Optimal kernel choice for kernel hypothesis testing

Arthur Gretton

Gatsby Computational Neuroscience Unit

MSR, Nov. 2012

First motivating question

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

First motivating question

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

First motivating question

- 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

First motivating question

- 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

First motivating question

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

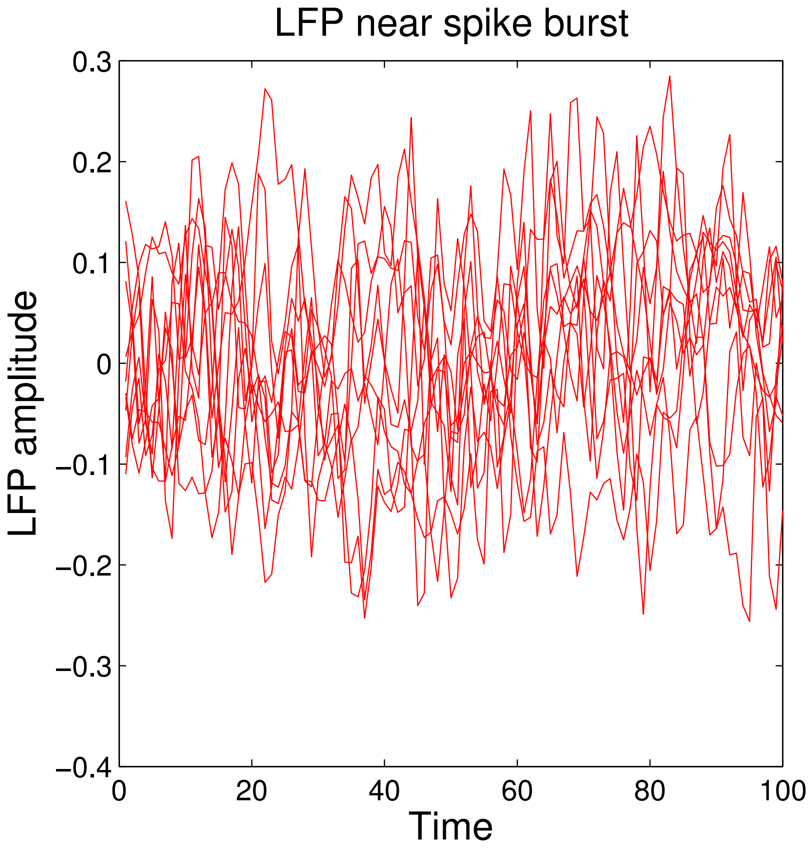
... no doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development...

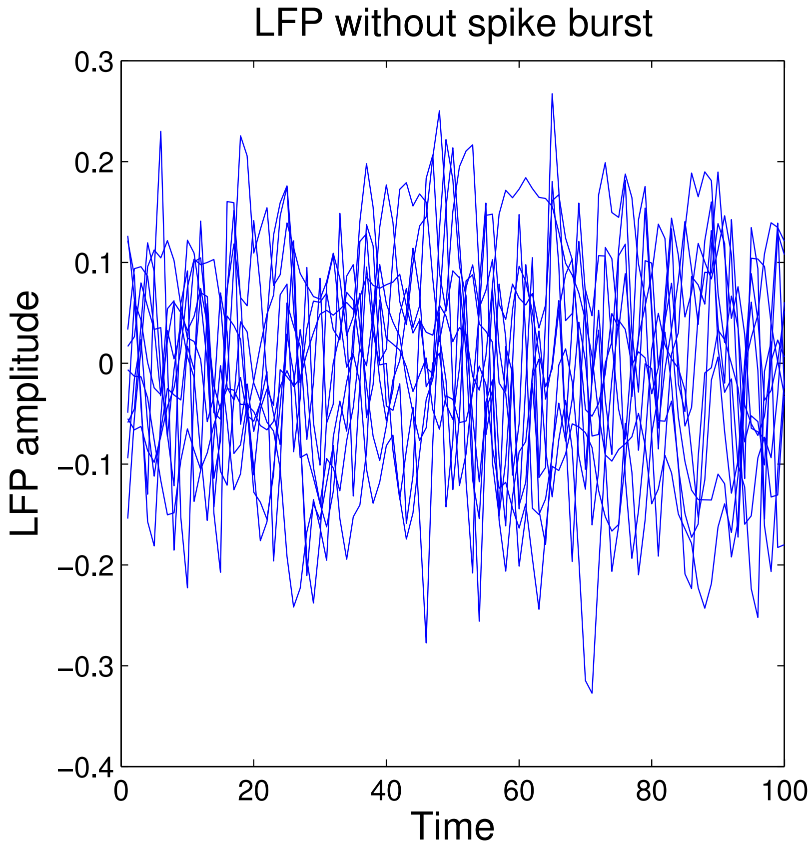
$\overset{?}{\Longleftrightarrow}$

... 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...

Second motivating question: differences in brain signals

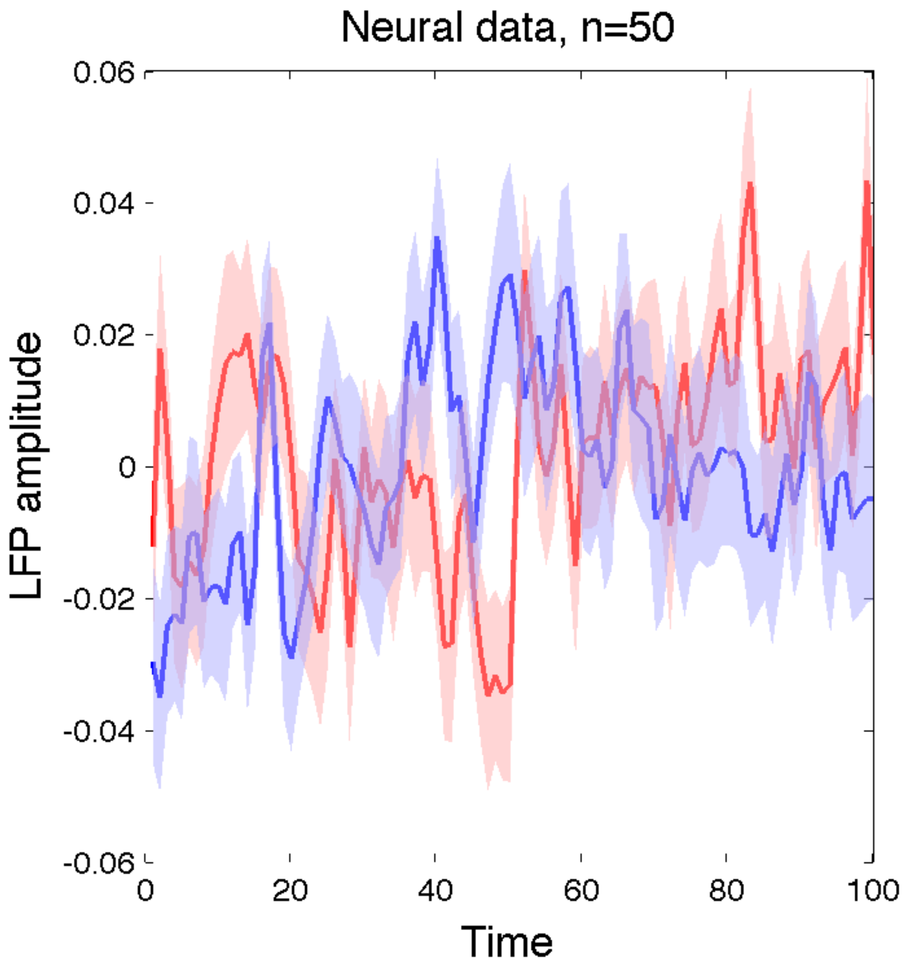
The problem: Do local field potential (LFP) signals change when measured near a spike burst?





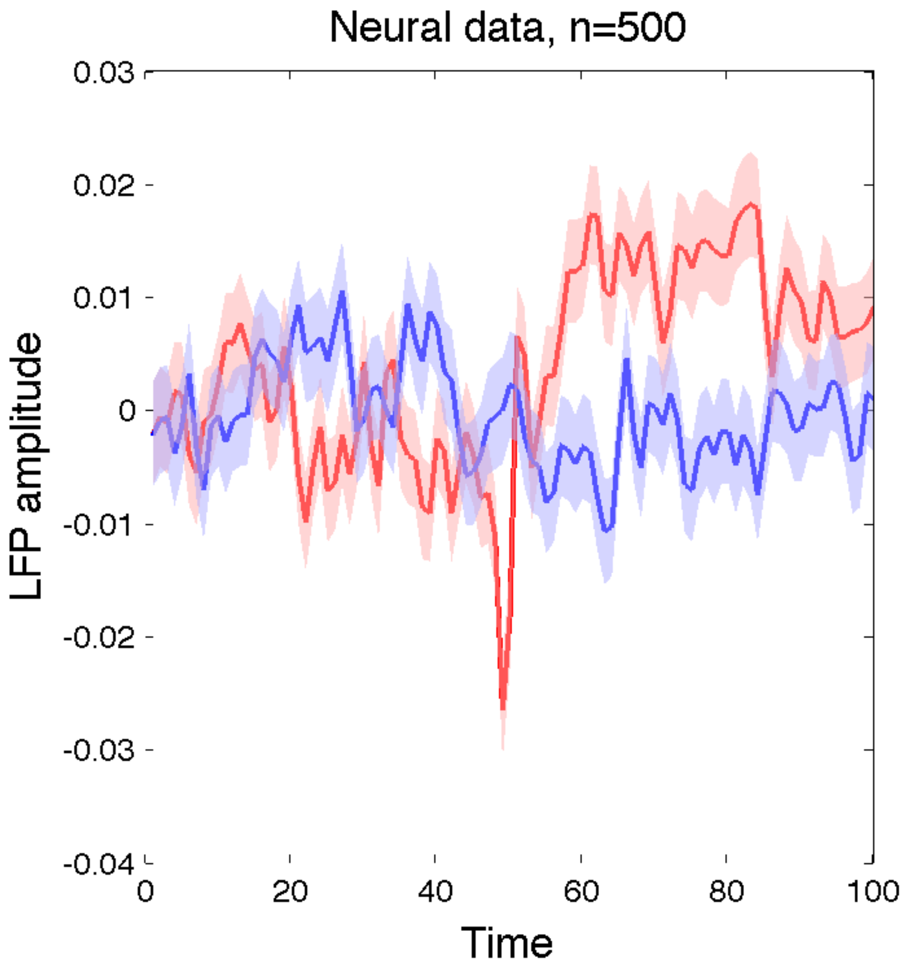
Second motivating question: differences in brain signals

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



Second motivating question: differences in brain signals

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



Overview

- Kernel metric on the space of probability measures: Maximum Mean Discrepancy $MMD(\mathbf{P}, \mathbf{Q})$
  - Distance between means of (nonlinear) features
  - Function revealing differences in distributions
  - Dependence detection: $\mathbf{P}_{xy}$ vs $\mathbf{P}_x\mathbf{P}_y$ using $MMD(\mathbf{P}_{xy}, \mathbf{P}_x\mathbf{P}_y)$

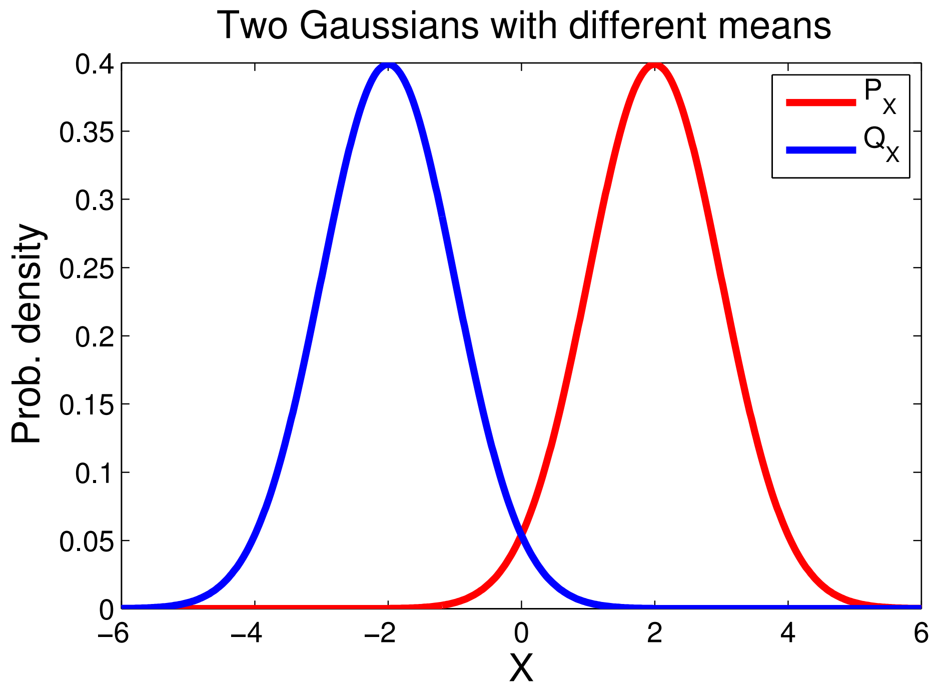
Overview

- Kernel metric on the space of probability measures: Maximum Mean Discrepancy $MMD(\mathbf{P}, \mathbf{Q})$
  - Distance between means of (nonlinear) features
  - Function revealing differences in distributions
  - Dependence detection: $\mathbf{P}_{xy}$ vs $\mathbf{P}_x\mathbf{P}_y$ using $MMD(\mathbf{P}_{xy}, \mathbf{P}_x\mathbf{P}_y)$
- Optimal kernel choice:
  - A criterion for kernel choice
  - What is a difficult testing problem?

