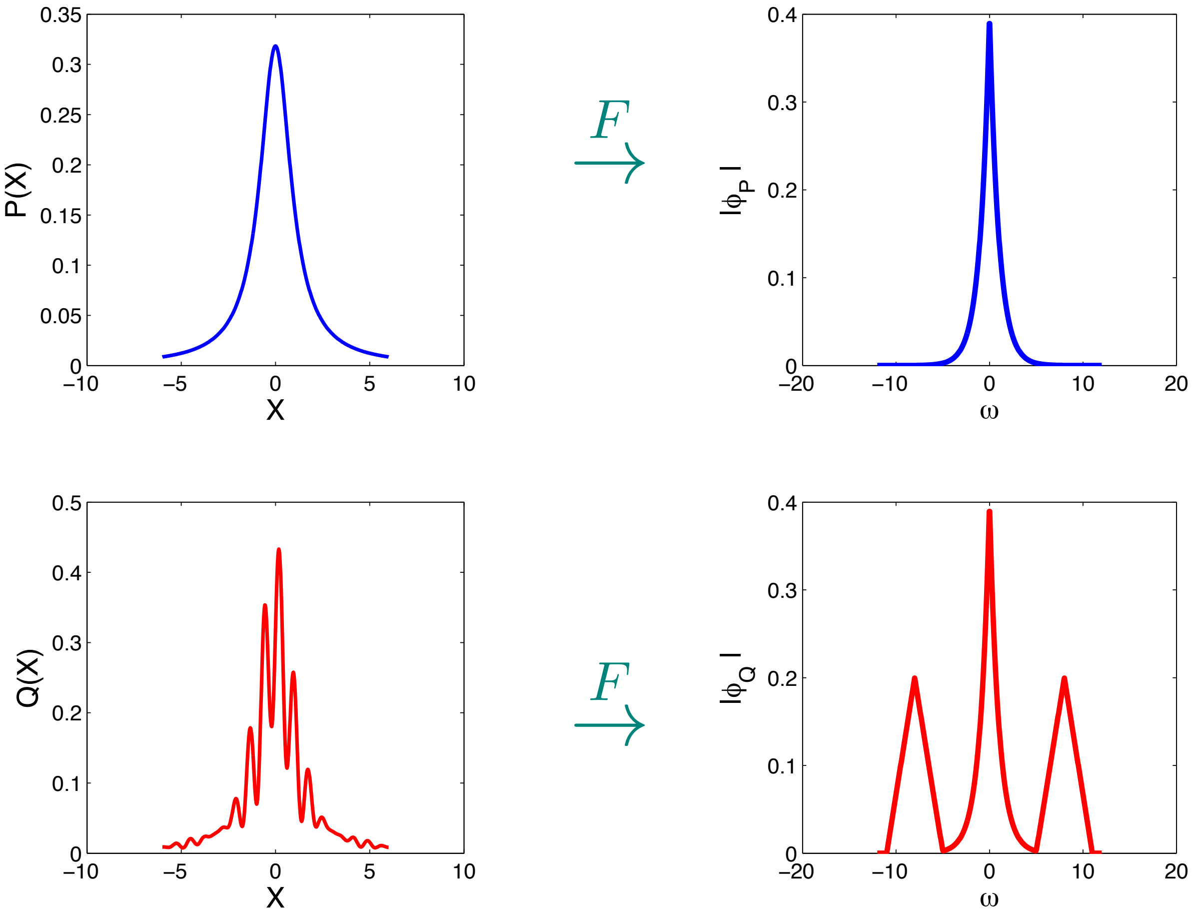
Example

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



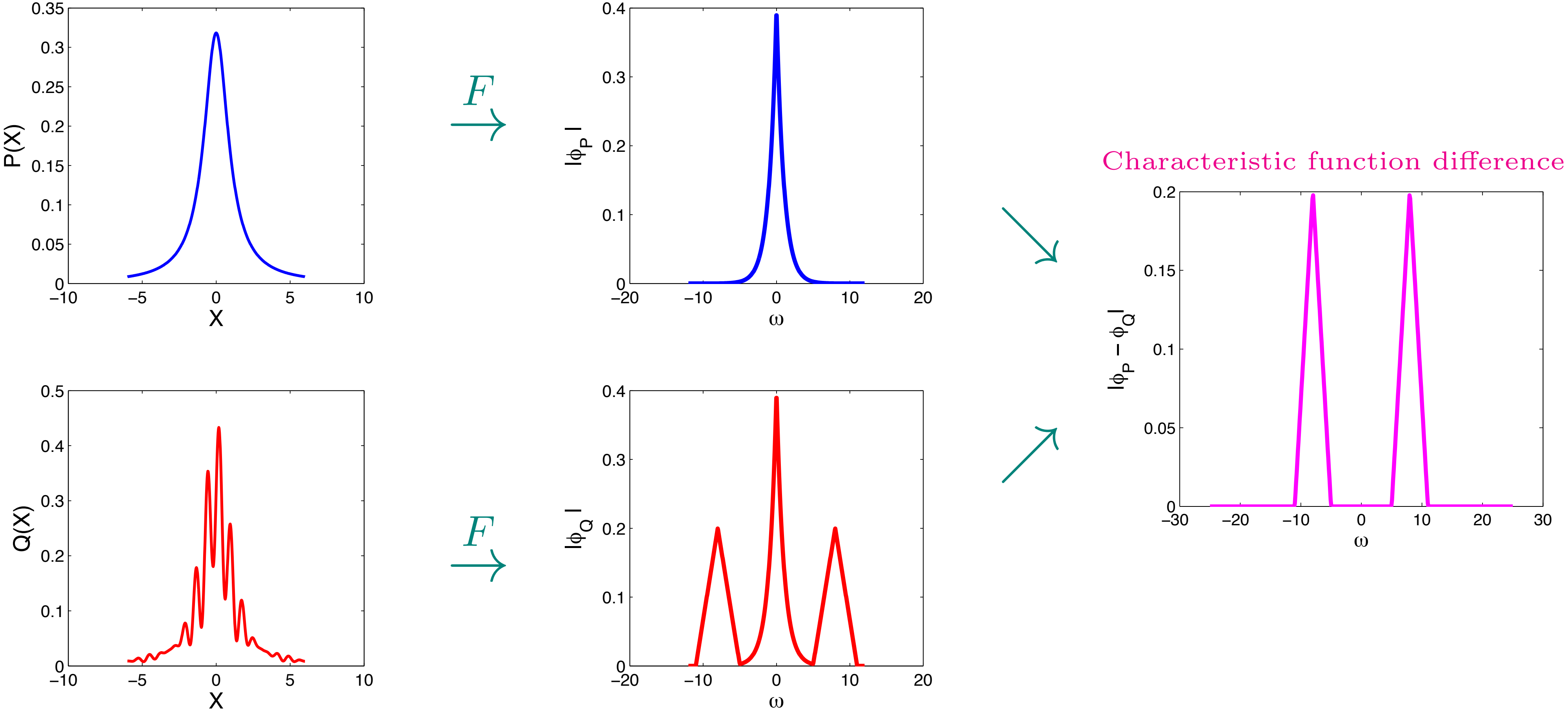
Example

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



Example

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

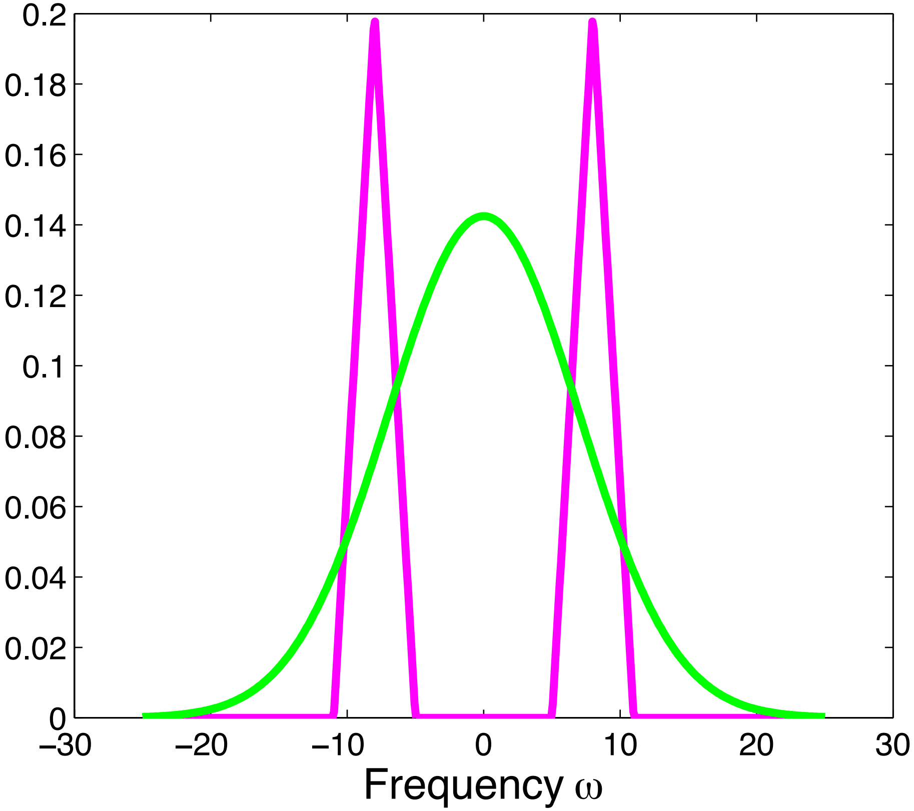


Example

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

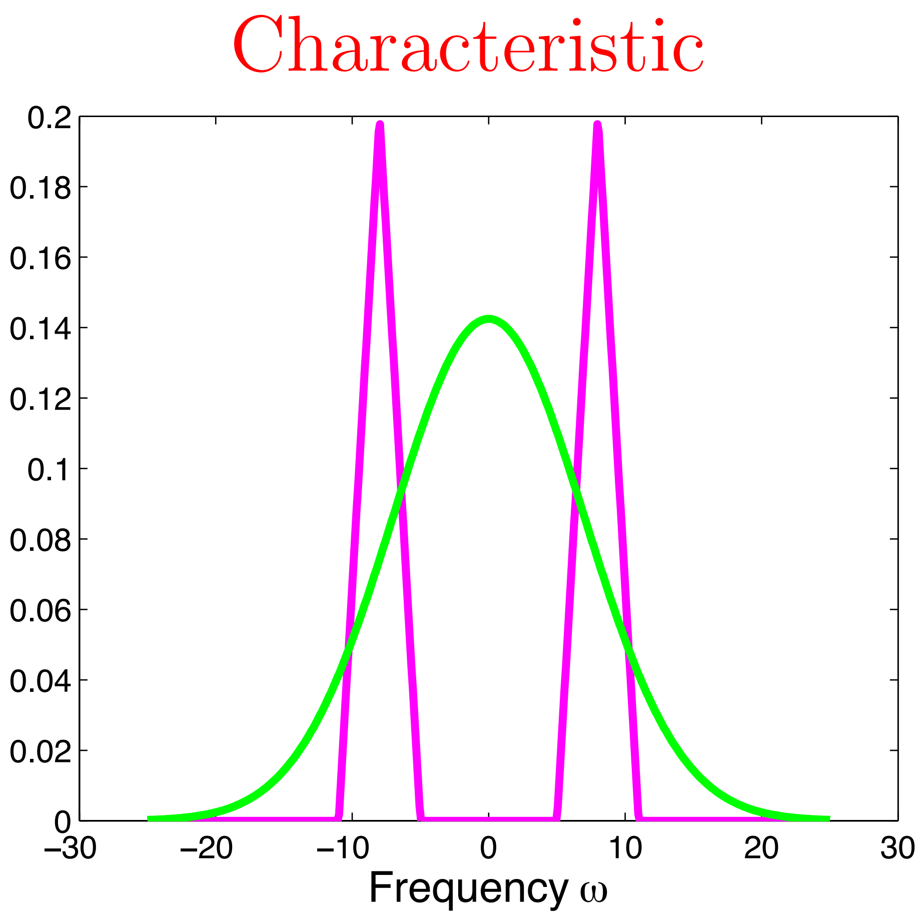
