Lecture 2: Mapping of Probabilities to RKHS and Applications

MLSS Arequipa, Peru, 2016

Arthur Gretton

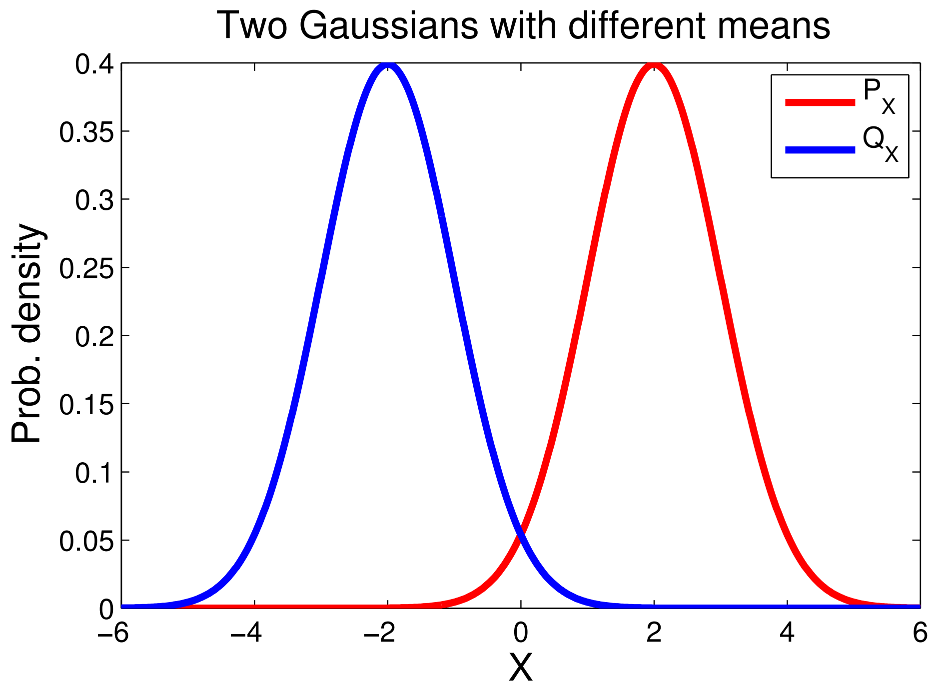
Gatsby Unit, CSML, UCL

Outline

- Kernel metric on the space of probability measures
  - Function revealing differences in distributions
  - Distance between means in space of features (RKHS)
  - Independence measure: features of joint minus product of marginals
- Characteristic kernels: feature space mappings of probabilities unique
- Two-sample, independence tests for (almost!) any data type
  - distributions on strings, images, graphs, groups (rotation matrices), semigroups,. . .

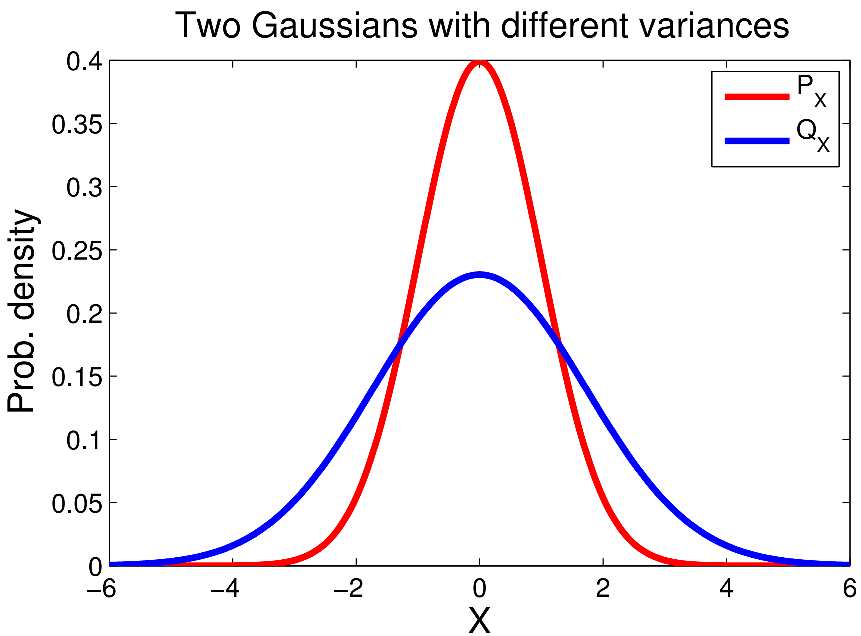
Feature mean difference

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



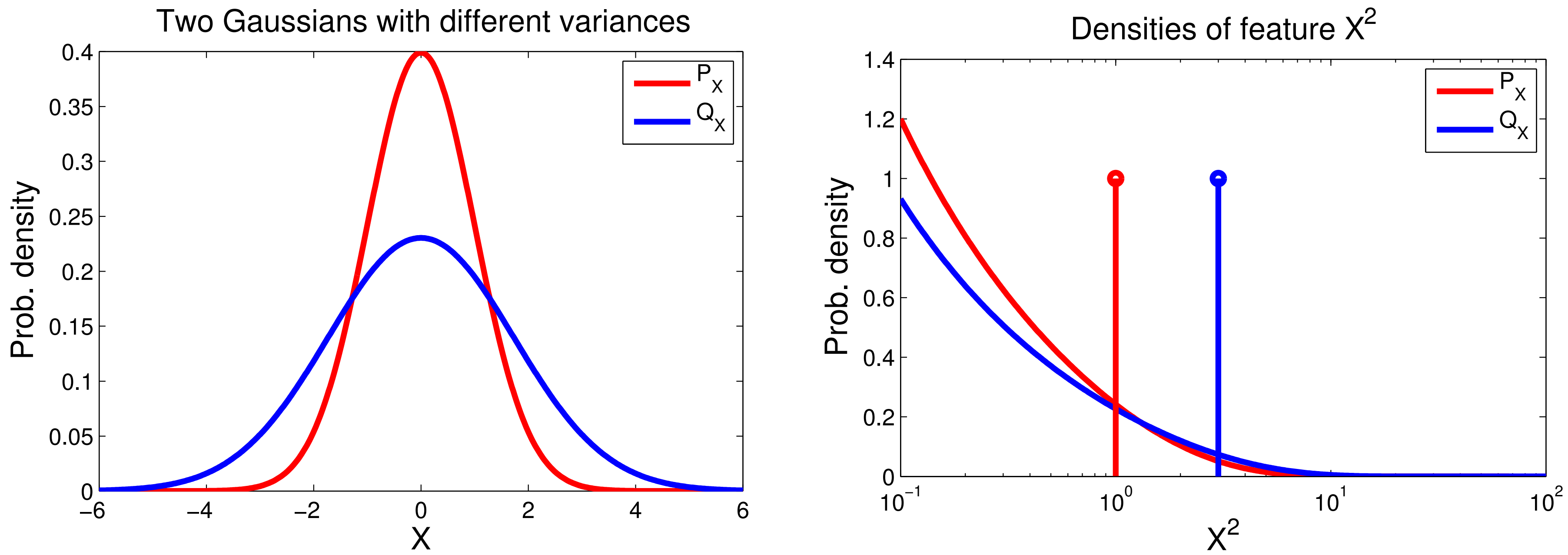
Feature mean difference

- 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 $\varphi(x) = x^2$



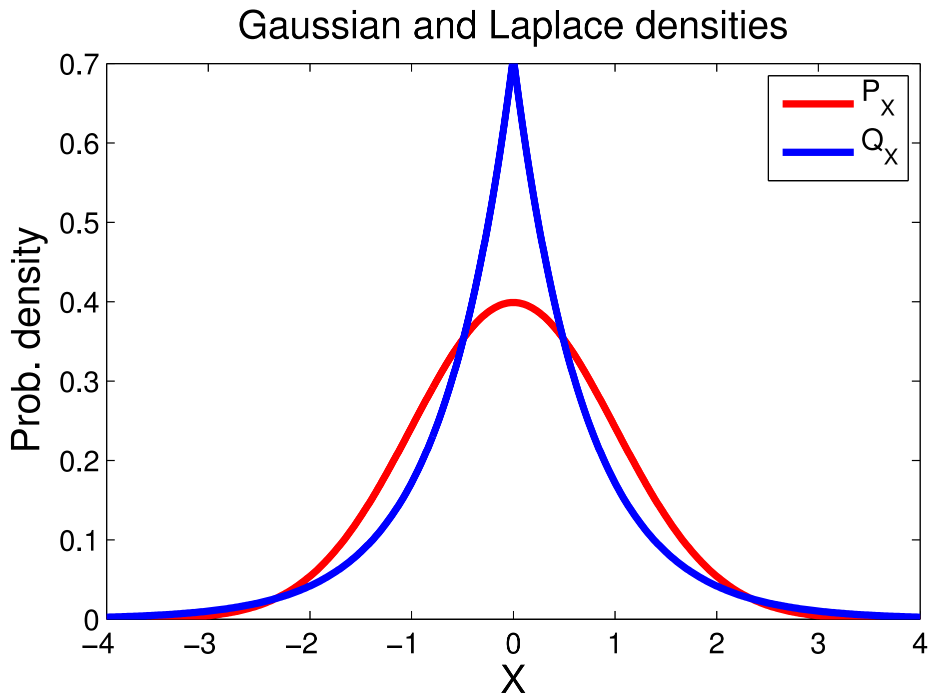
Feature mean difference

- 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



Probabilities in feature space: the mean trick

The reproducing property (kernel trick)

- Given $x \in \mathcal{X}$ for some set $\mathcal{X}$, define feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in \ell_2$$

- For positive definite $k(x, x')$,

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

- The reproducing property: $\forall f \in \mathcal{F}$,

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

Probabilities in feature space: the mean trick

The reproducing property (kernel trick)

- Given $x \in \mathcal{X}$ for some set $\mathcal{X}$, define feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_i(x) \ldots] \in \ell_2$$

- For positive definite $k(x, x')$,

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

- The reproducing property: $\forall f \in \mathcal{F}$,

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

The mean trick

- Given P a Borel probability measure on $\mathcal{X}$, define feature map $\mu_{\mathsf{P}} \in \mathcal{F}$

$$\mu_{\mathsf{P}} = [\ldots \mathbf{E}_{\mathsf{P}} [\varphi_i(\mathsf{x})] \ldots]$$

- For positive definite $k(x, x')$,

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

for $\mathsf{x} \sim \mathsf{P}$ and $\mathsf{y} \sim \mathsf{Q}$.

- The mean trick: (we call μ_{P} a mean/distribution embedding)

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

Does the feature space mean exist?

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

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

Does the feature space mean exist?

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

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

Yes: You can exchange expectation and innner product (i.e. $\varphi(\mathsf{x})$ is Bochner integrable [Steinwart and Christmann, 2008]) under the condition

$$\mathbf{E}_{\mathsf{P}} \|\varphi(\mathsf{x})\|_{\mathcal{F}} = \mathbf{E}_{\mathsf{P}} \sqrt{k(\mathsf{x}, \mathsf{x})} < \infty$$

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$MMD^2(\mathsf{P},\mathsf{Q}) = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|^2_{\mathcal{F}} = \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle_{\mathcal{F}} + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle_{\mathcal{F}}$$

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

(a)= within distrib. similarity, (b)= cross-distrib. similarity

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$MMD^2(\mathbf{P},\mathbf{Q}) = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}} = \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle_{\mathcal{F}} + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} - 2\langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}}$$

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

(a)= within distrib. similarity, (b)= cross-distrib. similarity

Proof:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$
\begin{aligned}
MMD^2(\mathbf{P},\mathbf{Q}) = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}} &= \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle_{\mathcal{F}} + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\
&= \underbrace{\mathbf{E}_{\mathbf{P}} k(x, x')}_{\text{(a)}} + \underbrace{\mathbf{E}_{\mathbf{Q}} k(y, y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

(a)= within distrib. similarity, (b)= cross-distrib. similarity

Proof:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$MMD^2(\mathbf{P},\mathbf{Q}) = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}} = \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle_{\mathcal{F}} + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}}$$

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

(a)= within distrib. similarity, (b)= cross-distrib. similarity

Proof:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$MMD^2(\mathbf{P},\mathbf{Q}) = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}} = \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle_{\mathcal{F}} + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} - 2\,\langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}}$$
$$= \underbrace{\mathbf{E}_{\mathbf{P}} k(x, x')}_{\text{(a)}} + \underbrace{\mathbf{E}_{\mathbf{Q}} k(y, y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)}_{\text{(b)}}$$

(a)= within distrib. similarity, (b)= cross-distrib. similarity

Proof:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$MMD^2(\mathsf{P},\mathsf{Q}) = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|^2_{\mathcal{F}} = \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle_{\mathcal{F}} + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle_{\mathcal{F}}$$

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

(a)= within distrib. similarity, (b)= cross-distrib. similarity

Proof:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$MMD^2(\mathsf{P},\mathsf{Q}) = \|\mu_\mathsf{P} - \mu_\mathsf{Q}\|^2_\mathcal{F} = \langle \mu_\mathsf{P}, \mu_\mathsf{P} \rangle_\mathcal{F} + \langle \mu_\mathsf{Q}, \mu_\mathsf{Q} \rangle_\mathcal{F} - 2\langle \mu_\mathsf{P}, \mu_\mathsf{Q} \rangle_\mathcal{F}$$

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

(a)= within distrib. similarity, (b)= cross-distrib. similarity

Proof:

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

The maximum mean discrepancy

The maximum mean discrepancy is the distance between feature means:

$$MMD^2(\mathbf{P},\mathbf{Q}) = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}} = \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle_{\mathcal{F}} + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle_{\mathcal{F}}$$
$$= \underbrace{\mathbf{E}_{\mathbf{P}} k(\mathsf{x},\mathsf{x}')}_{\text{(a)}} + \underbrace{\mathbf{E}_{\mathbf{Q}} k(\mathsf{y},\mathsf{y}')}_{\text{(a)}} - \underbrace{2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x},\mathsf{y})}_{\text{(b)}}$$

(a)= within distrib. similarity, (b)= cross-distrib. similarity

Unbiased empirical estimate of first term (quadratic time)

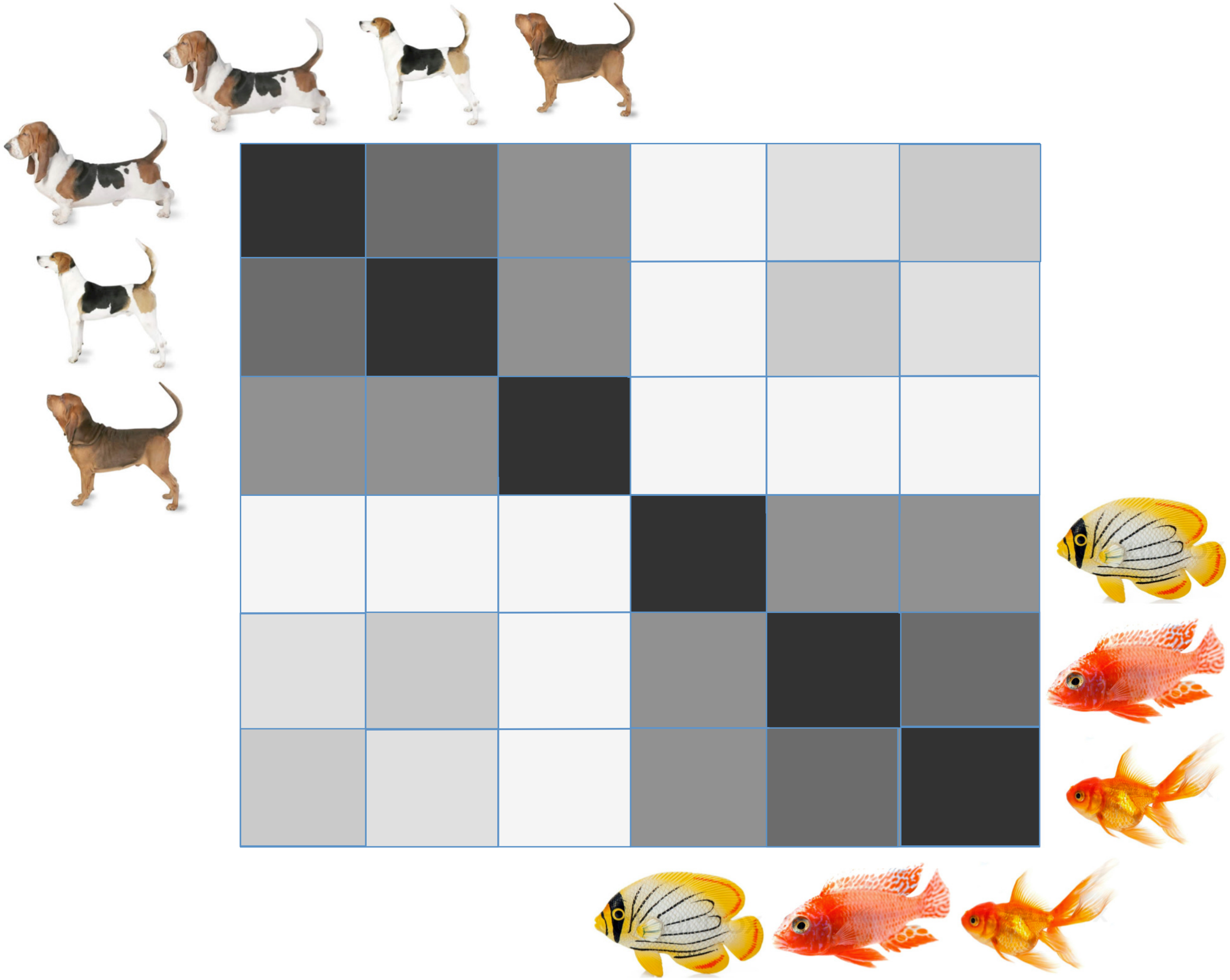
$$\widehat{\mathbb{E}}_{\mathbf{P}} k(x,x') = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j)$$

The maximum mean discrepancy

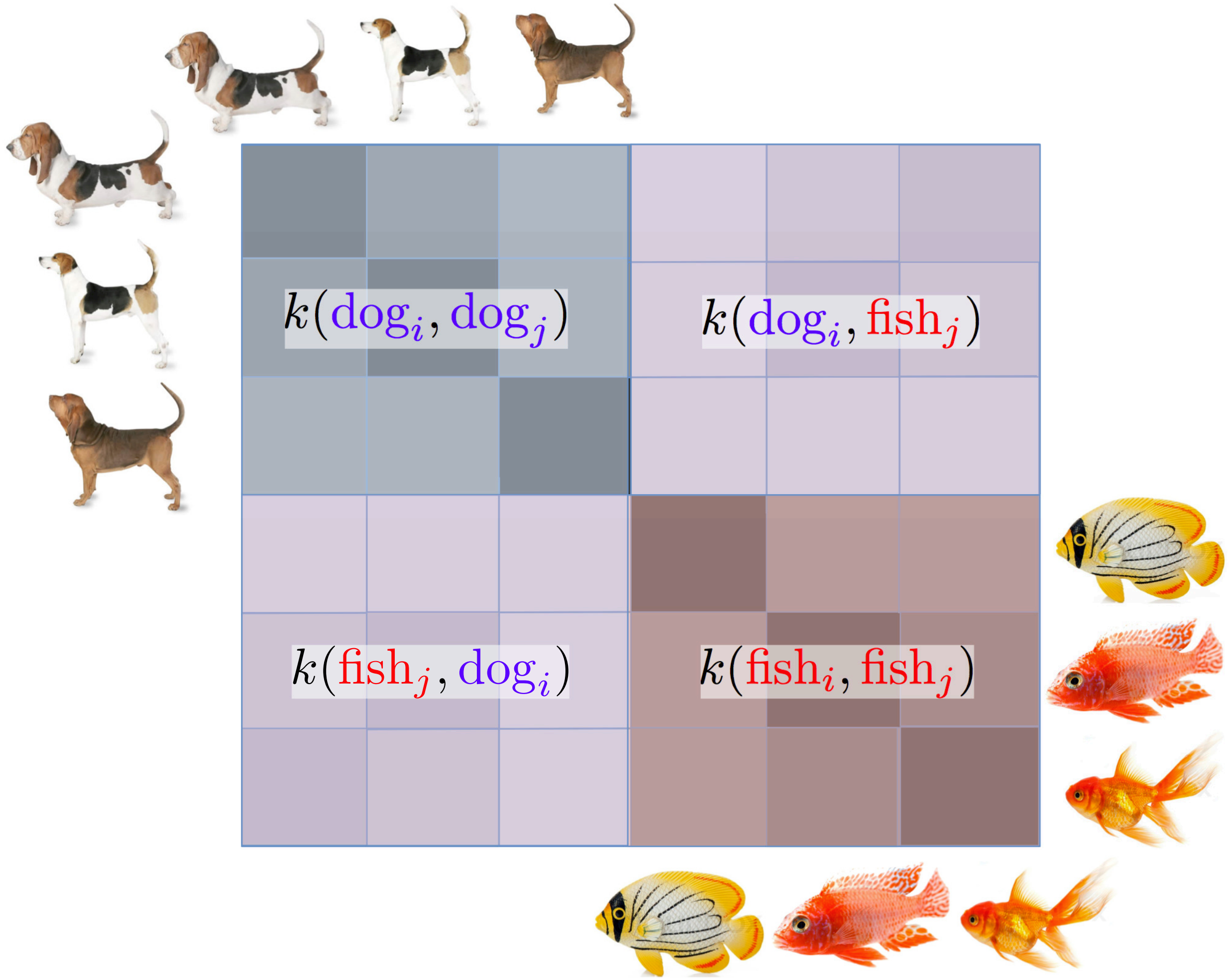
$$\sim P$$

$$\sim Q$$

The maximum mean discrepancy



The maximum mean discrepancy



$$\widehat{MMD}^2 = \overline{K_{P,P}} + \overline{K_{Q,Q}} - 2\overline{K_{P,Q}}$$

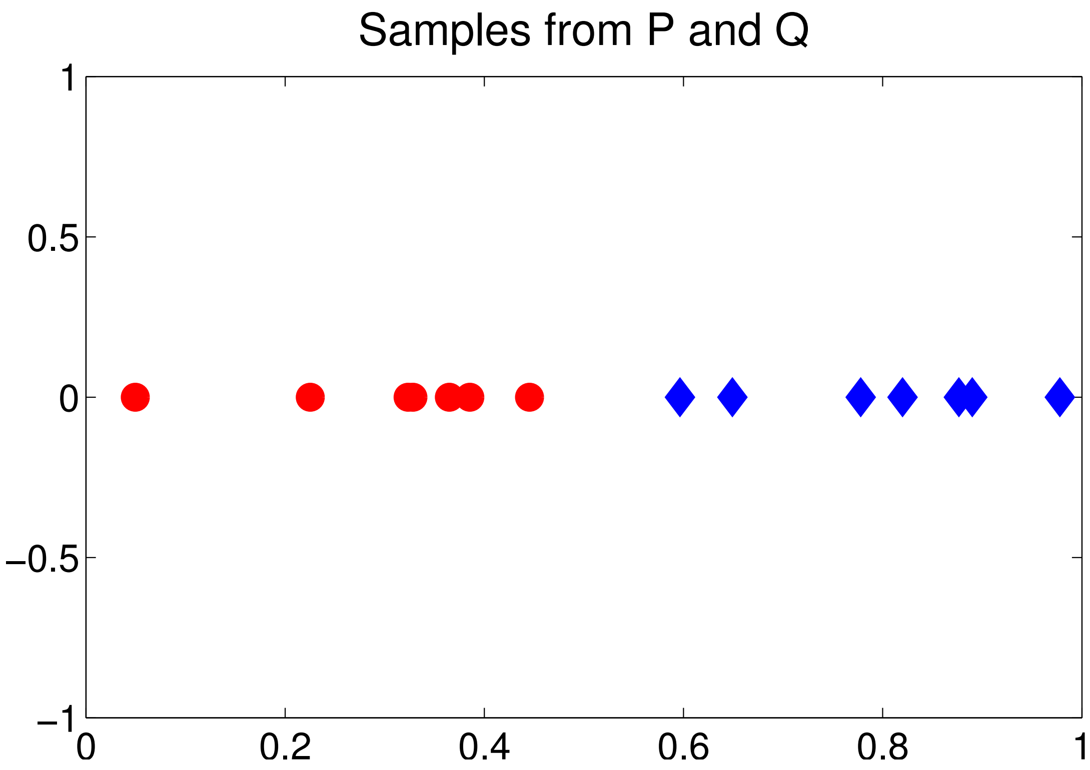
(diagonal terms removed from $K_{P,P}$ and $K_{Q,Q}$)

Function Showing Difference in Distributions

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

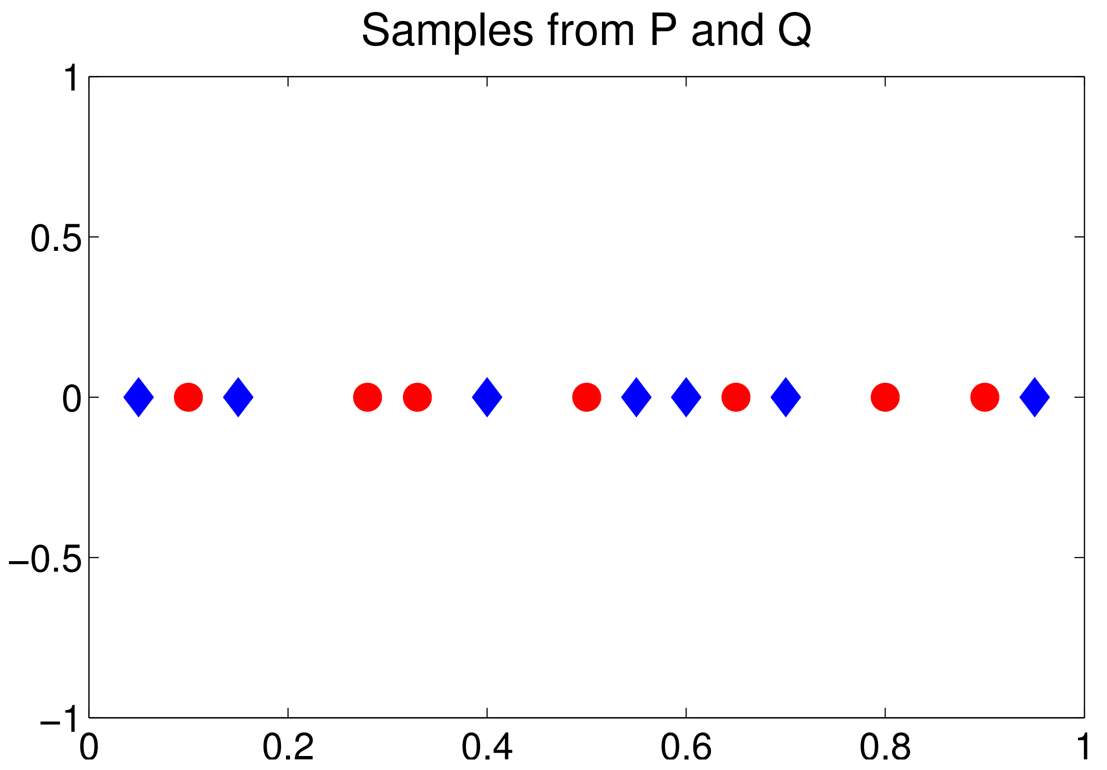
Function Showing Difference in Distributions

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



Function Showing Difference in Distributions

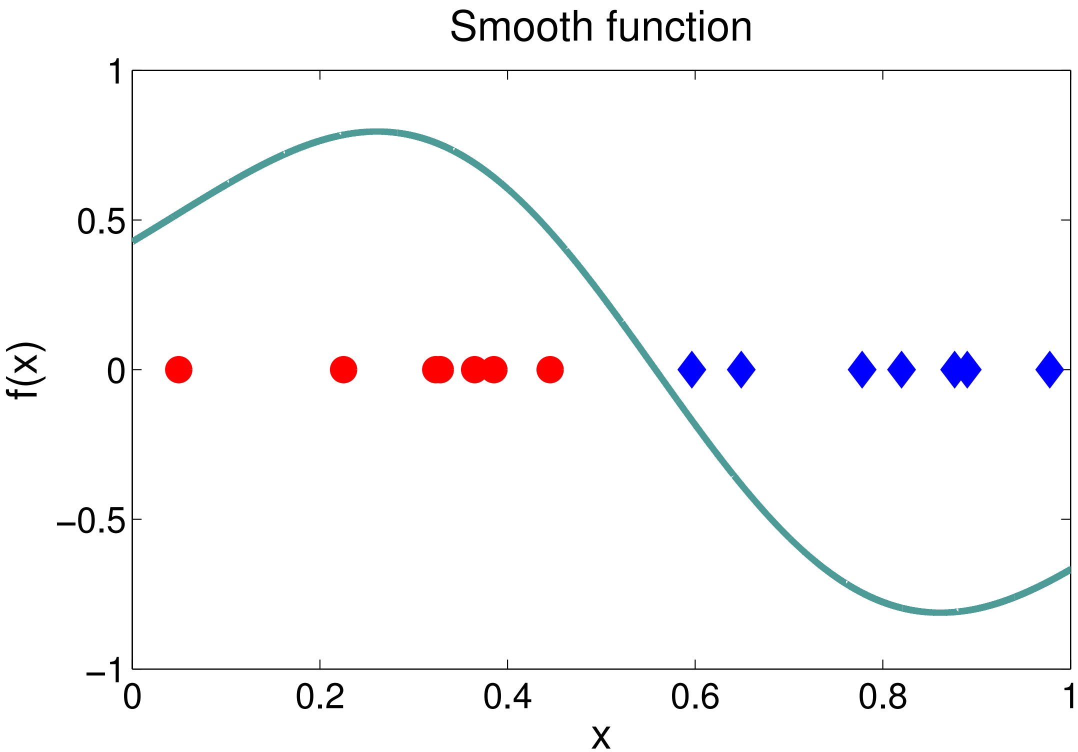
- Are **P** and **Q** different?