Kernel distance between distributions

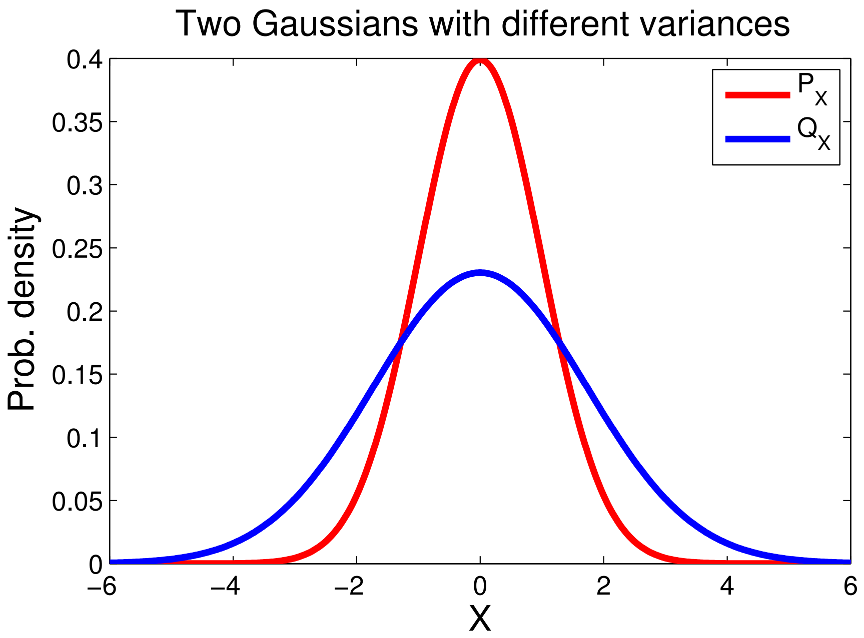
Feature mean difference

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



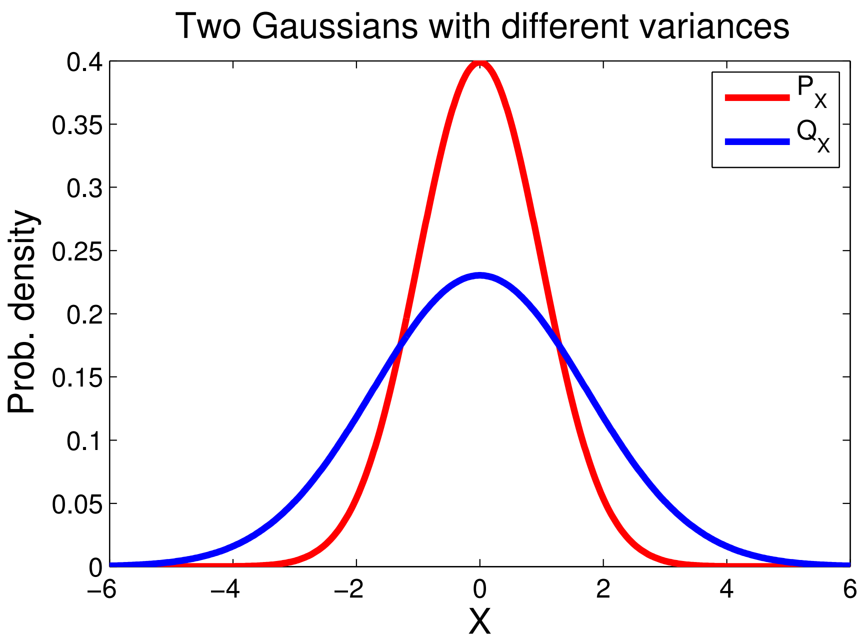
Feature mean difference

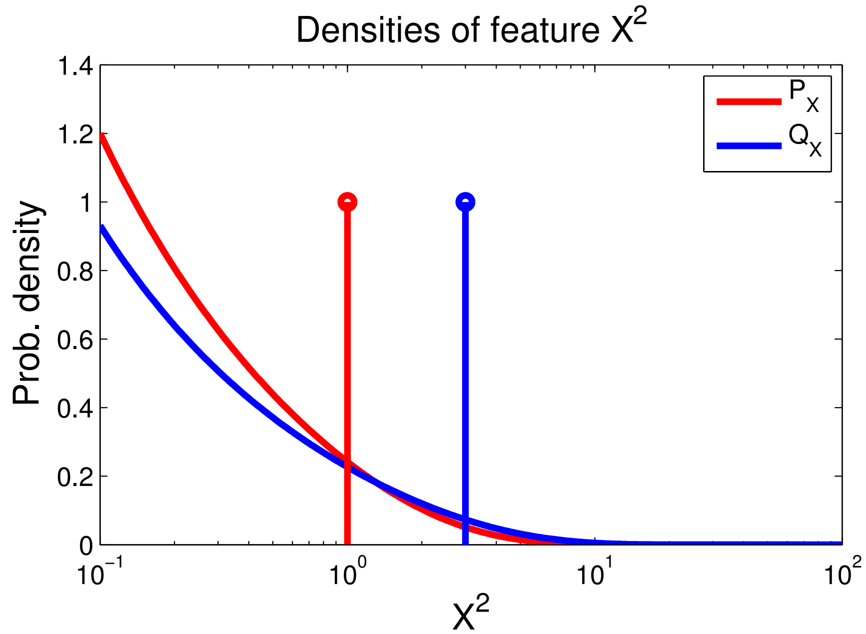
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

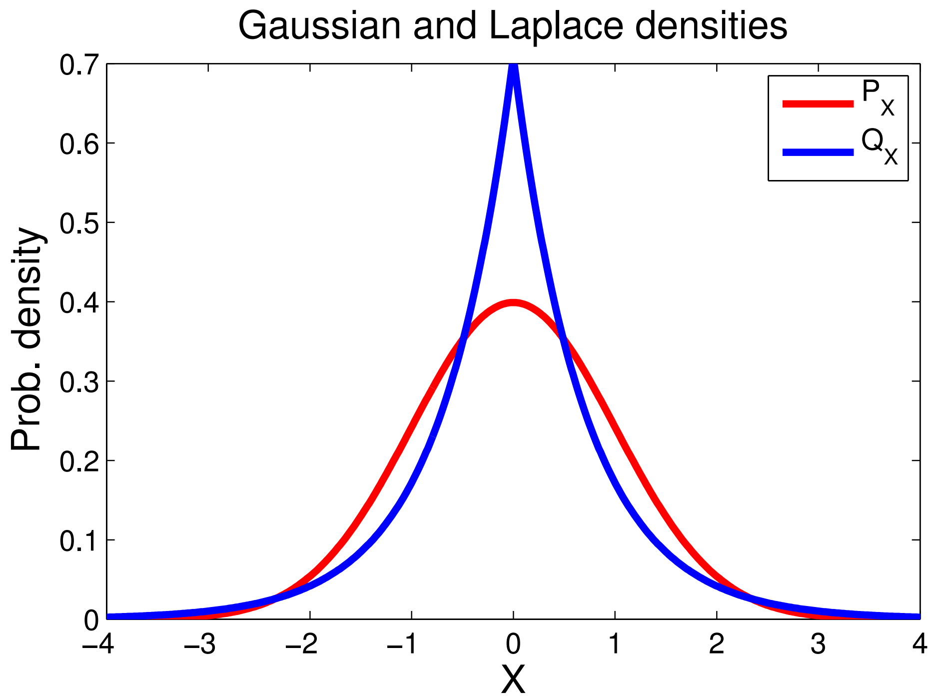
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi_x = x^2$





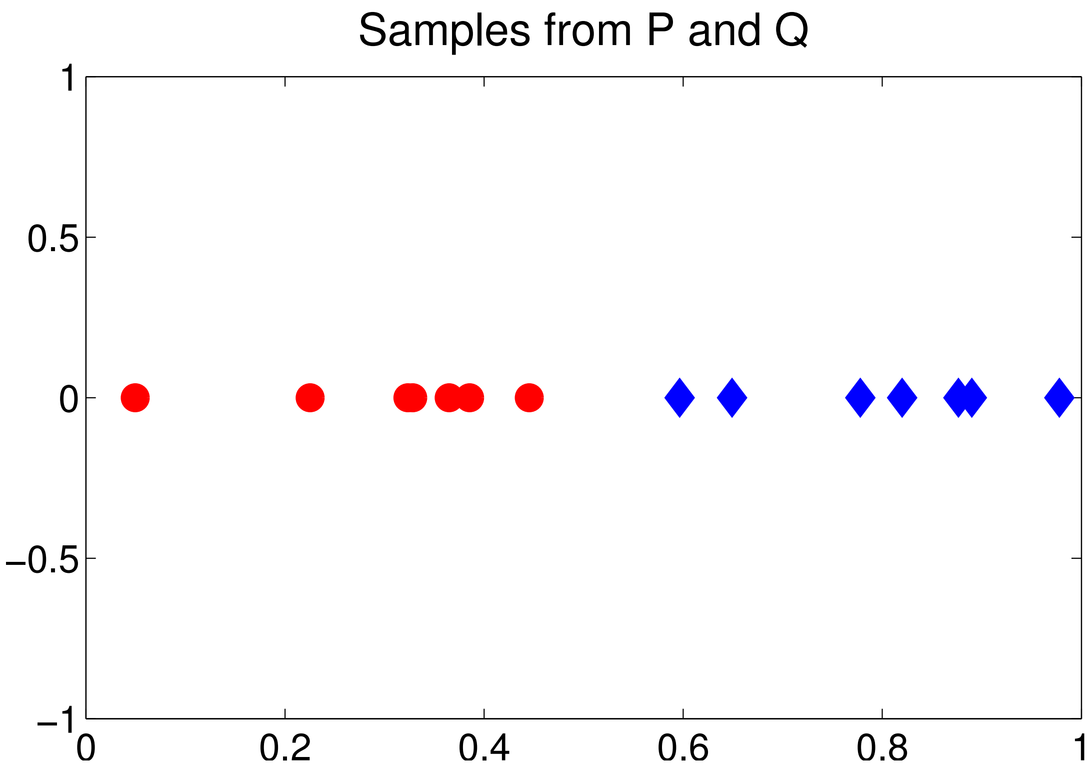
Feature mean difference

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



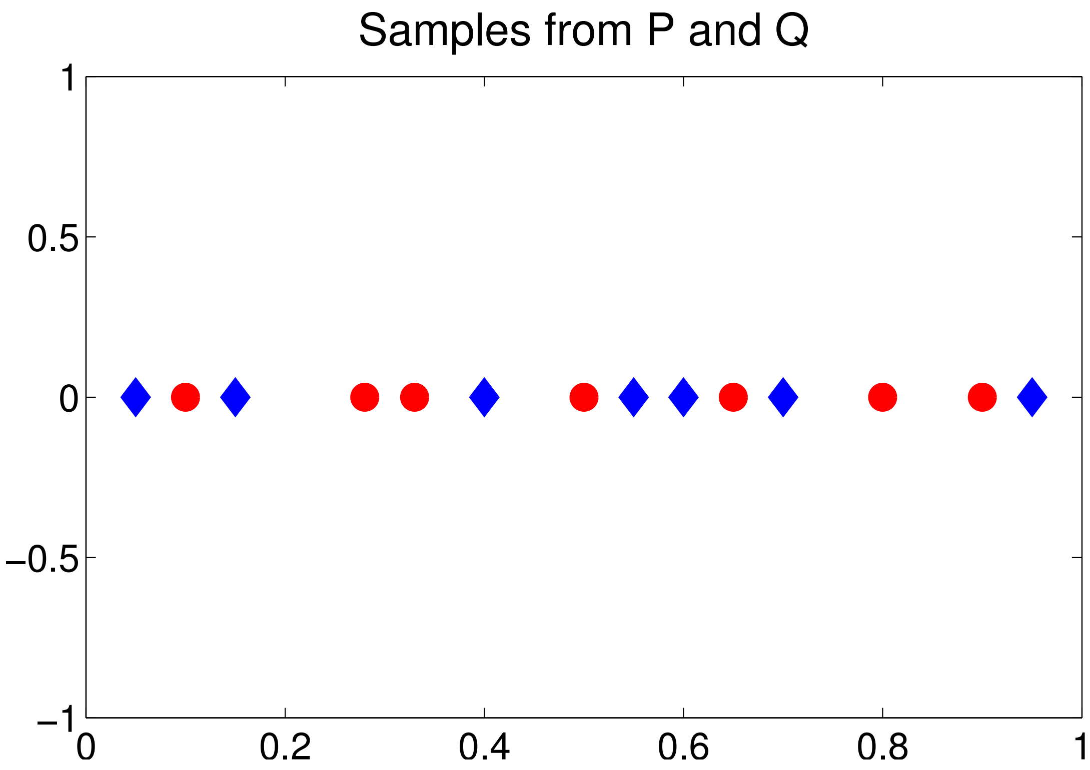
Function Showing Difference in Distributions

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



Function Showing Difference in Distributions

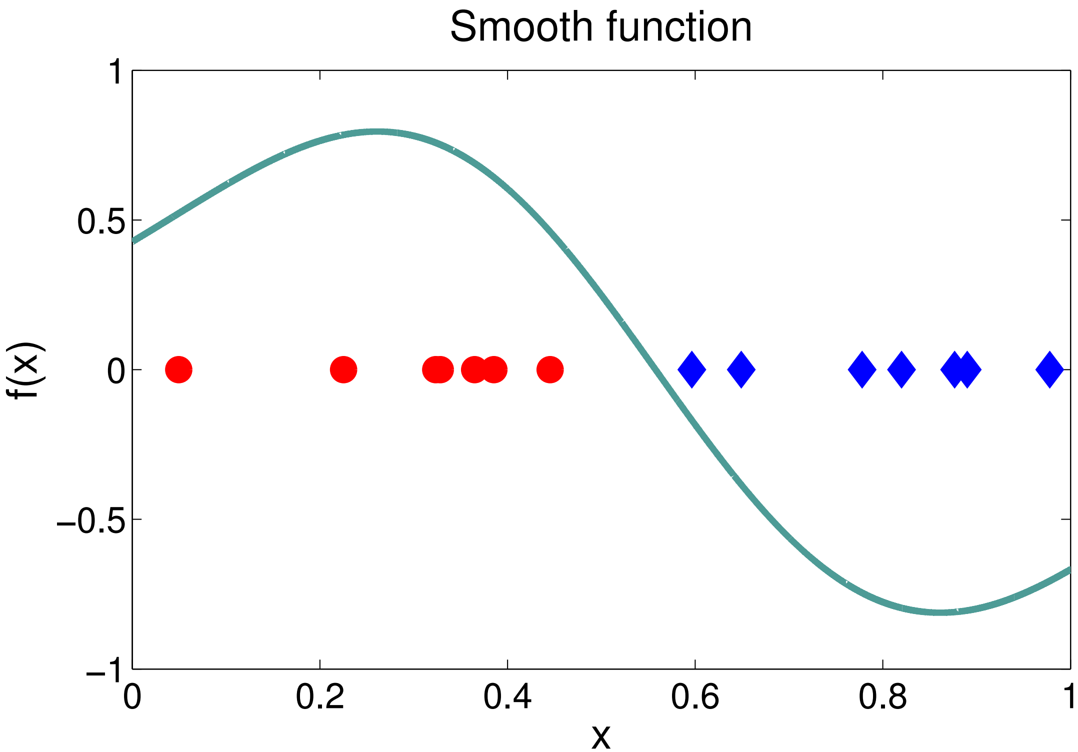
- Are **P** and **Q** different?