Gaussian kernel

Difference $|\phi_P - \phi_Q|$



Example

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

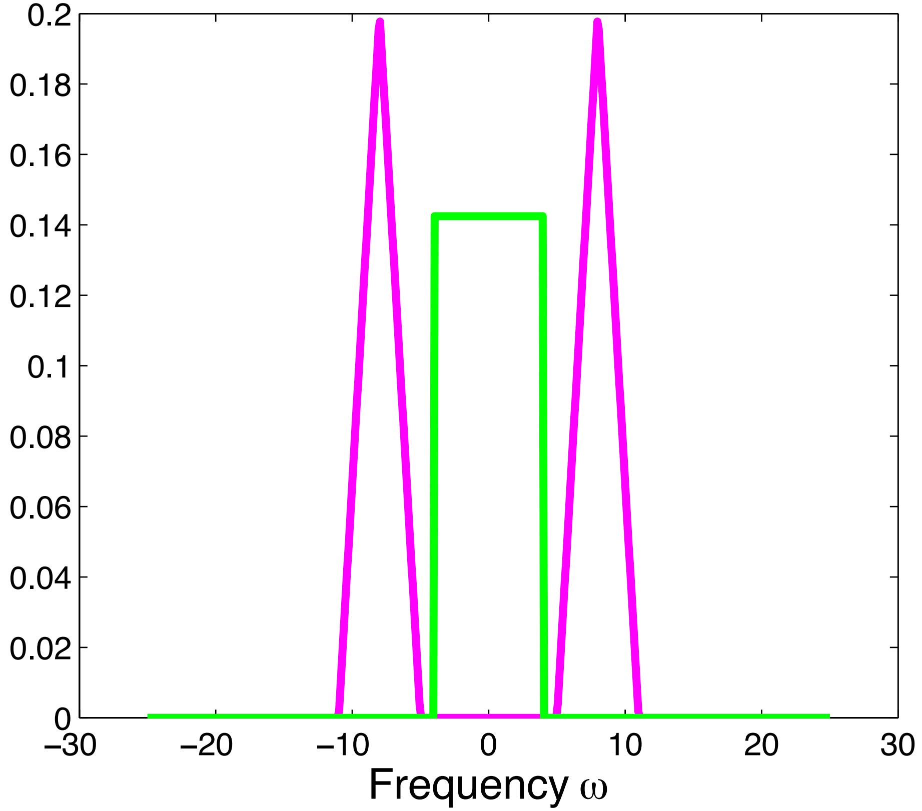


Example

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

Sinc kernel

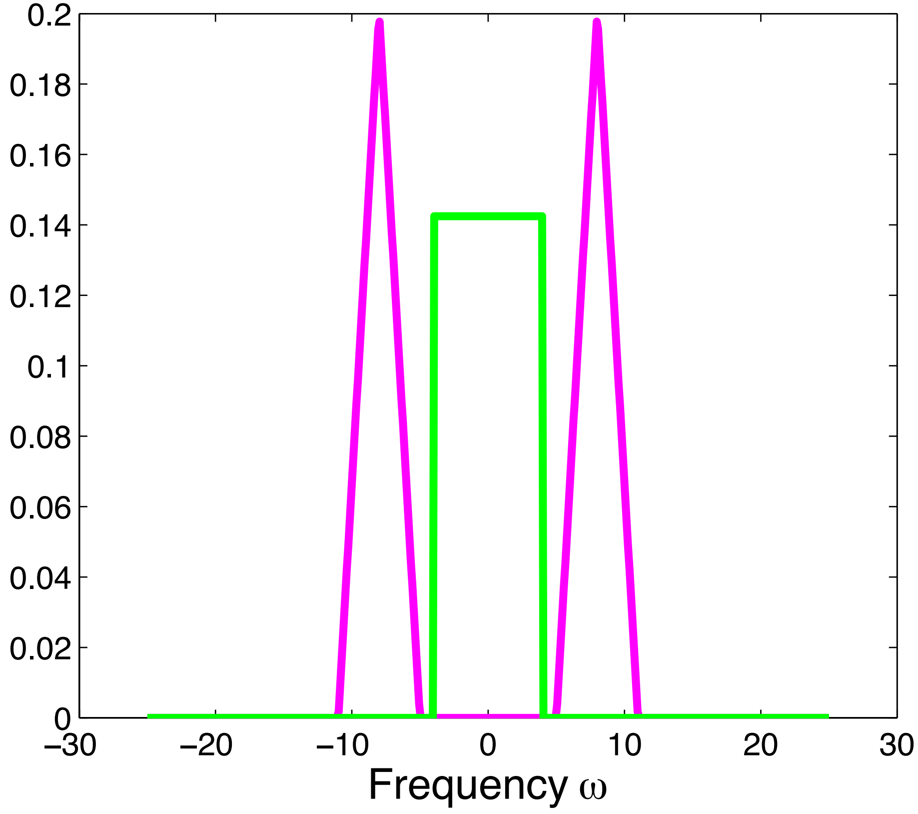
Difference $|\phi_P - \phi_Q|$



Example

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

NOT characteristic