Function Showing Difference in Distributions

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

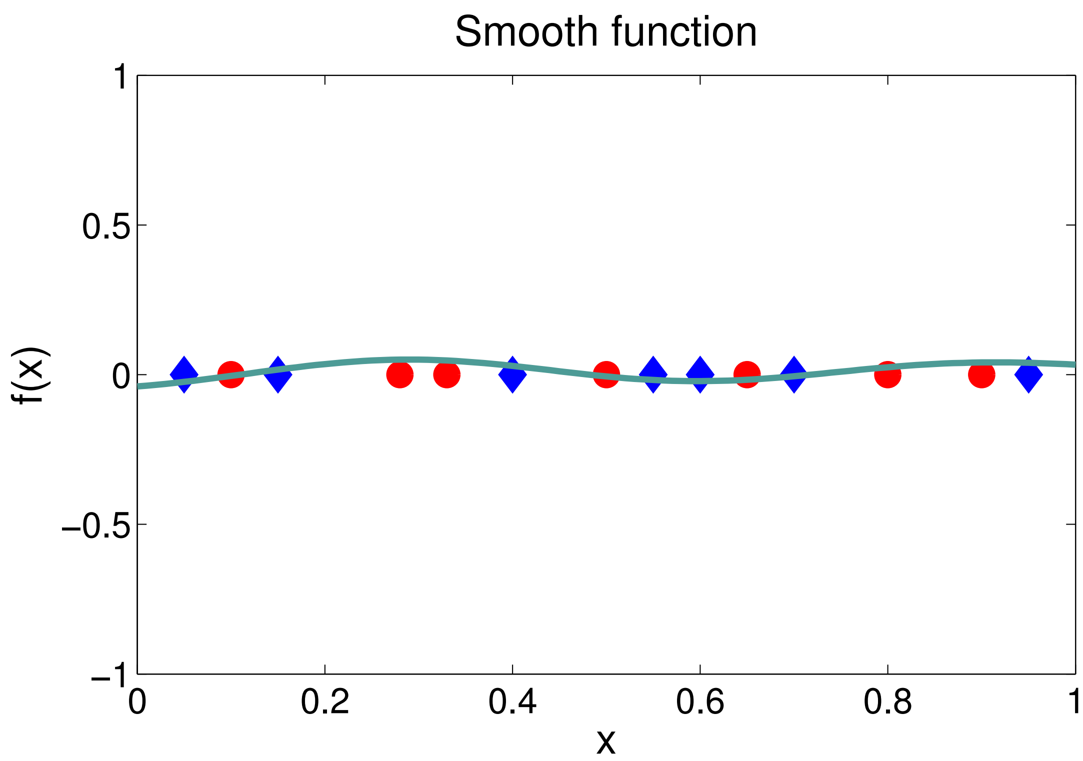
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathrm{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathrm{y}) \right].$$



Function Showing Difference in Distributions

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

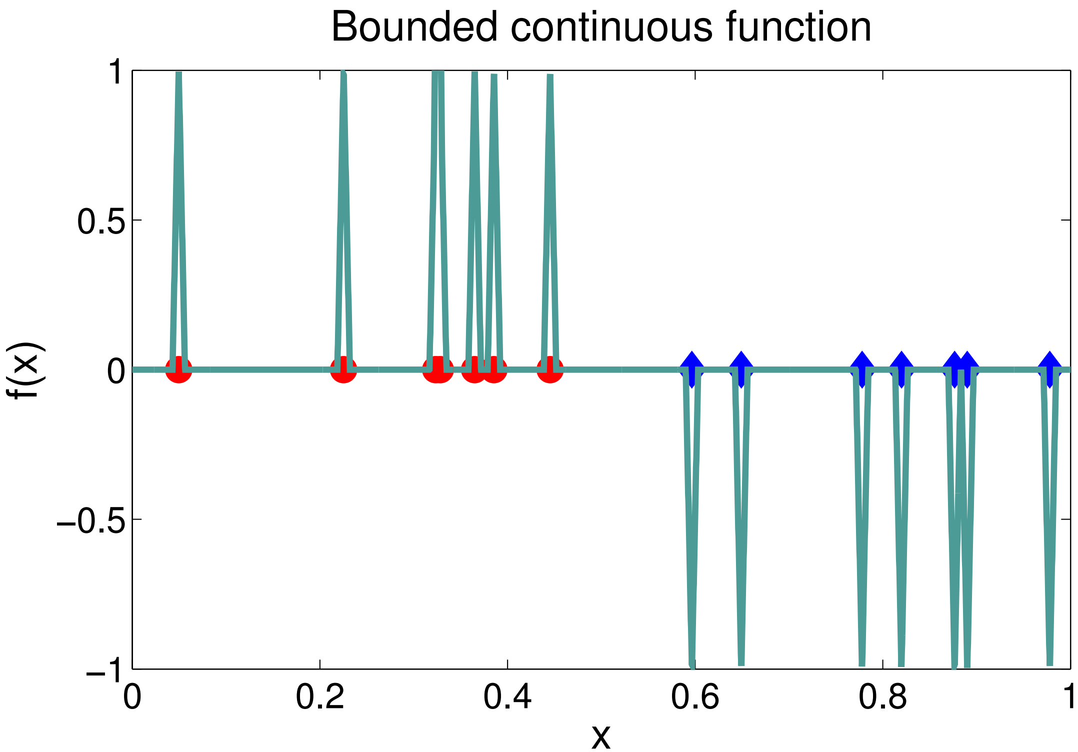
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

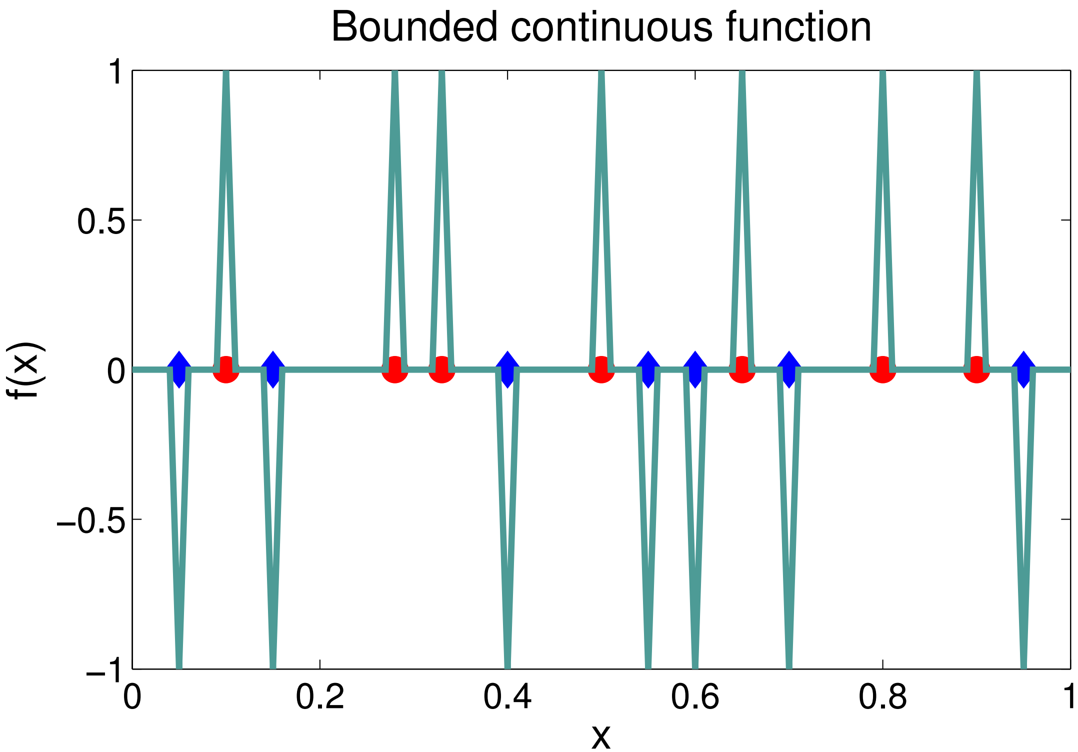
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(x) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(y)\right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

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

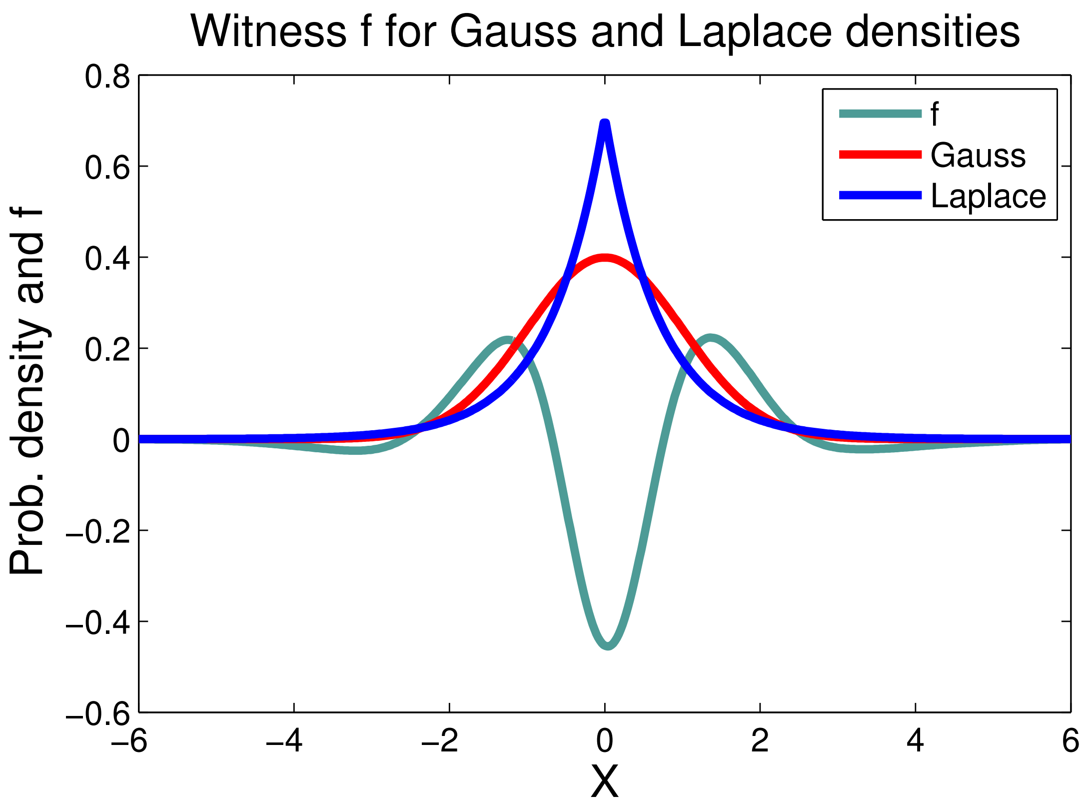


Function Showing Difference in Distributions

- 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 **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]
- $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$ when $F =$the unit ball in a characteristic RKHS $\mathcal{F}$ (coming soon!) [ISMB06, NIPS06a, NIPS07b, NIPS08a, JMLR10]

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 $\mathcal{F}$ (coming soon!) [ISMB06, NIPS06a, NIPS07b, NIPS08a, JMLR10]

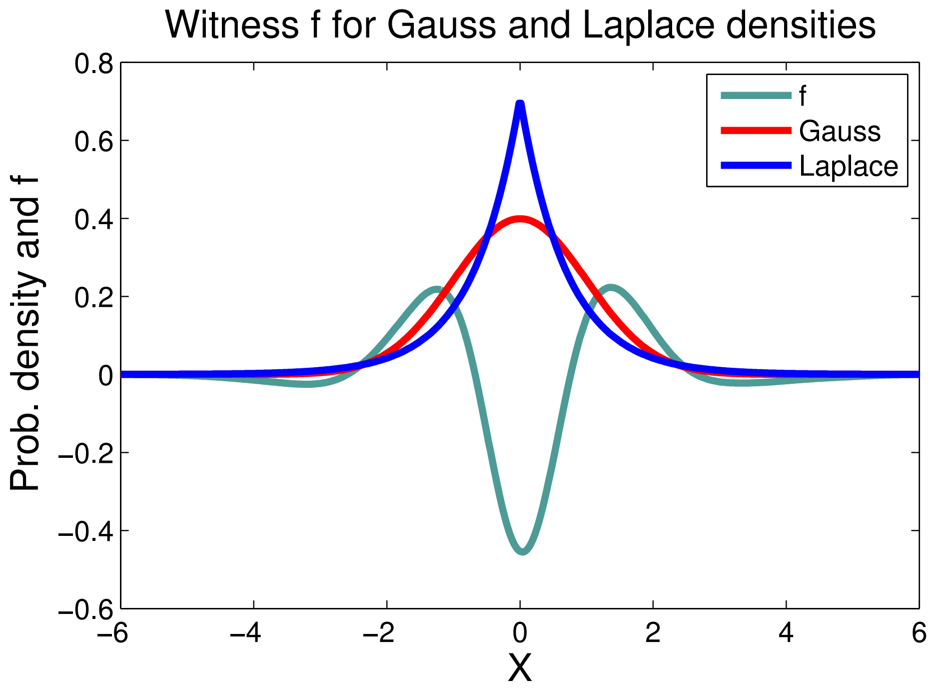
How do smooth functions relate to feature maps?

Function view vs feature mean view

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

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

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



Function view vs feature mean view

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

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

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]

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

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

$$= \sup_{f\in F}\langle f, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\rangle_{\mathcal{F}}$$

use

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

Function view vs feature mean view

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

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

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

$$= \sup_{f \in F} \langle f, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}}$$

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

use

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

since $F := \{f \in \mathcal{F} : \|f\| \leq 1\}$

Function view and feature view equivalent

MMD for independence: HSIC

- Dependence measure: the Hilbert Schmidt Independence Criterion [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

  Related to [Feuerverger, 1993]and [Székely and Rizzo, 2009, Székely et al., 2007]

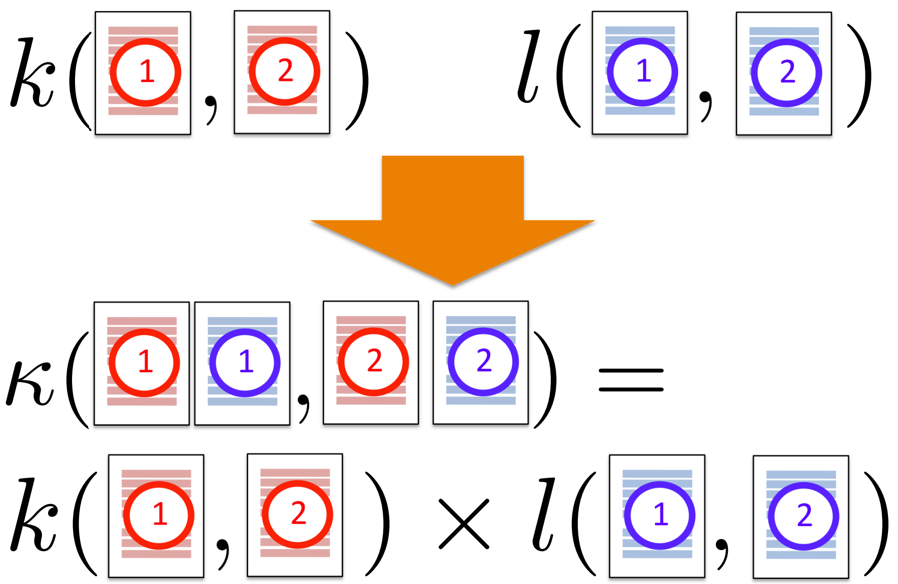
$$HSIC(\mathbf{P}_{XY}, \mathbf{P}_X \mathbf{P}_Y) := \left\| \mu_{\mathbf{P}_{XY}} - \mu_{\mathbf{P}_X \mathbf{P}_Y} \right\|^2$$

MMD for independence: HSIC

- Dependence measure: the Hilbert Schmidt Independence Criterion [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

  Related to [Feuerverger, 1993]and [Székely and Rizzo, 2009, Székely et al., 2007]

$$HSIC(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y) := \|\mu_{\mathbf{P}_{XY}} - \mu_{\mathbf{P}_X\mathbf{P}_Y}\|^2$$



MMD for independence: HSIC

- Dependence measure: the Hilbert Schmidt Independence Criterion [ALT05, NIPS07a, ALT07, ALT08, JMLR10]

  Related to [Feuerverger, 1993]and [Székely and Rizzo, 2009, Székely et al., 2007]

$$HSIC(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y) := \left\| \mu_{\mathbf{P}_{XY}} - \mu_{\mathbf{P}_X\mathbf{P}_Y} \right\|^2$$

HSIC using expectations of kernels:

Define RKHS $\mathcal{F}$ on $\mathcal{X}$ with kernel k, RKHS $\mathcal{G}$ on $\mathcal{Y}$ with kernel l. Then

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

HSIC: empirical estimate and intuition

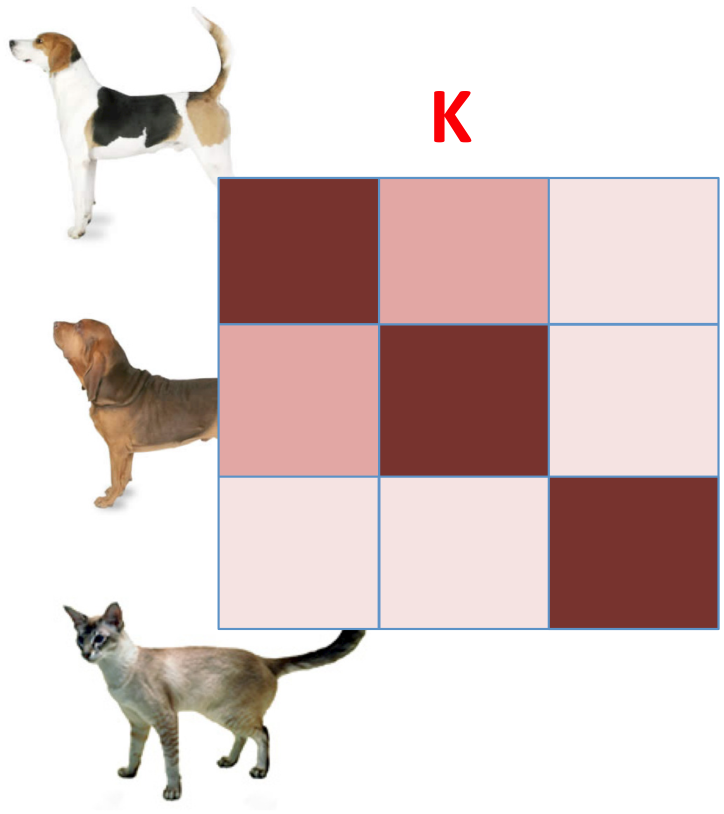
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose. They need a significant amount of exercise and mental stimulation.

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

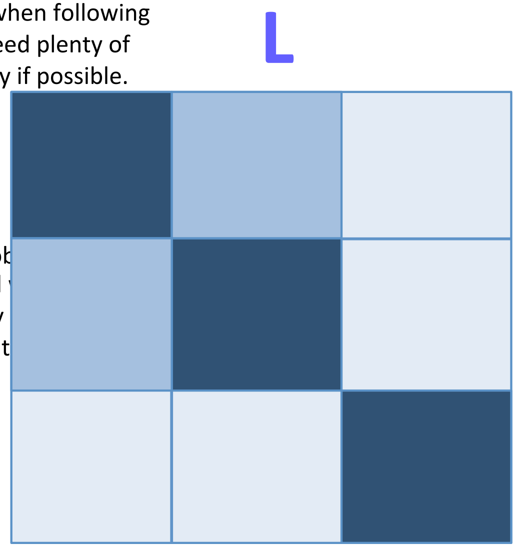
Text from dogtime.com and petfinder.com

HSIC: empirical estimate and intuition



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

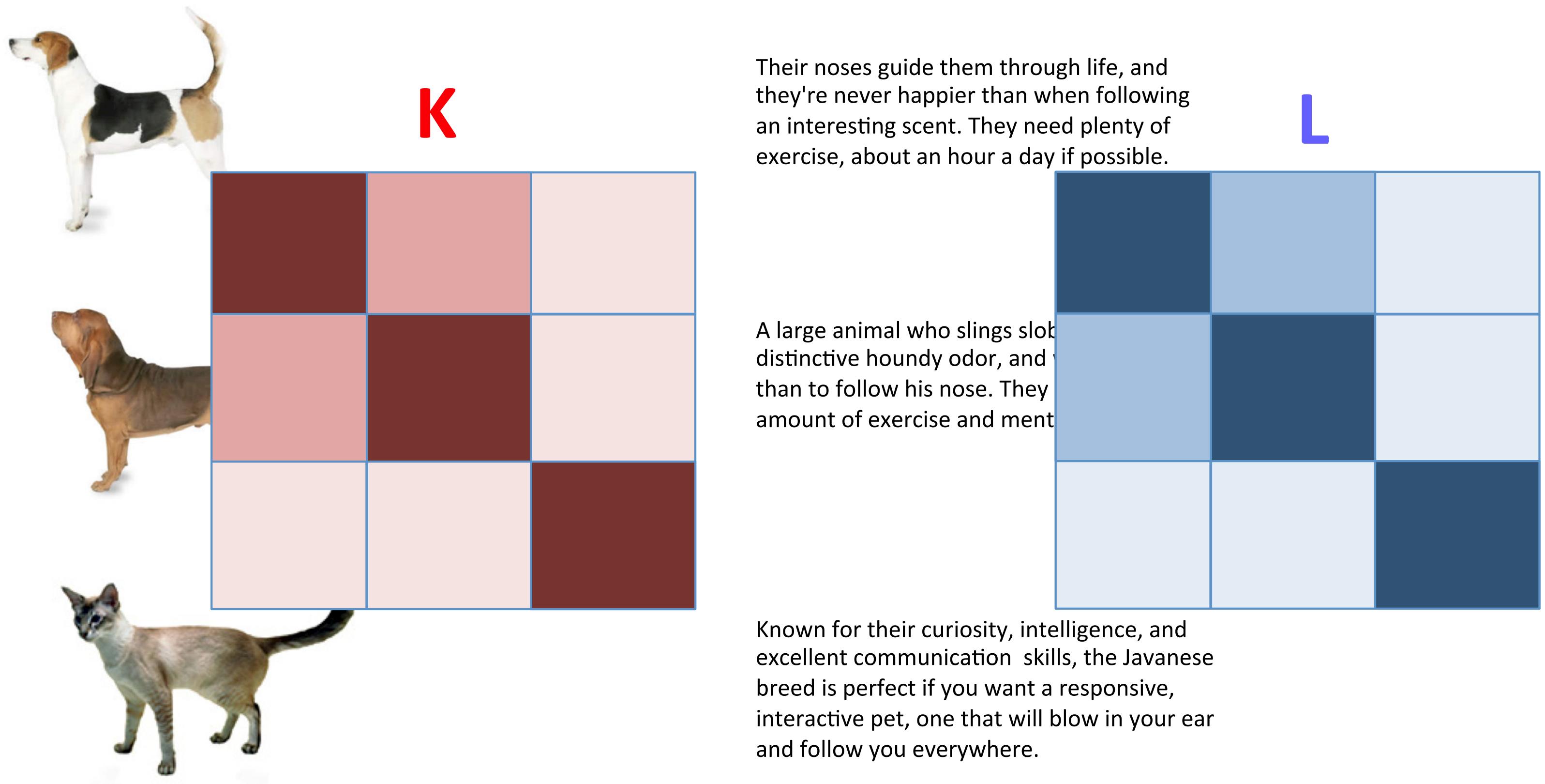
A large animal who slings slob
distinctive houndy odor, and
than to follow his nose. They
amount of exercise and ment



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

Text from dogtime.com and petfinder.com

HSIC: empirical estimate and intuition



Text from dogtime.com and petfinder.com

Empirical $HSIC(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y)$:

$$\frac{1}{n^2}\left(HKH \circ HLH\right)_{++}$$

Characteristic kernels (Via Fourier, on the torus $\mathbb{T}$)

Characteristic Kernels (via Fourier)

Reminder:

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

In the next slides:

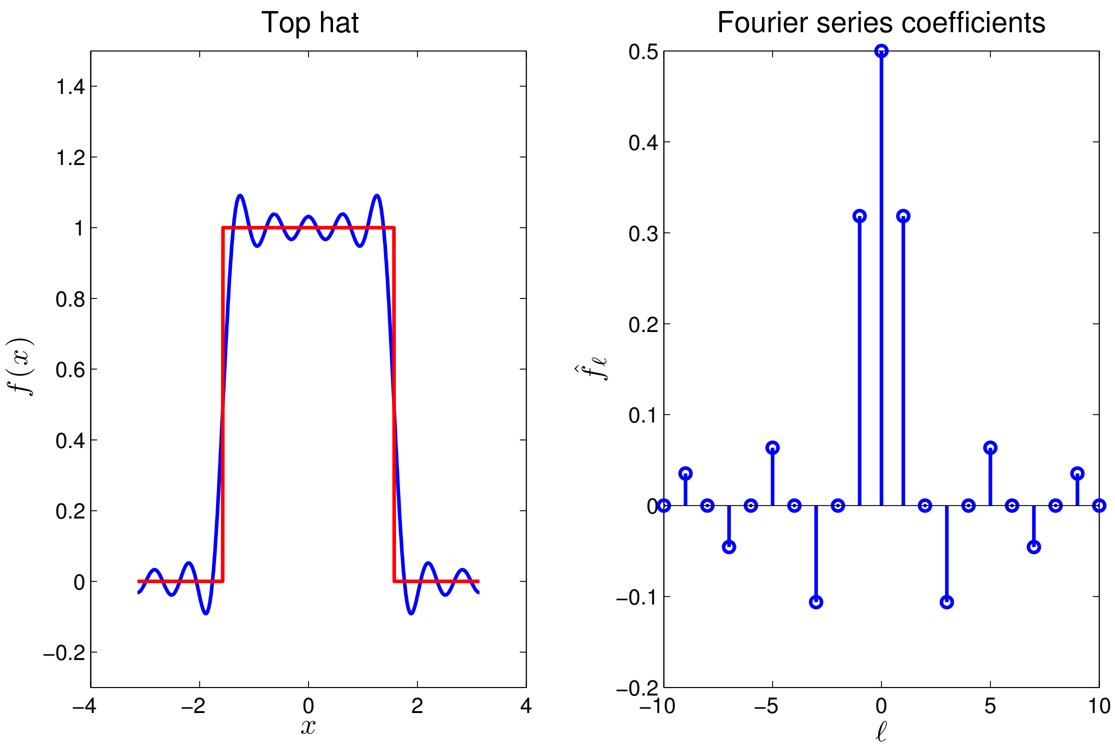
1. Characteristic property on $[-\pi, \pi]$ with periodic boundary
2. Characteristic property on $\mathbb{R}^d$

Characteristic Kernels (via Fourier)

Reminder: Fourier series

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

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



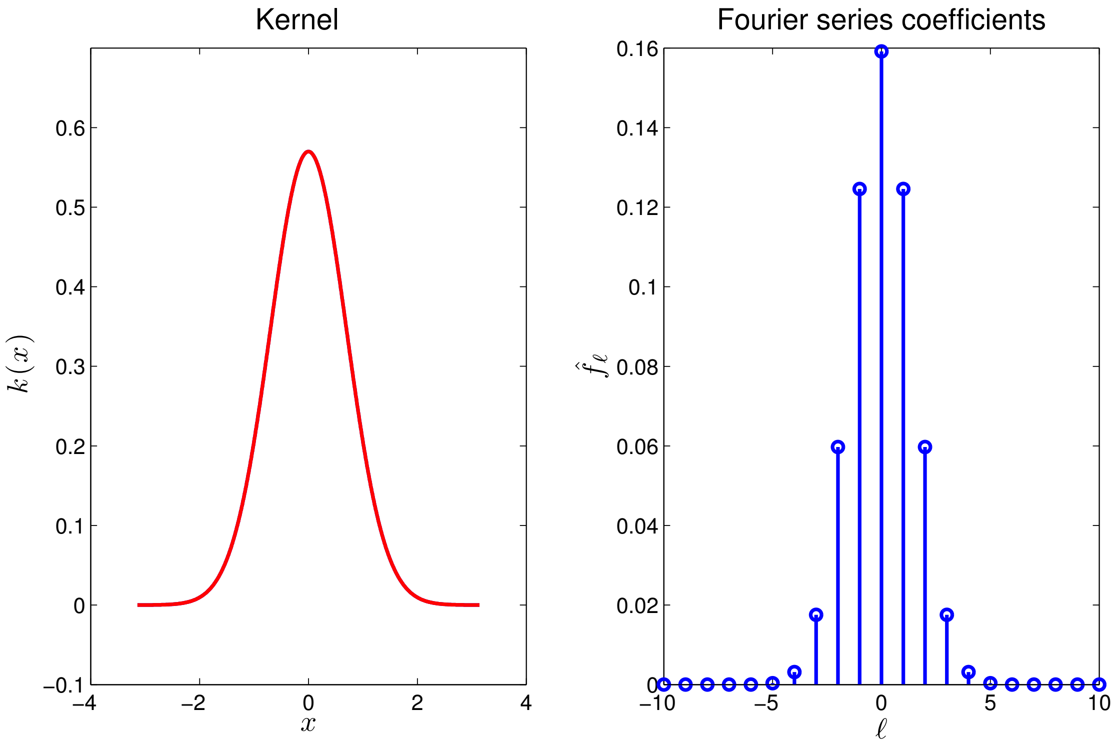
Characteristic Kernels (via Fourier)

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., $$k(x) = \frac{1}{2\pi}\vartheta\left(\frac{x}{2\pi}, \frac{\imath\sigma^2}{2\pi}\right), \qquad \hat{k}_\ell = \frac{1}{2\pi}\exp\left(\frac{-\sigma^2\ell^2}{2}\right).$$

ϑ is the Jacobi theta function, close to Gaussian when σ^2 sufficiently narrower than $[-\pi, \pi]$.