Function Showing Difference in Distributions

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

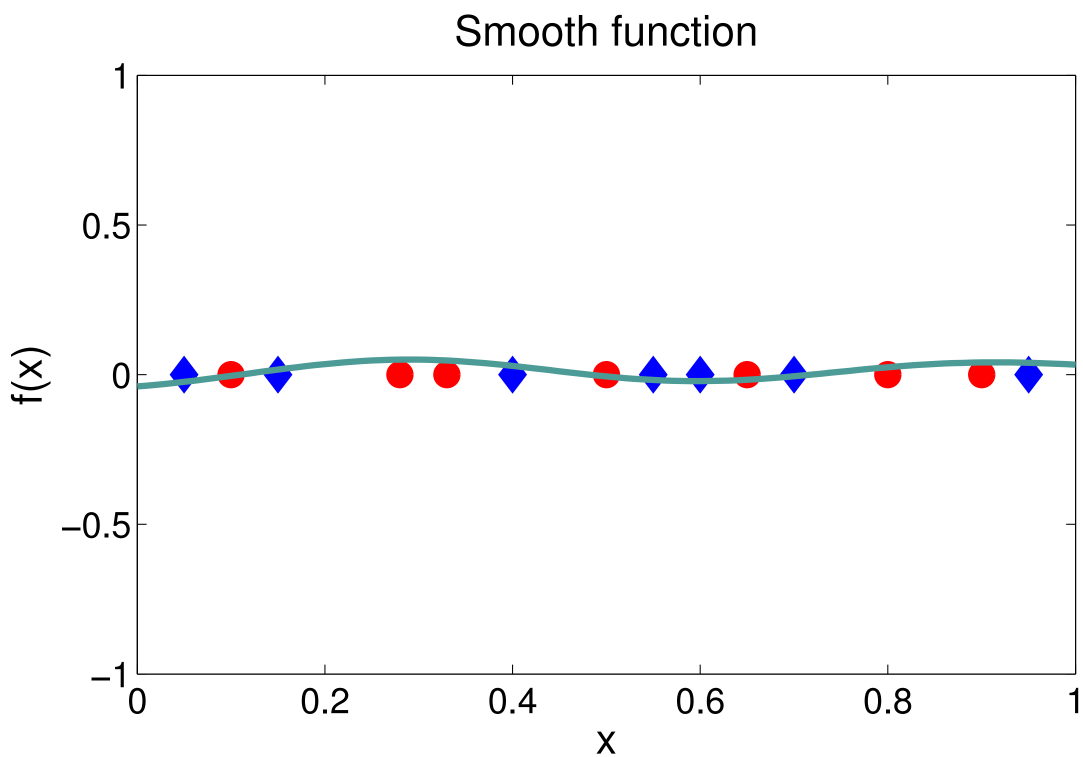
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

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

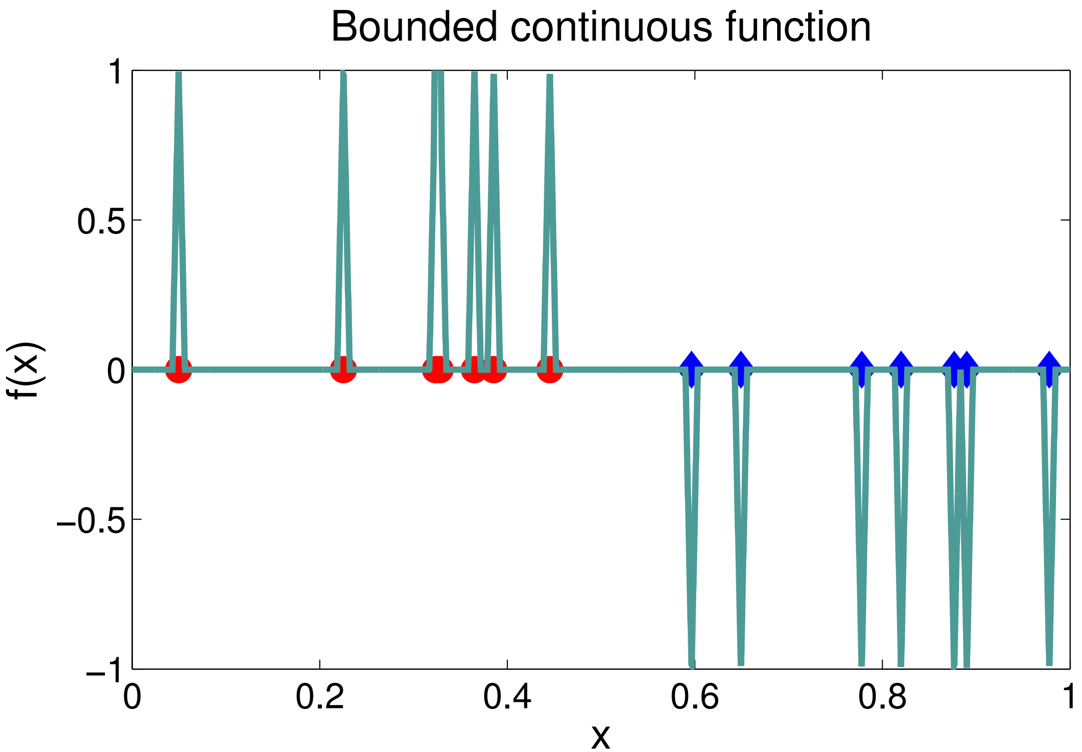
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathrm{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathrm{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

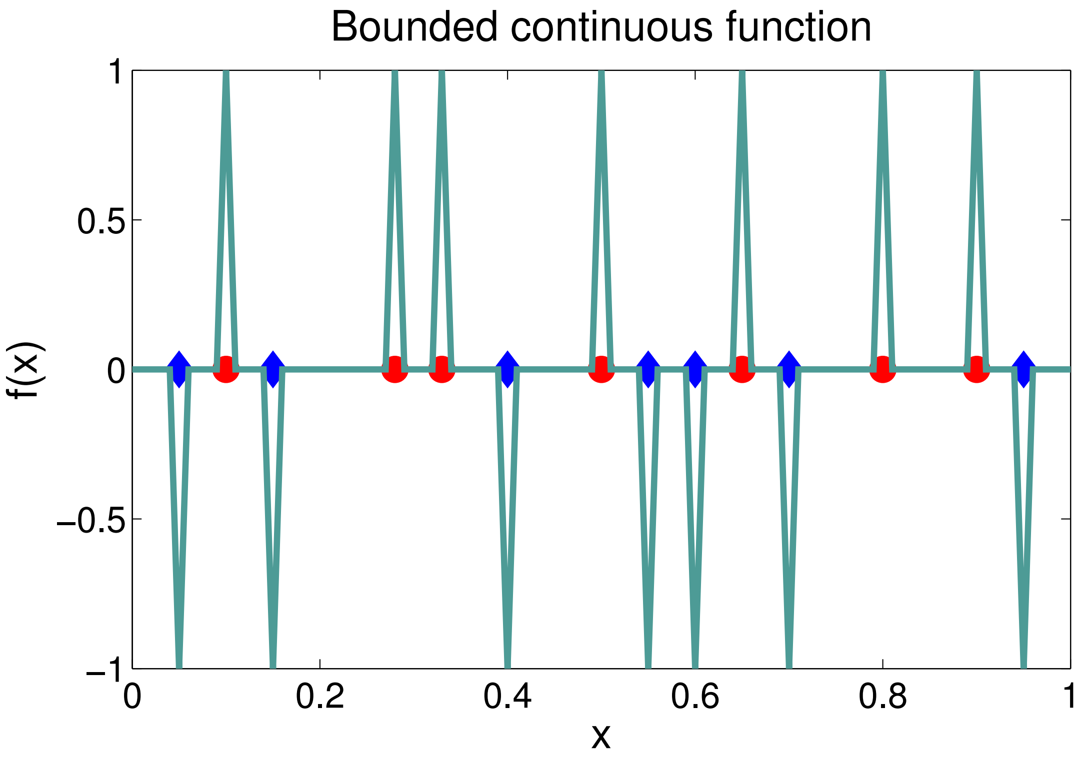
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

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

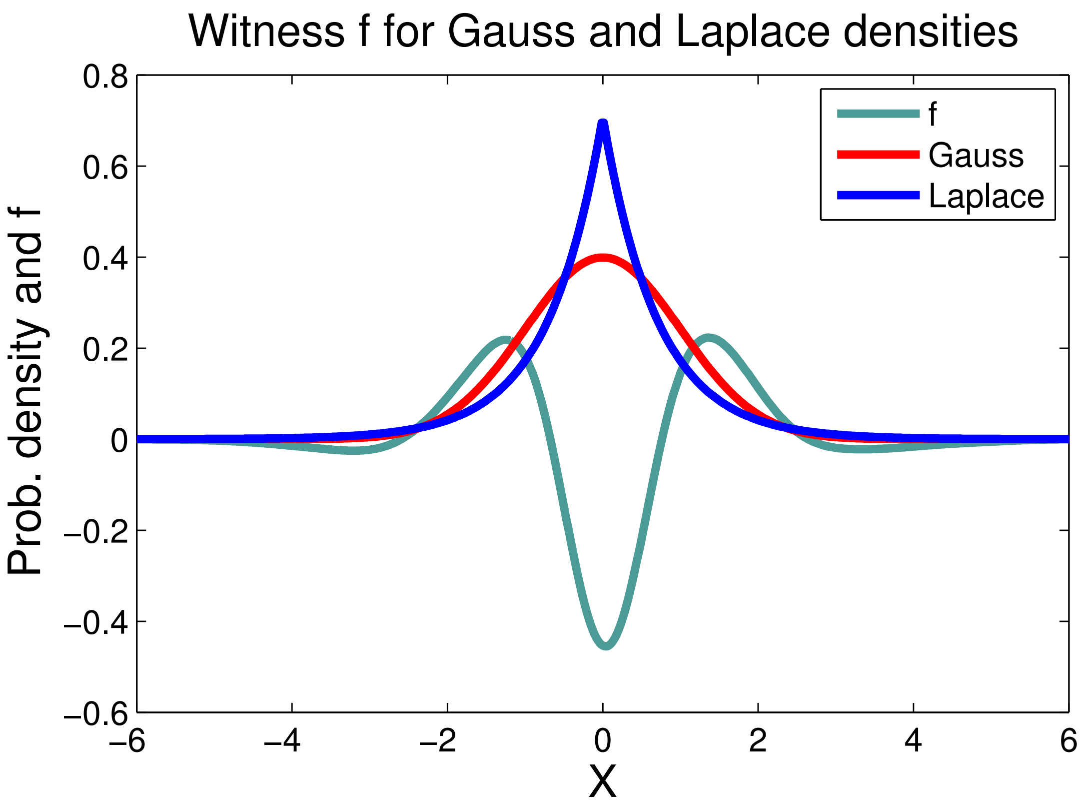


Function Showing Difference in Distributions

- 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 $\mathbf{P}$ 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]

Function Showing Difference in Distributions

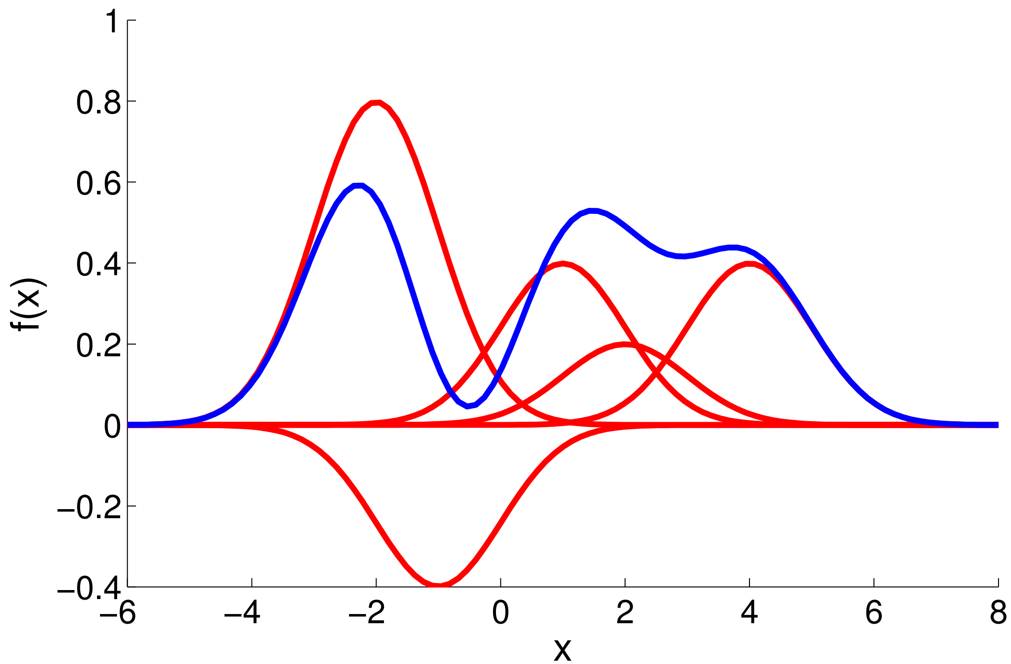
- Maximum mean discrepancy: smooth function for $\mathbf{P}$ vs $\mathbf{Q}$

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

- 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 $\mathcal{F}$ [Gretton et al., 2007, Sriperumbudur et al., 2010, Gretton et al., 2012]

Functions in the RKHS

- $\mathcal{F}$ RKHS from $\mathcal{X}$ to $\mathbb{R}$ with positive definite kernel $k(x_i, x_j)$
- $\mathcal{F} = \overline{\text{span}\{k(x, \cdot) | x \in \mathcal{X}\}}$
  - Example: $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$ for arbitrary $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, $x_i \in \mathcal{X}$.