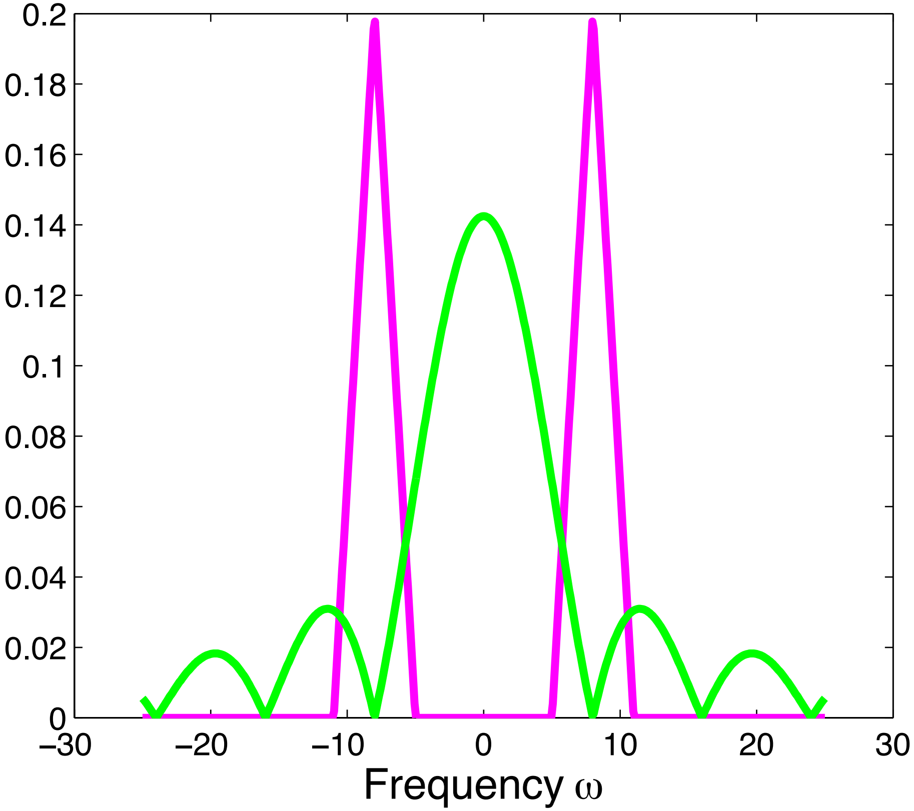


Example

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

Triangle (B-spline) kernel

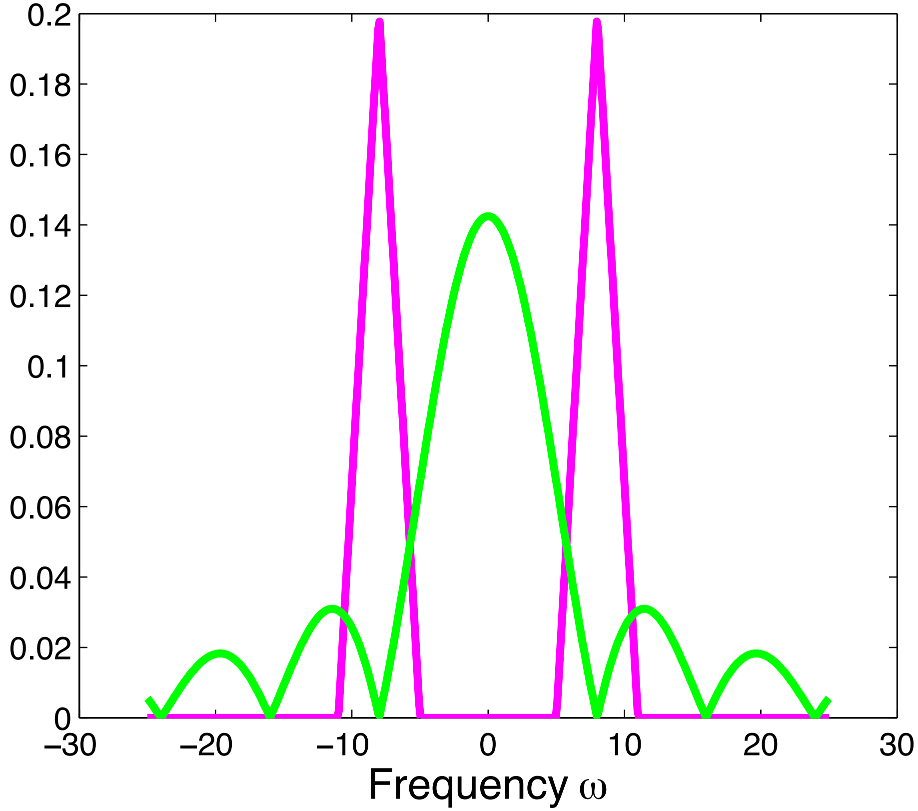
Difference $|\phi_P - \phi_Q|$



Example

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

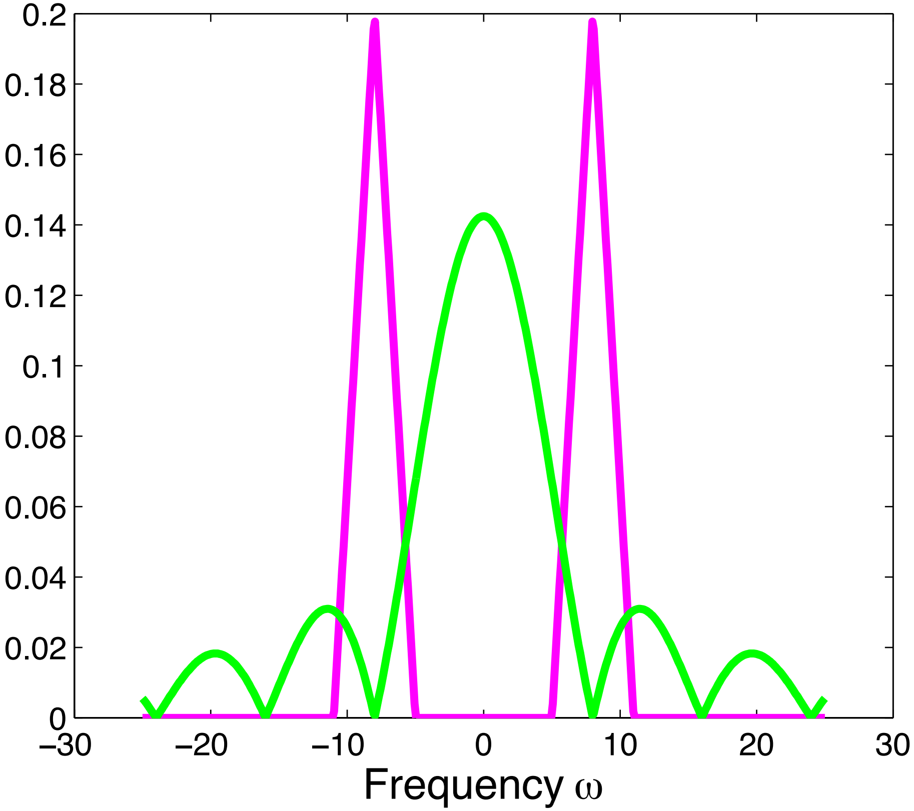
???



Example

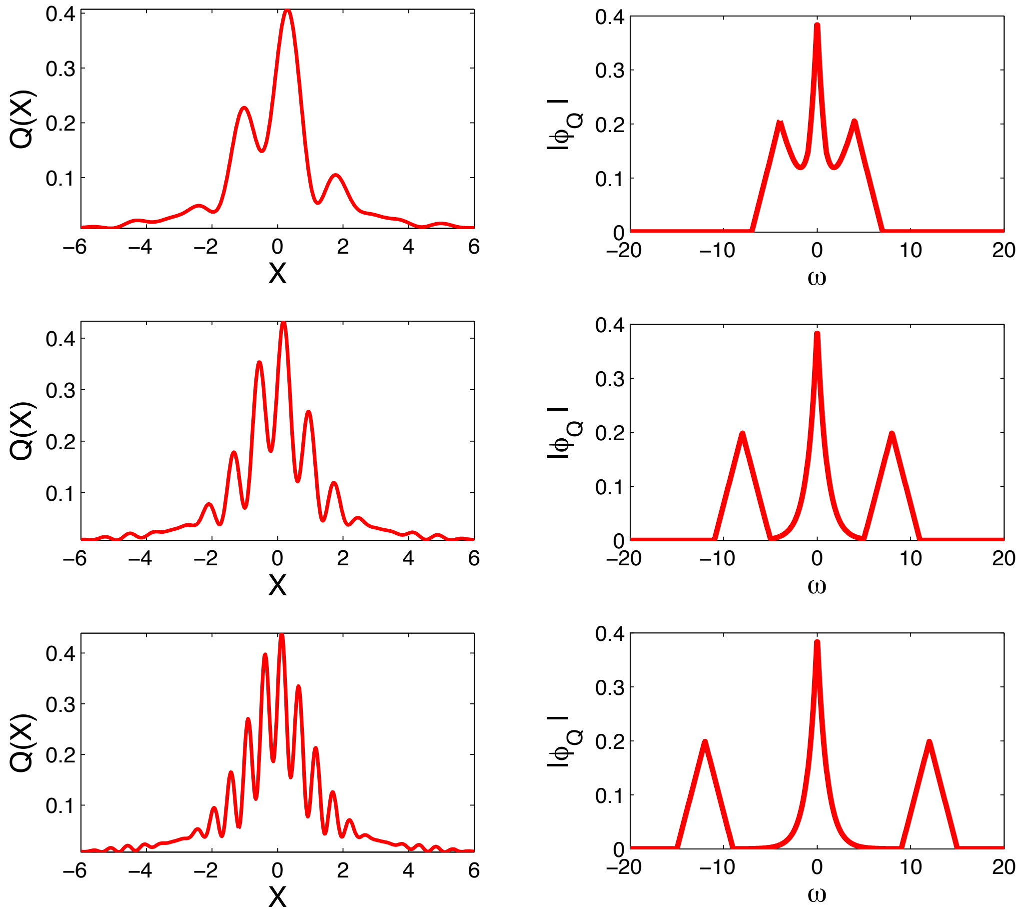
- Example: **P** differs from **Q** at (roughly) one frequency