Characteristic Kernels (via Fourier)

Maximum mean embedding via Fourier series:

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

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

Characteristic Kernels (via Fourier)

Maximum mean embedding via Fourier series:

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

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

- MMD can be written in terms of Fourier series:

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

A simpler Fourier expression for MMD

- From previous slide,

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

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

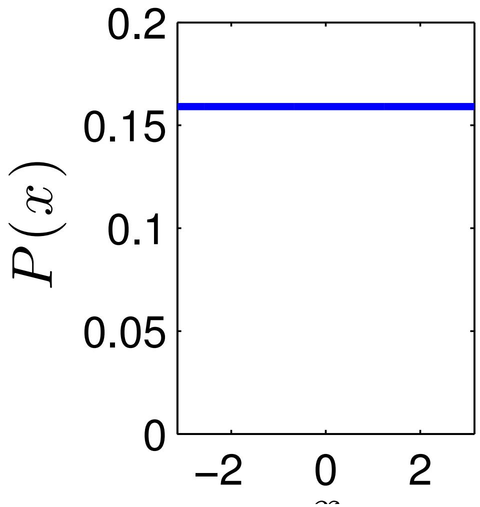
$$\|f\|_{\mathcal{F}}^2 = \langle f, f \rangle_{\mathcal{F}} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}.$$

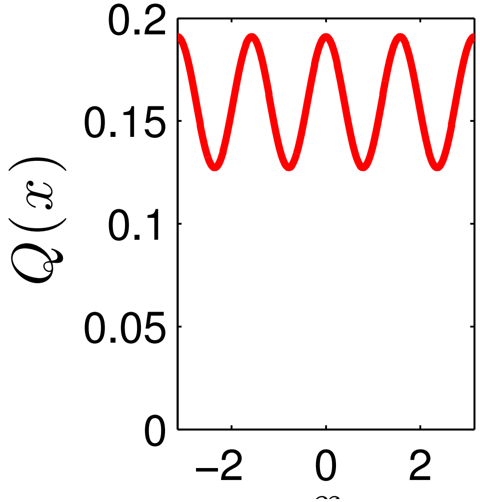
- Simple, interpretable expression for squared MMD:

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

Example

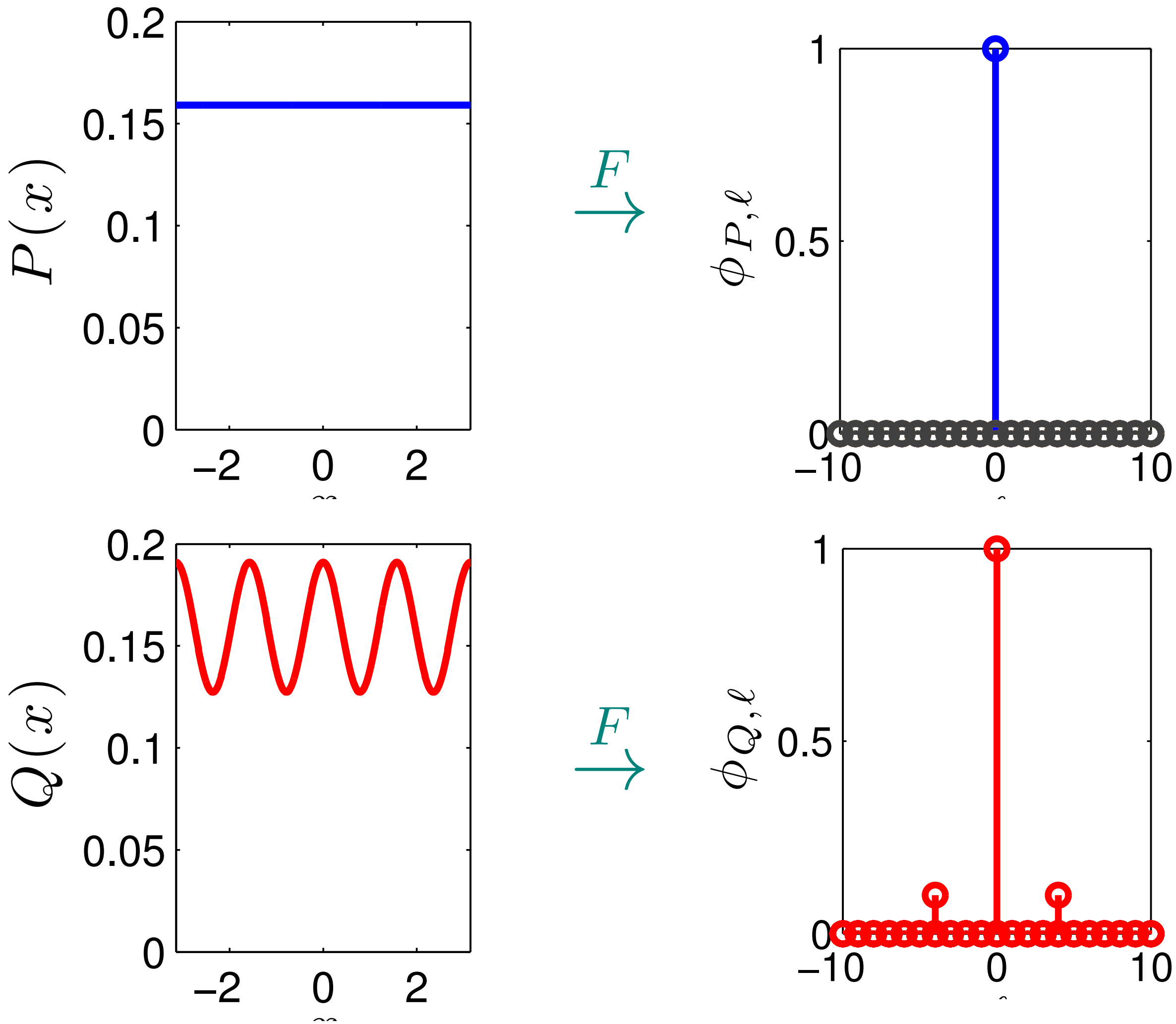
- Example: **P** differs from **Q** at one frequency





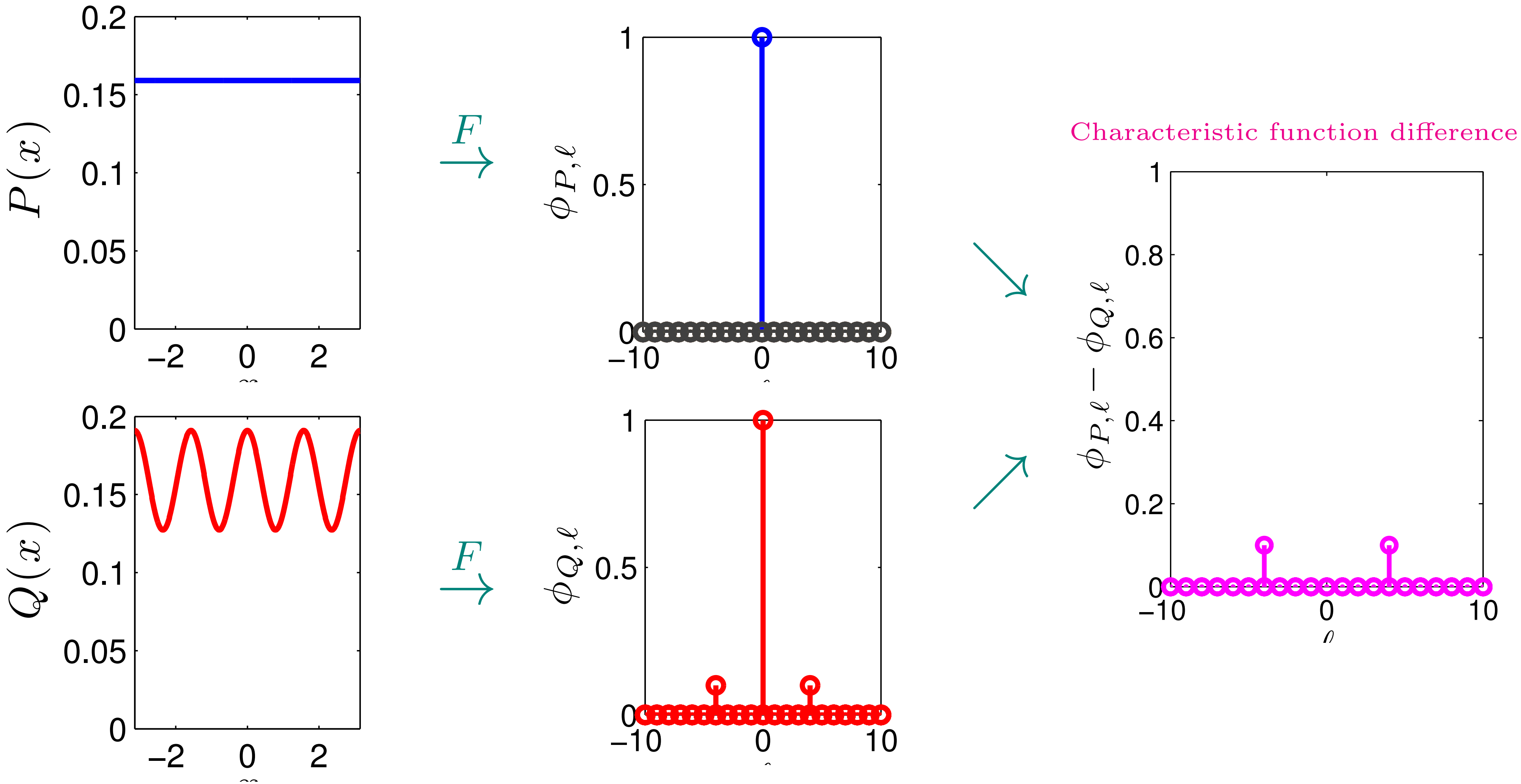
Characteristic Kernels (2)

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



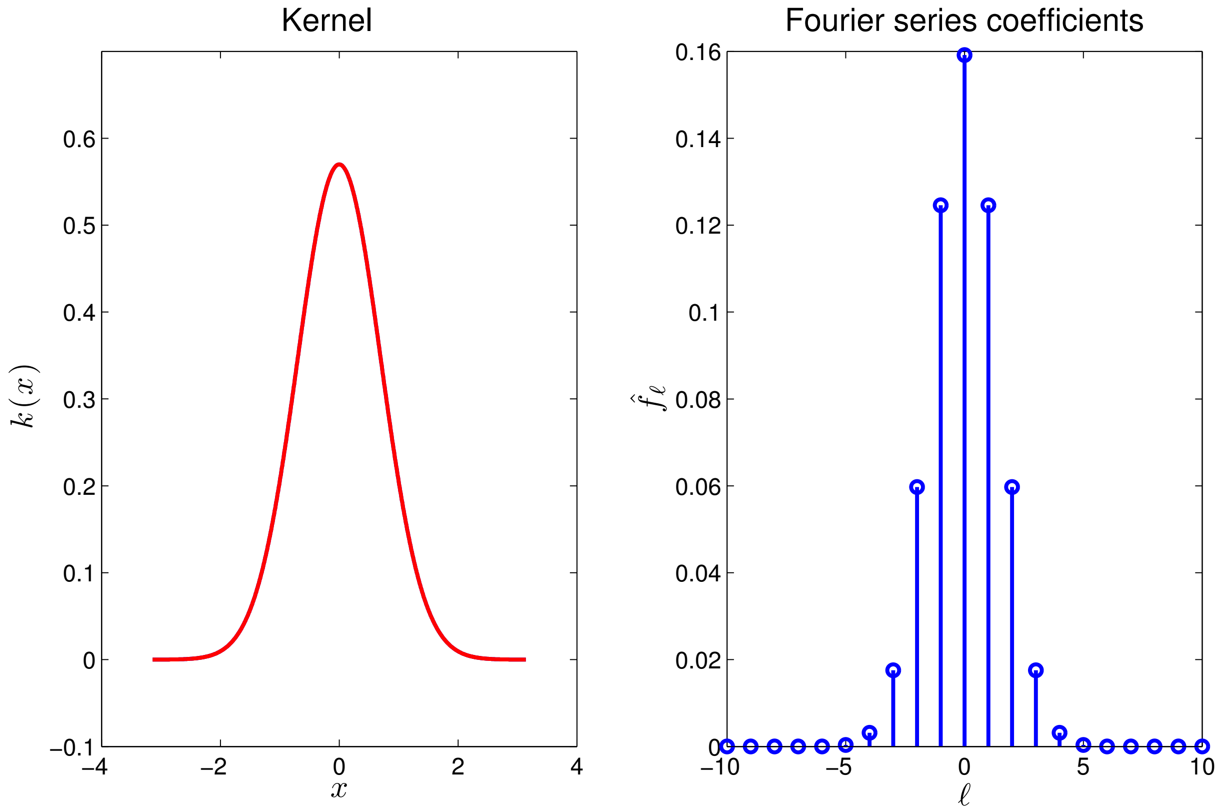
Characteristic Kernels (2)

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



Example

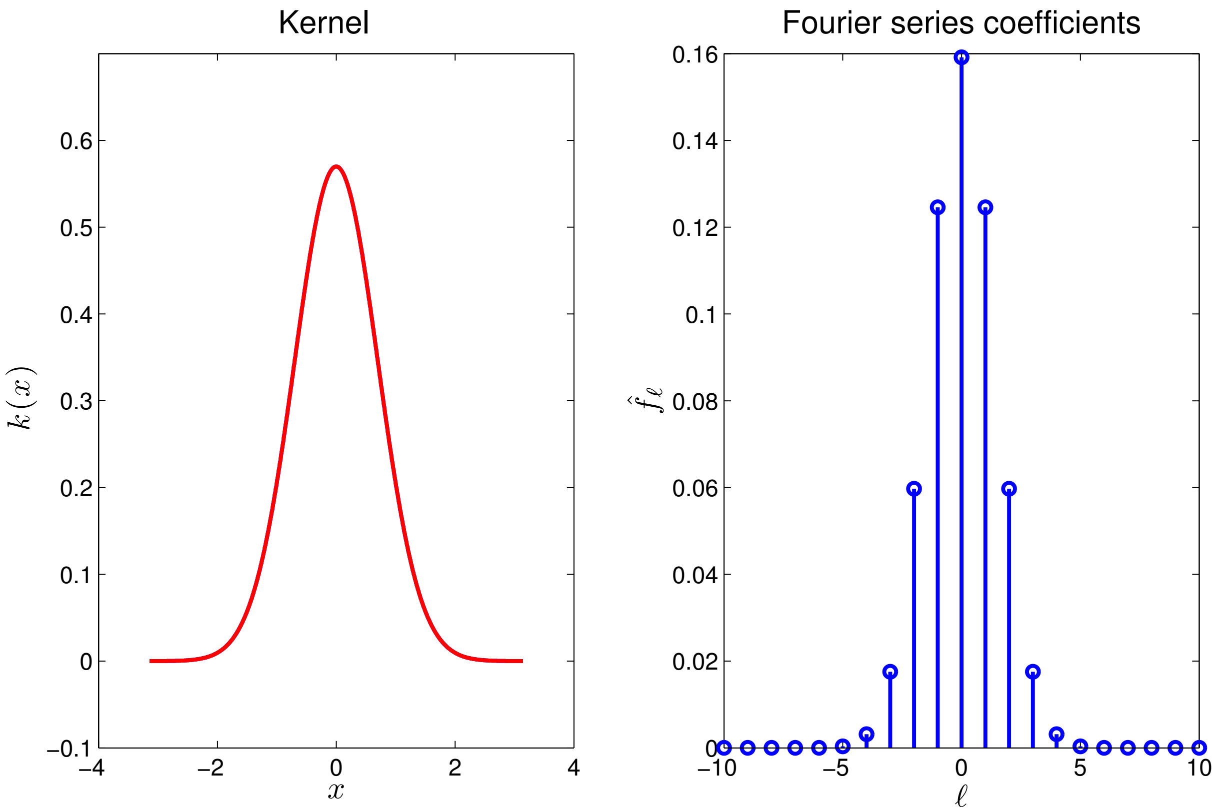
Is the Gaussian-spectrum kernel characteristic?



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

Example

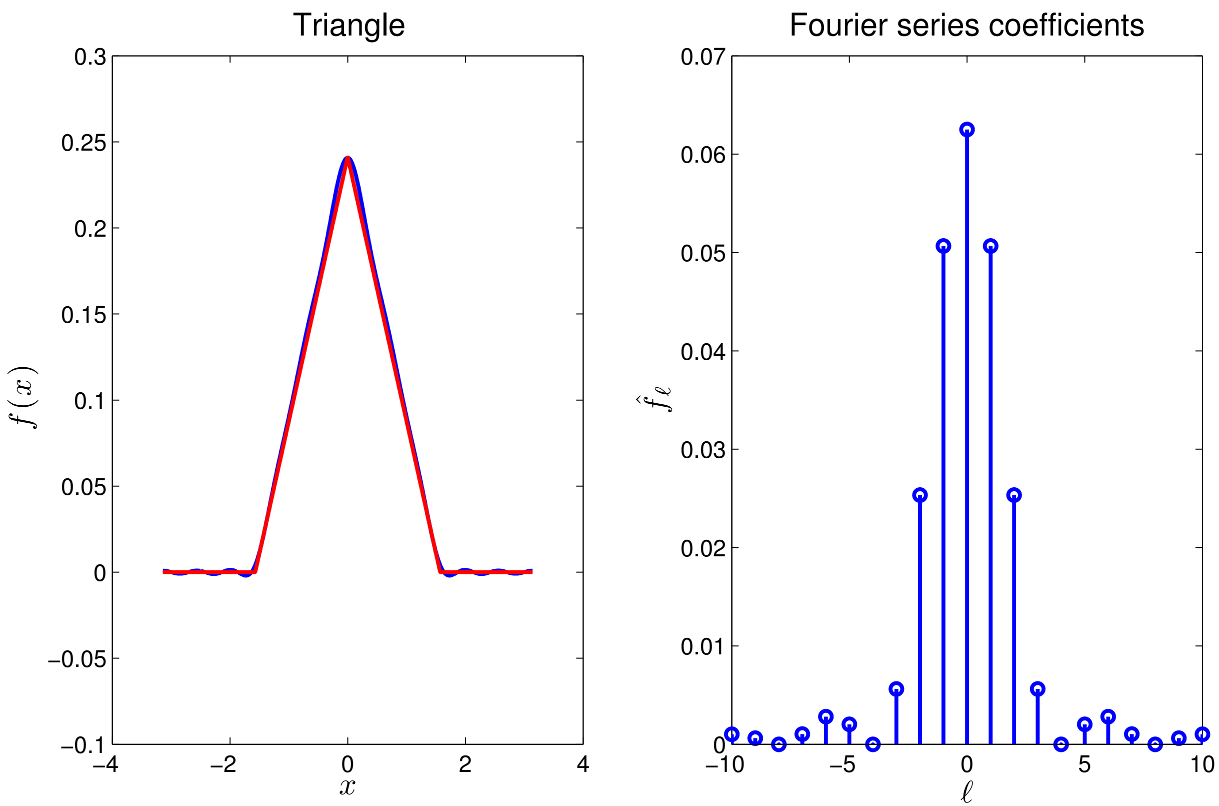
Is the Gaussian-spectrum kernel characteristic? YES



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

Example

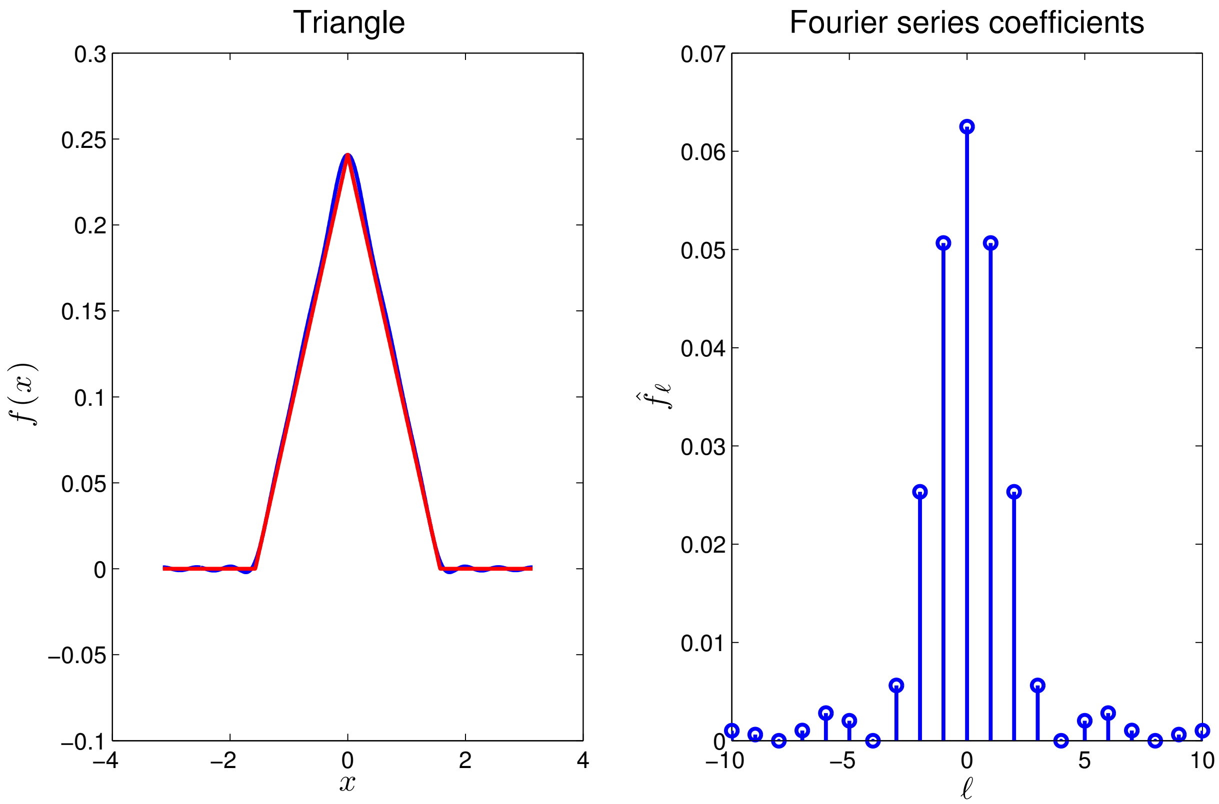
Is the triangle kernel characteristic?