The RKHS as feature map

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

$$\varphi_x^{(p)} = \left[\begin{array}{ccc} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{array}\right] \qquad \varphi_x^{(g)} = \exp\left(-\lambda \left\| x - \cdot \right\|^2\right)$$

The RKHS as feature map

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

$$\varphi_x^{(p)} = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad \varphi_x^{(g)} = \exp\left(-\lambda \|x - \cdot\|^2\right)$$

- Inner product between feature maps:

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

The RKHS as feature map

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

$$\varphi_x^{(p)} = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 \sqrt{2} \end{bmatrix} \qquad \varphi_x^{(g)} = \exp\left(-\lambda \left\| x - \cdot \right\|^2\right)$$

- Inner product between feature maps:

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

- In general,

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

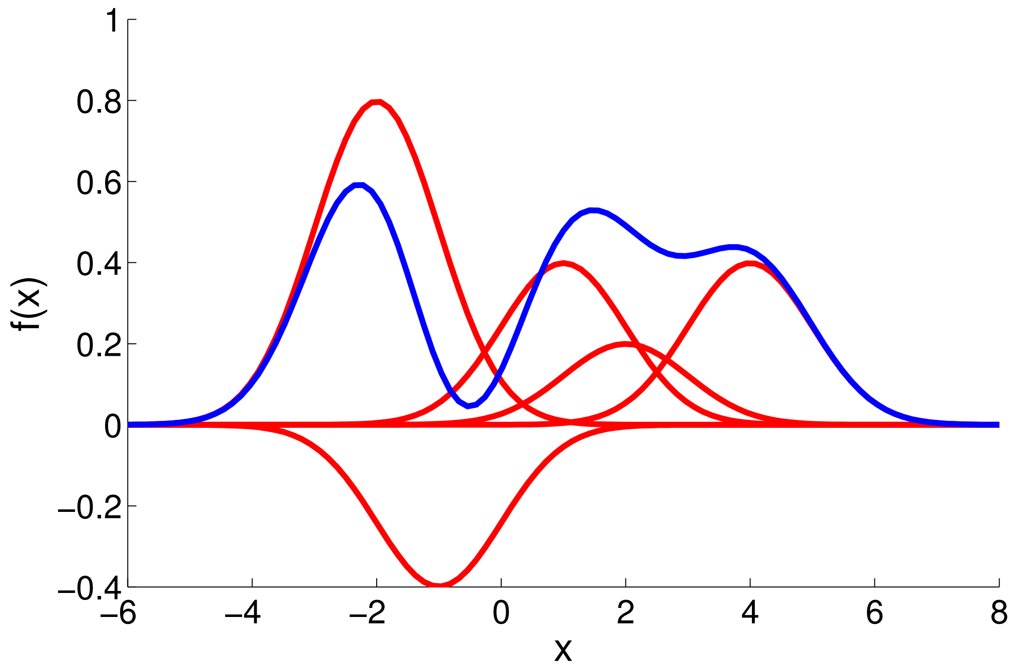
  for positive definite $k(x, y)$

Kernels are inner products of feature maps

The RKHS as feature map

- Function in RKHS:

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

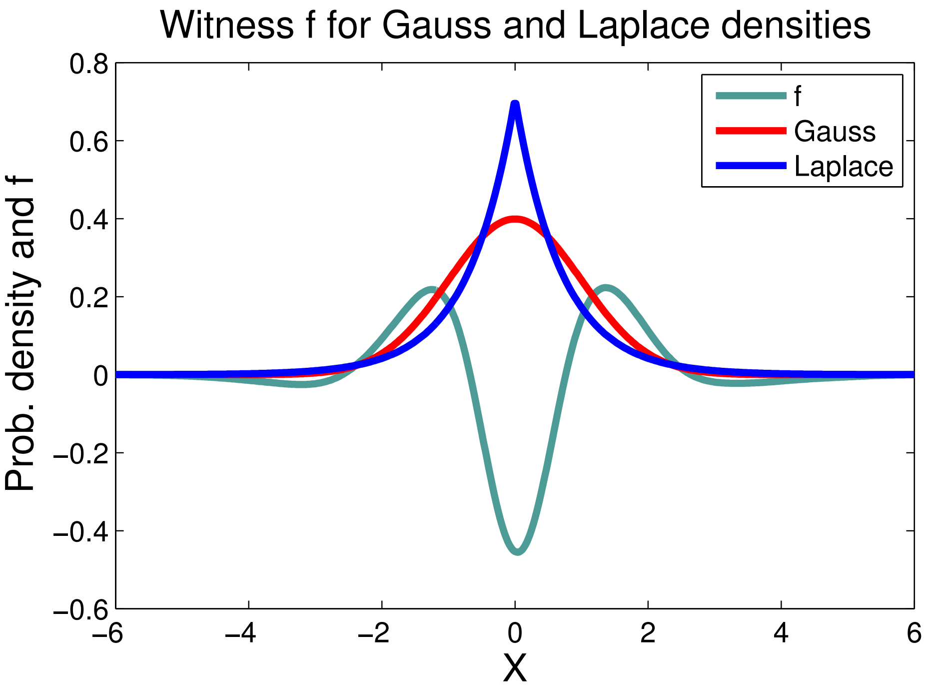


Function view vs feature mean view

- 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]

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

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

use

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

$$=: \langle \mu_{\mathbf{P}}, f \rangle_{\mathcal{F}}$$

Function view vs feature mean view

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

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

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

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

use

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

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

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

Function view and feature view equivalent

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

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

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

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

Statistical hypothesis testing

Statistical test using MMD

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

Statistical test using MMD

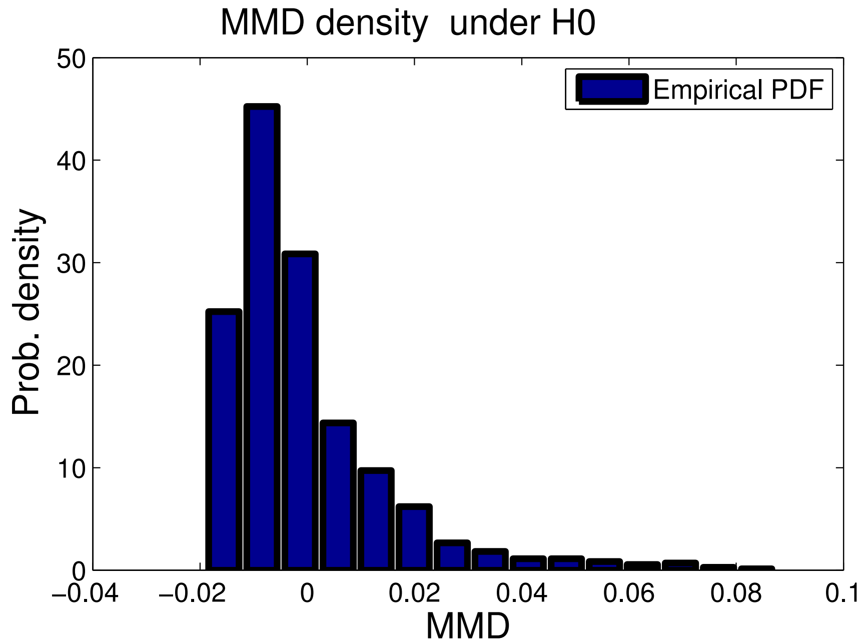
- Two hypotheses:
  - H_0: null hypothesis ($\mathbf{P} = \mathbf{Q}$)
  - H_1: alternative hypothesis ($\mathbf{P} \neq \mathbf{Q}$)
- Observe samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ from $\mathbf{P}$ and $\boldsymbol{y}$ from $\mathbf{Q}$
- If empirical $\widehat{\mathrm{MMD}}^2$ is
  - "far from zero": reject H_0
  - "close to zero": accept H_0

Statistical test using MMD

- When $\mathbf{P} = \mathbf{Q}$, U-statistic degenerate: [Gretton et al., 2007, 2012]
- Distribution is

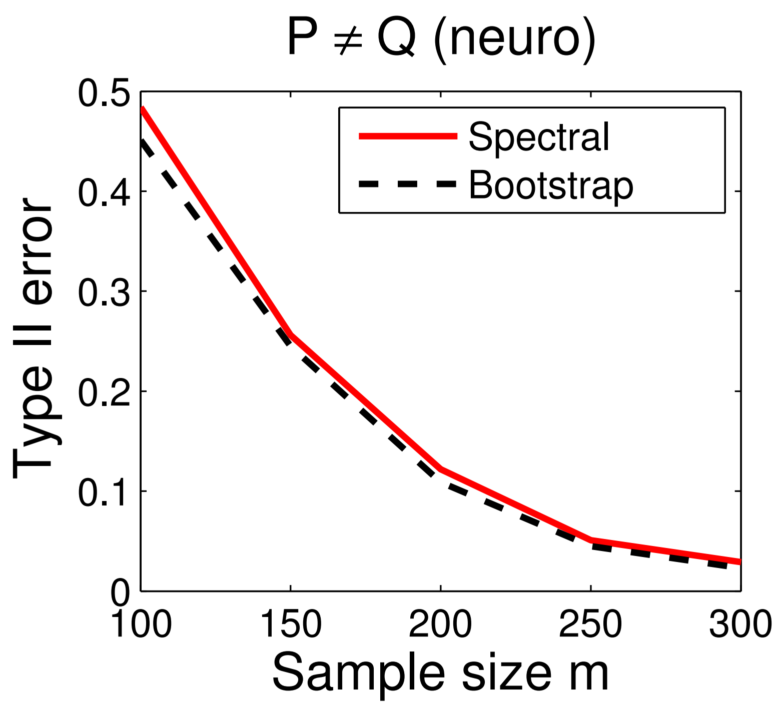
$$m\widehat{\mathrm{MMD}}^2 \sim \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2 \right]$$

- 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

- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\widehat{\text{MMD}}^2 > T) \leq \alpha$
- Bootstrap for empirical CDF [Arcones and Giné, 1992]
- Pearson curves by matching first four moments [Johnson et al., 1994]
- Large deviation bounds [Hoeffding, 1963, McDiarmid, 1989]
- Consistent test using kernel eigenspectrum [Gretton et al., 2009]