Characteristic

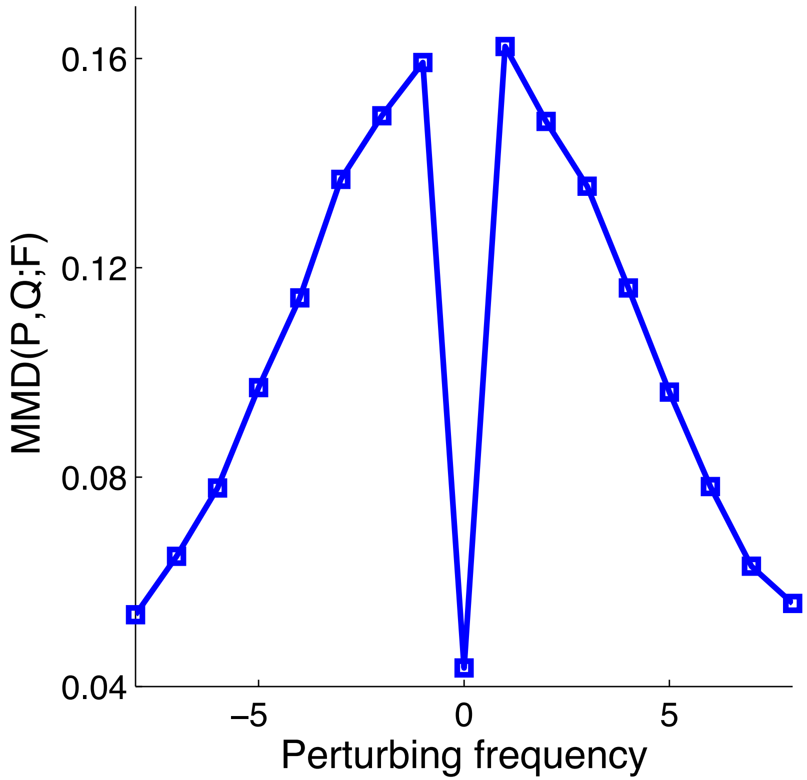


Choosing the kernel

- Gaussian kernel example



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



Why does MMD decay with increasing peturbation freq.?

- Fourier series argument (notationally easier, for periodic domains only):

$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \left\| \sum_{\ell=-\infty}^{\infty} \left[(\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell}) \hat{k}_\ell \right] \exp(\imath \ell x) \right\|_{\mathcal{F}}$$

- The squared norm of a function f in $\mathcal{F}$ is:

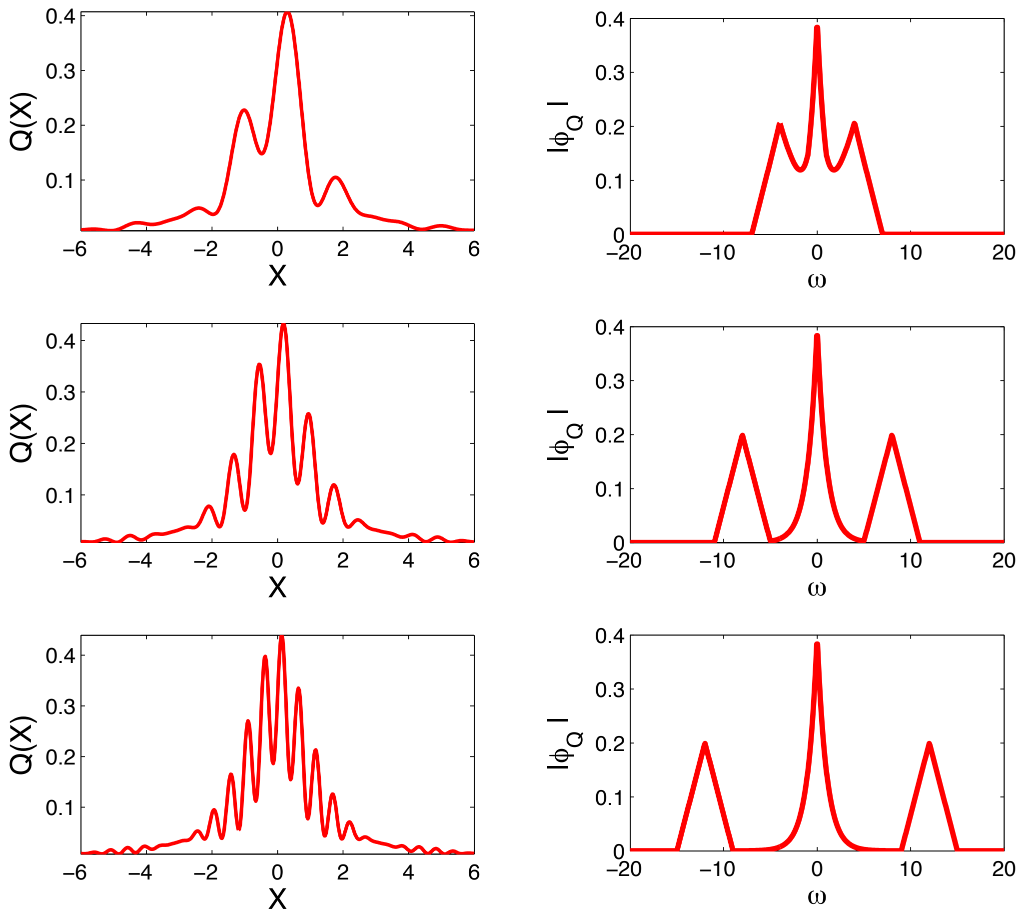
$$\|f\|_{\mathcal{F}}^2 = \langle f, f \rangle_{\mathcal{F}} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}.$$

- Squared MMD is

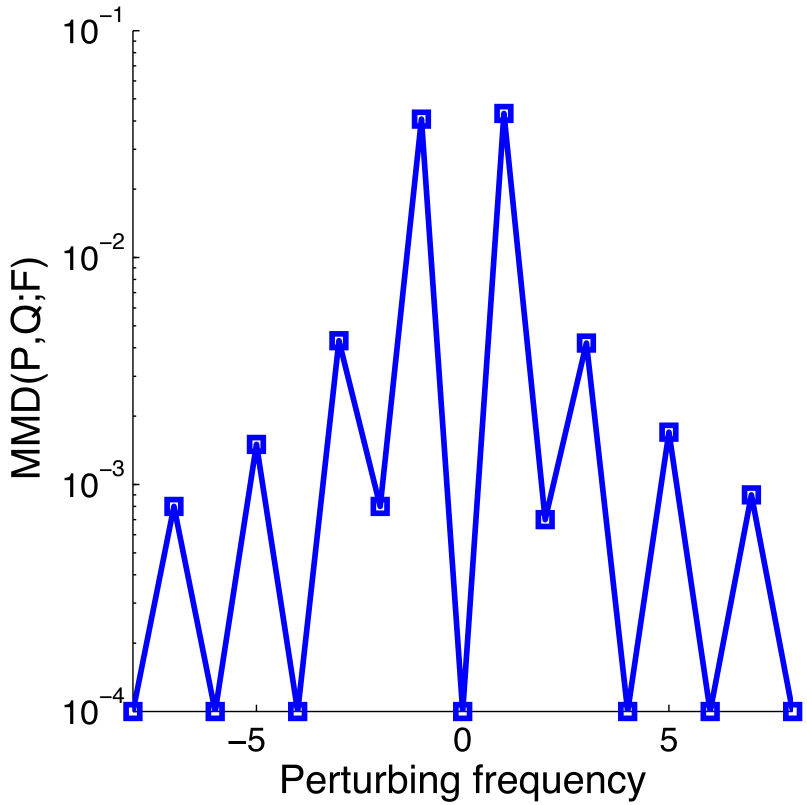
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = \sum_{l=-\infty}^{\infty} \frac{[|\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell}|\hat{k}_\ell]^2}{\hat{k}_\ell} = \sum_{l=-\infty}^{\infty} |\phi_{\mathbf{P},\ell} - \phi_{\mathbf{Q},\ell}|\hat{k}_\ell$$

Choosing the kernel

- B-spline kernel example



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



Summary: Characteristic Kernels

- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic for prob. measures on $\mathbb{R}^d$ if and only if $\text{supp}(\Lambda) = \mathbb{R}^d$ [COLT08, JMLR10]

Summary: Characteristic Kernels

- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic for prob. measures on $\mathbb{R}^d$ if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ [COLT08, JMLR10]
  - Corollary: continuous, compactly supported k characteristic