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

Example

Is the triangle kernel characteristic? NO



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

Characteristic kernels (Via Fourier, on $\mathbb{R}^d$)

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:

$$\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}; F) = \int |\phi_{\mathbf{P}}(\omega) - \phi_{\mathbf{Q}}(\omega)|^2 \, d\Lambda(\omega)$$

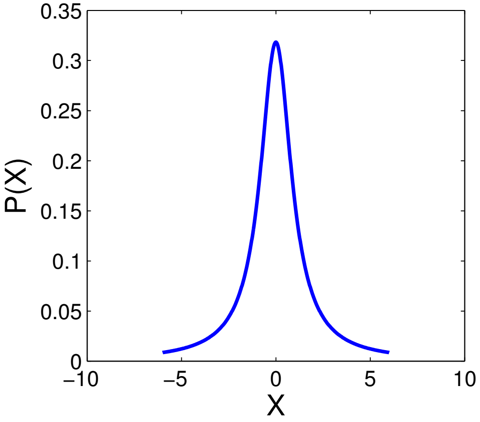
$\phi_{\mathbf{P}}$ characteristic function of $\mathbf{P}$

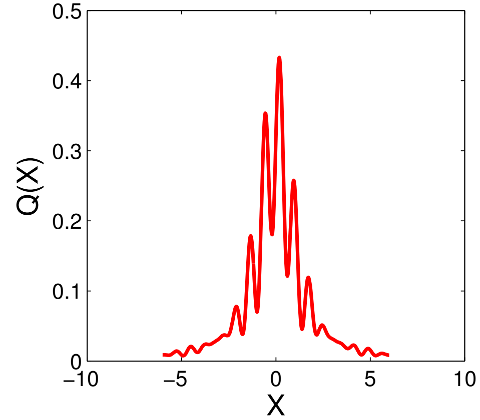
Proof: Using Bochner's theorem (a)... and Fubini's theorem (b)

$$\begin{aligned}
\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}) &:= \mathbb{E}_{\mathbf{P}} k(\mathsf{x} - \mathsf{x}') + \mathbb{E}_{\mathbf{Q}} k(\mathsf{y} - \mathsf{y}') - 2\mathbb{E}_{\mathbf{P},\mathbf{Q}} k(\mathsf{x}, \mathsf{y}) \\
&= \int \int \left[k(s - t) \, d(\mathbf{P} - \mathbf{Q})(s) \right] d(\mathbf{P} - \mathbf{Q})(t) \\
&\overset{(a)}{=} \int \int \int_{\mathbb{R}^d} e^{-i(s-t)^T \omega} \, d\Lambda(\omega) \, d(\mathbf{P} - \mathbf{Q})(s) \, d(\mathbf{P} - \mathbf{Q})(t) \\
&\overset{(b)}{=} \int \int_{\mathbb{R}^d} e^{-ix^T \omega} \, d(\mathbf{P} - \mathbf{Q})(s) \int_{\mathbb{R}^d} e^{iy^T \omega} \, d(\mathbf{P} - \mathbf{Q})(t) \, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} |\phi_{\mathbf{P}}(\omega) - \phi_{\mathbf{Q}}(\omega)|^2 \, d\Lambda(\omega)
\end{aligned}$$

Example

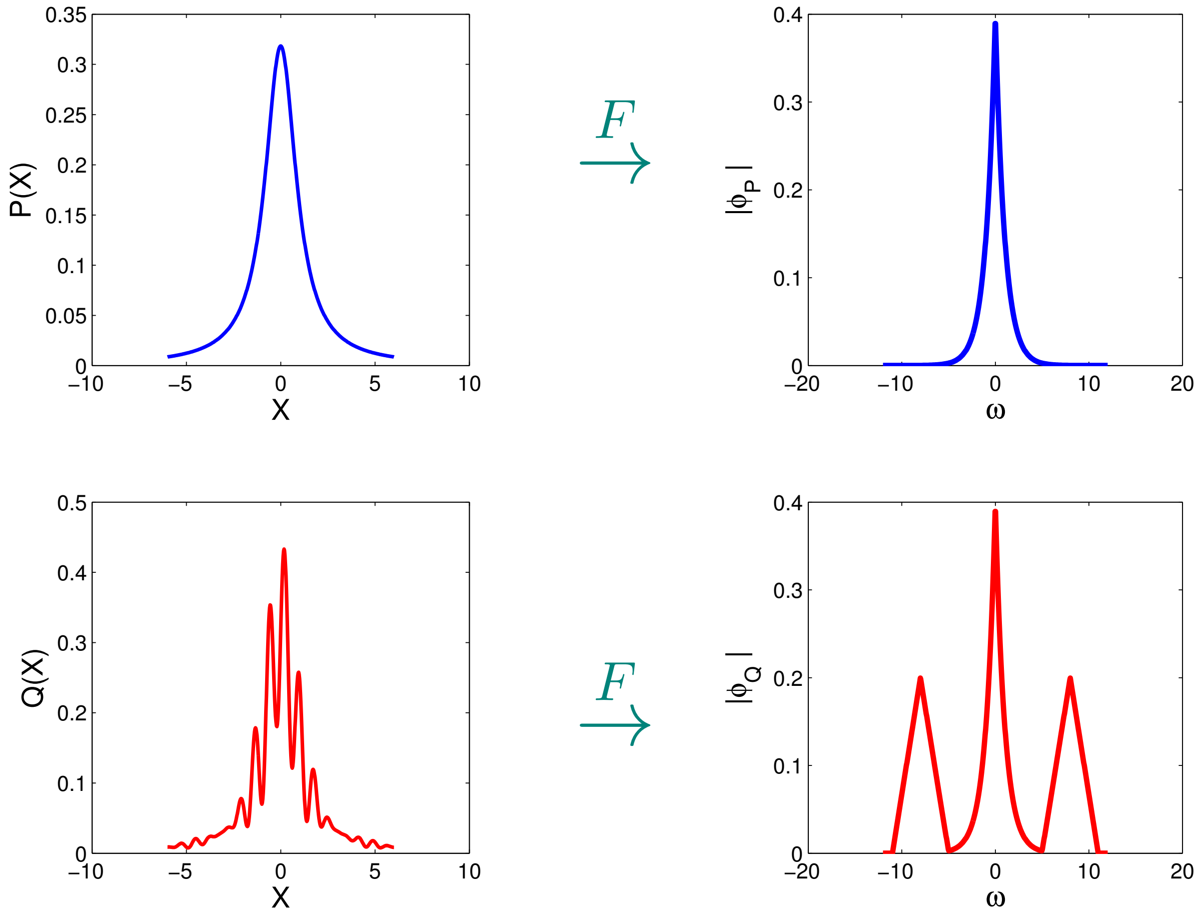
- Example: **P** differs from **Q** at (roughly) one frequency





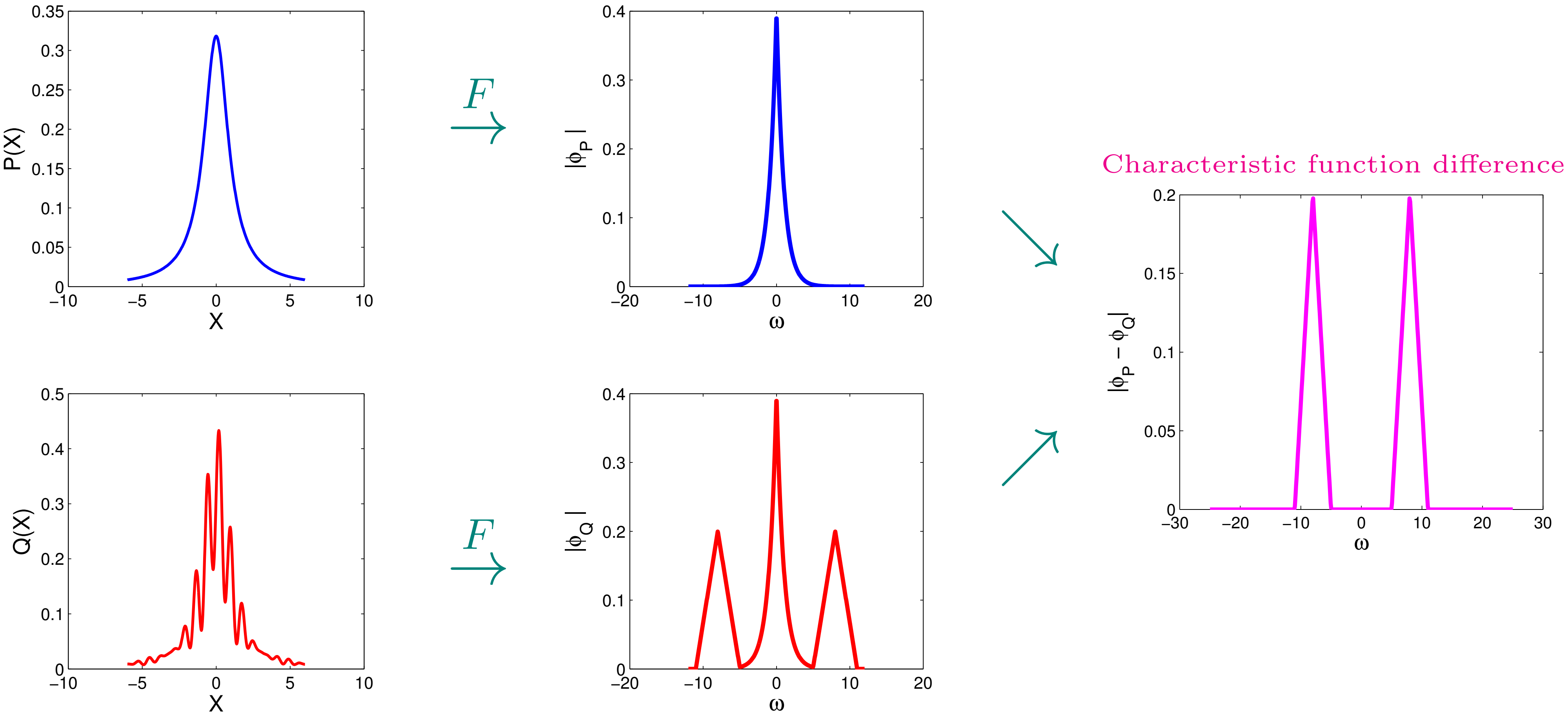
Example

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



Example

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

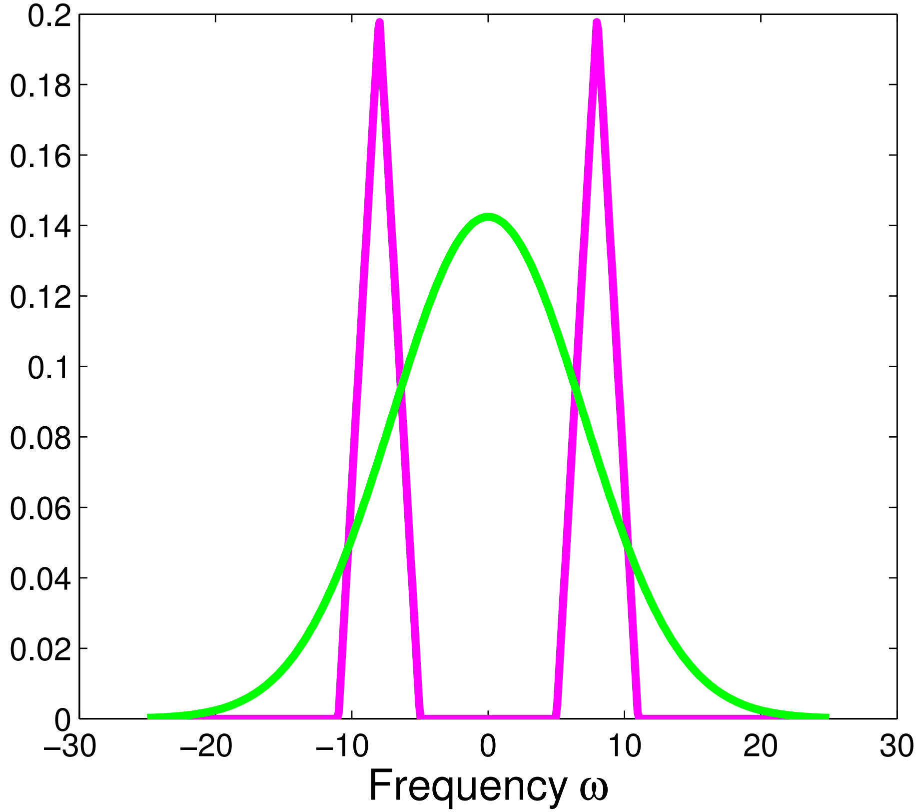


Example

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

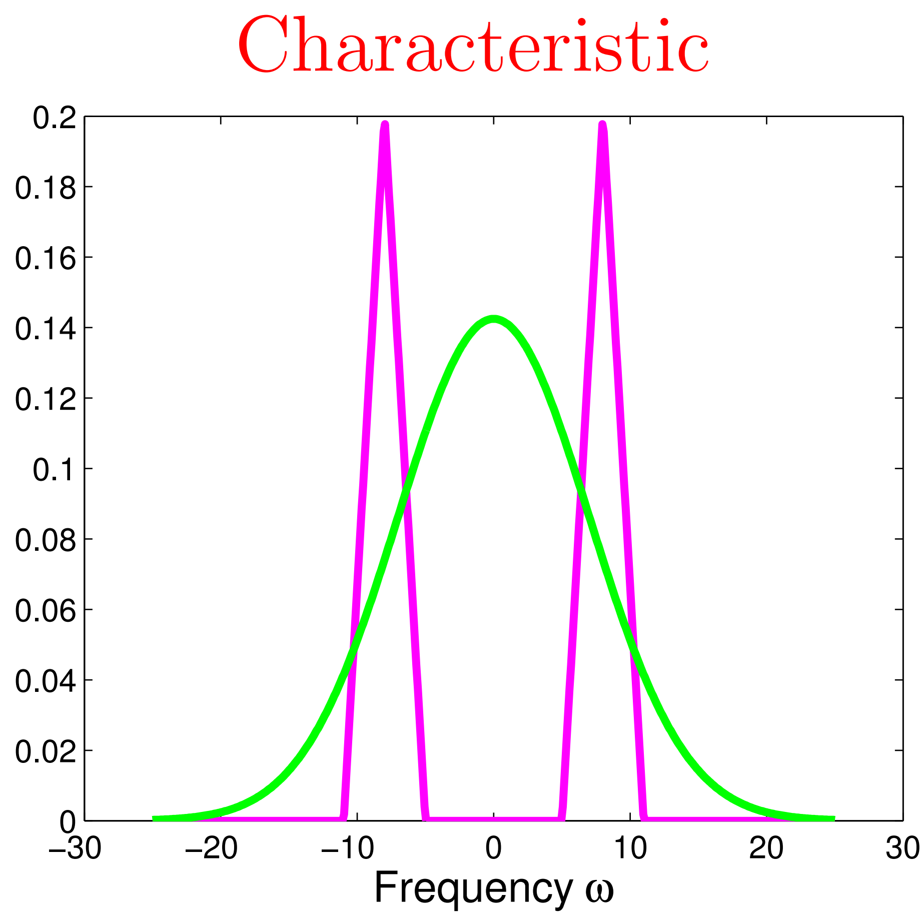
Exponentiated quadratic kernel

Difference $|\phi_P - \phi_Q|$



Example

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

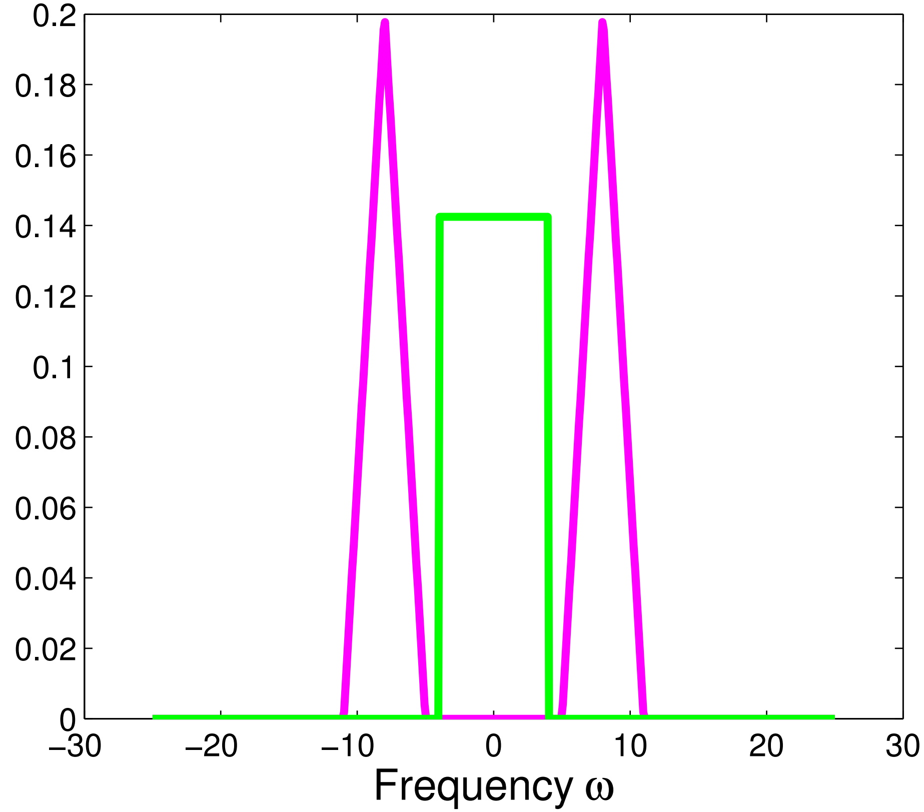


Example

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

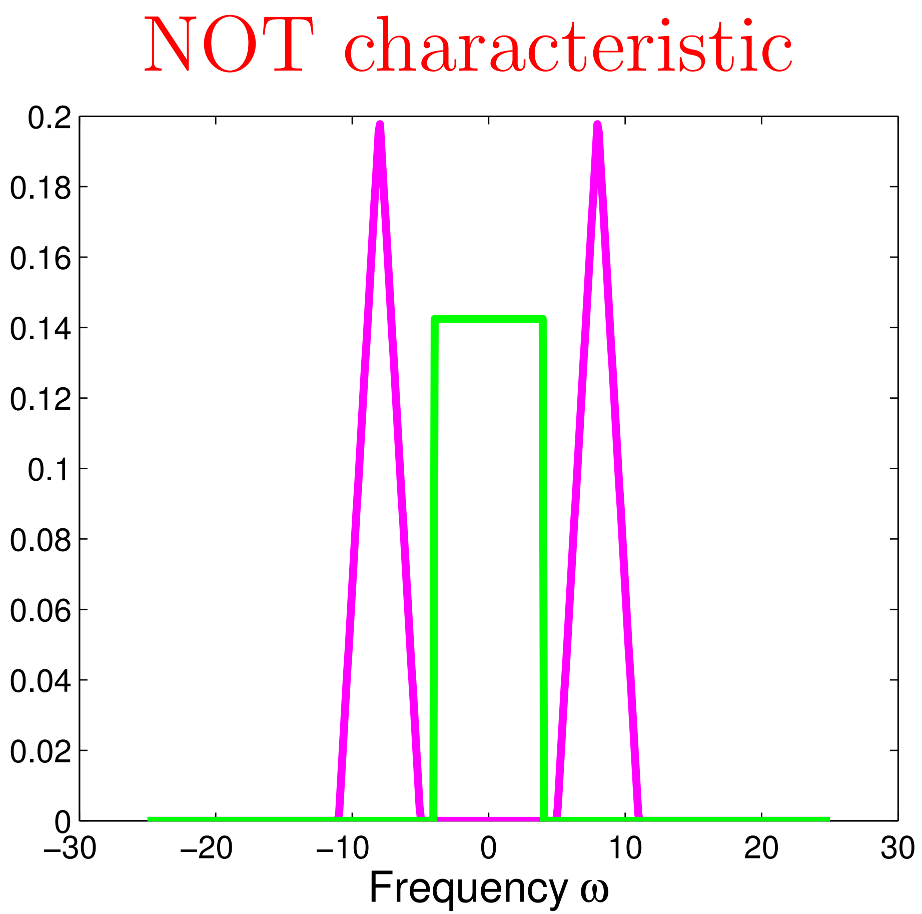
Sinc kernel

Difference $|\phi_P - \phi_Q|$



Example

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

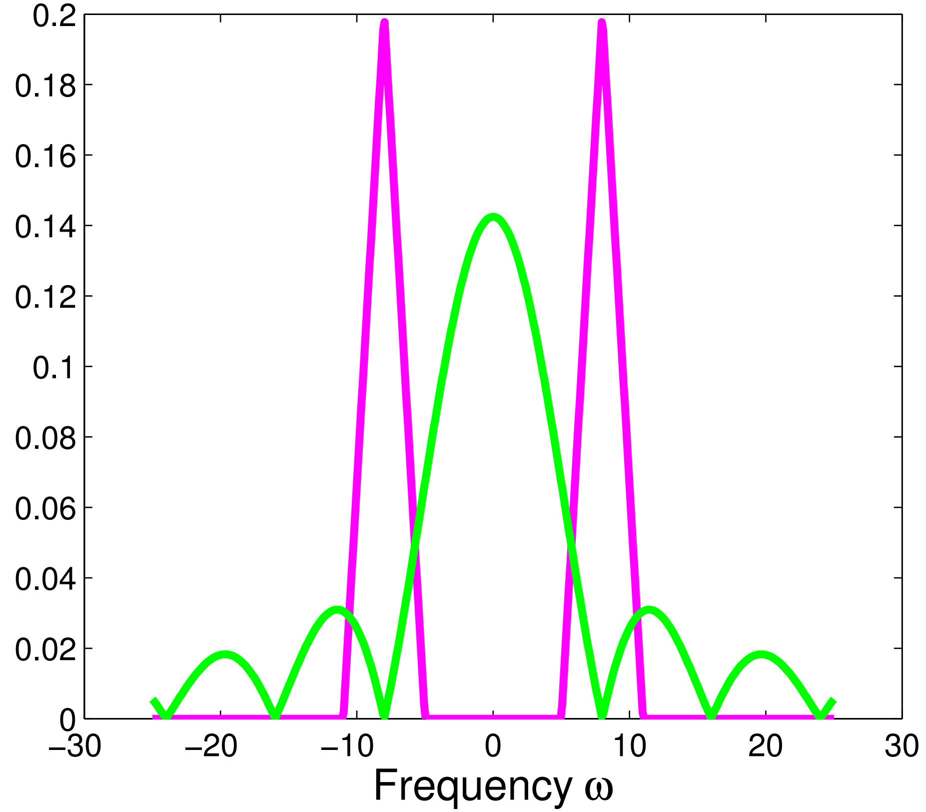


Example

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

Triangle (B-spline) kernel

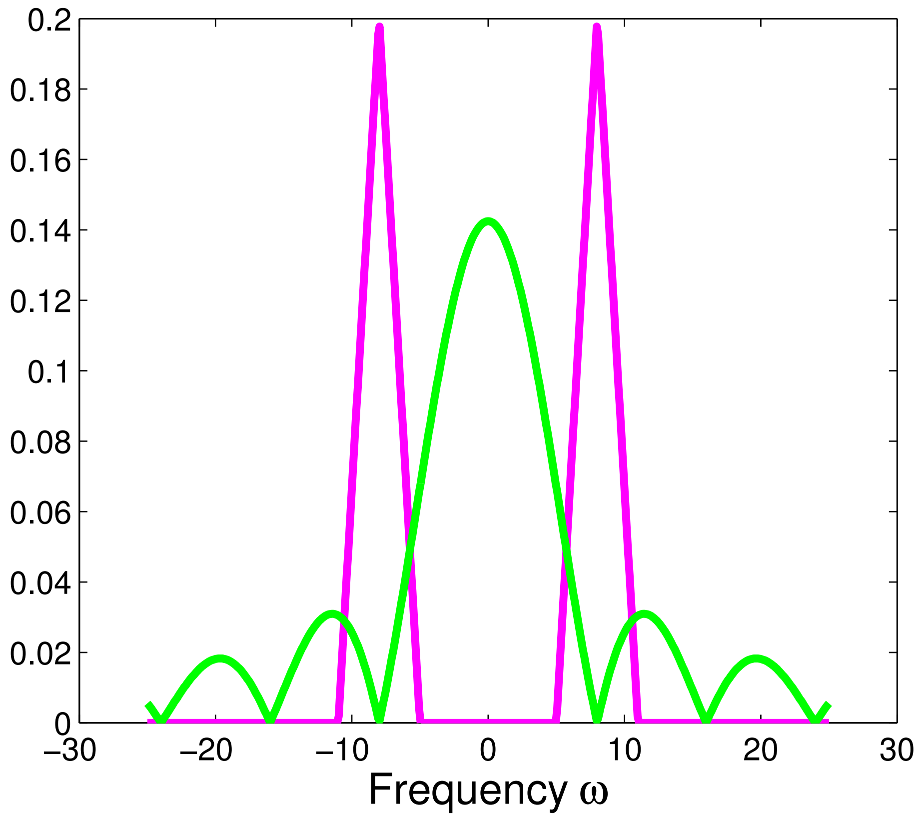
Difference $|\phi_P - \phi_Q|$



Example

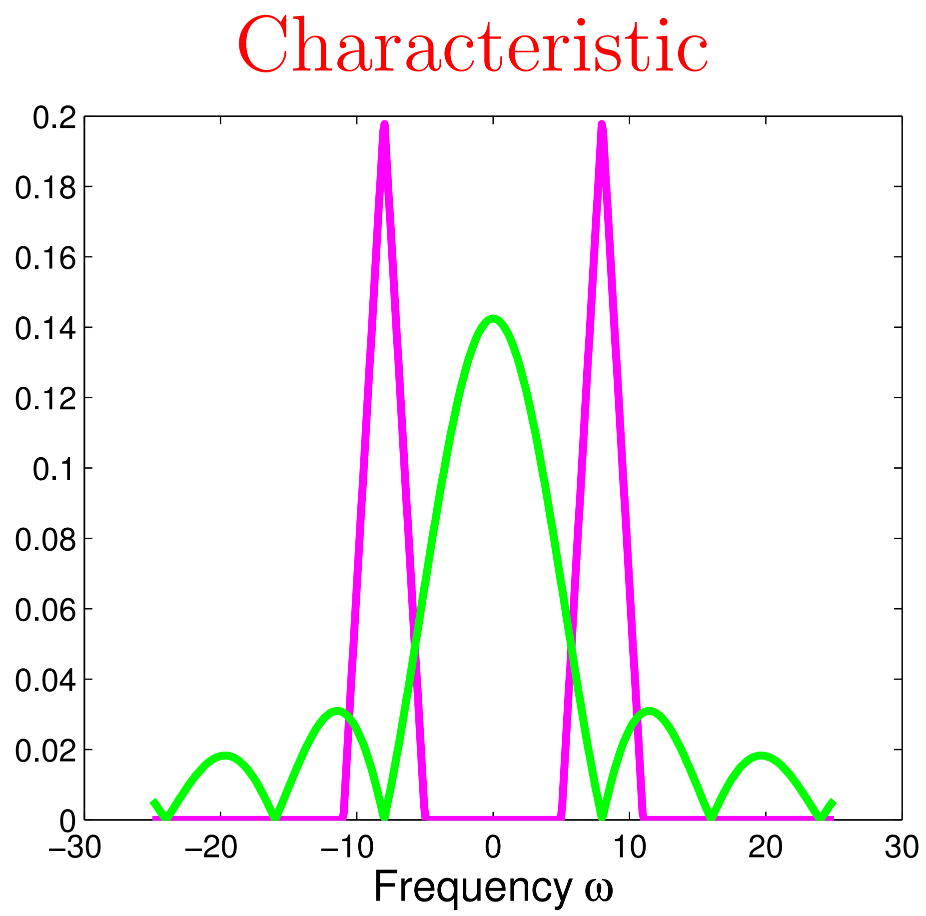
- Example: **P** differs from **Q** at (roughly) one frequency

???



Example

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



Summary: Characteristic Kernels

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

Main theorem: A translation invariant k characteristic for prob. measures on $\mathbb{R}^d$ if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ (i.e. support zero on at most a countable set) [COLT08, JMLR10]

Corollary: continuous, compactly supported k characteristic (since Fourier spectrum $\Lambda(\omega)$ cannot be zero on an interval). 1-D proof sketch from [Mallat, 1999, Theorem 2.6] proof on $\mathbb{R}^d$ via distribution theory in [Sriperumbudur et al., 2010, Corollary 10 p. 1535]

k characteristic iff $\text{supp}(\Lambda) = \mathbb{R}^d$

Proof: $\text{supp}\{\Lambda\} = \mathbb{R}^d \implies$ k characteristic:

Recall Fourier definition of MMD:

$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}) = \int_{\mathbb{R}^d} |\phi_{\mathbf{P}}(\omega) - \phi_{\mathbf{Q}}(\omega)|^2 \, d\Lambda(\omega).$$

Characteristic functions $\phi_{\mathbf{P}}(\omega)$ and $\phi_{\mathbf{Q}}(\omega)$ uniformly continuous, hence their difference cannot be non-zero only on a countable set.

Map $\phi_{\mathbf{P}}$ uniformly continuous: $\forall \epsilon > 0, \quad \exists \delta > 0$ such that $\forall(\omega_1, \omega_2) \in \Omega$ for which $d(\omega_1, \omega_2) < \delta$, we have $d(\phi_{\mathbf{P}}(\omega_1), \phi_{\mathbf{P}}(\omega_2)) < \epsilon$. **Uniform:** δ depends only on ϵ, not on ω_1, ω_2.

k characteristic iff $\text{supp}(\Lambda) = \mathbb{R}^d$

Proof: k characteristic $\implies$ $\text{supp}\{\Lambda\} = \mathbb{R}^d$:

Proof by contrapositive.

Given $\text{supp}\{\Lambda\} \subsetneq \mathbb{R}^d$, hence $\exists$ open interval U such that $\Lambda(\omega)$ zero on U.