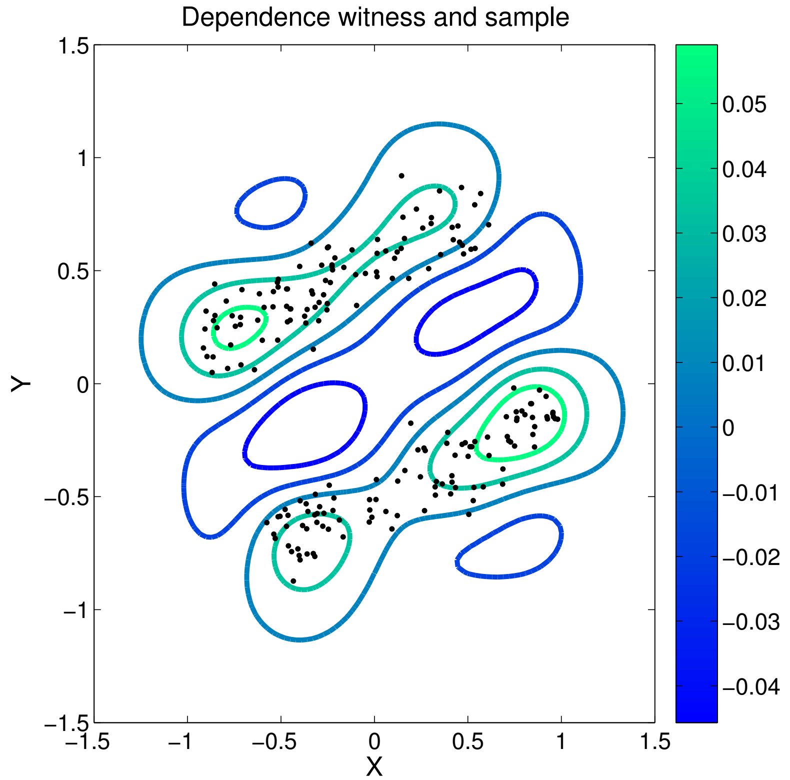


MMD for independence

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

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

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

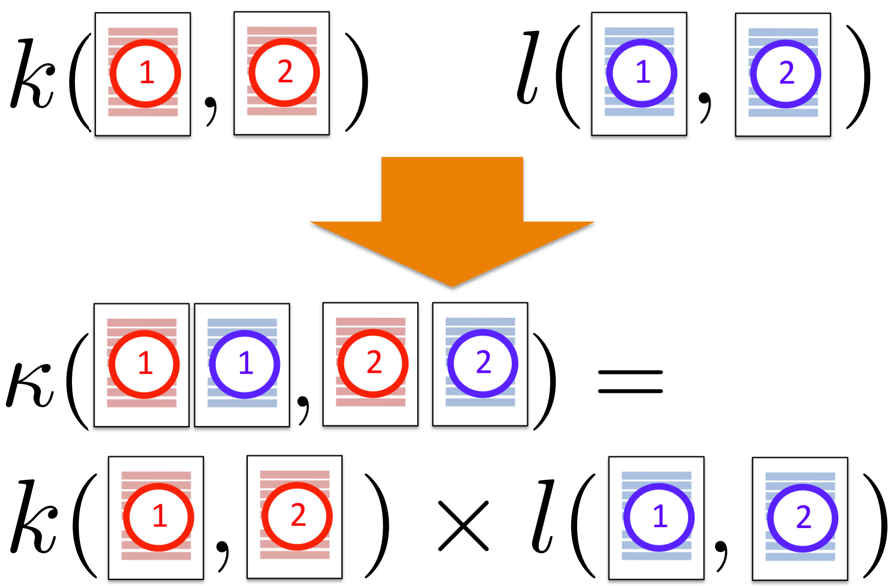


MMD for independence

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

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

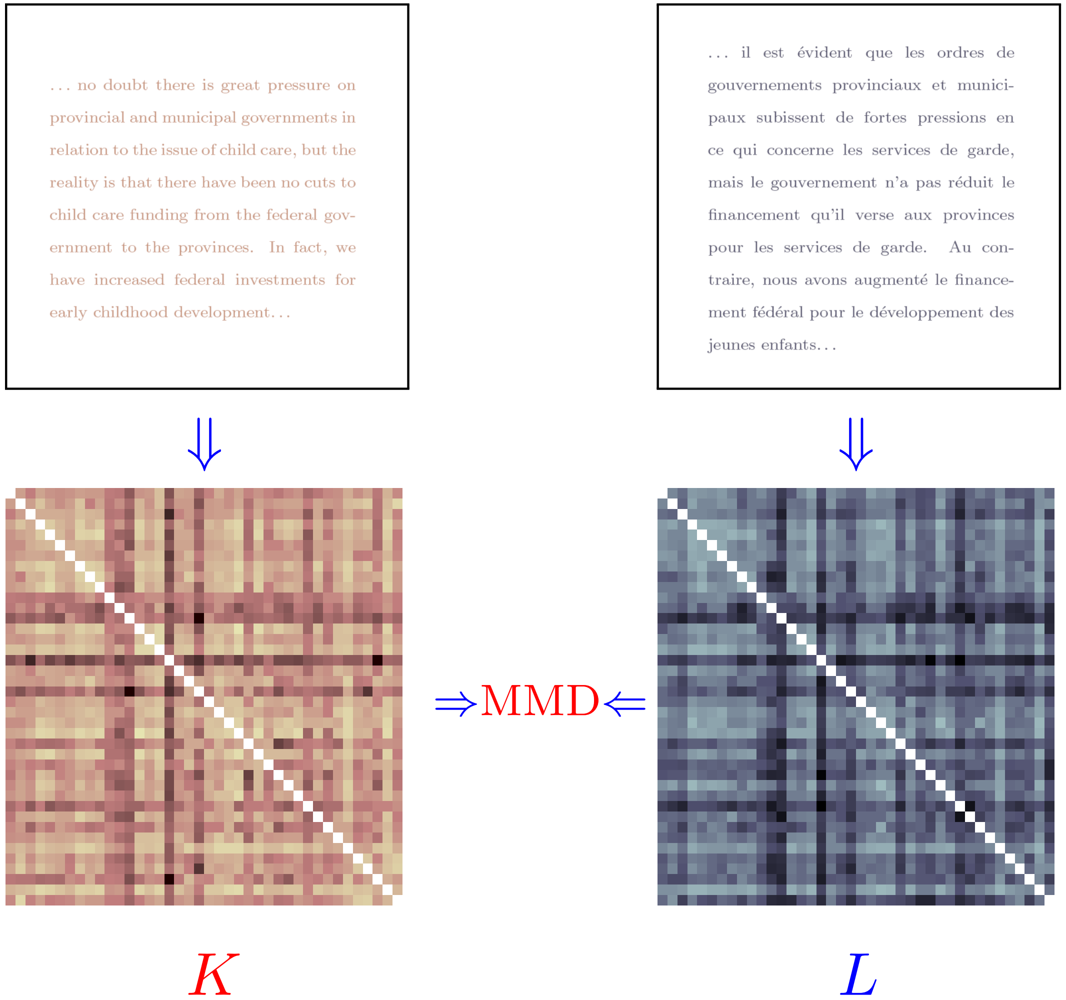
$$= \|\mu_{\mathsf{P}_{XY}} - \mu_{\mathsf{P}_X \mathsf{P}_Y}\|^2_{\mathcal{F} \times \mathcal{G}} := MMD(\mathsf{P}_{XY}, \mathsf{P}_X \mathsf{P}_Y)$$



Experiment: dependence testing for translation

- Translation example: [NIPS07b] Canadian Hansard (agriculture)
- 5-line extracts, k-spectrum kernel, $k = 10$, repetitions=300, sample size 10
- Empirical $MMD(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y)$:

$$\frac{1}{m^2}\text{trace}(\mathbf{KHLH})$$



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

Part 2: optimal kernel choice for two-sample tests

Empirical estimate of MMD: more detail

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

Empirical estimate of MMD: more detail

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

MMD in terms of kernels:

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

Empirical estimate of MMD: more detail

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

MMD in terms of kernels:

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

Empirical estimate of MMD: more detail

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

MMD in terms of kernels:

$$\begin{aligned}
\mathrm{MMD}^2 = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|^2_{\mathcal{F}} \quad &= \quad \langle \mu_{\mathsf{P}} - \mu_{\mathsf{Q}}, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} \\
&= \quad \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle \\
&= \quad \langle \mathbf{E}_{\mathsf{P}} \varphi_x, \mathbf{E}_{\mathsf{P}} \varphi_x \rangle + \ldots
\end{aligned}$$

Empirical estimate of MMD: more detail

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

MMD in terms of kernels:

$$\begin{aligned}
\mathrm{MMD}^2 = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|_{\mathcal{F}}^2 \quad &= \quad \langle \mu_{\mathsf{P}} - \mu_{\mathsf{Q}}, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} \\
&= \quad \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle \\
&= \quad \langle \mathbf{E}_{\mathsf{P}} \varphi_x, \mathbf{E}_{\mathsf{P}} \varphi_x \rangle + \ldots \\
&= \quad \mathbf{E}_{\mathsf{P}} \langle \varphi_x, \varphi_{x'} \rangle + \ldots
\end{aligned}$$

Empirical estimate of MMD: more detail

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

MMD in terms of kernels:

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

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

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

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)

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

Quadratic time estimate of MMD

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

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)

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

New, linear time estimate:

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

Linear time MMD

Shorter expression with explicit k dependence:

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

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:

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

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,

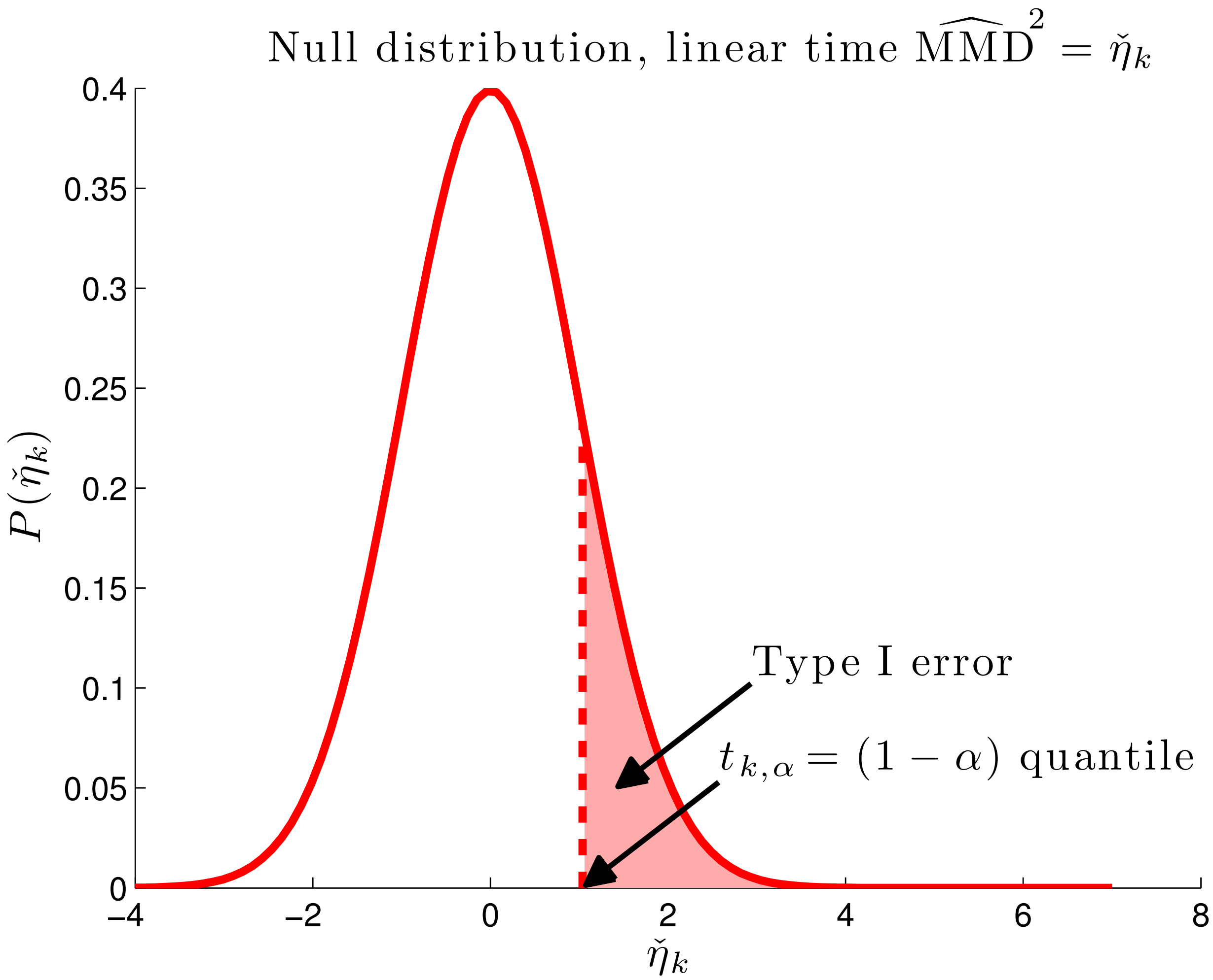
$$m^{1/2}\left(\check{\eta}_k - \eta_k(p,q)\right) \xrightarrow{D} \mathcal{N}(0, 2\sigma_k^2)$$

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

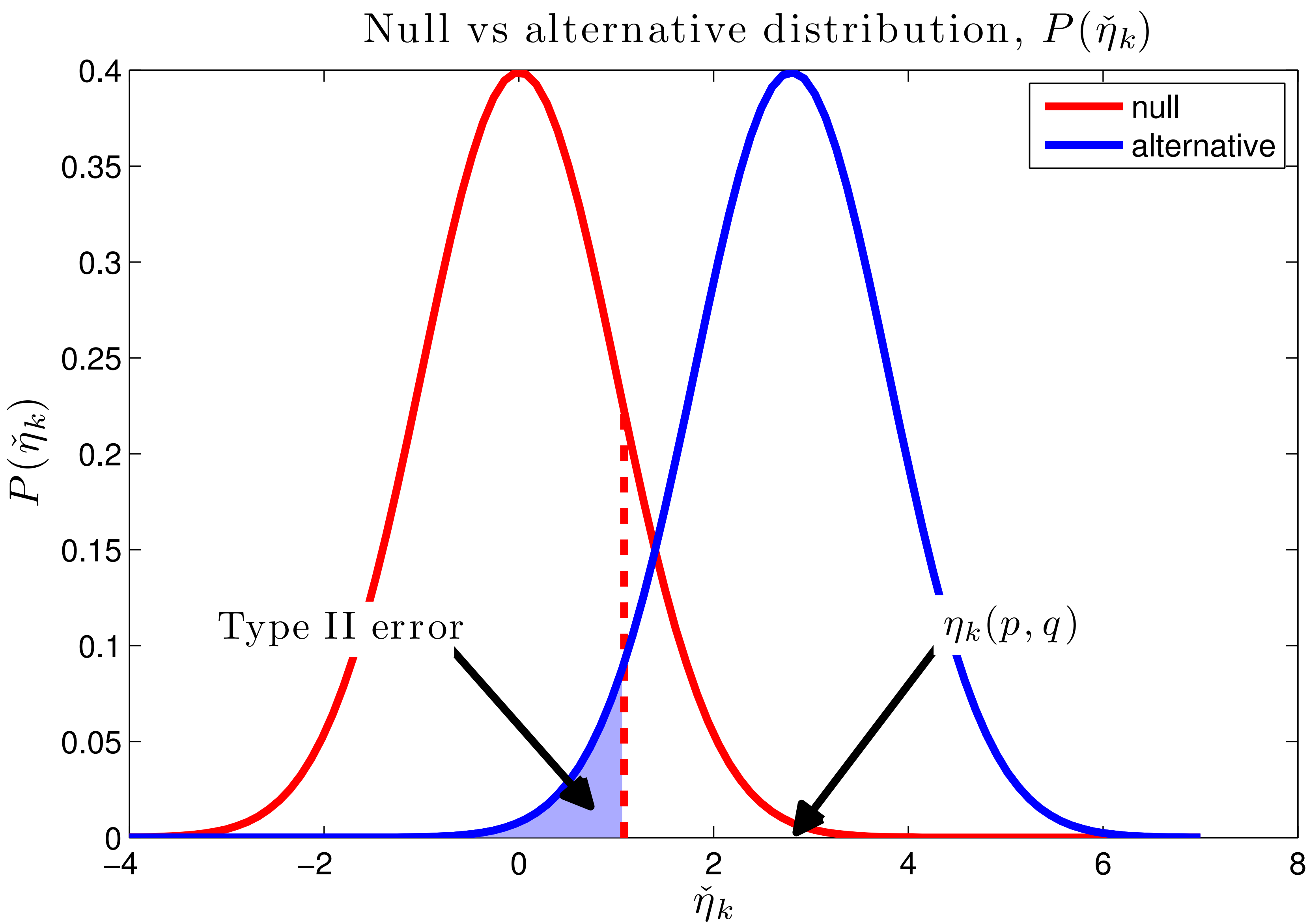
Hypothesis test

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

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

Denote:

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

Quantities for test:

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

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 $\mathrm{cov}(h)$.

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?)