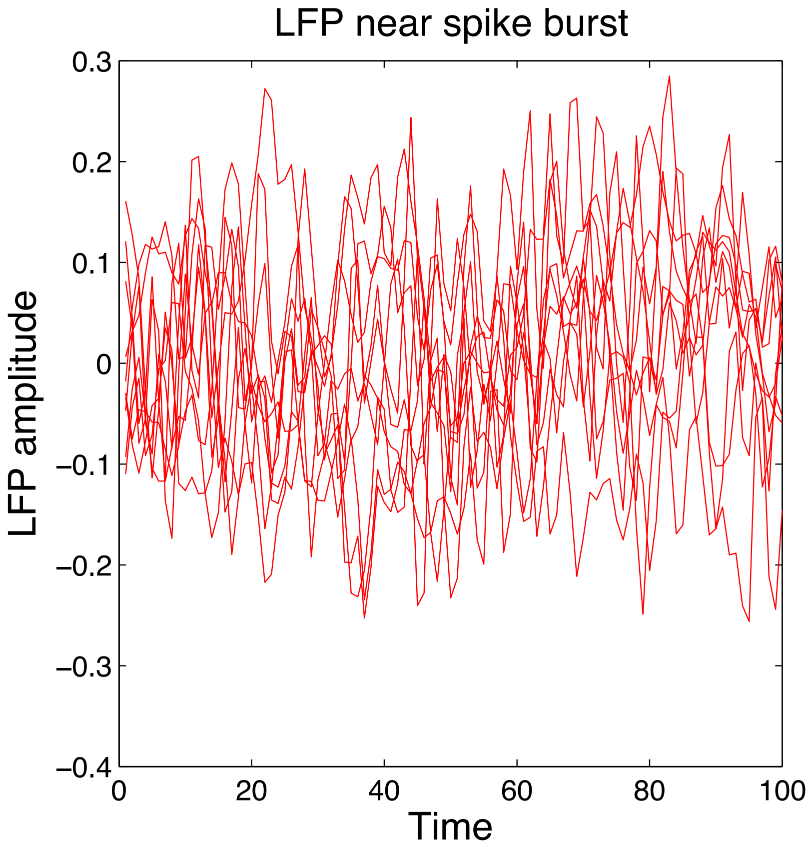
Summary: Characteristic Kernels

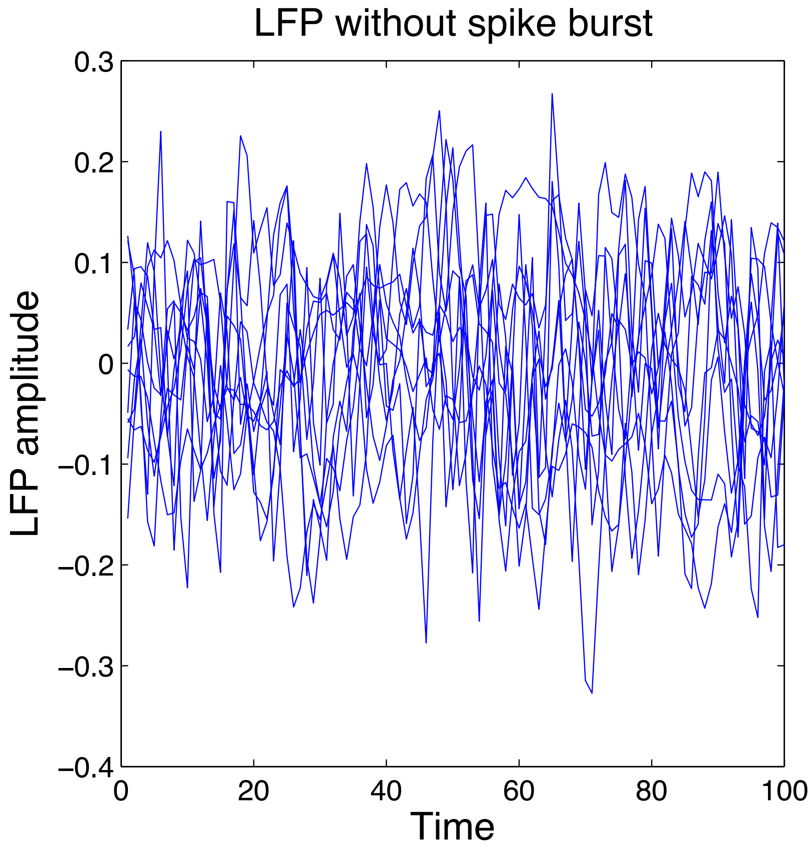
- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic for prob. measures on $\mathbb{R}^d$ if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ [COLT08, JMLR10]
  - Corollary: continuous, compactly supported k characteristic
- Similar reasoning wherever extensions of Bochner's theorem exist: [NIPS08a]
  - Locally compact Abelian groups (periodic domains, as we saw)
  - Compact, non-Abelian groups (orthogonal matrices)
  - The semigroup $\mathbb{R}_n^+$ (histograms)

Statistical hypothesis testing

Reminder: detecting differences in brain signals

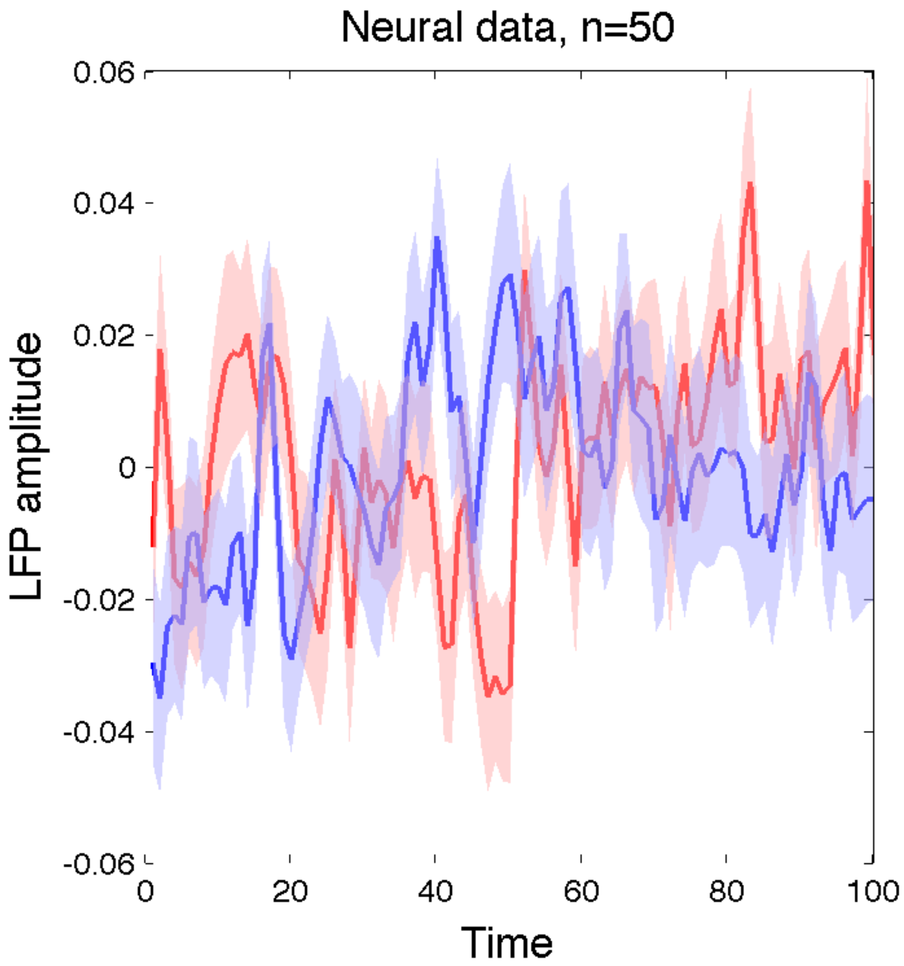
The problem: Do local field potential (LFP) signals change when measured near a spike burst?





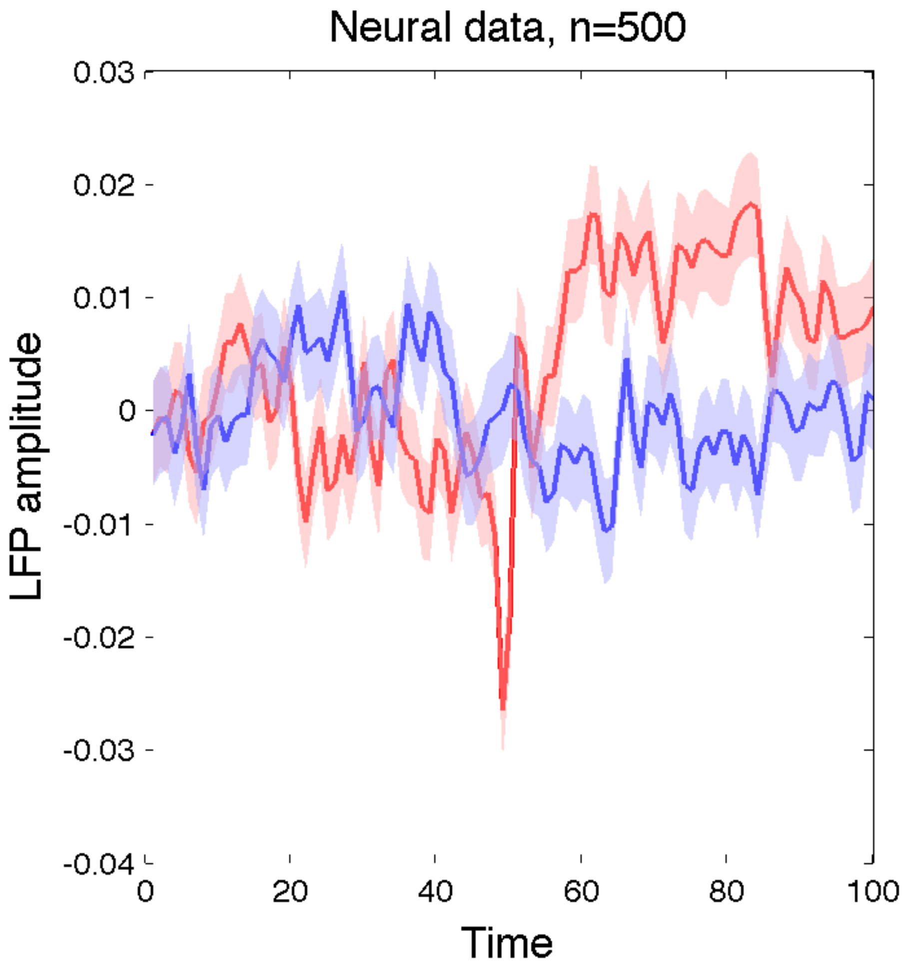
Reminder: detecting differences in brain signals

The problem: Do local field potential (LFP) signals change when measured near a spike burst?