Construct densities $p(x)$, $q(x)$ such that $\phi_{\mathbf{P}}, \phi_{\mathbf{Q}}$ differ only inside U

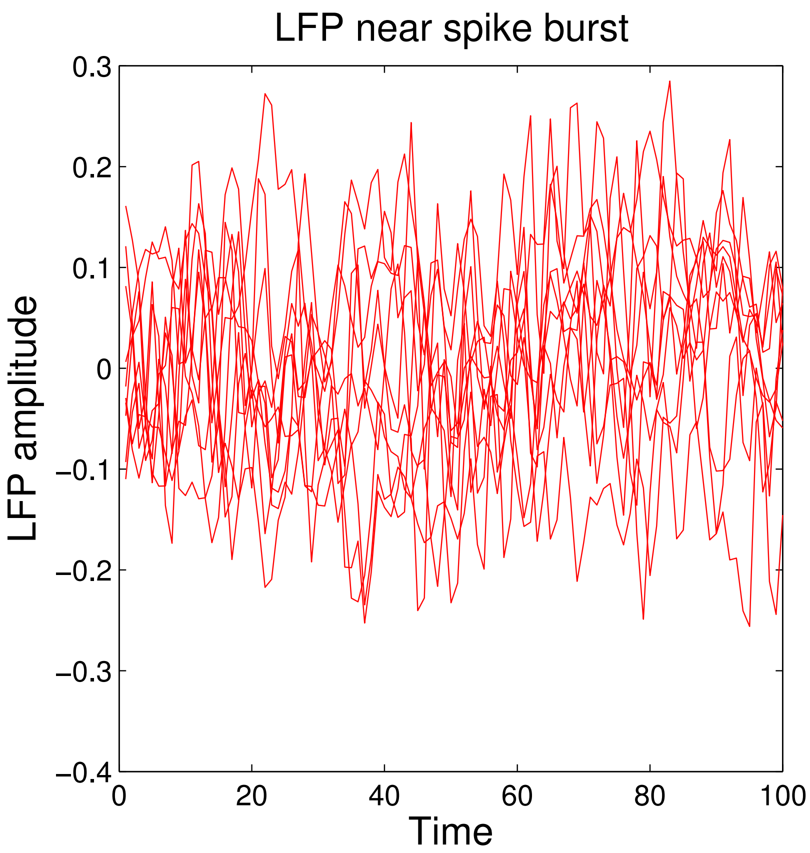
Further extensions

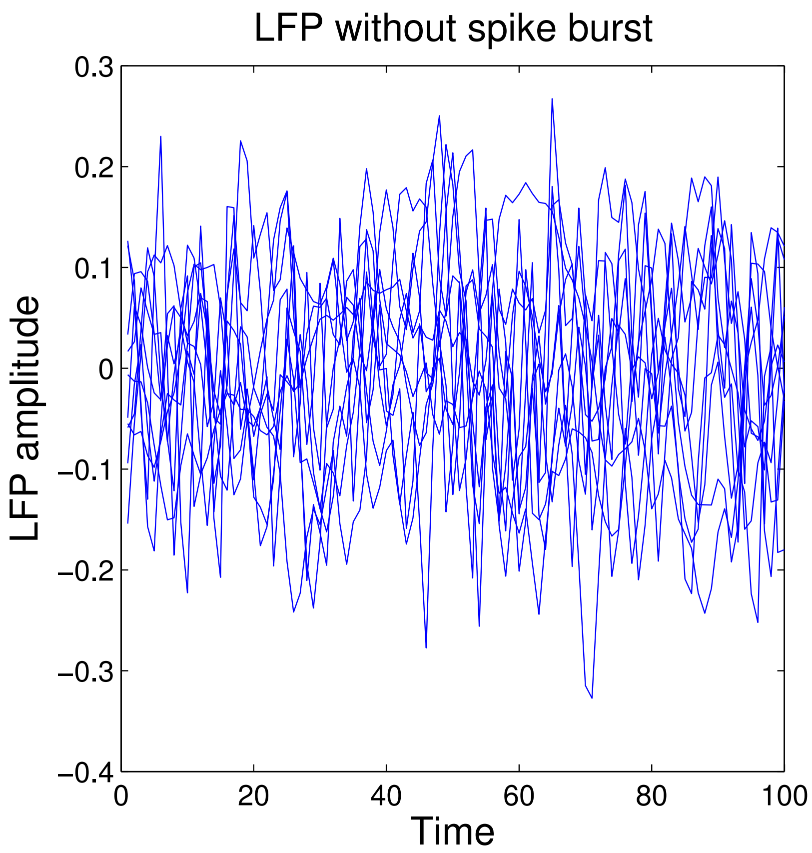
- Similar reasoning wherever extensions of Bochner's theorem exist: [Fukumizu et al., 2009]
  - Locally compact Abelian groups (periodic domains, as we saw)
  - Compact, non-Abelian groups (orthogonal matrices)
  - The semigroup $\mathbb{R}_n^+$ (histograms)
- Related kernel statistics: Fisher statistic [Harchaoui et al., 2008](zero iff $\mathbf{P} = \mathbf{Q}$ for characteristic kernels), other distances [Zhou and Chellappa, 2006](not yet shown to establish whether $\mathbf{P} = \mathbf{Q}$), energy distances

Statistical hypothesis testing

Motivating question: differences in brain signals

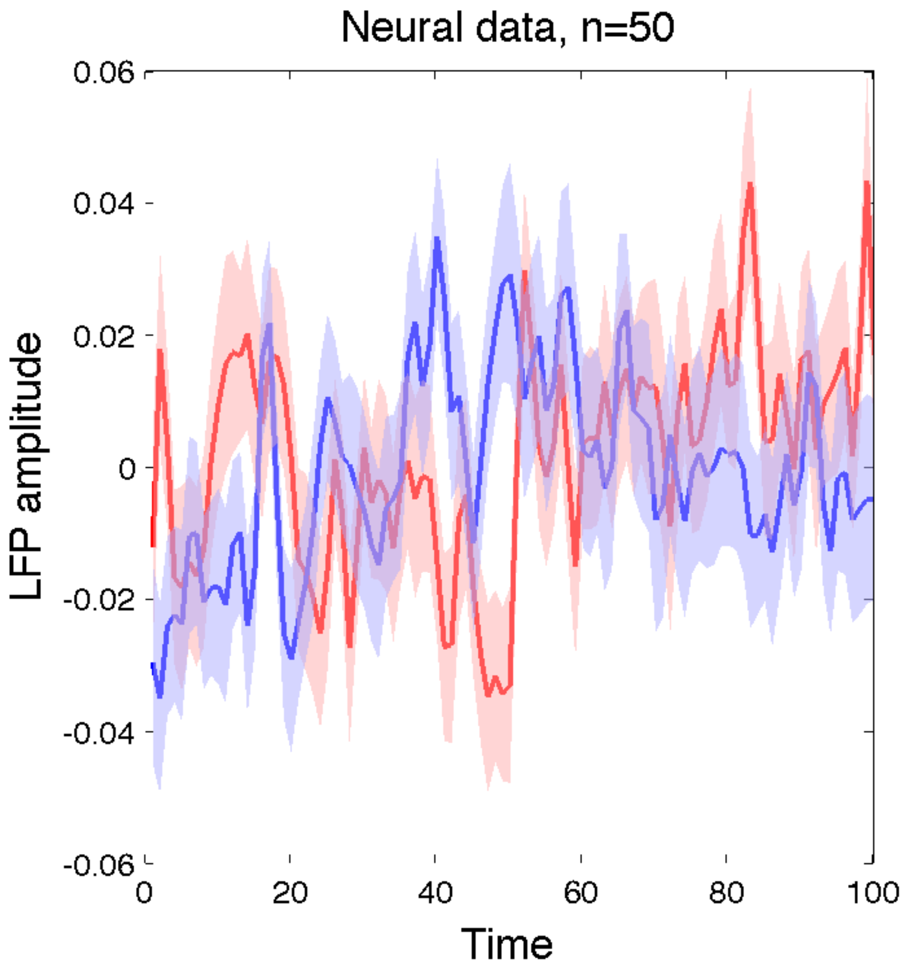
The problem: Do local field potential (LFP) signals change when measured near a spike burst?





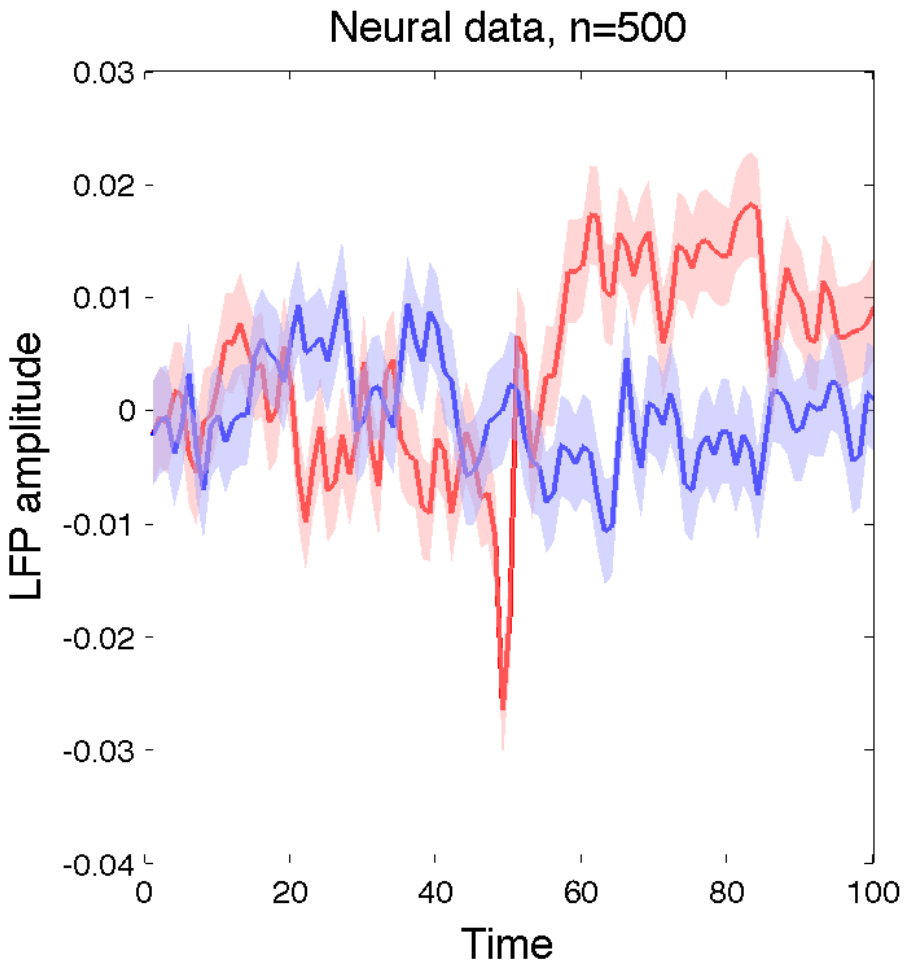
Motivating question: differences in brain signals

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



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

Statistical test using MMD (2)

- “far from zero” vs “close to zero” - threshold?
- One answer: asymptotic distribution of $\widehat{\text{MMD}}^2$

Statistical test using MMD (2)

- "far from zero" vs "close to zero" - threshold?
- One answer: asymptotic distribution of $\widehat{\mathrm{MMD}}^2$
- An unbiased empirical estimate (quadratic cost):

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{h((x_i, y_i), (x_j, y_j))}$$

Statistical test using MMD (2)

- "far from zero" vs "close to zero" - threshold?
- One answer: asymptotic distribution of $\widehat{\mathrm{MMD}}^2$
- An unbiased empirical estimate (quadratic cost):

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} \underbrace{k(x_i, x_j) - k(x_i, y_j) - k(y_i, x_j) + k(y_i, y_j)}_{h((x_i, y_i), (x_j, y_j))}$$

- When $\mathbf{P} \neq \mathbf{Q}$, asymptotically normal

$$(\sqrt{n}) \left(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2 \right) \sim \mathcal{N}(0, \sigma_u^2)$$

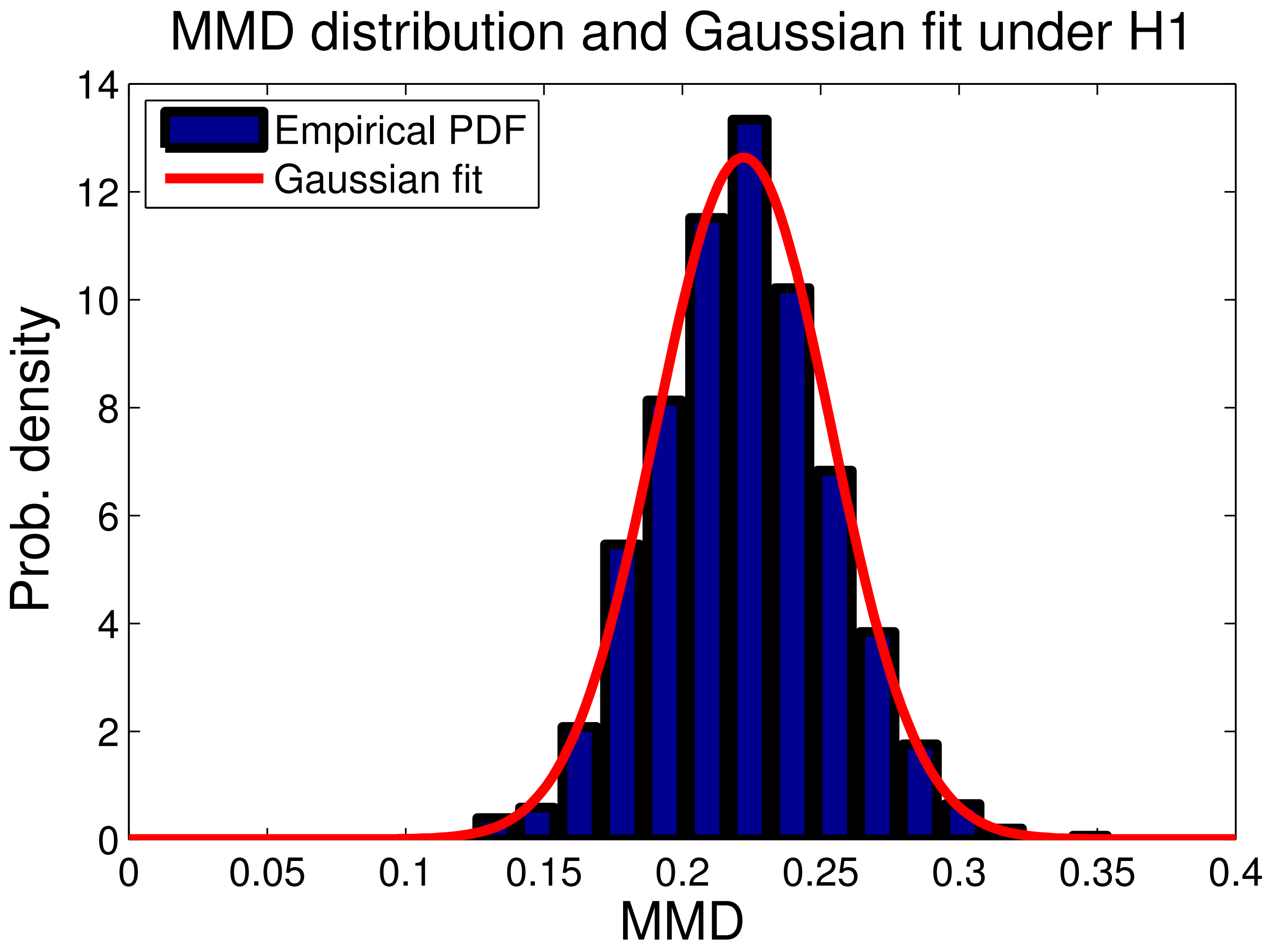
[Hoeffding, 1948, Serfling, 1980]

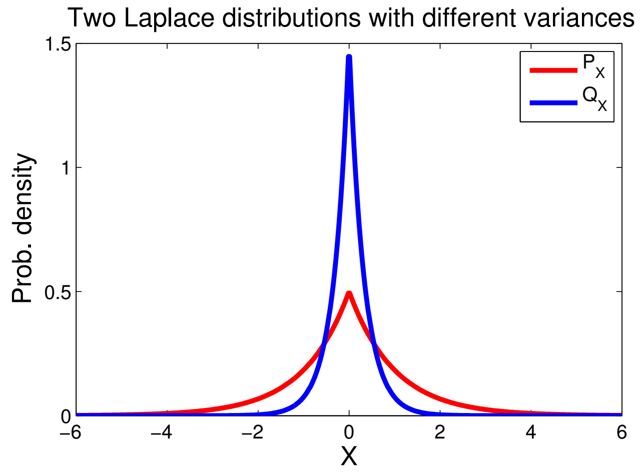
- Expression for the variance: $z_i := (x_i, y_i)$

$$\sigma_u^2 = 4 \left(\mathbb{E}_{\mathsf{z}} \left[(\mathbb{E}_{\mathsf{z}'} h(\mathsf{z}, \mathsf{z}'))^2 \right] - \left[\mathbb{E}_{\mathsf{z}, \mathsf{z}'} (h(\mathsf{z}, \mathsf{z}')) \right]^2 \right)$$

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

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

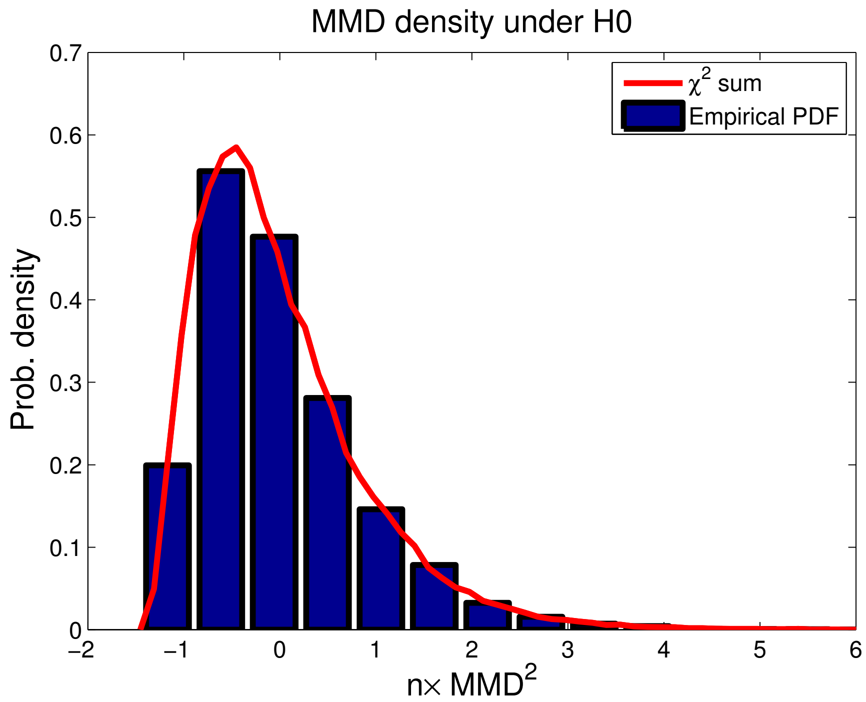
- where
  - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
  - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathbf{P}(x) = \lambda_i \psi_i(x')$

Statistical test using MMD (4)

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

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

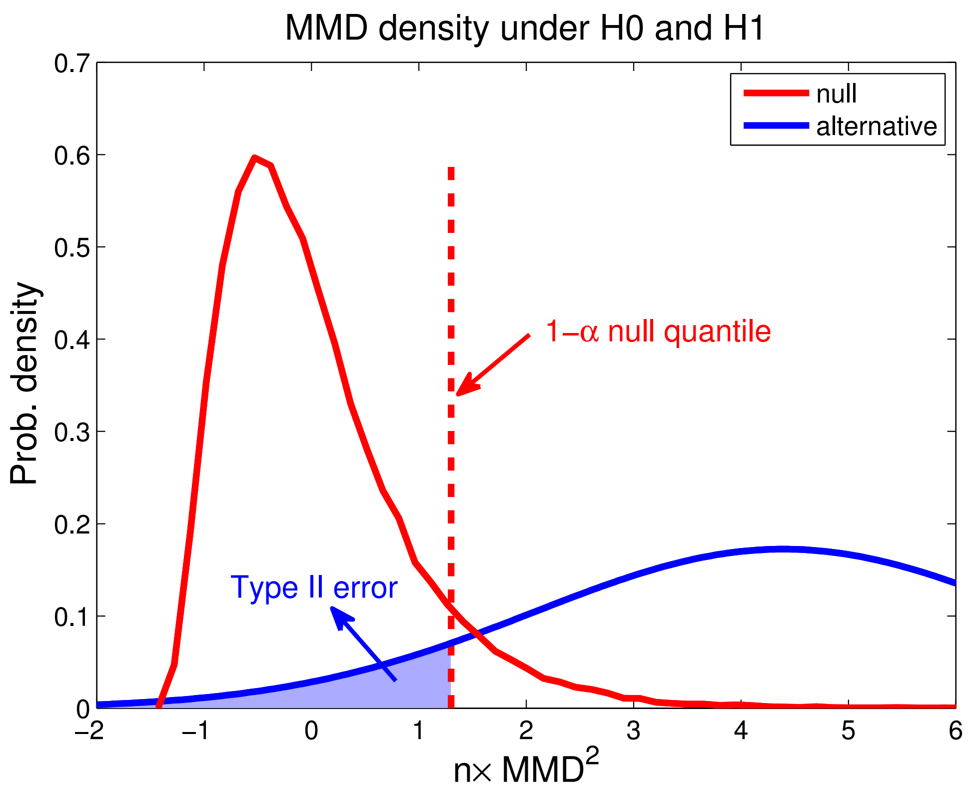
- where
  - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
  - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathbf{P}(x) = \lambda_i \psi_i(x')$



Statistical test using MMD (5)

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

$$\widehat{MMD}^2 = \overline{K_{P,P}} + \overline{K_{Q,Q}} - 2\overline{K_{P,Q}}$$



Statistical test using MMD (5)

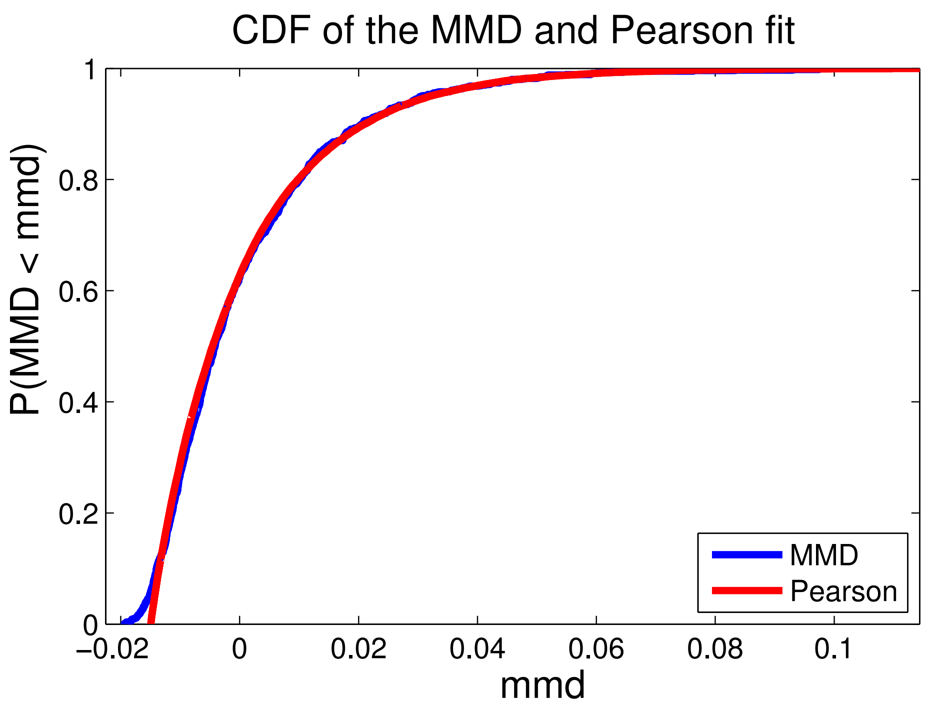
- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\text{MMD} > T) \leq 0.05$

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, Alba Fernández et al., 2008]
- 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, Alba Fernández et al., 2008]
- Pearson curves by matching first four moments [Johnson et al., 1994]
- Large deviation bounds [Hoeffding, 1963, McDiarmid, 1989]
- Consistent test using kernel eigenspectrum [NIPS09b]



Approximate null distribution of $\widehat{MMD}$ via permutation

Empirical MMD:

$$w = (\underbrace{1, 1, 1, \ldots 1}_{n}, \underbrace{-1 \ldots, -1, -1, -1}_{n})^\top$$

$$\frac{1}{n^2} \sum \left(\begin{bmatrix} K_{P,P} & K_{P,Q} \\ K_{Q,P} & K_{Q,Q} \end{bmatrix} \odot \left[ww^\top \right] \right) \approx \widehat{\text{MMD}}^2$$

Approximate null distribution of $\widehat{MMD}$ via permutation

Permuted case: [Alba Fernández et al., 2008]

$$w = (\underbrace{1, -1, 1, \ldots 1}_{n}, \underbrace{-1 \ldots, 1, -1, -1}_{n})^\top$$

(equal number of $+1$ and -1)

$$\frac{1}{n^2} \sum \left(\begin{bmatrix} K_{P,P} & K_{P,Q} \\ K_{Q,P} & K_{Q,Q} \end{bmatrix} \odot \left[w w^\top \right] \right) = [?]$$

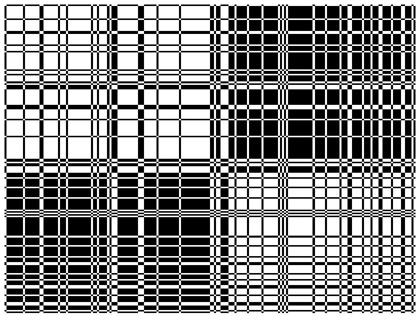
Approximate null distribution of $\widehat{MMD}$ via permutation

Permuted case: [Alba Fernández et al., 2008]

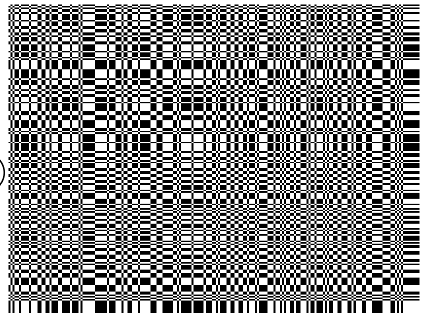
$$w = (\underbrace{1, -1, 1, \ldots 1}_{n}, \underbrace{-1 \ldots, 1, -1, -1}_{n})^\top$$

(equal number of +1 and −1)

$$\frac{1}{n^2} \sum \left(\begin{bmatrix} K_{P,P} & K_{P,Q} \\ K_{Q,P} & K_{Q,Q} \end{bmatrix} \odot \left[ww^\top \right] \right) = [?]$$



$\odot$



$=$

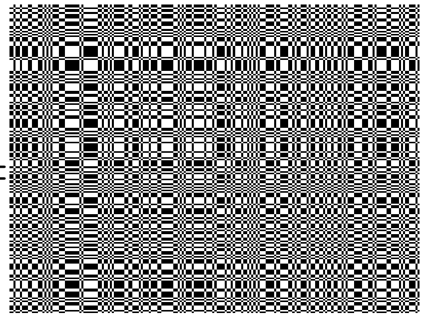
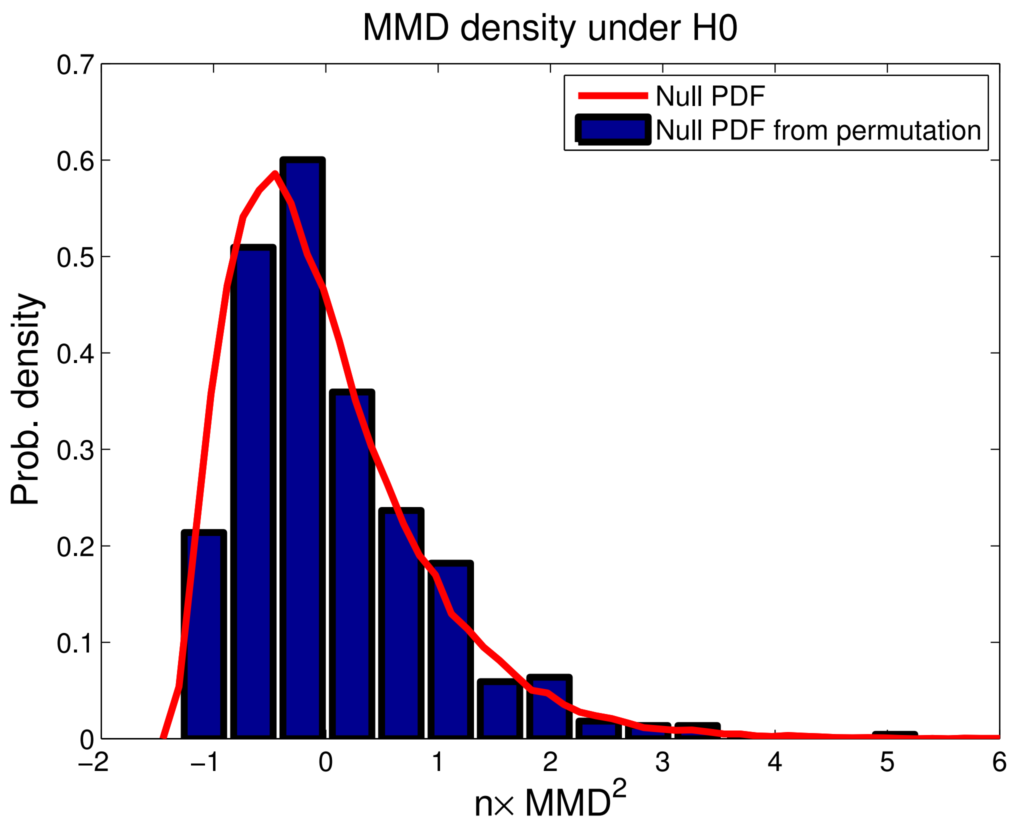


Figure thanks to Kacper Chwialkowski.