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

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:

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

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

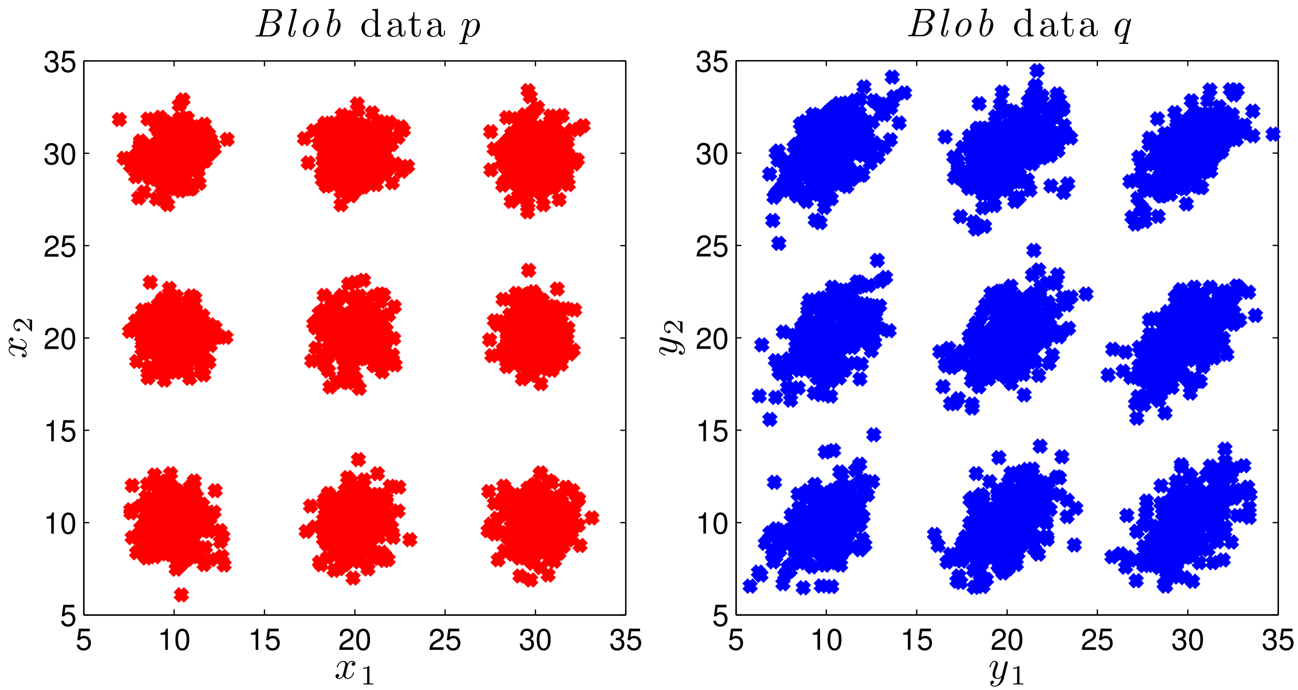
Blobs: data

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

Blobs: data

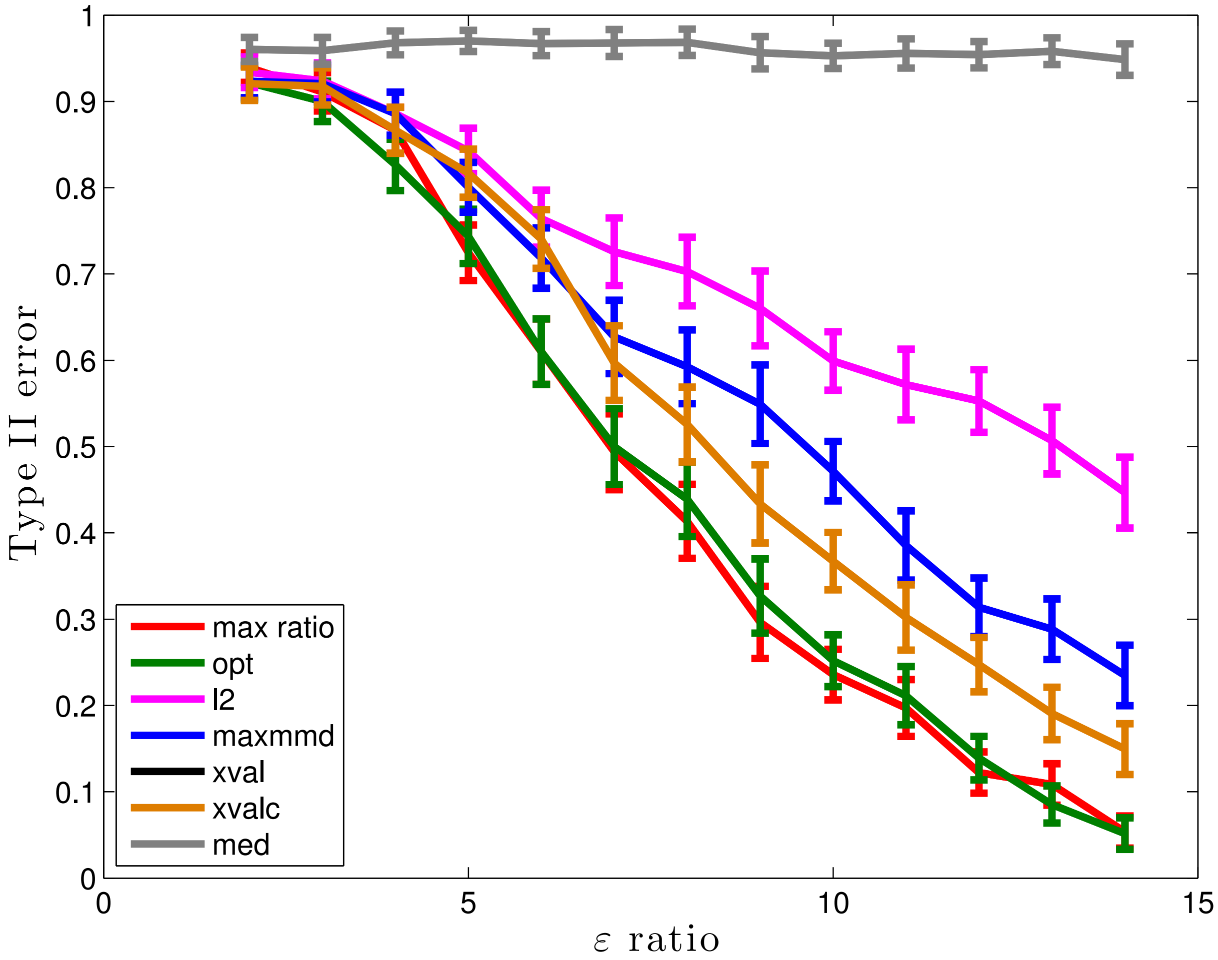
Difficult problems: lengthscale of the *difference* in distributions not the same as that of the distributions.

We distinguish a field of Gaussian blobs with different covariances.



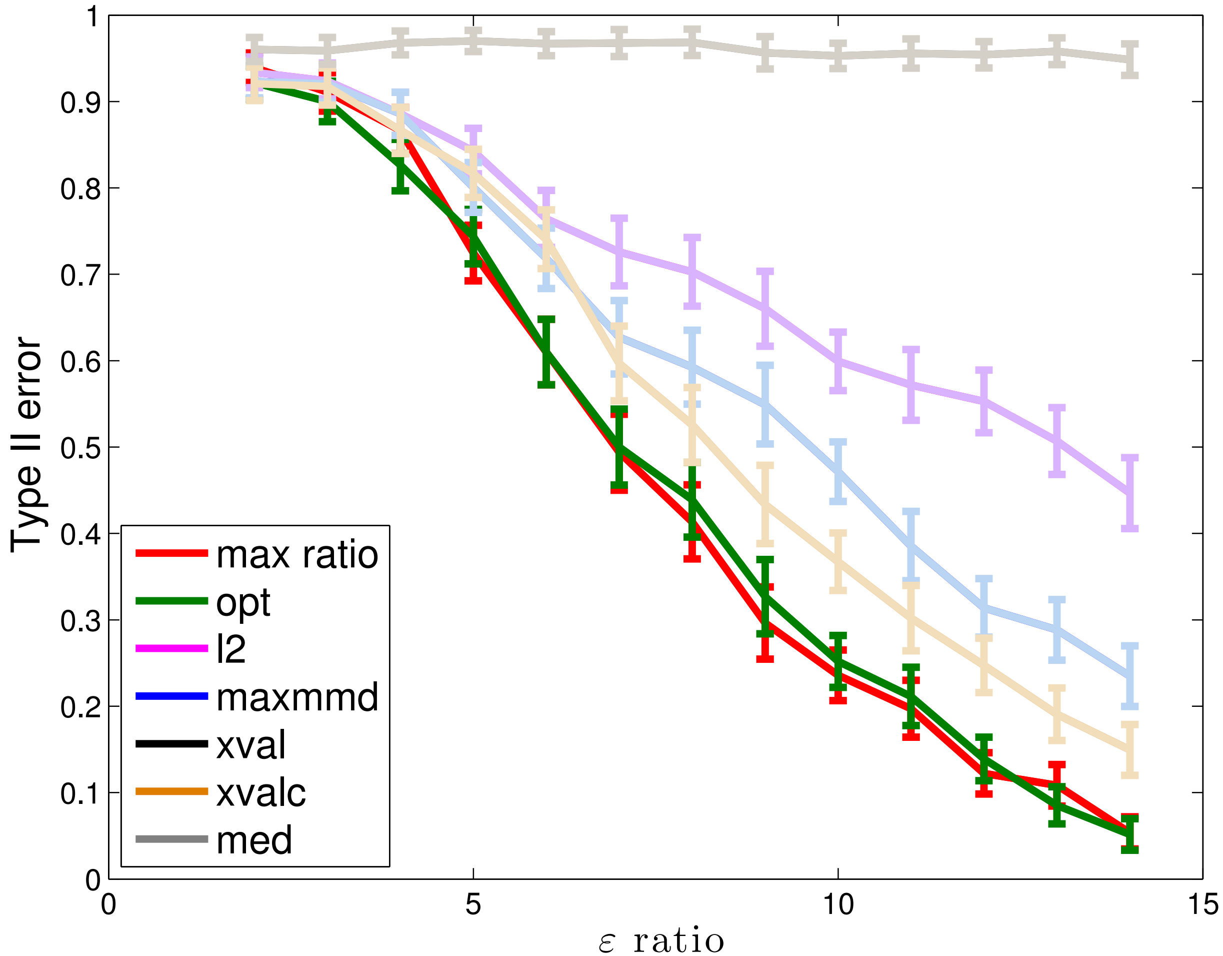
Ratio $\varepsilon = 3.2$ of largest to smallest eigenvalues of blobs in q.

Blobs: results



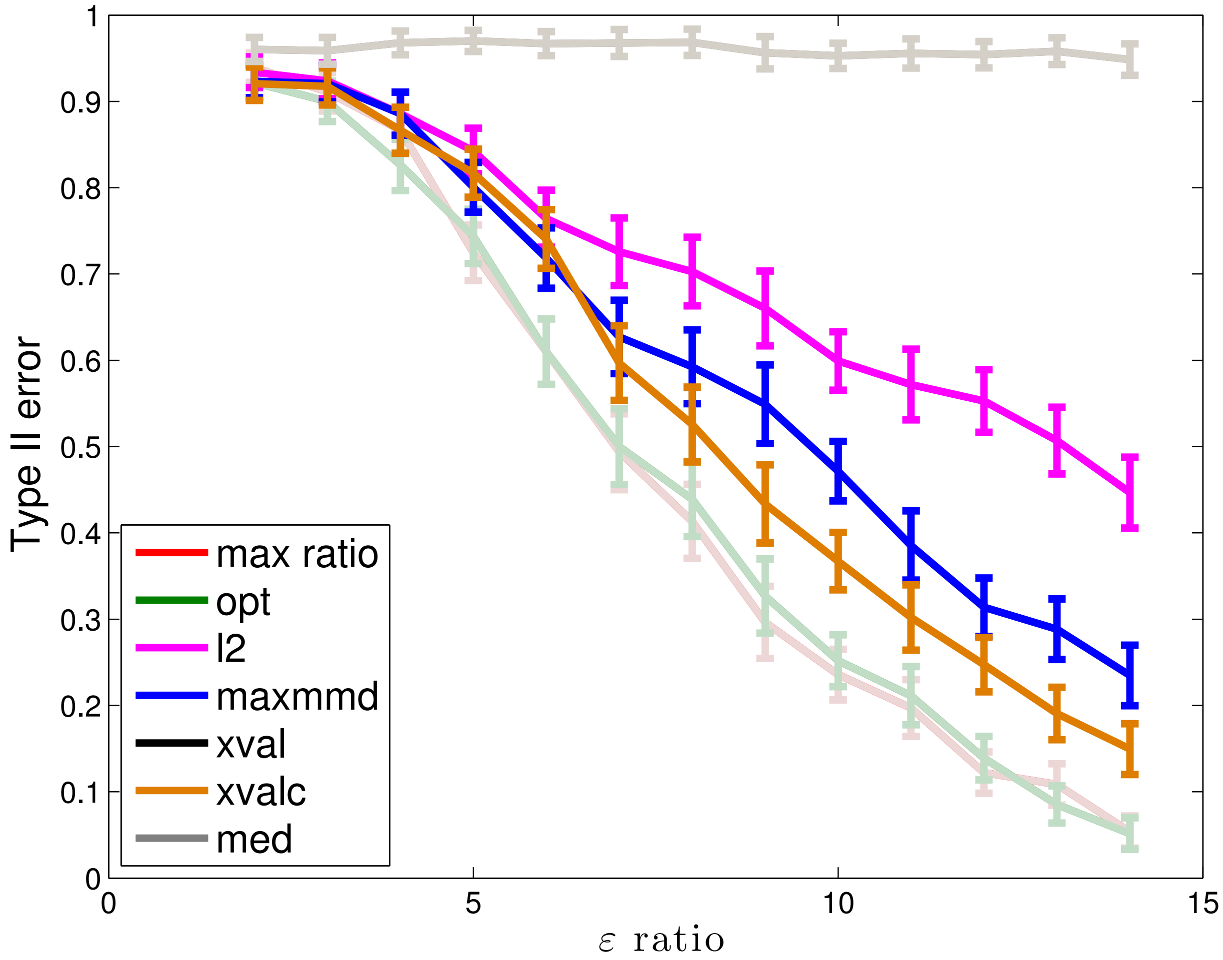
Parameters: $m = 10,000$ (for training and test). Ratio ε of largest to smallest eigenvalues of blobs in q. Results are average over 617 trials.