Reminder: detecting differences in brain signals

The problem: Do local field potential (LFP) signals change when measured near a spike burst?



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 ($\mathsf{P} = \mathsf{Q}$)
  - H_1: alternative hypothesis ($\mathsf{P} \neq \mathsf{Q}$)
- Observe samples $\boldsymbol{x} := \{x_1, \ldots, x_n\}$ from P and $\boldsymbol{y}$ from 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 (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):

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

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):

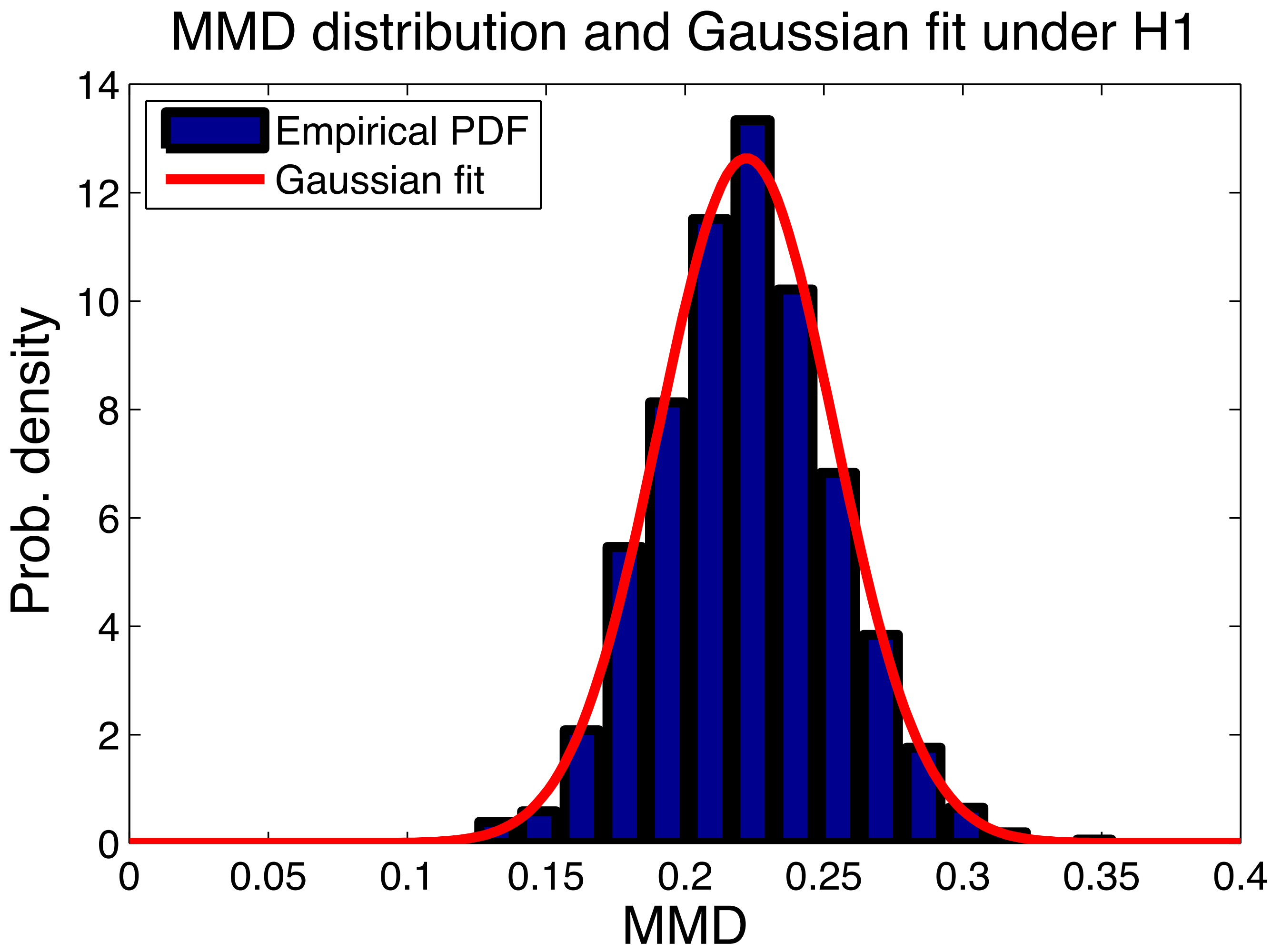
$$\mathrm{MMD}(\boldsymbol{x}, \boldsymbol{y}; F) = \frac{1}{n(n-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{h((x_i, y_i), (x_j, y_j))}$$

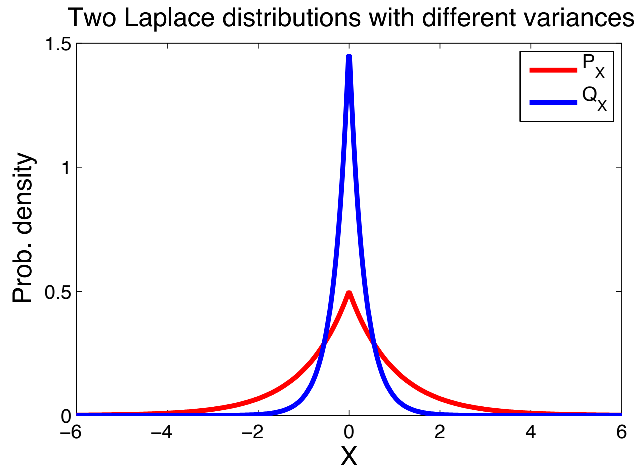
- When $\mathbf{P} \neq \mathbf{Q}$, asymptotically normal [Hoeffding, 1948, Serfling, 1980]
- Expression for the variance: $z_i := (x_i, y_i)$

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

Statistical test using MMD (3)

- Example: laplace distributions with different variance





Statistical test using MMD (4)

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

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

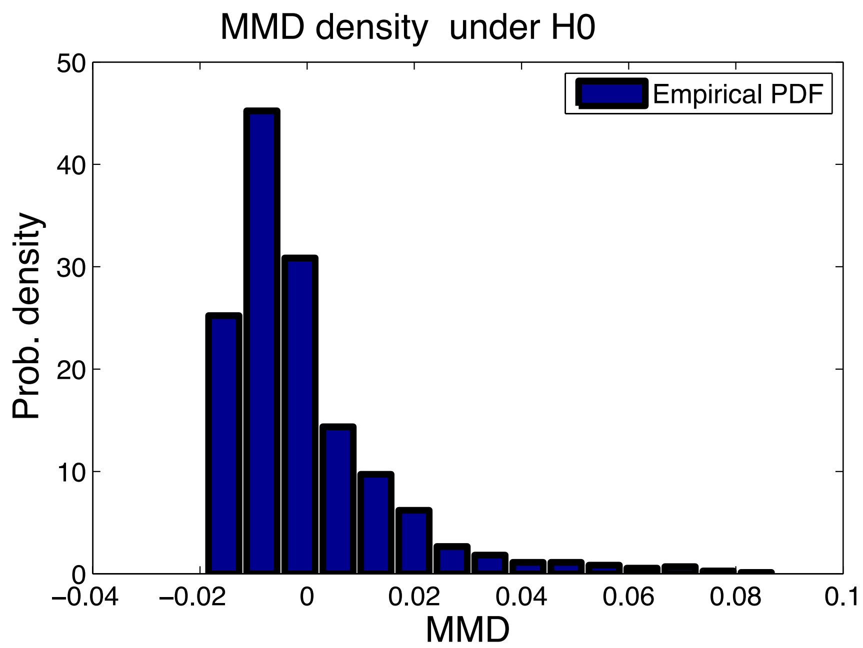
- where
  - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
  - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathbf{P}(x) = \lambda_i \psi_i(x')$

Statistical test using MMD (4)

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

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

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



Statistical test using MMD (5)

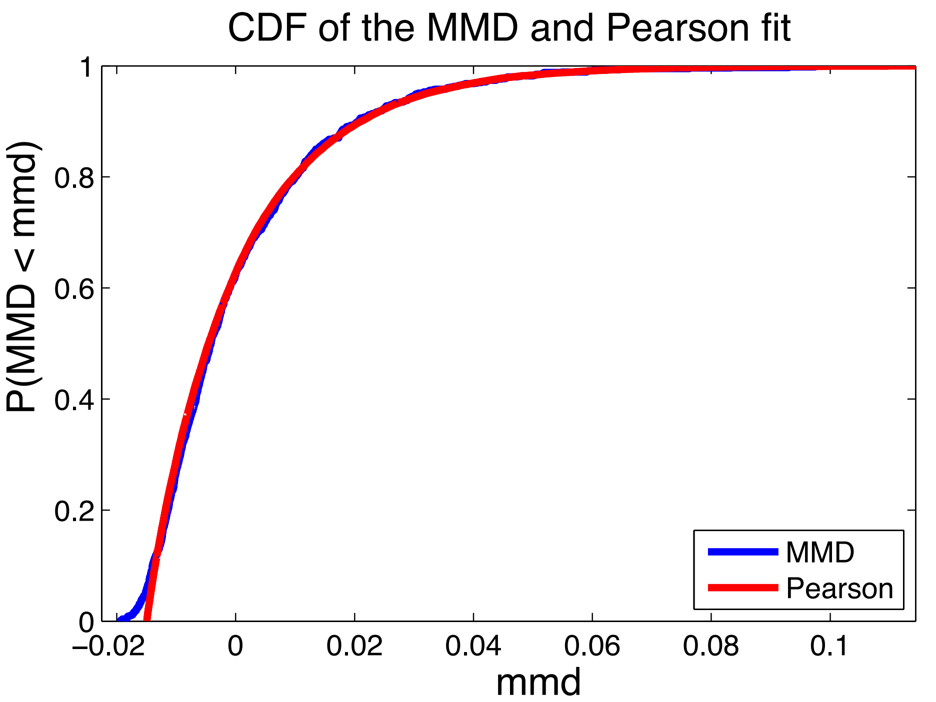
- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\mathrm{MMD} > T) \leq 0.05$