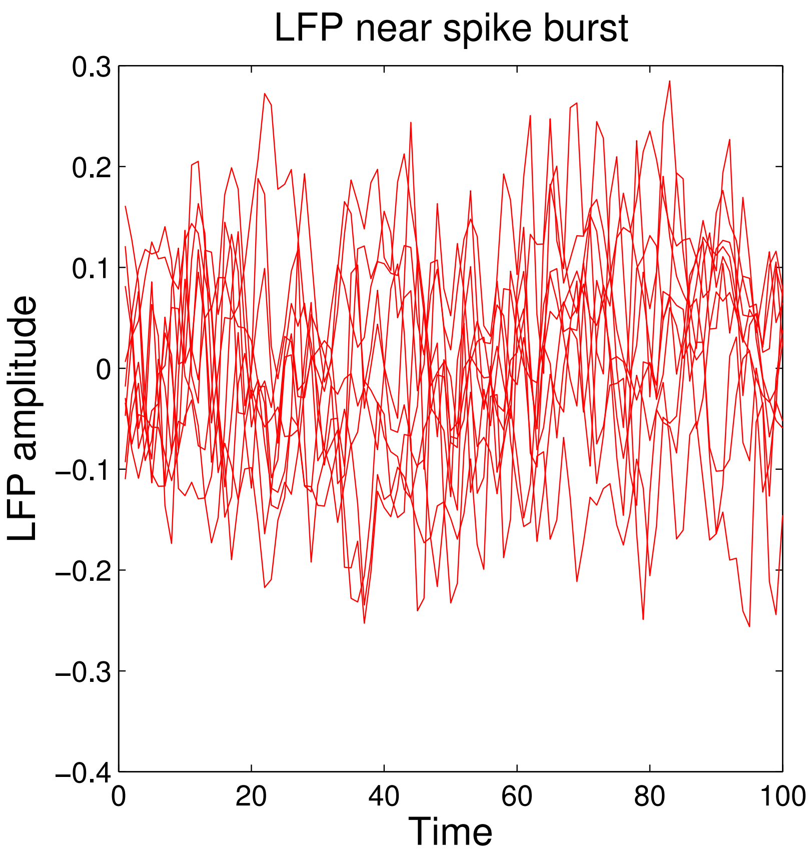
Approximate null distribution of $\widehat{MMD}^2$ via permutation

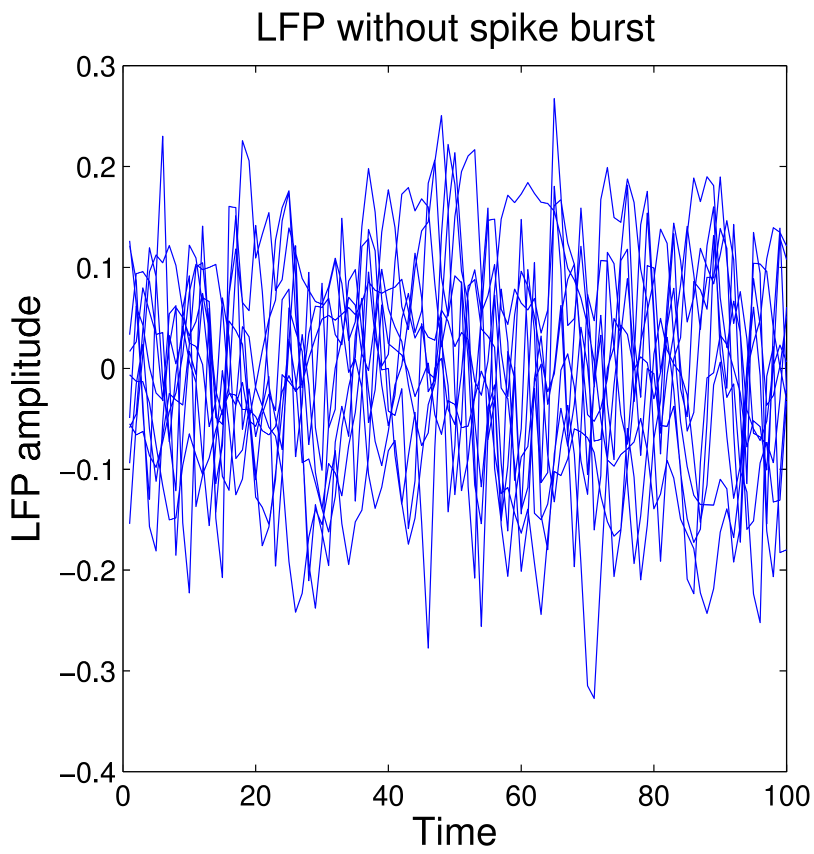
$$\widehat{MMD}_p^2 \approx \frac{1}{n^2} \sum \left(\begin{bmatrix} K_{P,P} & K_{P,Q} \\ K_{Q,P} & K_{Q,Q} \end{bmatrix} \odot \left[w w^\top \right] \right)$$



Detecting differences in brain signals

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





Neuro data: consistent test w/o permutation

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

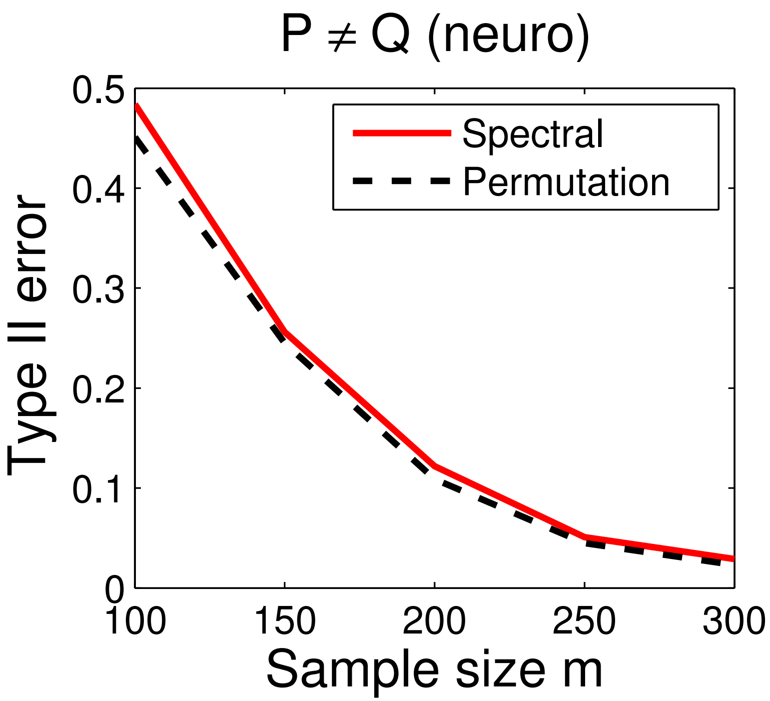
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}}$$

- 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 permutation

Hypothesis testing with HSIC

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

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

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

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

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

- First two moments [NIPS07b]

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

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

Statistical testing with HSIC

- Given $\mathbf{P} = \mathbf{P}_x \mathbf{P}_y$, 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

Are the French text extracts translations of English?

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.

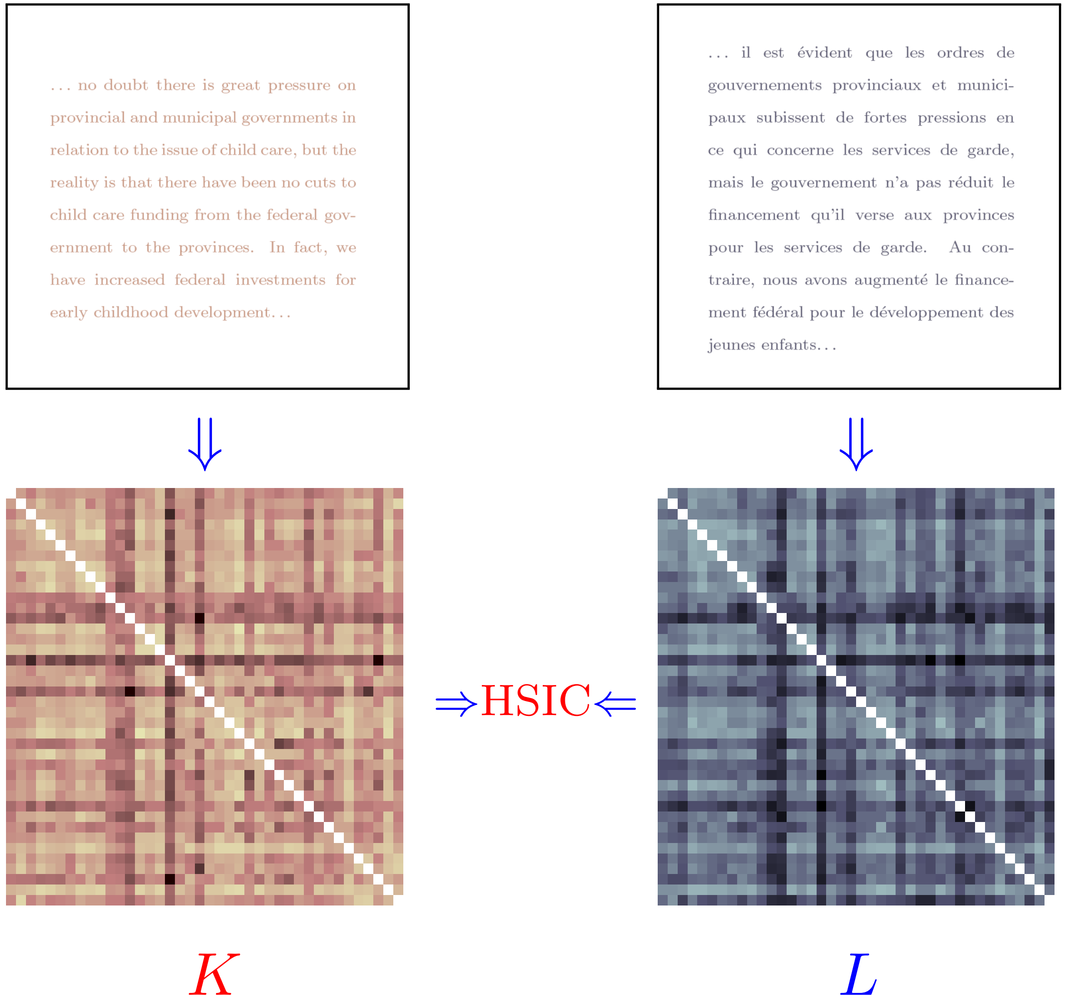
...

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



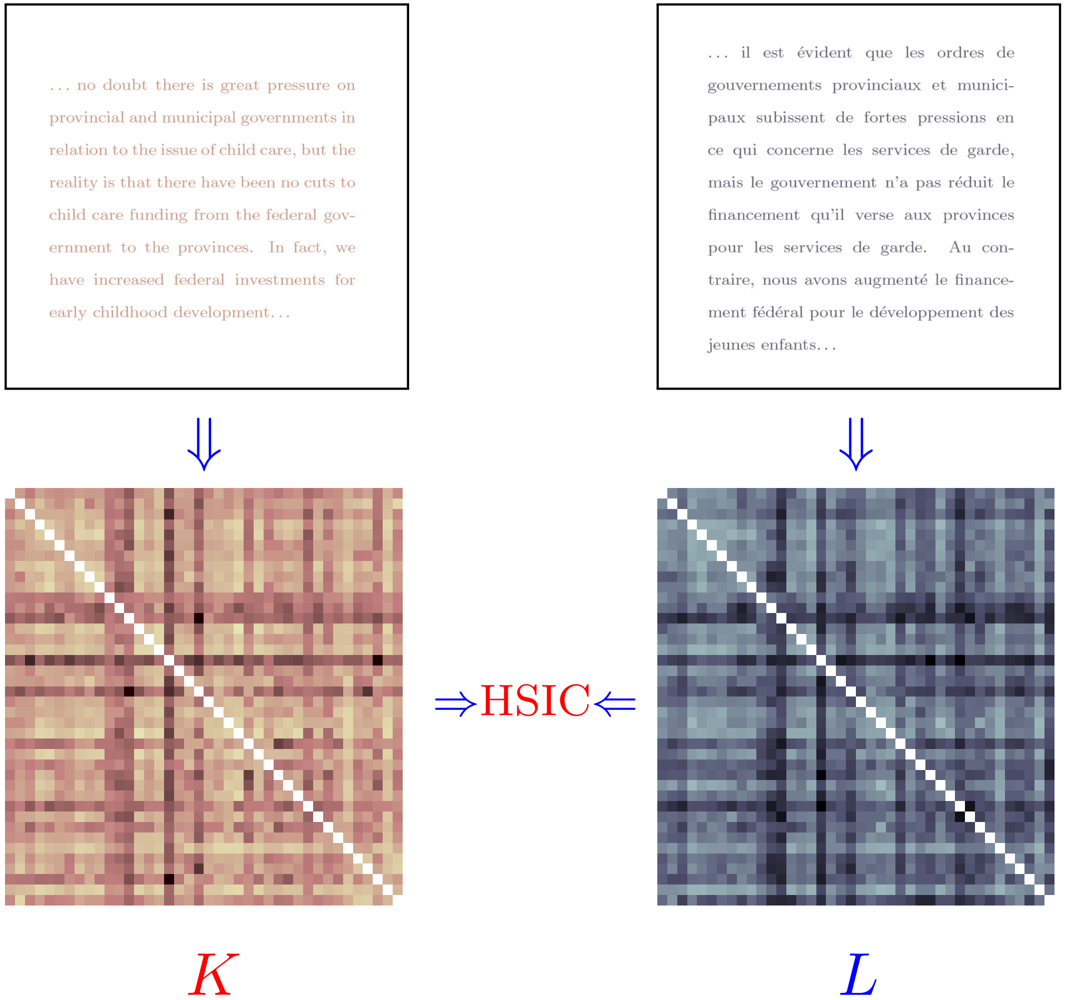
- k-spectrum kernel: average Type II error 0 ($\alpha = 0.05$)

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



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

Kernel two-sample tests for big data, optimal kernel choice

Quadratic time estimate of MMD

$$\mathrm{MMD}^2 = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|^2_{\mathcal{F}} = \mathbf{E}_{\mathsf{P}} k(x, x') + \mathbf{E}_{\mathsf{Q}} k(y, y') - 2\mathbf{E}_{\mathsf{P},\mathsf{Q}} k(x, y)$$

Quadratic time estimate of MMD

$$\mathrm{MMD}^2 = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_{\mathsf{P}} k(x, x') + \mathbf{E}_{\mathsf{Q}} k(y, y') - 2\mathbf{E}_{\mathsf{P},\mathsf{Q}} k(x, y)$$

Given i.i.d. $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_m\}$ from $\mathbf{P}, \mathbf{Q}$, respectively:

The earlier estimate: (quadratic time)

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

Quadratic time estimate of MMD

$$\text{MMD}^2 = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 = \mathbf{E}_{\mathbf{P}} k(x, x') + \mathbf{E}_{\mathbf{Q}} k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)$$

Given i.i.d. $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_m\}$ from $\mathbf{P}$, $\mathbf{Q}$, respectively:

The earlier estimate: (quadratic time)

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

New, linear time estimate:

$$\begin{aligned}
\widehat{\mathbb{E}}_{\mathbf{P}} k(x, x') &= \frac{2}{m} \left[k(x_1, x_2) + k(x_3, x_4) + \ldots \right] \\
&= \frac{2}{m} \sum_{i=1}^{m/2} k(x_{2i-1}, x_{2i})
\end{aligned}$$

Linear time MMD

Shorter expression with explicit k dependence:

$$\mathrm{MMD}^2 =: \eta_k(p,q) = \mathbb{E}_{xx'yy'} h_k(x,x',y,y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x,x',y,y') = k(x,x') + k(y,y') - k(x,y') - k(x',y),$$

and $v := [x,\ x',\ y,\ y']$.

Linear time MMD

Shorter expression with explicit k dependence:

$$\text{MMD}^2 =: \eta_k(p, q) = \mathbb{E}_{xx'yy'} h_k(x, x', y, y') =: \mathbb{E}_v h_k(v),$$

where

$$h_k(x, x', y, y') = k(x, x') + k(y, y') - k(x, y') - k(x', y),$$

and $v := [x,\ x',\ y,\ y']$.

The linear time estimate again:

$$\check{\eta}_k = \frac{2}{m} \sum_{i=1}^{m/2} h_k(v_i),$$

where $v_i := [x_{2i-1},\ x_{2i},\ y_{2i-1},\ y_{2i}]$ and

$$h_k(v_i) := k(x_{2i-1}, x_{2i}) + k(y_{2i-1}, y_{2i}) - k(x_{2i-1}, y_{2i}) - k(x_{2i}, y_{2i-1})$$

Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given m, hence...
- ...a much less powerful test for a given m

Linear time vs quadratic time MMD

Disadvantages of linear time MMD vs quadratic time MMD

- Much higher variance for a given m, hence...
- ...a much less powerful test for a given m

Advantages of the linear time MMD vs quadratic time MMD

- Very simple asymptotic null distribution (a Gaussian, vs an infinite weighted sum of χ^2)
- Both test statistic and threshold computable in $O(m)$, with storage $O(1)$.
- Given unlimited data, a given Type II error can be attained with less computation

Asymptotics of linear time MMD

By central limit theorem,

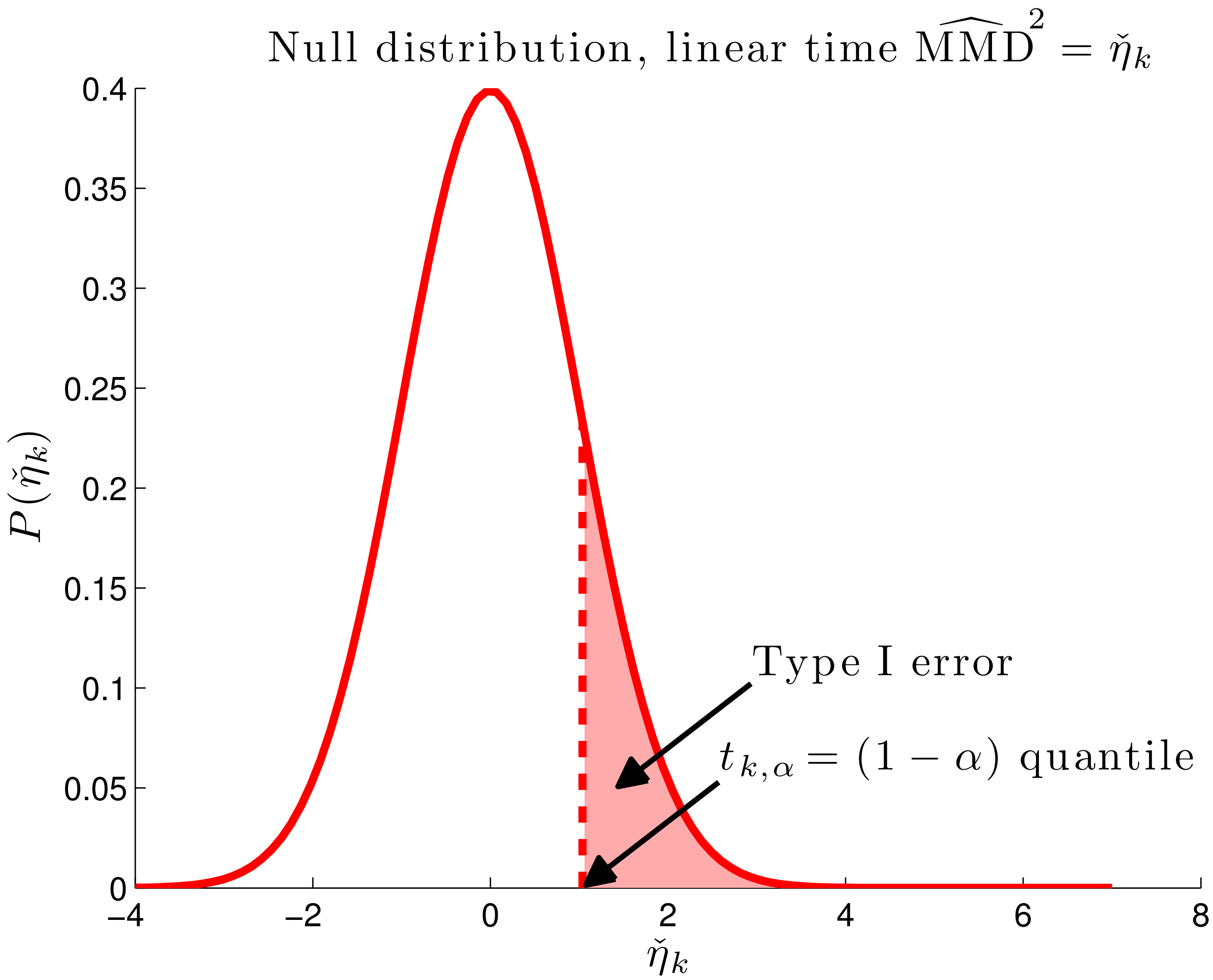
$$m^{1/2}\left(\check{\eta}_k - \eta_k(p,q)\right) \xrightarrow{D} \mathcal{N}(0, 2\sigma_k^2)$$

- assuming $0 < \mathbb{E}(h_k^2) < \infty$ (true for bounded k)
- $\sigma_k^2 = \mathbb{E}_v h_k^2(v) - \left[\mathbb{E}_v(h_k(v))\right]^2.$

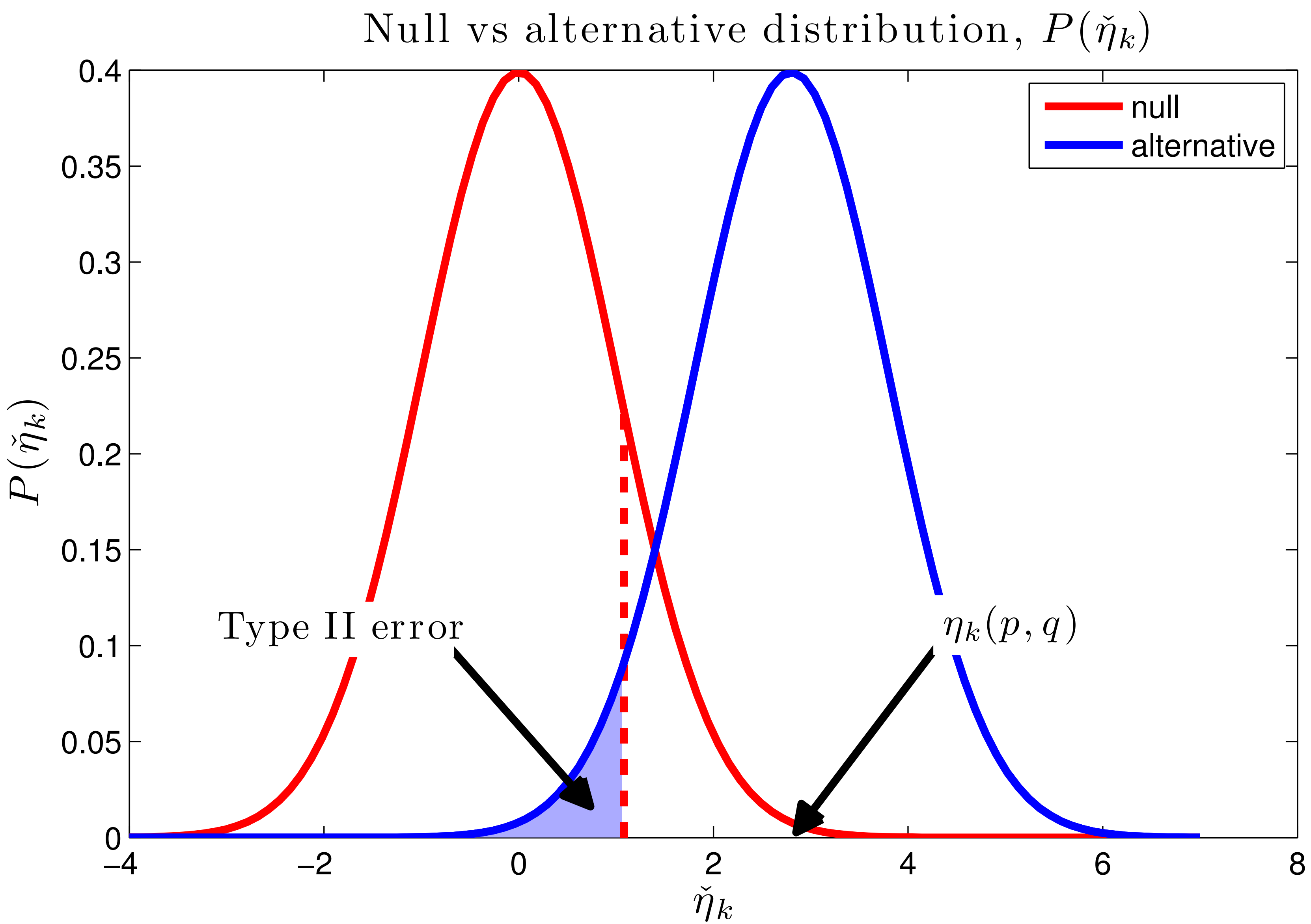
Hypothesis test

Hypothesis test of asymptotic level α:

$t_{k,\alpha} = m^{-1/2}\sigma_k\sqrt{2}\Phi^{-1}(1-\alpha)$ where Φ^{-1} is inverse CDF of $\mathcal{N}(0,1)$.



Type II error



The best kernel: minimizes Type II error

Type II error: $\check{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p,q) > 0$.

Prob. of a Type II error:

$$P(\check{\eta}_k < t_{k,\alpha}) = \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\eta_k(p,q)\sqrt{m}}{\sigma_k\sqrt{2}}\right)$$

where Φ is a Normal CDF.

The best kernel: minimizes Type II error

Type II error: $\check{\eta}_k$ falls below the threshold $t_{k,\alpha}$ and $\eta_k(p,q) > 0$.

Prob. of a Type II error:

$$P(\check{\eta}_k < t_{k,\alpha}) = \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\eta_k(p,q)\sqrt{m}}{\sigma_k\sqrt{2}}\right)$$

where Φ is a Normal CDF.

Since Φ monotonic, best kernel choice to minimize Type II error prob. is:

$$k_* = \arg\max_{k \in \mathcal{K}} \eta_k(p,q)\sigma_k^{-1},$$

where $\mathcal{K}$ is the family of kernels under consideration.

Learning the best kernel in a family

Define the family of kernels as follows:

$$\mathcal{K} := \left\{ k \ : \ k = \sum_{u=1}^{d} \beta_u k_u, \ \|\beta\|_1 = D, \ \beta_u \geq 0, \ \forall u \in \{1, \ldots, d\} \right\}.$$

Properties: if at least one $\beta_u > 0$

- all $k \in \mathcal{K}$ are valid kernels,
- If all k_u charateristic then k characteristic

Test statistic

The squared MMD becomes

$$\eta_k(p, q) = \|\mu_k(p) - \mu_k(q)\|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p, q),$$

where $\eta_u(p, q) := \mathbb{E}_v h_u(v)$.

Test statistic

The squared MMD becomes

$$\eta_k(p,q) = \|\mu_k(p) - \mu_k(q)\|^2_{\mathcal{F}_k} = \sum_{u=1}^{d} \beta_u \eta_u(p,q),$$

where $\eta_u(p,q) := \mathbb{E}_v h_u(v)$.

Denote:

- $\beta = (\beta_1, \beta_2, \ldots, \beta_d)^\top \in \mathbb{R}^d$,
- $h = (h_1, h_2, \ldots, h_d)^\top \in \mathbb{R}^d$,
  - $h_u(x, x', y, y') = k_u(x, x') + k_u(y, y') - k_u(x, y') - k_u(x', y)$
- $\eta = \mathbb{E}_v(h) = (\eta_1, \eta_2, \ldots, \eta_d)^\top \in \mathbb{R}^d$.

Quantities for test:

$$\eta_k(p,q) = \mathbb{E}(\beta^\top h) = \beta^\top \eta \qquad \sigma_k^2 := \beta^\top \mathrm{cov}(h) \beta.$$

Optimization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \qquad\qquad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top \left(\hat{Q} + \lambda_m I\right)\beta},$$

$\hat{Q}$ is empirical estimate of $\mathrm{cov}(h)$.

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\check{\eta}_k, \check{\sigma}_k$ on data to be tested (why?)

Optimization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Empirical test parameters:

$$\hat{\eta}_k = \beta^\top \hat{\eta} \qquad\qquad \hat{\sigma}_{k,\lambda} = \sqrt{\beta^\top \left(\hat{Q} + \lambda_m I\right)\beta},$$

$\hat{Q}$ is empirical estimate of $\text{cov}(h)$.

Note: $\hat{\eta}_k, \hat{\sigma}_{k,\lambda}$ computed on training data, vs $\check{\eta}_k, \check{\sigma}_k$ on data to be tested (why?)