Blobs: results



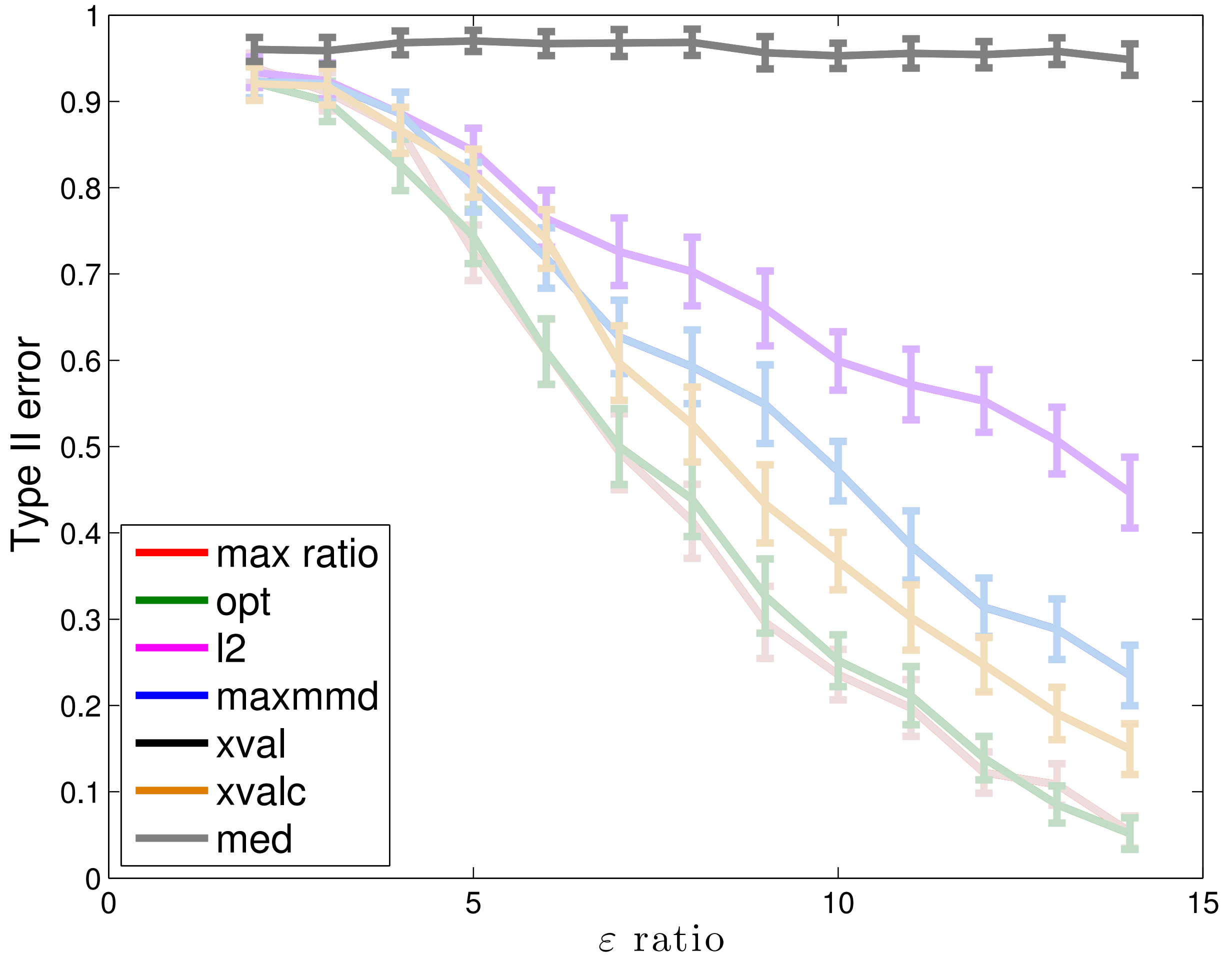
Optimize ratio $\eta_k(p,q)\sigma_k^{-1}$

Blobs: results



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

Blobs: results



Median heuristic

Feature selection: data

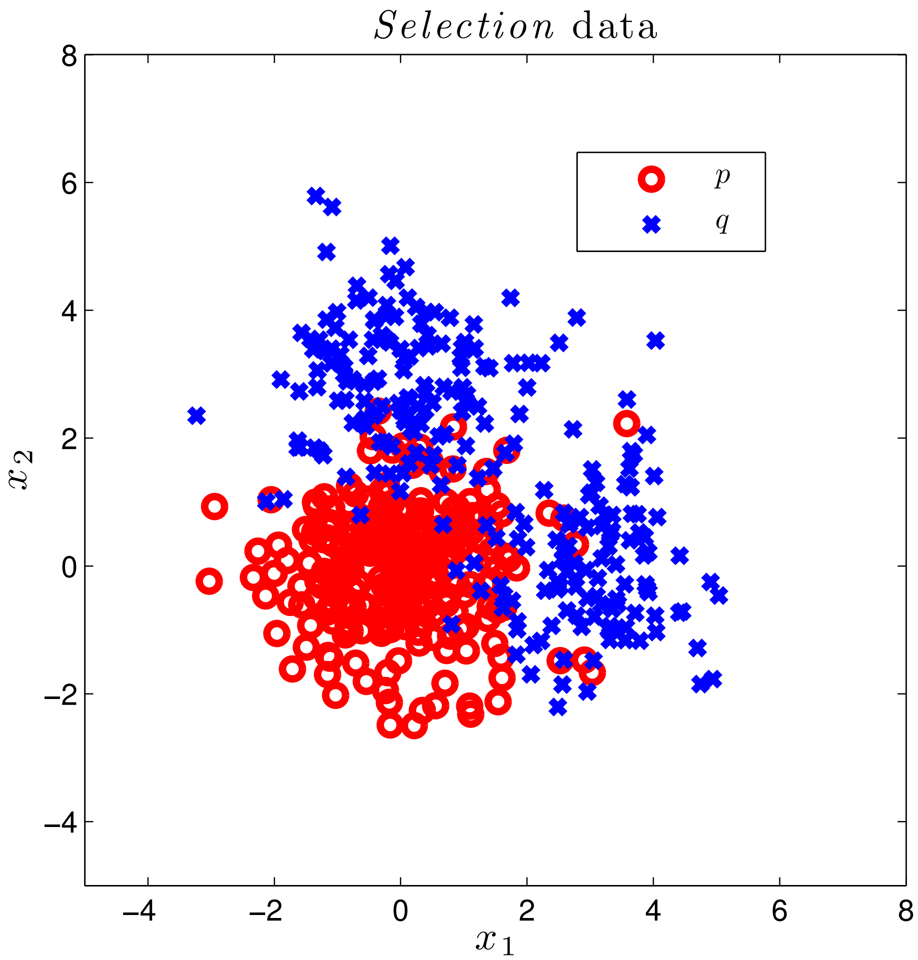
Idea: no single best kernel.

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

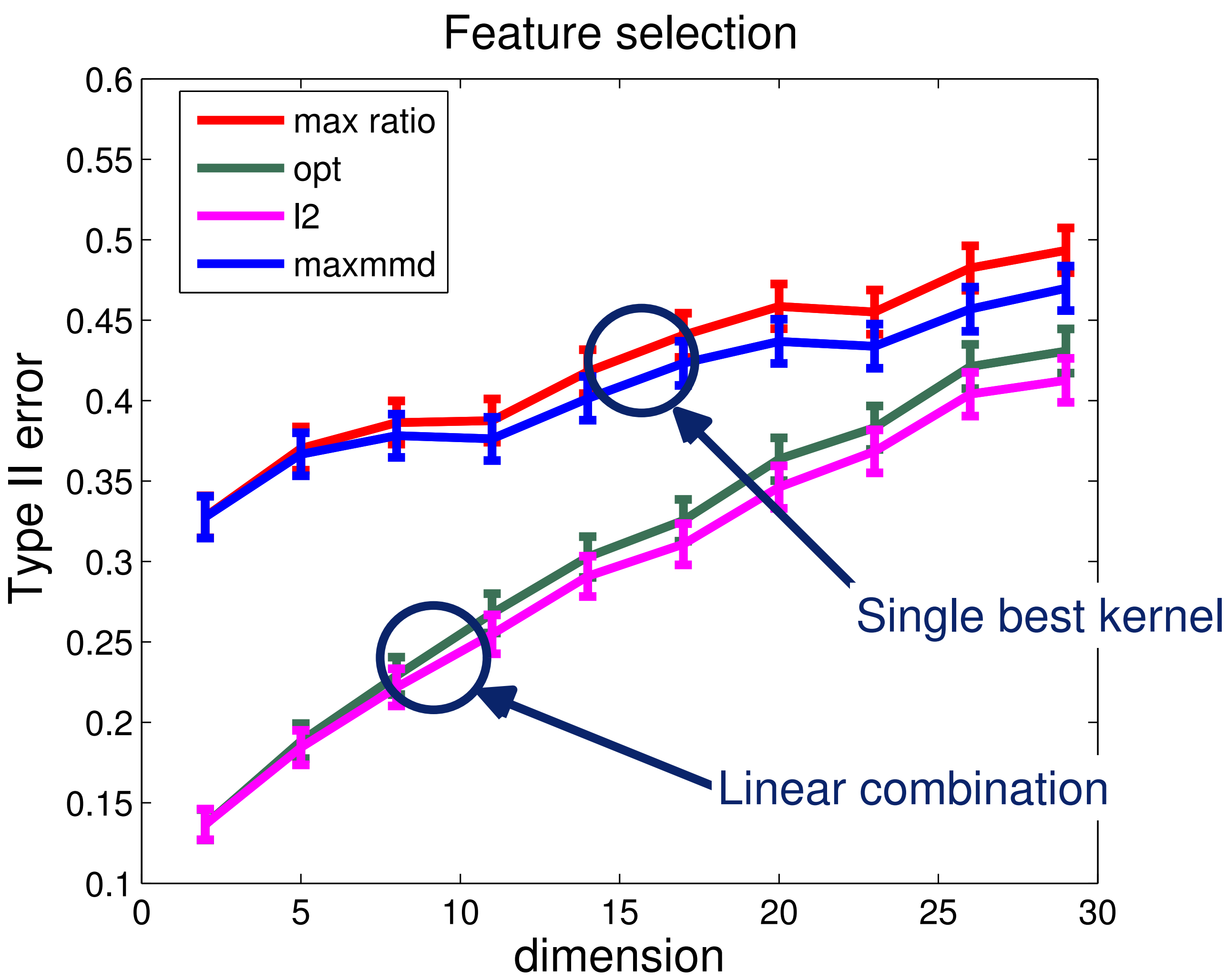
Feature selection: data

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

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

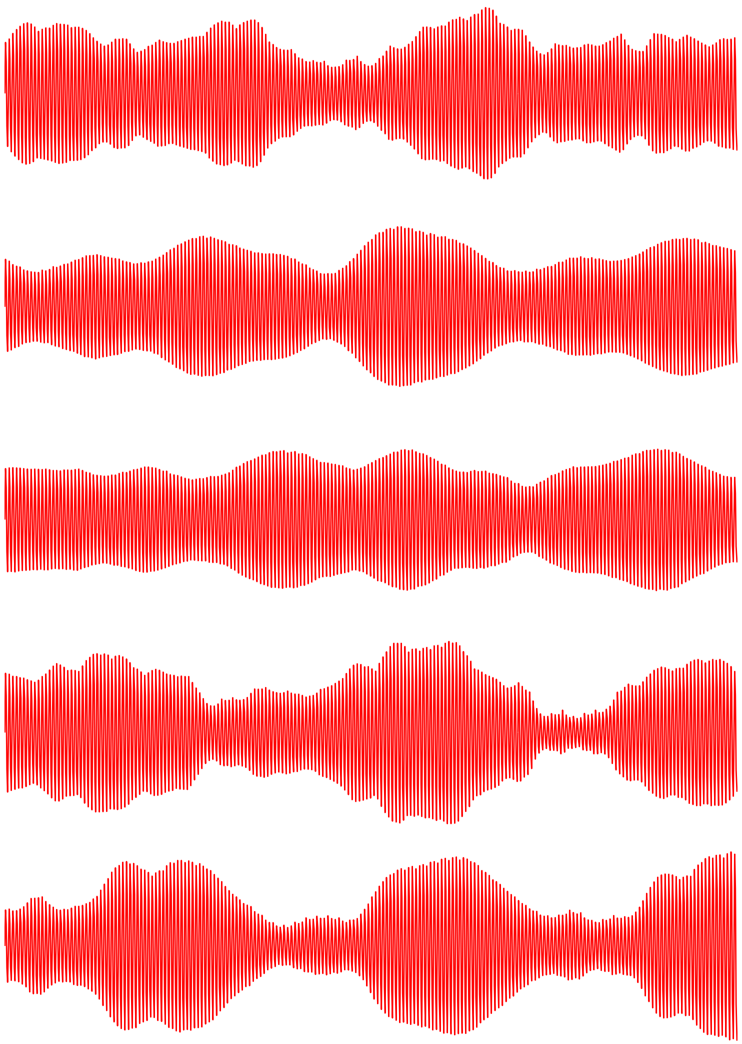
- ω_c: carrier frequency
- $a = 0.2$ is signal scaling, $l = 2$ is offset