Statistical test using MMD (5)

- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\text{MMD} > T) \leq 0.05$
- Permutation 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]
- Consistent test using kernel eigenspectrum [NIPS09b]

Statistical test using MMD (5)

- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\mathrm{MMD} > T) \leq 0.05$
- Permutation 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]
- Consistent test using kernel eigenspectrum [NIPS09b]



Consistent test w/o bootstrap (not examinable)

- Maximum mean discrepancy (MMD): distance between **P** and **Q**

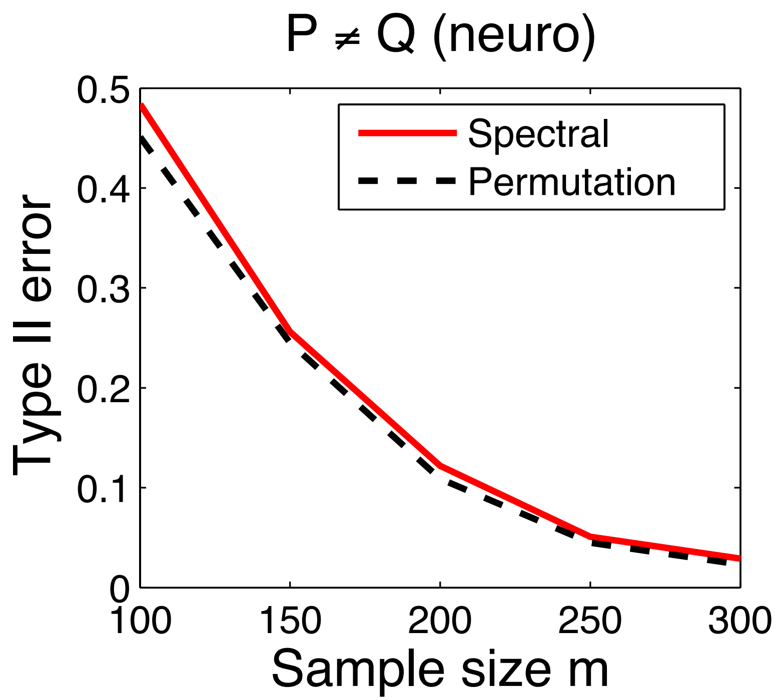
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}}$$

- Is $\widehat{\mathrm{MMD}}$ significantly > 0?

- $\mathbf{P} = \mathbf{Q}$, null distrib. of $\widehat{\mathrm{MMD}}$:

$$n\widehat{\mathrm{MMD}} \xrightarrow[D]{} \sum_{l=1}^{\infty} \lambda_l (z_l^2 - 2),$$

  - λ_l is lth eigenvalue of kernel $\tilde{k}(x_i, x_j)$



Use Gram matrix spectrum for $\hat{\lambda}_l$: consistent test without bootstrap

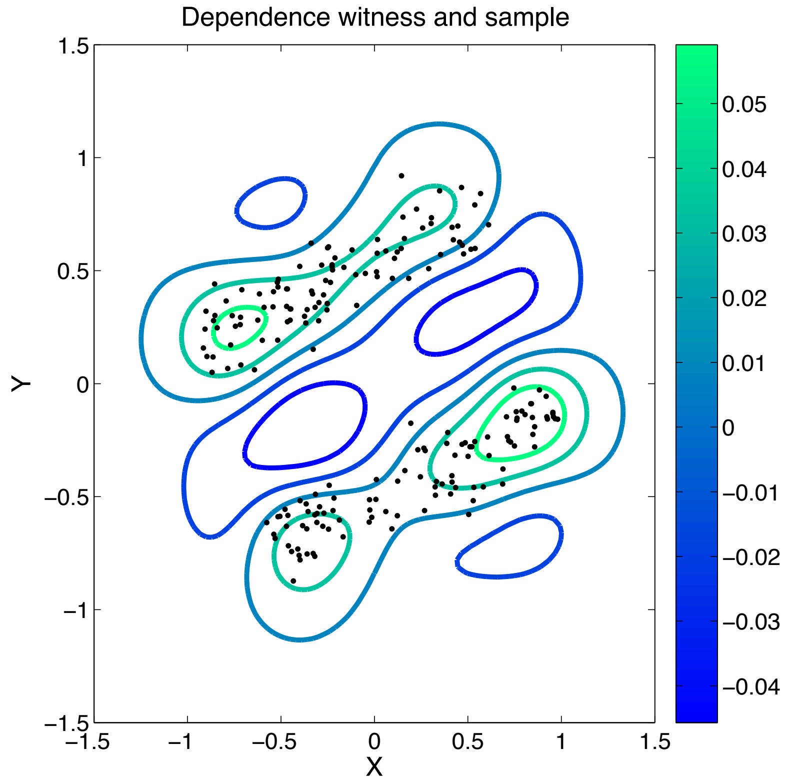
Kernel dependence measures

Reminder: MMD can be used as a dependence measure

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

$$\left(\sup_f \left[\mathbf{E}_{\mathsf{P}_{XY}} f - \mathbf{E}_{\mathsf{P}_X \mathsf{P}_Y} f\right]\right)^2 = \sup_{\|f\| \leq 1} \langle f, \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \rangle^2_{\mathcal{F} \times \mathcal{G}}$$

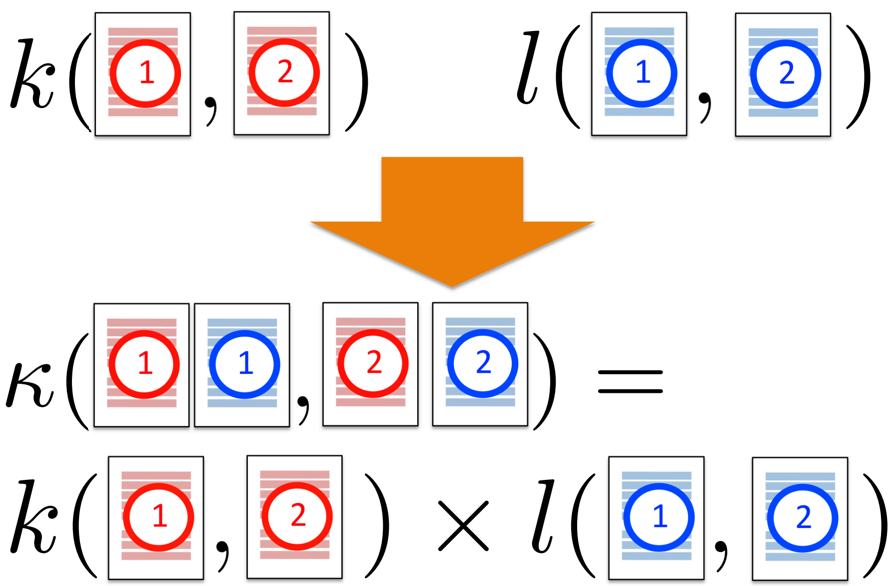
$$= \|\mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y}\|^2_{\mathcal{F} \times \mathcal{G}} := HSIC(\mathsf{P}_{XY}, \mathsf{P}_X \mathsf{P}_Y)$$



Reminder: MMD can be used as a dependence measure

- Dependence measure: [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

$$
\begin{aligned}
\left(\sup_{f}\left[\mathbf{E}_{\mathsf{P}_{XY}} f - \mathbf{E}_{\mathsf{P}_X \mathsf{P}_Y} f\right]\right)^2 &= \sup_{\|f\|\leq 1} \left\langle f, \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\rangle^2_{\mathcal{F}\times\mathcal{G}} \\
= \left\| \mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y} \right\|^2_{\mathcal{F}\times\mathcal{G}} &:= HSIC(\mathsf{P}_{XY}, \mathsf{P}_X \mathsf{P}_Y)
\end{aligned}
$$



Distribution of HSIC at independence

- (Biased) empirical HSIC a v-statistic

$$HSIC_b = \frac{1}{n^2}\mathrm{trace}(KHLH)$$

  – Statistical testing: How do we find when this is larger enough that the null hypothesis $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$ is unlikely?

  – Formally: given $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$, what is the threshold T such that $\mathbf{P}(\mathrm{HSIC} > T) < \alpha$ for small α?