Objective:

$$\begin{aligned}
\hat{\beta}^* &= \arg\max_{\beta \succeq 0} \; \hat{\eta}_k(p,q)\hat{\sigma}_{k,\lambda}^{-1} \\
&= \arg\max_{\beta \succeq 0} \; \left(\beta^\top \hat{\eta}\right)\left(\beta^\top \left(\hat{Q} + \lambda_m I\right)\beta\right)^{-1/2} \\
&=: \alpha(\beta; \hat{\eta}, \hat{Q})
\end{aligned}$$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\quad \beta \succeq 0 \quad$ s.t. $\quad \alpha(\beta; \hat{\eta}, \hat{Q}) > 0.$

Thus: $\quad \alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\quad \beta \succeq 0 \quad$ s.t. $\quad \alpha(\beta; \hat{\eta}, \hat{Q}) > 0.$

Thus: $\quad \alpha(\hat{\beta}^*; \hat{\eta}, \hat{Q}) > 0$

Solve easier problem: $\quad \hat{\beta}^* = \arg\max_{\beta \succeq 0} \alpha^2(\beta; \hat{\eta}, \hat{Q}).$

Quadratic program:

$$\min\{\beta^\top \left(\hat{Q} + \lambda_m I\right) \beta : \beta^\top \hat{\eta} = 1,\ \beta \succeq 0\}$$

Optmization of ratio $\eta_k(p,q)\sigma_k^{-1}$

Assume: $\hat{\eta}$ has at least one positive entry

Then there exists $\quad \beta \succeq 0 \quad$ s.t. $\quad \alpha(\beta;\hat{\eta},\hat{Q}) > 0.$

Thus: $\quad \alpha(\hat{\beta}^*;\hat{\eta},\hat{Q}) > 0$

Solve easier problem: $\quad \hat{\beta}^* = \arg\max_{\beta\succeq 0} \alpha^2(\beta;\hat{\eta},\hat{Q}).$

Quadratic program:

$$\min\{\beta^\top\left(\hat{Q}+\lambda_m I\right)\beta : \beta^\top\hat{\eta}=1,\ \beta\succeq 0\}$$

What if $\hat{\eta}$ has no positive entries?

Test procedure

1. Split the data into testing and training.
2. On the training data:
   (a) Compute $\hat{\eta}_u$ for all $k_u \in \mathcal{K}$
   (b) If at least one $\hat{\eta}_u > 0$, solve the QP to get β^*, else choose random kernel from $\mathcal{K}$
3. On the test data:
   (a) Compute $\check{\eta}_{k^*}$ using $k^* = \sum_{u=1}^{d} \beta^* k_u$
   (b) Compute test threshold $\check{t}_{\alpha,k^*}$ using $\check{\sigma}_{k^*}$
4. Reject null if $\check{\eta}_{k^*} > \check{t}_{\alpha,k^*}$

Convergence bounds

Assume bounded kernel, σ_k, bounded away from 0.
If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left|\sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1}\right| = O_P\left(m^{-1/3}\right).$$

Convergence bounds

Assume bounded kernel, σ_k, bounded away from 0.

If $\lambda_m = \Theta(m^{-1/3})$ then

$$\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| = O_P\left(m^{-1/3}\right).$$

Idea:

$$\begin{aligned}
&\left| \sup_{k \in \mathcal{K}} \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \sup_{k \in \mathcal{K}} \eta_k \sigma_k^{-1} \right| \\
&\leq \sup_{k \in \mathcal{K}} \left| \hat{\eta}_k \hat{\sigma}_{k,\lambda}^{-1} - \eta_k \sigma_{k,\lambda}^{-1} \right| + \sup_{k \in \mathcal{K}} \left| \eta_k \sigma_{k,\lambda}^{-1} - \eta_k \sigma_k^{-1} \right| \\
&\leq \frac{\sqrt{d}}{D\sqrt{\lambda_m}} \left(C_1 \sup_{k \in \mathcal{K}} |\hat{\eta}_k - \eta_k| + C_2 \sup_{k \in \mathcal{K}} |\hat{\sigma}_{k,\lambda} - \sigma_{k,\lambda}| \right) + C_3 D^2 \lambda_m,
\end{aligned}$$

Experiments

Competing approaches

- Median heuristic
- Max. MMD: choose $k_u \in \mathcal{K}$ with the largest $\hat{\eta}_u$
  - same as maximizing $\beta^\top \hat{\eta}$ subject to $\|\beta\|_1 \leq 1$
- ℓ_2 statistic: maximize $\beta^\top \hat{\eta}$ subject to $\|\beta\|_2 \leq 1$
- Cross validation on training set

Also compare with:

- Single kernel that maximizes ratio $\eta_k(p, q)\sigma_k^{-1}$

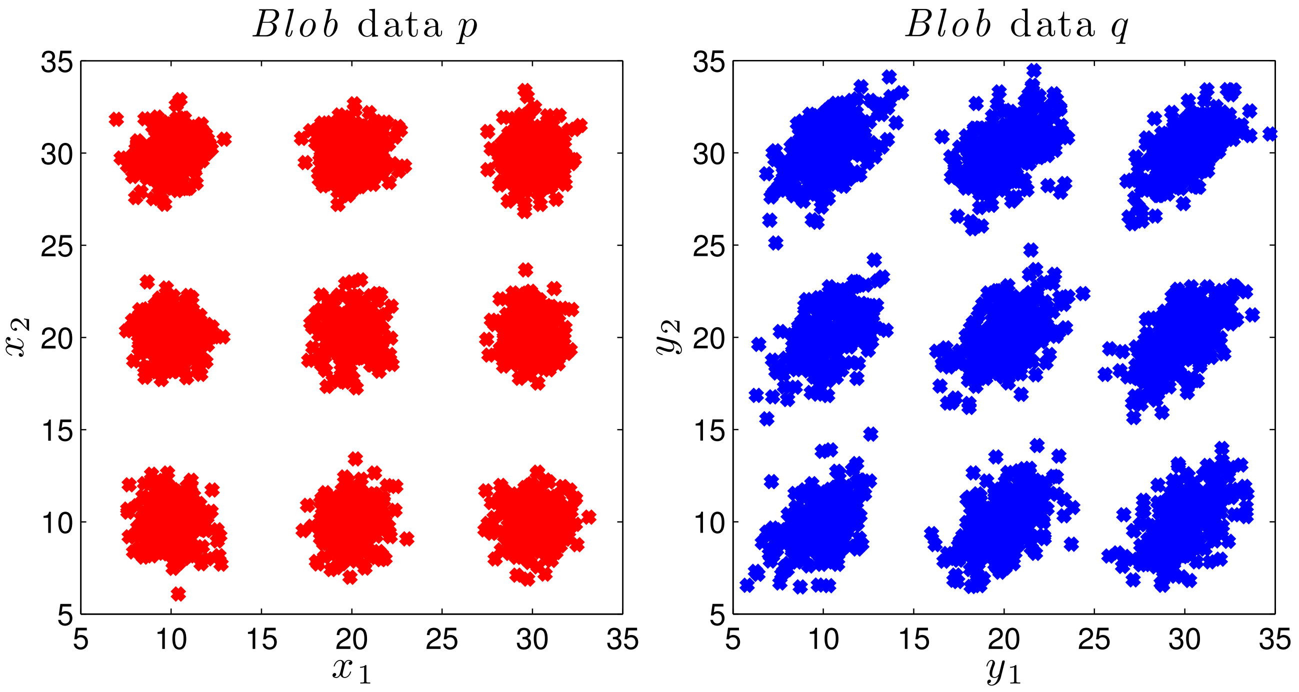
Blobs: data

Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

Blobs: data

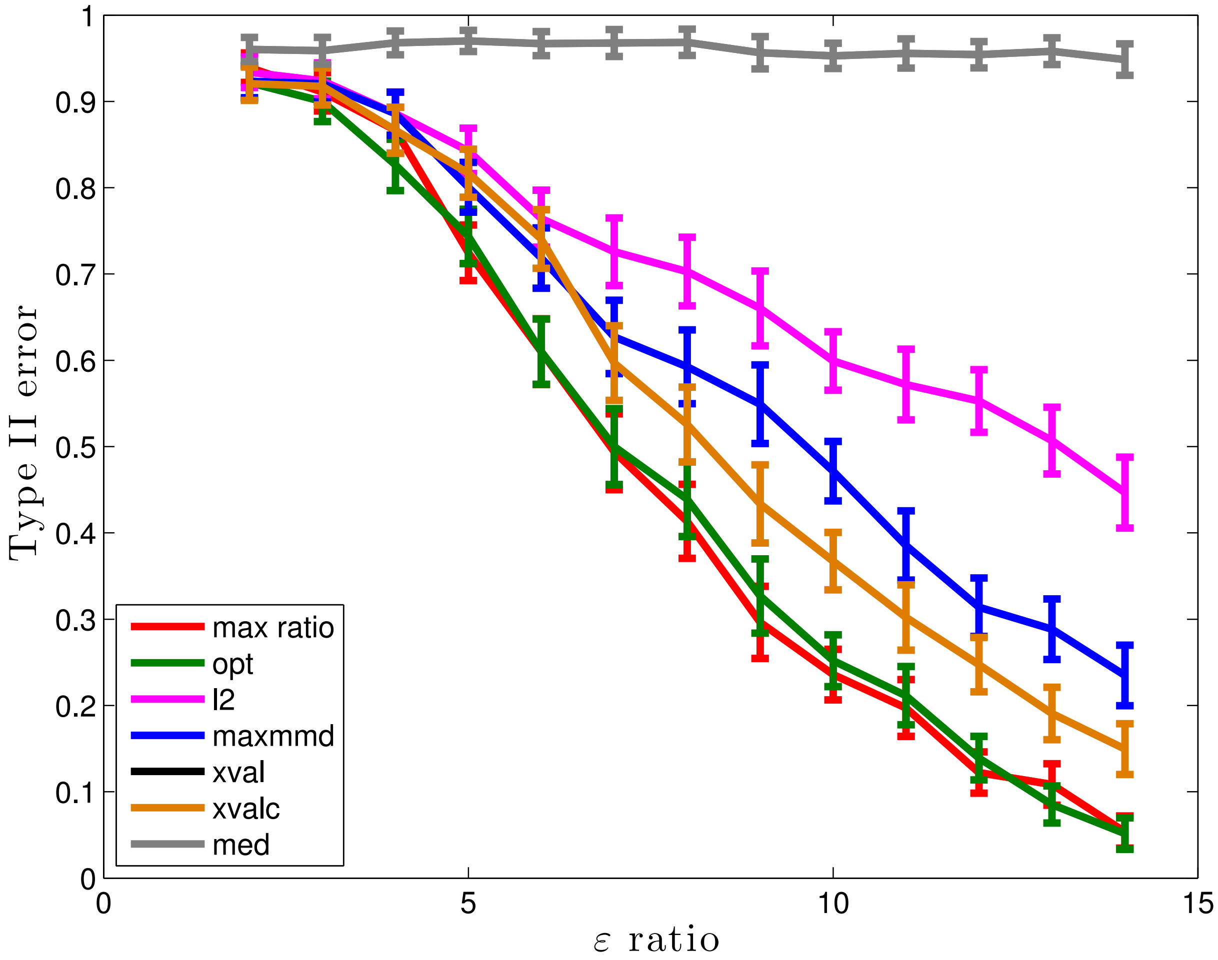
Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

We distinguish a field of Gaussian blobs with different covariances.



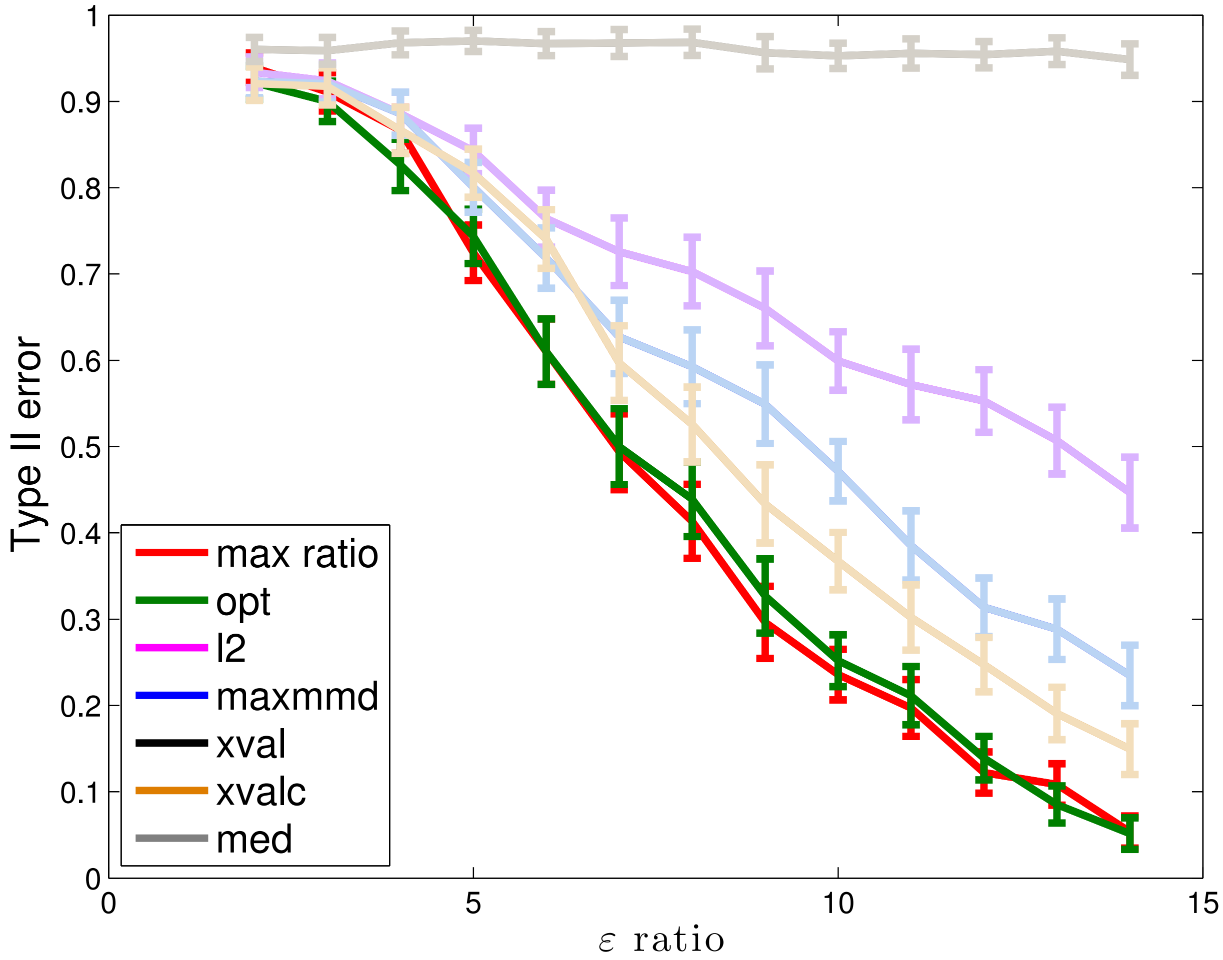
Ratio $\varepsilon = 3.2$ of largest to smallest eigenvalues of blobs in q.

Blobs: results



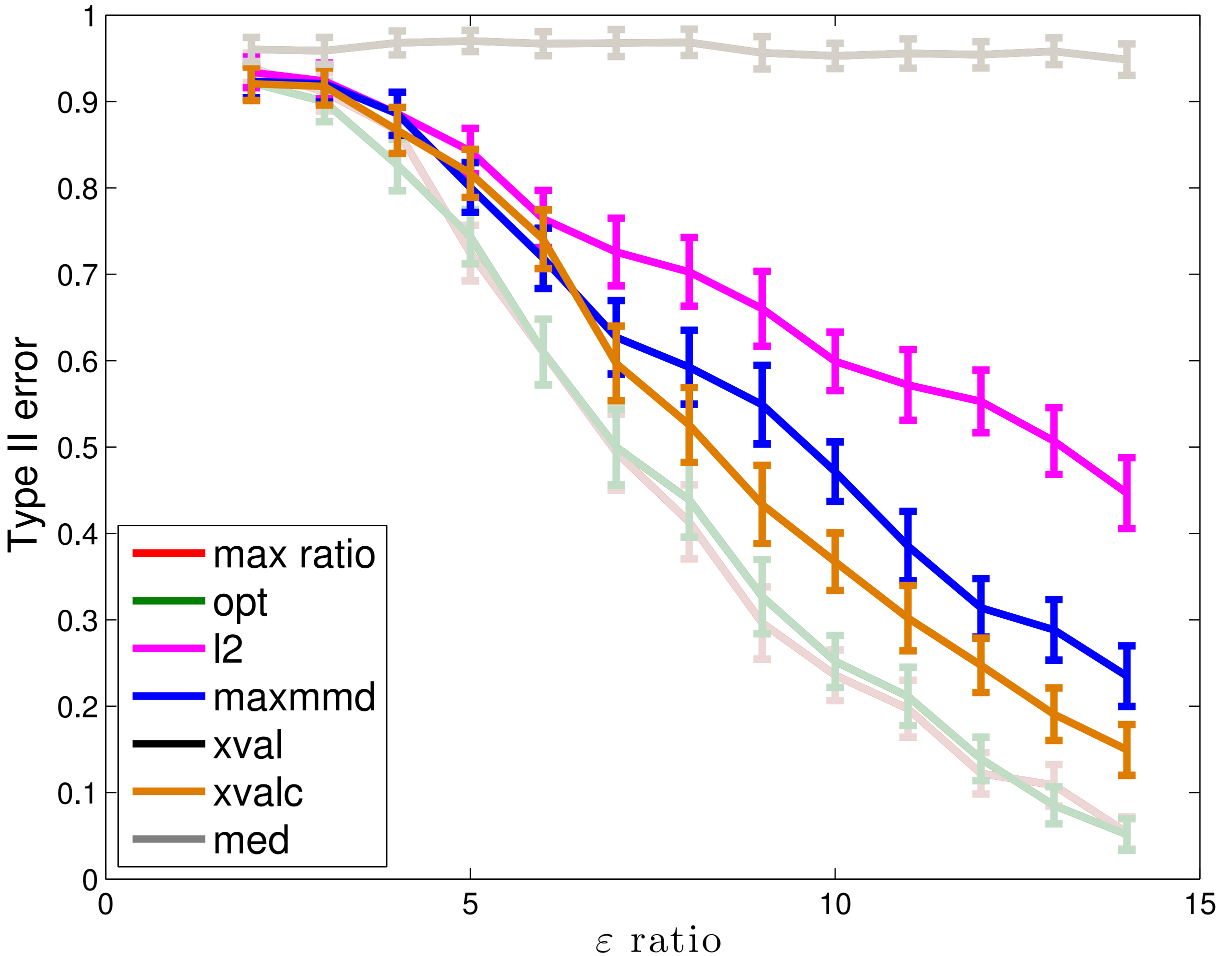
Parameters: $m = 10,000$ (for training and test). Ratio ε of largest to smallest eigenvalues of blobs in q. Results are average over 617 trials.

Blobs: results



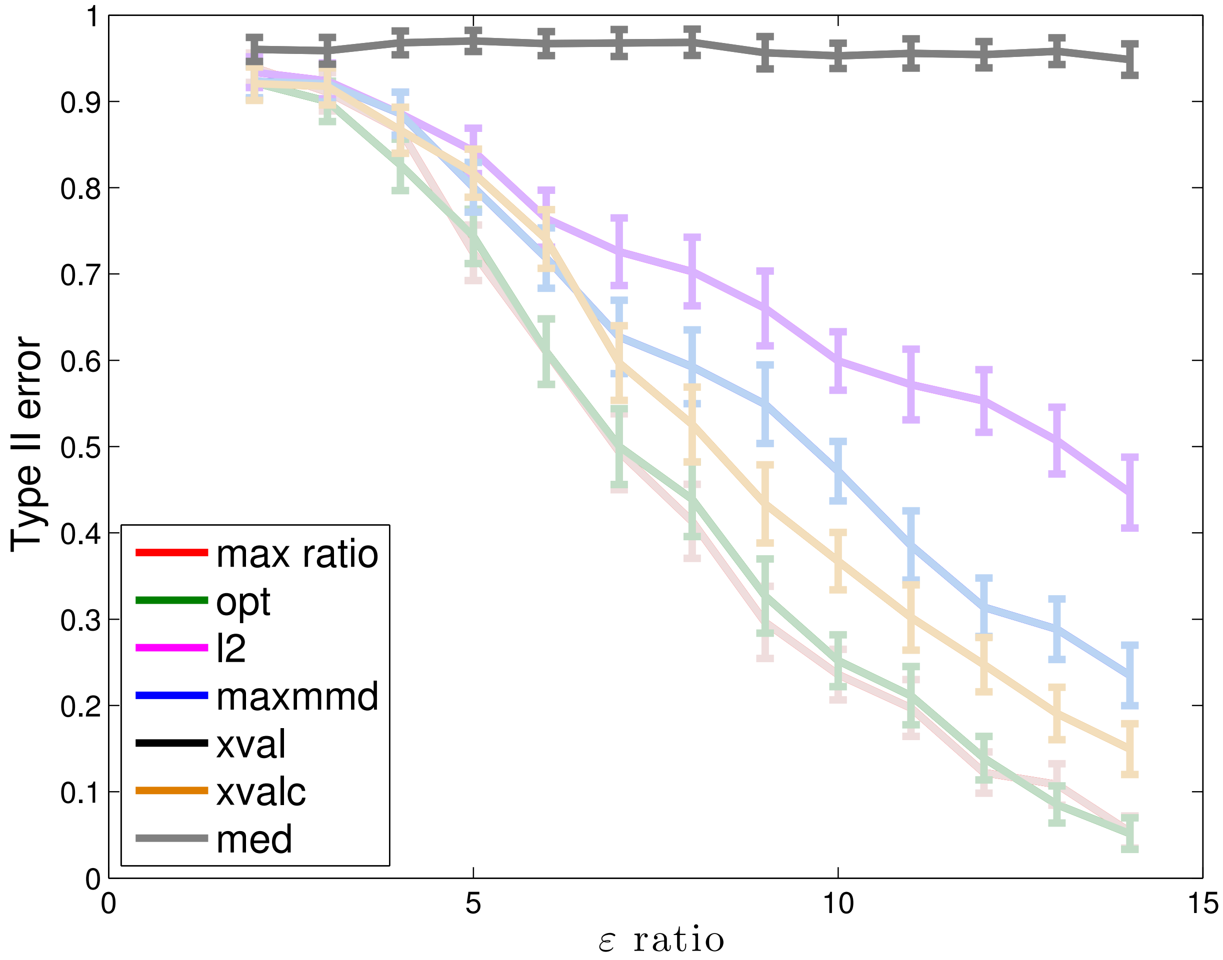
Optimize ratio $\eta_k(p,q)\sigma_k^{-1}$

Blobs: results



Maximize $\eta_k(p, q)$ with β constraint

Blobs: results



Median heuristic

Feature selection: data

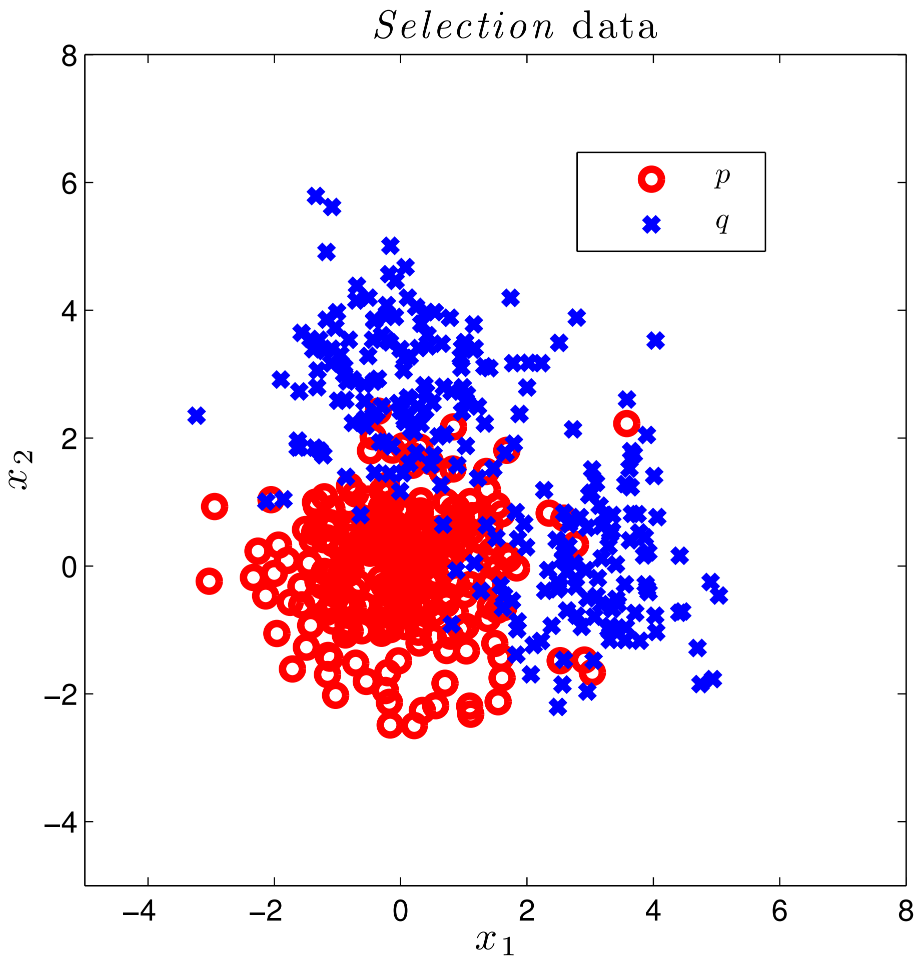
Idea: no single best kernel.

Each of the k_u are univariate (along a single coordinate)

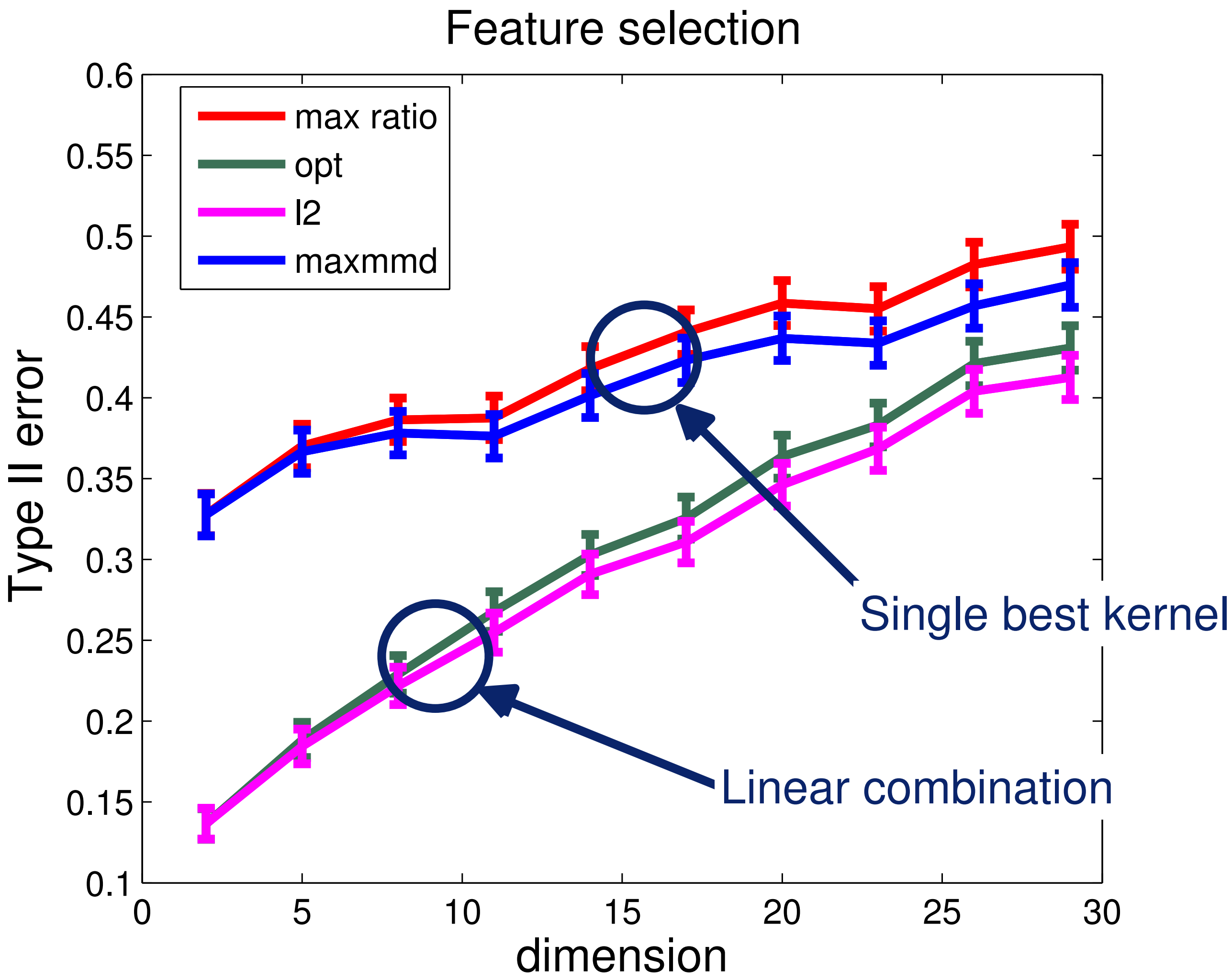
Feature selection: data

Idea: no single best kernel.