Distribution of HSIC at independence

- (Biased) empirical HSIC a v-statistic

$$HSIC_b = \frac{1}{n^2}\mathrm{trace}(KHLH)$$

- Associated U-statistic degenerate when $\mathbf{P} = \mathbf{P}_x\mathbf{P}_y$ [Serfling, 1980]:

$$n\mathrm{HSIC}_b \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \qquad z_l \sim \mathcal{N}(0,1)\mathrm{i.i.d.}$$

$$\lambda_l \psi_l(z_j) = \int h_{ijqr}\psi_l(z_i)dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv}$$

Distribution of HSIC at independence

- (Biased) empirical HSIC a v-statistic

$$HSIC_b = \frac{1}{n^2}\mathrm{trace}(KHLH)$$

- Associated U-statistic degenerate when $\mathbf{P} = \mathbf{P}_x\mathbf{P}_y$ [Serfling, 1980]:

$$n\mathrm{HSIC}_b \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \qquad z_l \sim \mathcal{N}(0,1)\text{i.i.d.}$$

$$\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv}$$

- First two moments [NIPS07b]

$$\mathbf{E}(\mathrm{HSIC}_b) = \frac{1}{n}\mathrm{Tr}C_{xx}\mathrm{Tr}C_{yy}$$

$$\mathrm{var}(\mathrm{HSIC}_b) = \frac{2(n-4)(n-5)}{(n)_4}\|C_{xx}\|_{\mathrm{HS}}^2\|C_{yy}\|_{\mathrm{HS}}^2 + \mathrm{O}(n^{-3}).$$

Statistical testing with HSIC

- Given $\mathbf{P} = \mathbf{P}_{\mathsf{x}}\mathbf{P}_{\mathsf{y}}$, what is the threshold T such that $\mathbf{P}(\text{HSIC} > T) < \alpha$ for small α?
- Null distribution via permutation [Feuerverger, 1993]
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^{n}$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold T is $1 - \alpha$ quantile of empirical CDF

Statistical testing with HSIC

- Given $\mathsf{P} = \mathsf{P}_\mathsf{x}\mathsf{P}_\mathsf{y}$, what is the threshold T such that $\mathsf{P}(\text{HSIC} > T) < \alpha$ for small α?
- Null distribution via permutation [Feuerverger, 1993]
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold T is $1 - \alpha$ quantile of empirical CDF
- Approximate null distribution via moment matching [Kankainen, 1995]:

$$n\text{HSIC}_b(Z) \sim \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$$

where

$$\alpha = \frac{(\mathbf{E}(\text{HSIC}_b))^2}{\text{var}(\text{HSIC}_b)}, \quad \beta = \frac{\text{var}(\text{HSIC}_b)}{n\mathbf{E}(\text{HSIC}_b)}.$$

Experiment: dependence testing for translation

- (Biased) empirical HSIC:

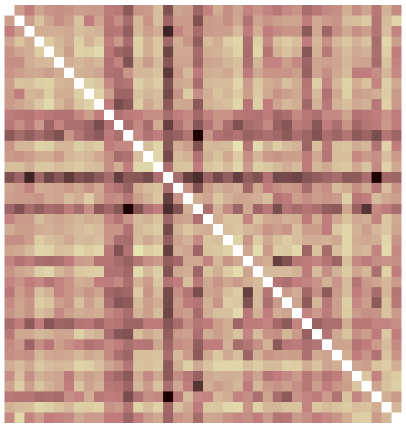
$$HSIC_b = \frac{1}{n^2}\text{trace}(KHLH)$$

- Translation example: [NIPS07b]
  Canadian Hansard
  (agriculture)
- 5-line extracts,
  k-spectrum kernel, $k = 10$,
  repetitions=300,
  sample size 10

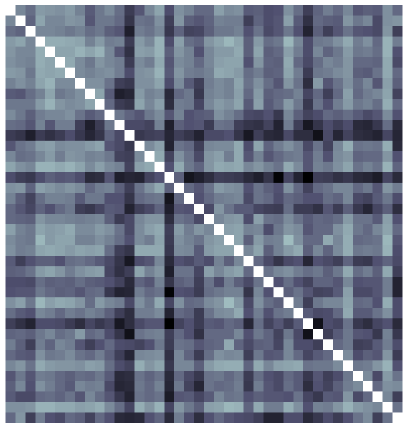
. . . no doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development. . .

. . . il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants. . .

$\Downarrow$ $\Downarrow$



$\Rightarrow$HSIC$\Leftarrow$



K L

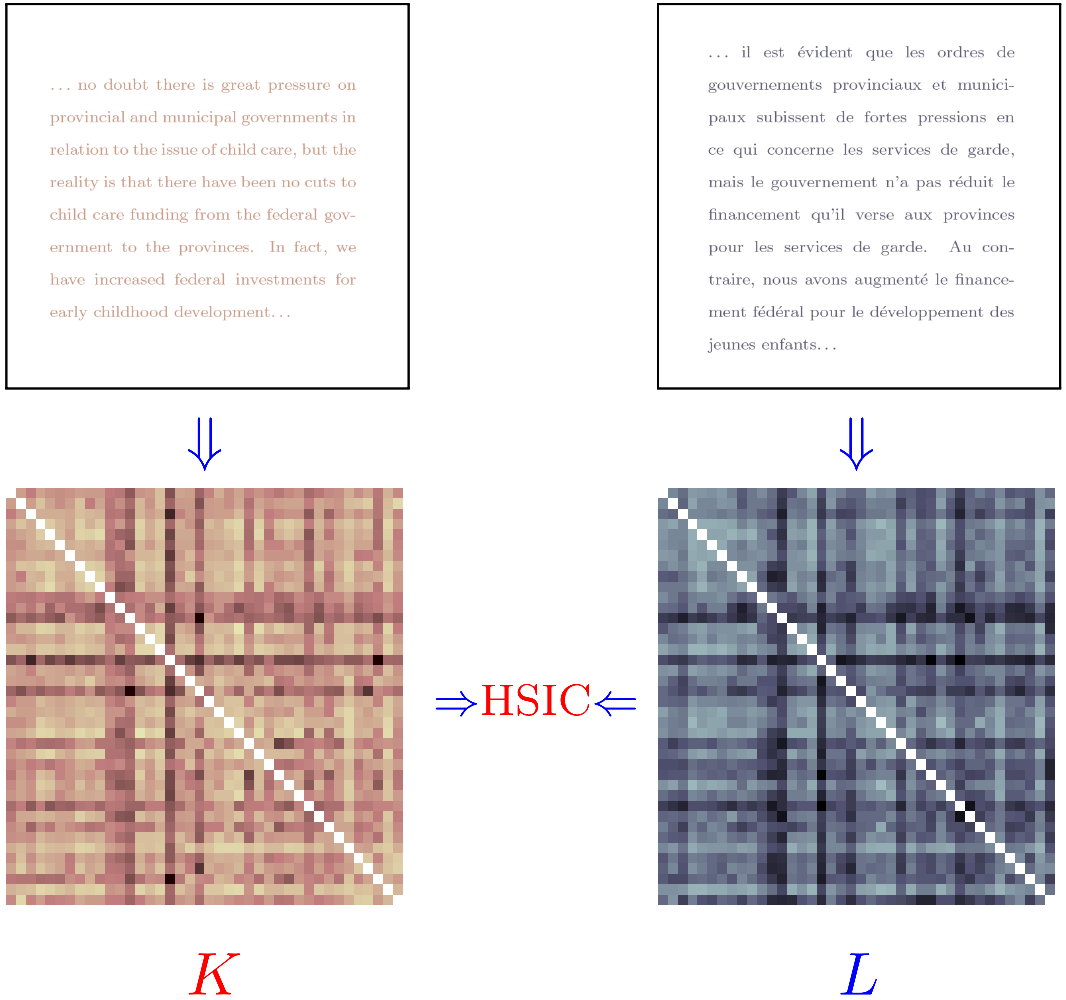
- k-spectrum kernel: average Type II error 0 ($\alpha = 0.05$)

Experiment: dependence testing for translation

- (Biased) empirical HSIC:

$$HSIC_b = \frac{1}{n^2}\mathrm{trace}(KHLH)$$

- Translation example: [NIPS07b]
  Canadian Hansard
  (agriculture)
- 5-line extracts,
  k-spectrum kernel, $k = 10$,
  repetitions=300,
  sample size 10



- k-spectrum kernel: average Type II error 0 ($\alpha = 0.05$)
- Bag of words kernel: average Type II error 0.18

Advanced topics

- Energy distance and the MMD
- Two-sample testing for big data
- Testing three-way interactions
- Bayesian inference without models

Summary

- MMD a distance between distributions [ISMB06, NIPS06a, JMLR10, JMLR12a]
  - high dimensionality
  - non-euclidean data (strings, graphs)
  - Nonparametric hypothesis tests
- Measure and test independence [ALT05, NIPS07a, NIPS07b, ALT08, JMLR10, JMLR12a]
- Characteristic RKHS: MMD a metric [NIPS07b, COLT08, NIPS08a]
  - Easy to check: does spectrum cover $\mathbb{R}^d$

Co-authors

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

Selected references

Characteristic kernels and mean embeddings:

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

Two-sample, independence, conditional independence tests:

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

Energy distance, relation to kernel distances

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

Three way interaction

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

Selected references (continued)

Conditional mean embedding, RKHS-valued regression:

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

Kernel Bayes rule:

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

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.

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

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

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

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

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

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 1.* John Wiley and Sons, 2nd edition, 1994.

A. Kankainen. *Consistent Testing of Total Independence Based on the Empirical Characteristic Function.* PhD thesis, University of Jyväskylä, 1995.

R. Lyons. Distance covariance in metric spaces. arXiv:1106.5758, to appear in Ann. Probab., June 2011.

C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

T. Read and N. Cressie. *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis*. Springer-Verlag, New York, 1988.

R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

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

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

G. Székely and M. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 4(3):1233–1303, 2009.

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