Each of the k_u are univariate (along a single coordinate)



Feature selection: results



$m = 10,000$, average over 5000 trials

Amplitude modulated signals

Given an audio signal $s(t)$, an amplitude modulated signal can be defined

$$u(t) = \sin(\omega_c t)\,[a\,s(t) + l]$$

- ω_c: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset

Amplitude modulated signals

Given an audio signal $s(t)$, an amplitude modulated signal can be defined

$$u(t) = \sin(\omega_c t)\,[a\,s(t) + l]$$

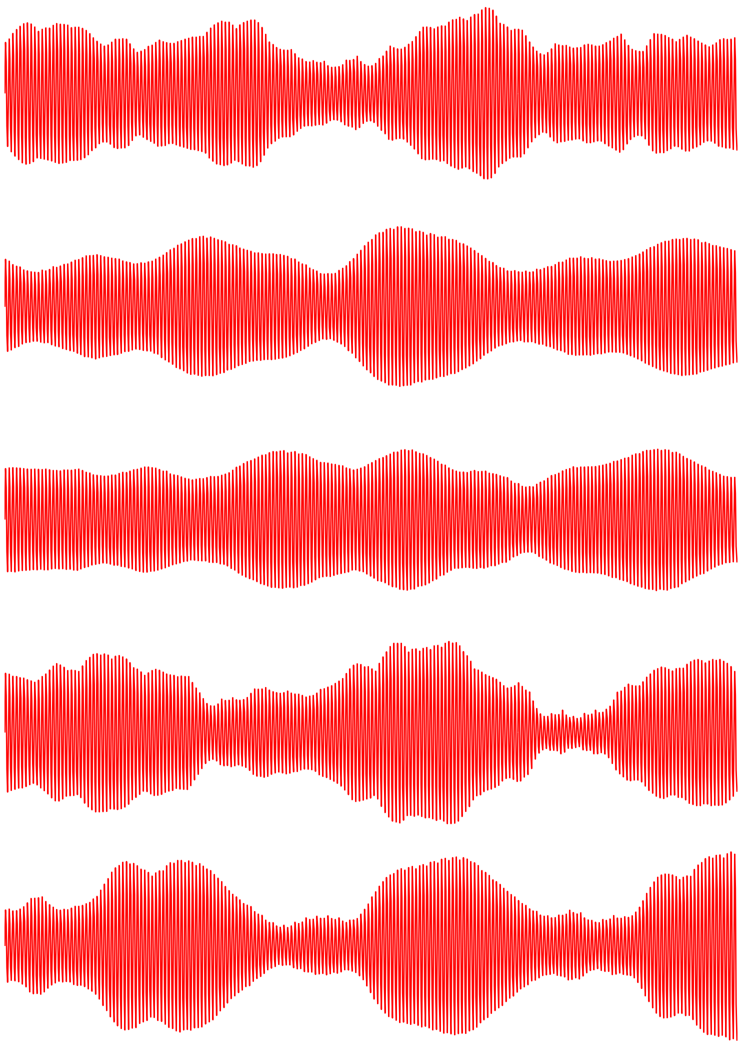
- ω_c: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset

Two amplitude modulated signals from same artist (in this case, Magnetic Fields).

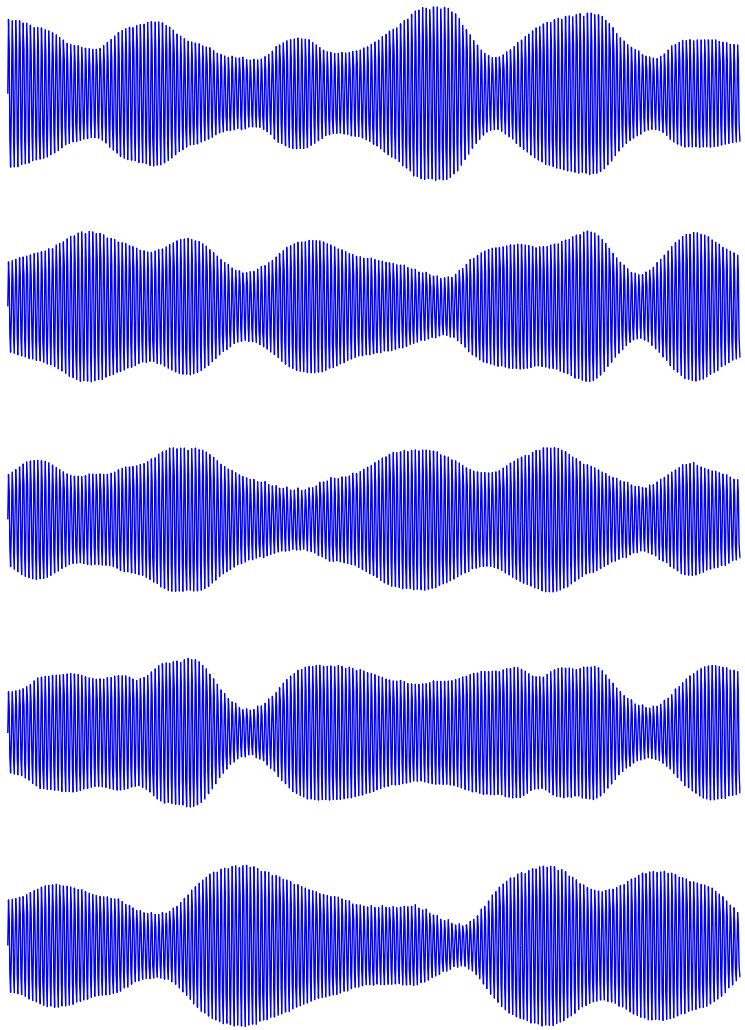
- Music sampled at 8KHz (very low)
- Carrier frequency is 24kHz
- AM signal observed at 120kHz
- Samples are extracts of length $N = 1000$, approx. 0.01 sec (very short).
- Total dataset size is 30,000 samples from each of p, q.

Amplitude modulated signals

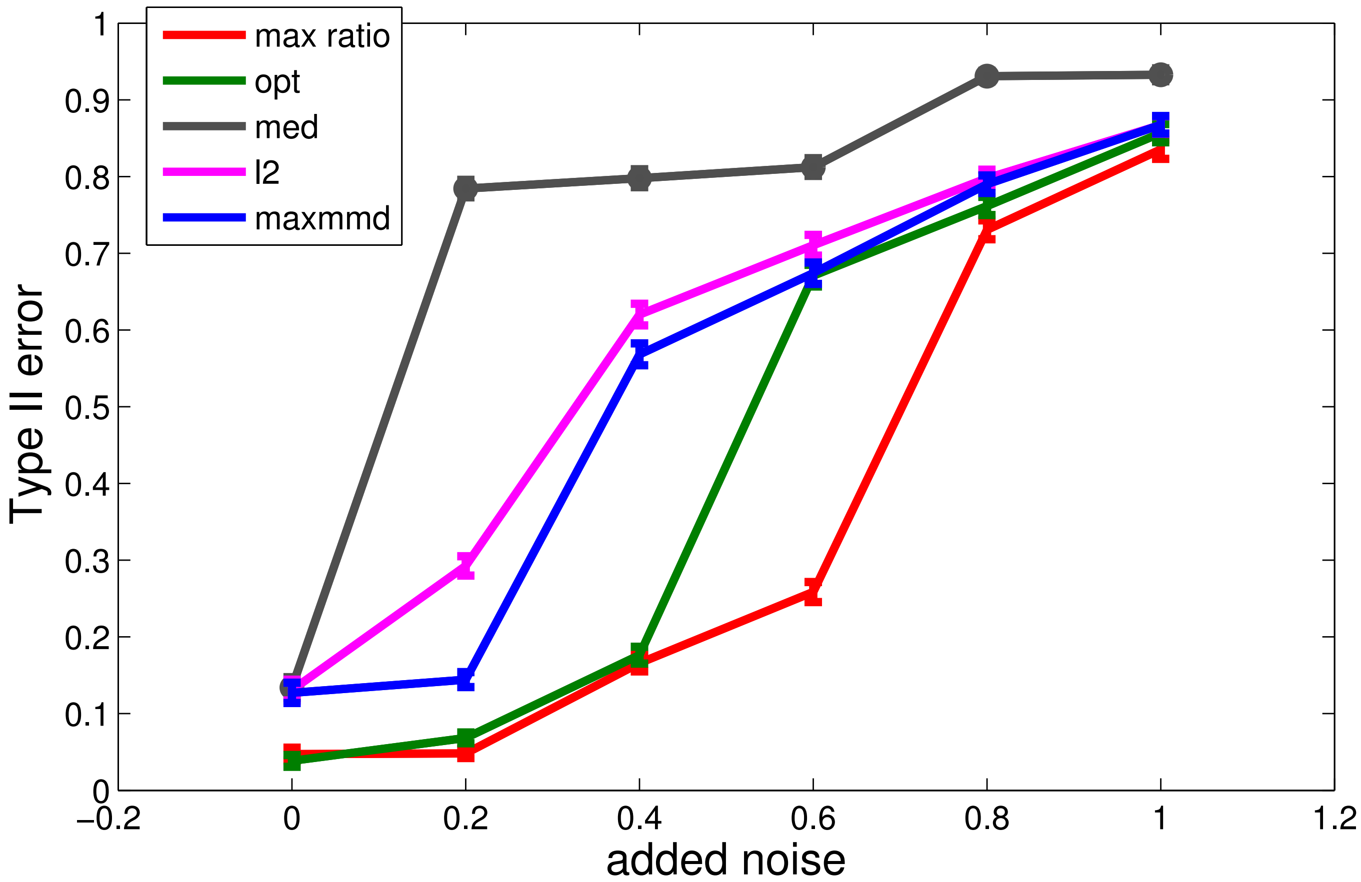
Samples from P



Samples from Q



Results: AM signals



$m = 10,000$ (for training and test) and scaling $a = 0.5$. Average over 4124 trials. Gaussian noise added.

Observations on kernel choice

- It is possible to choose the best kernel for a kernel two-sample test
- Kernel choice matters for “difficult” problems, where the distributions differ on a lengthscale different to that of the data.
- Ongoing work:
  - quadratic time statistic
  - avoid training/test split

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

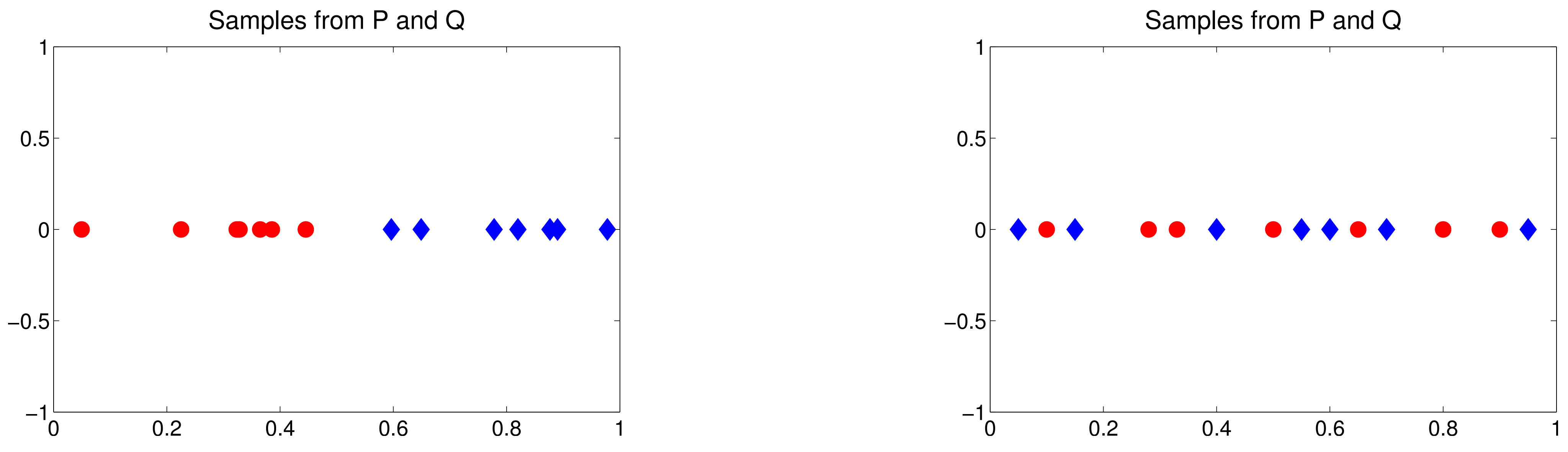
Local departures from the null

What is a hard testing problem?

Local departures from the null

What is a hard testing problem?

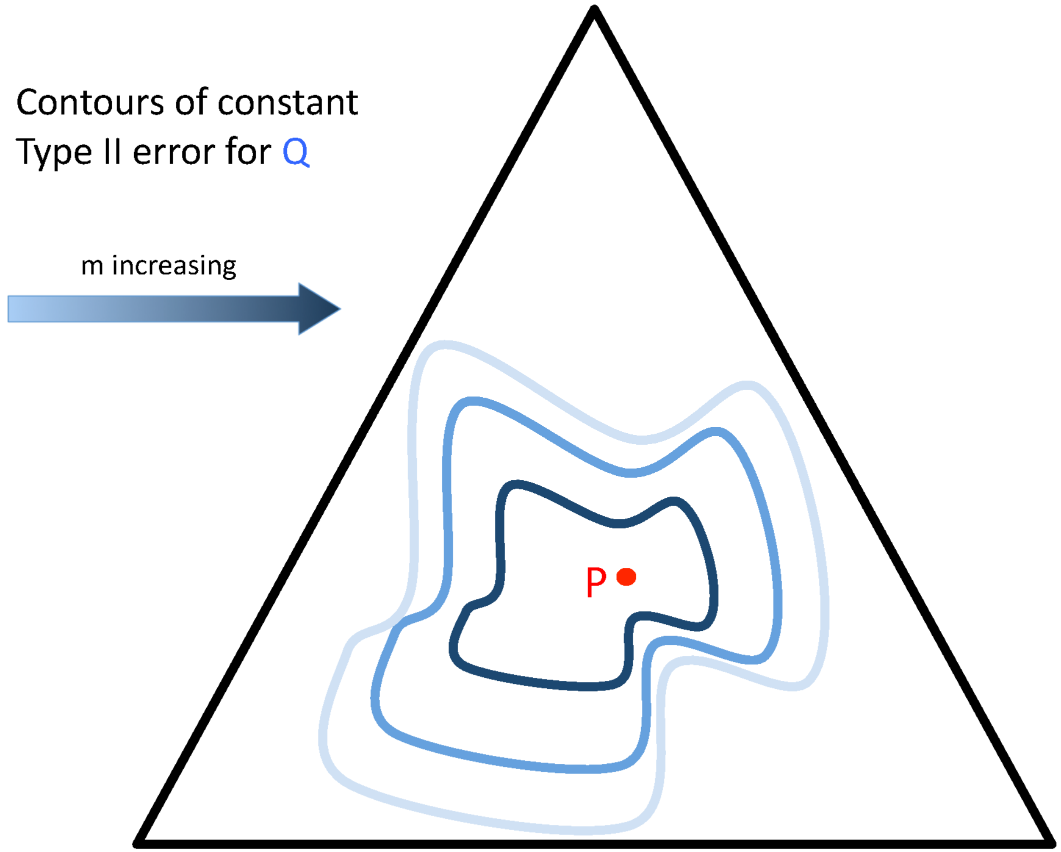
- First version: for fixed m, "closer" **P** and **Q** have higher Type II error



Local departures from the null

What is a hard testing problem?

- As m increases, distinguish "closer" **P** and **Q** with fixed Type II error



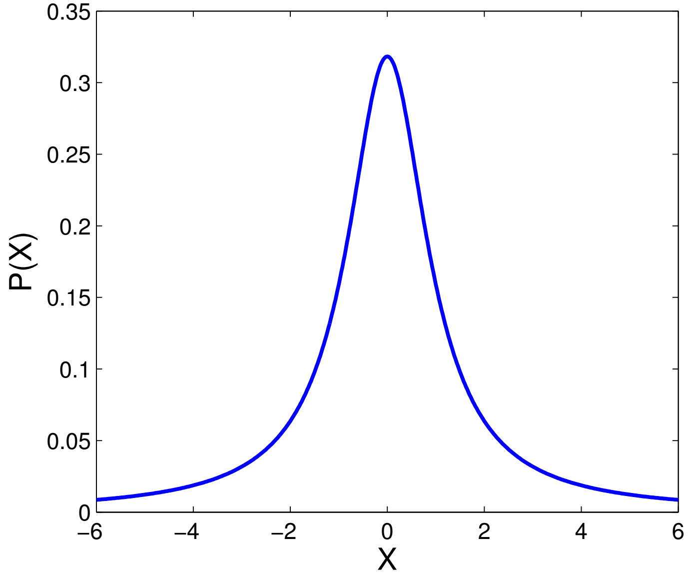
Local departures from the null

What is a hard testing problem?

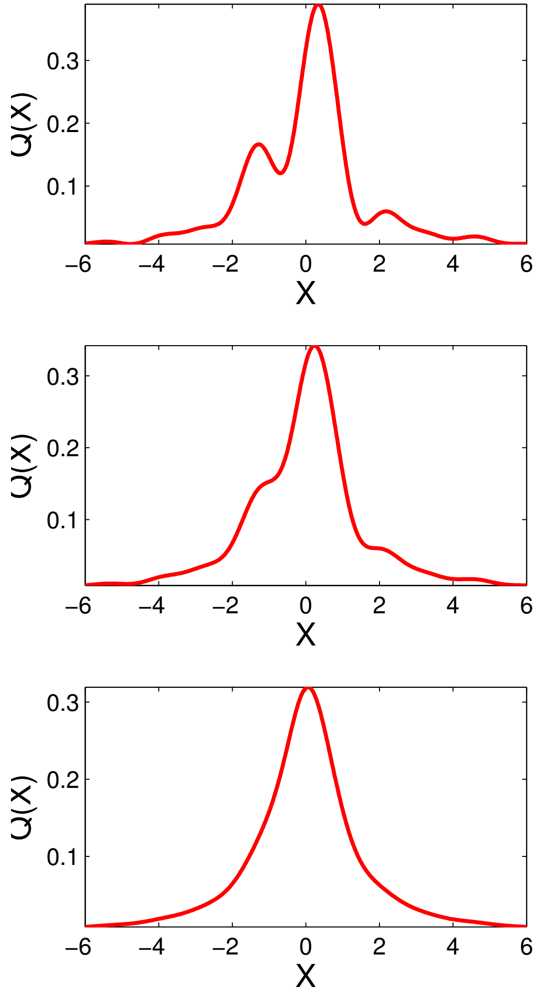
- As m increases, distinguish "closer" **P** and **Q** with fixed Type II error
- Example: $f_{\mathbf{P}}$ and $f_{\mathbf{Q}}$ probability densities, $f_{\mathbf{Q}} = f_{\mathbf{P}} + \delta g$, where $\delta \in \mathbb{R}$, g some *fixed* function such that $f_{\mathbf{Q}}$ is a valid density
  - If $\delta \sim m^{-1/2}$, Type II error approaches a constant

More general local departures from null

- Example: $f_{\mathbf{P}}$ and $f_{\mathbf{Q}}$ probability densities, $f_{\mathbf{Q}} = f_{\mathbf{P}} + \delta g$, where $\delta \in \mathbb{R}$, g some *fixed* function such that $f_{\mathbf{Q}}$ is a valid density



VS



Local departures from the null

What is a hard testing problem?

- As we see more samples m, distinguish "closer" $\mathbf{P}$ and $\mathbf{Q}$ with same Type II error
- Example: $f_{\mathbf{P}}$ and $f_{\mathbf{Q}}$ probability densities, $f_{\mathbf{Q}} = f_{\mathbf{P}} + \delta g$, where $\delta \in \mathbb{R}$, g some *fixed* function such that $f_{\mathbf{Q}}$ is a valid density
  - If $\delta \sim m^{-1/2}$, Type II error approaches a constant
- ...but **other choices also possible** – how to characterize them all?

Local departures from the null

What is a hard testing problem?

- As we see more samples m, distinguish "closer" $\mathbf{P}$ and $\mathbf{Q}$ with same Type II error
- Example: $f_{\mathbf{P}}$ and $f_{\mathbf{Q}}$ probability densities, $f_{\mathbf{Q}} = f_{\mathbf{P}} + \delta g$, where $\delta \in \mathbb{R}$, g some *fixed* function such that $f_{\mathbf{Q}}$ is a valid density
  - If $\delta \sim m^{-1/2}$, Type II error approaches a constant
- ...but **other choices also possible** – how to characterize them all?

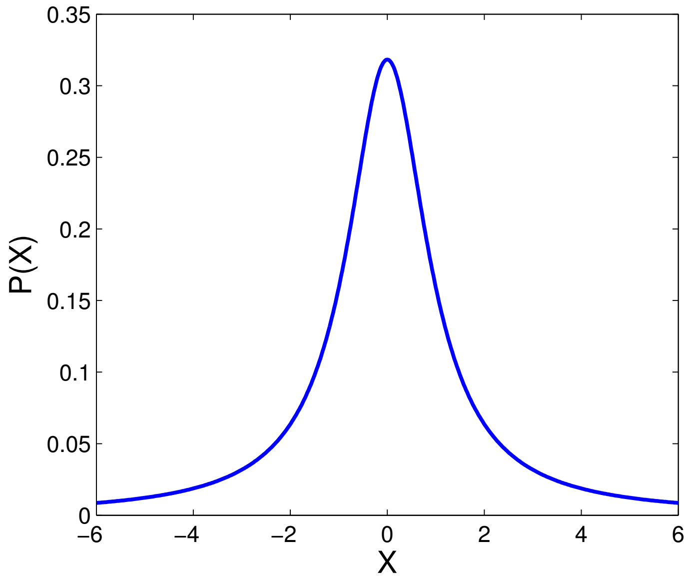
General characterization of local departures from $\mathcal{H}_0$:

- Write $\mu_{\mathbf{Q}} = \mu_{\mathbf{P}} + g_m$, where $g_m \in \mathcal{F}$ chosen such that $\mu_{\mathbf{P}} + g_m$ a valid distribution embedding
- Minimum distinguishable distance [JMLR12]

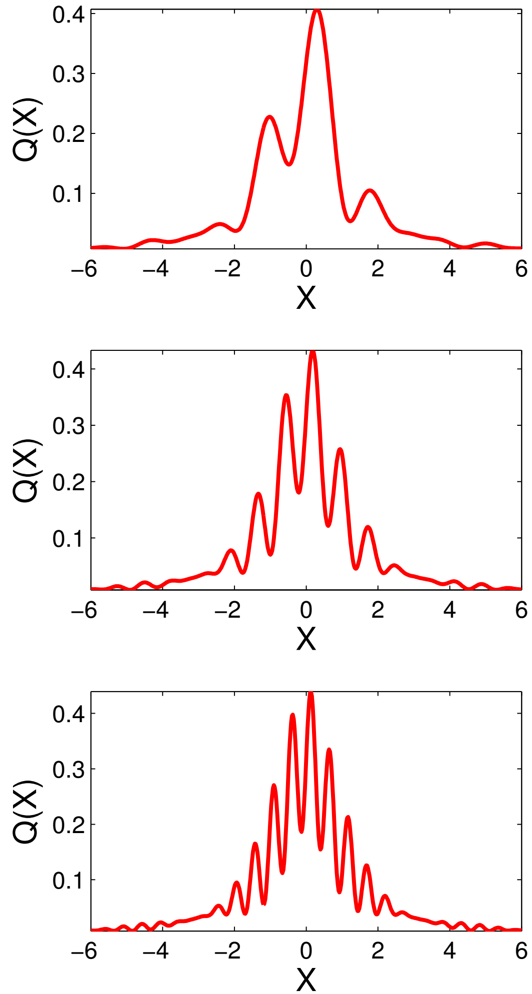
$$\|g_m\|_{\mathcal{F}} = cm^{-1/2}$$

More general local departures from null

- More advanced example of a local departure from the null
- Recall: $\mu_{\mathbf{Q}} = \mu_{\mathbf{P}} + g_m$, and $\|g_m\|_{\mathcal{F}} = cm^{-1/2}$



VS



Kernels vs kernels

- How does MMD relate to Parzen density estimate? [Anderson et al., 1994]

$$\hat{f}_{\mathbf{P}}(x) = \frac{1}{m}\sum_{i=1}^{m}\kappa\left(x_i - x\right), \text{ where } \kappa \text{ satisfies } \int_{\mathcal{X}}\kappa\left(x\right)dx = 1 \text{ and } \kappa\left(x\right) \geq 0.$$

Kernels vs kernels

- How does MMD relate to Parzen density estimate? [Anderson et al., 1994]

$$\hat{f}_{\mathbf{P}}(x) = \frac{1}{m}\sum_{i=1}^{m} \kappa\left(x_i - x\right), \text{ where } \kappa \text{ satisfies } \int_{\mathcal{X}} \kappa\left(x\right) dx = 1 \text{ and } \kappa\left(x\right) \geq 0.$$

- L_2 distance between Parzen density estimates:

$$D_2(\hat{f}_{\mathbf{P}}, \hat{f}_{\mathbf{Q}})^2 = \int \left[\frac{1}{m}\sum_{i=1}^{m} \kappa(x_i - z) - \frac{1}{m}\sum_{i=1}^{m} \kappa(y_i - z)\right]^2 dz$$

$$= \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i - x_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(y_i - y_j) - \frac{2}{m^2}\sum_{i,j=1}^{m} k(x_i - y_j),$$

where $k(x-y) = \int \kappa(x-z)\kappa(y-z)dz$

Kernels vs kernels

- How does MMD relate to Parzen density estimate? [Anderson et al., 1994]

$$\hat{f}_{\mathbf{P}}(x) = \frac{1}{m}\sum_{i=1}^{m} \kappa\left(x_i - x\right), \text{ where } \kappa \text{ satisfies } \int_{\mathcal{X}} \kappa\left(x\right) dx = 1 \text{ and } \kappa\left(x\right) \geq 0.$$

- L_2 distance between Parzen density estimates:

$$D_2(\hat{f}_{\mathbf{P}}, \hat{f}_{\mathbf{Q}})^2 = \int \left[\frac{1}{m}\sum_{i=1}^{m} \kappa(x_i - z) - \frac{1}{m}\sum_{i=1}^{m} \kappa(y_i - z)\right]^2 dz$$

$$= \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i - x_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} k(y_i - y_j) - \frac{2}{m^2}\sum_{i,j=1}^{m} k(x_i - y_j),$$

where $k(x - y) = \int \kappa(x - z)\kappa(y - z)dz$

- $f_{\mathbf{Q}} = f_{\mathbf{P}} + \delta g$, minimum distance to discriminate $f_{\mathbf{P}}$ from $f_{\mathbf{Q}}$ is $\delta = (m)^{-1/2} h_m^{-d/2}$, where h_m is width of κ.

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 $\mathrm{MMD}\{\mathbf{P}, \mathbf{Q}; F\} = 0$ iff $\mathbf{P} = \mathbf{Q}$

Characteristic Kernels (via universality)

Proof:

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

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

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

Characteristic Kernels (via universality)

Proof:

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

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

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

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 $\mathrm{MMD}\{\mathsf{P}, \mathsf{Q}; F\} = 0$ 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 $\mathsf{P} = \mathsf{Q}$.

References

V. Alba Fernández, M. Jiménez-Gamero, and J. Muñoz Garcia. A test for the two-sample problem based on empirical characteristic functions. *Comput. Stat. Data An.*, 52:3730–3748, 2008.

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.

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.

K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schoelkopf. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems 21*, pages 473–480, Red Hook, NY, 2009. Curran Associates Inc.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems 20*, pages 609–616. MIT Press, Cambridge, MA, 2008.

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.

S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 2nd edition, 1999.

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.

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

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

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

Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Information Science and Statistics. Springer, 2008.

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.

S. K. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):917–929, 2006.