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

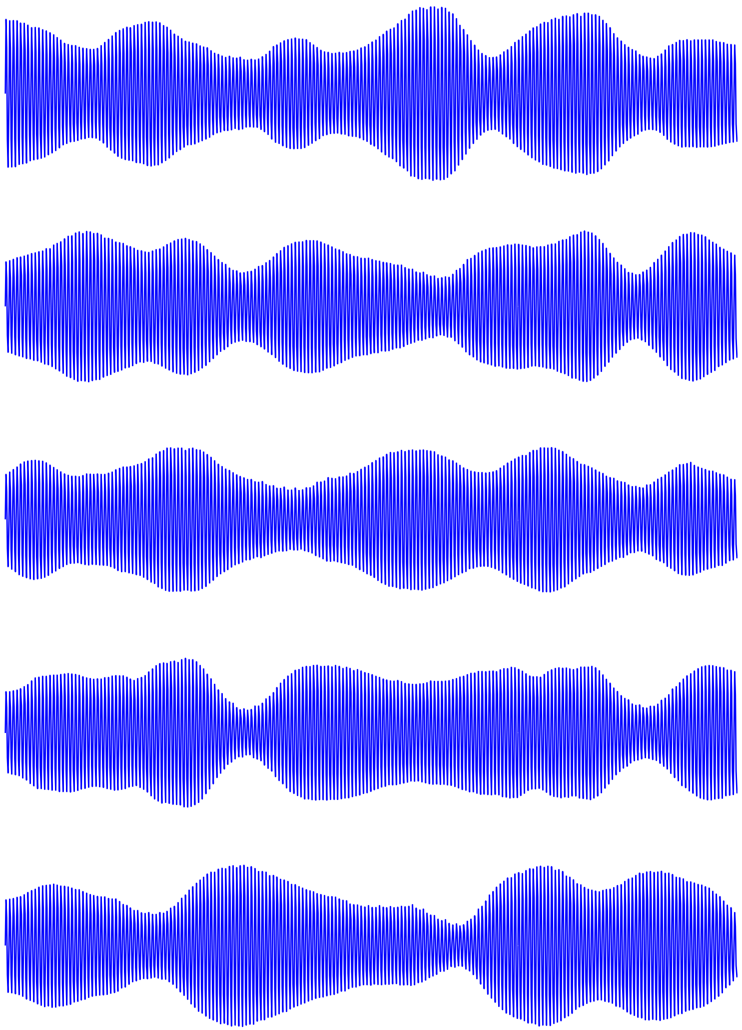
- Music sampled at 8KHz (very low)
- Carrier frequency is 24kHz
- AM signal observed at 120kHz
- Samples are extracts of length $N = 1000$, approx. 0.01 sec (very short).
- Total dataset size is 30,000 samples from each of p, q.

Amplitude modulated signals

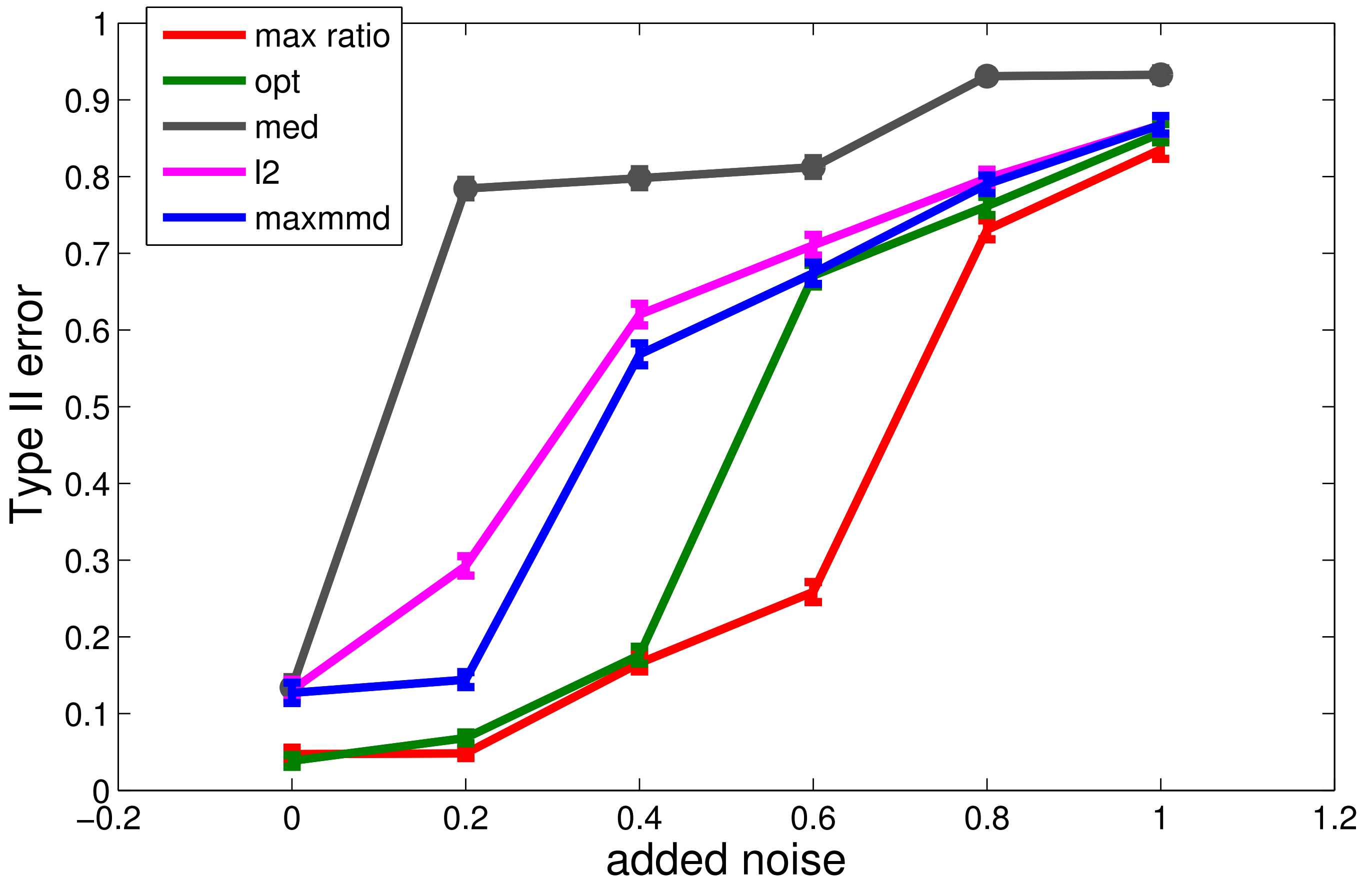
Samples from P



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.

Conclusions

- 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

Co-authors

- **From Gatsby/UCL:**
  - Bharath Sriperumbudur
  - Dino Sejdinovic
  - Heiko Strathmann
  - Massimiliano Pontil
- **External:**
  - Sivaraman Balakrishnan, CMU
  - Kenji Fukumizu, ISM

Empirical estimate of MMD

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

Empirical estimate of MMD

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

MMD in terms of kernels:

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

Empirical estimate of MMD

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

MMD in terms of kernels:

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

Empirical estimate of MMD

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

MMD in terms of kernels:

$$\begin{aligned} \mathrm{MMD}^2 = \|\mu_{\mathsf{P}} - \mu_{\mathsf{Q}}\|_{\mathcal{F}}^2 \quad &= \quad \langle \mu_{\mathsf{P}} - \mu_{\mathsf{Q}}, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle_{\mathcal{F}} \\ &= \quad \langle \mu_{\mathsf{P}}, \mu_{\mathsf{P}} \rangle + \langle \mu_{\mathsf{Q}}, \mu_{\mathsf{Q}} \rangle - 2 \langle \mu_{\mathsf{P}}, \mu_{\mathsf{Q}} \rangle \\ &= \quad \langle \mathbf{E}_{\mathsf{P}} \varphi_x, \mathbf{E}_{\mathsf{P}} \varphi_x \rangle + \ldots \end{aligned}$$

Empirical estimate of MMD

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

MMD in terms of kernels:

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

Empirical estimate of MMD

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

MMD in terms of kernels:

$$\begin{aligned}
\mathrm{MMD}^2 = \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|_{\mathcal{F}}^2 \quad &= \quad \langle \mu_{\mathbf{P}} - \mu_{\mathbf{Q}}, \mu_{\mathbf{P}} - \mu_{\mathbf{Q}} \rangle_{\mathcal{F}} \\
&= \quad \langle \mu_{\mathbf{P}}, \mu_{\mathbf{P}} \rangle + \langle \mu_{\mathbf{Q}}, \mu_{\mathbf{Q}} \rangle - 2 \langle \mu_{\mathbf{P}}, \mu_{\mathbf{Q}} \rangle \\
&= \quad \langle \mathbf{E}_{\mathbf{P}} \varphi_x, \mathbf{E}_{\mathbf{P}} \varphi_x \rangle + \ldots \\
&= \quad \mathbf{E}_{\mathbf{P}} \langle \varphi_x, \varphi_{x'} \rangle + \ldots \\
&= \quad \mathbf{E}_{\mathbf{P}} k(x, x') + \mathbf{E}_{\mathbf{Q}} k(y, y') - 2\mathbf{E}_{\mathbf{P},\mathbf{Q}} k(x, y)
\end{aligned}$$

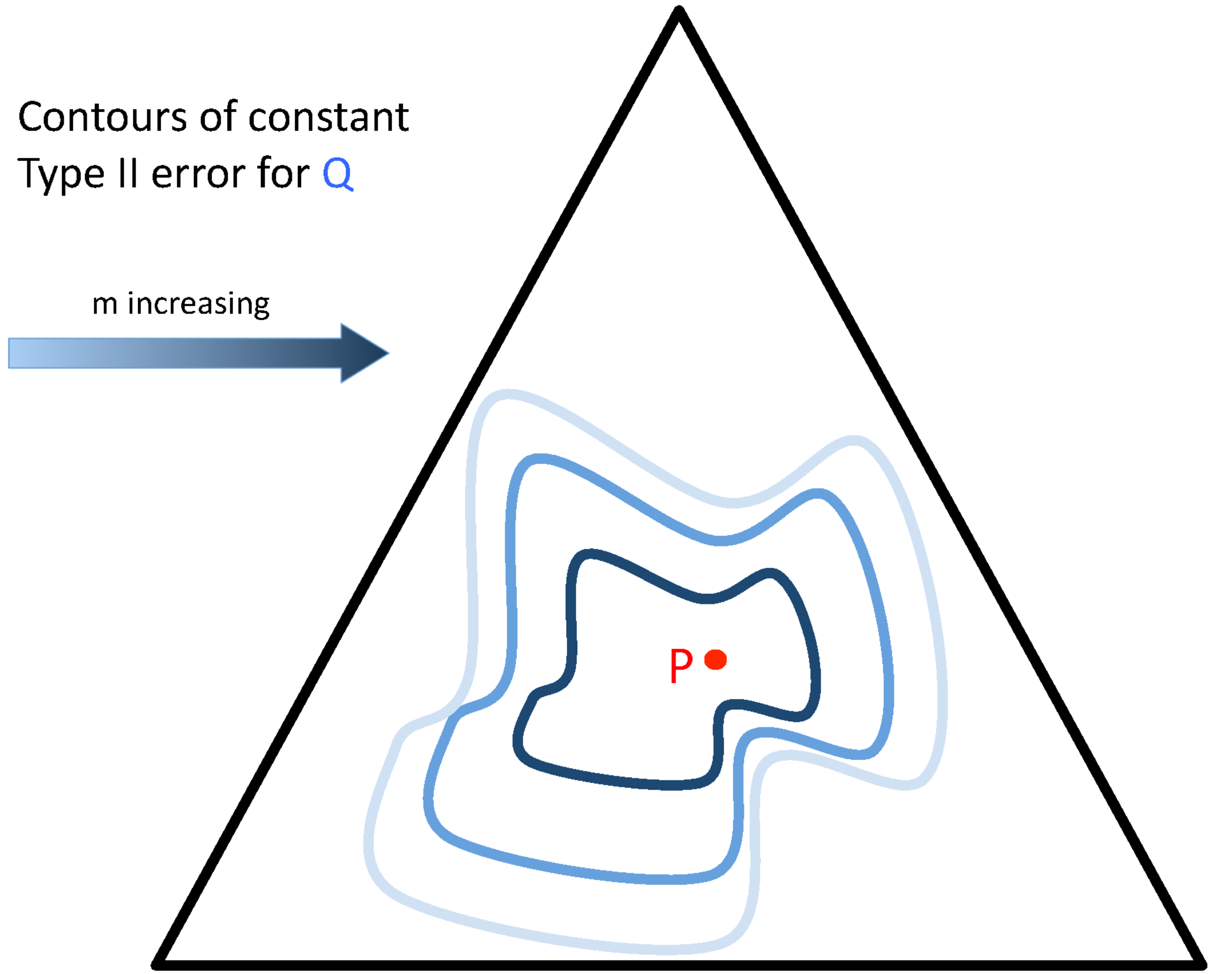
Local departures from the null

What is a hard testing problem?

Local departures from the null

What is a hard testing problem?

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



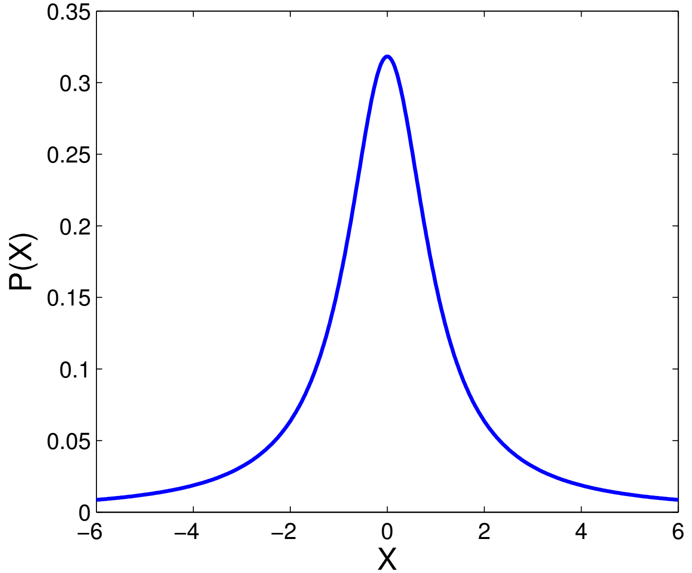
Local departures from the null

What is a hard testing problem?

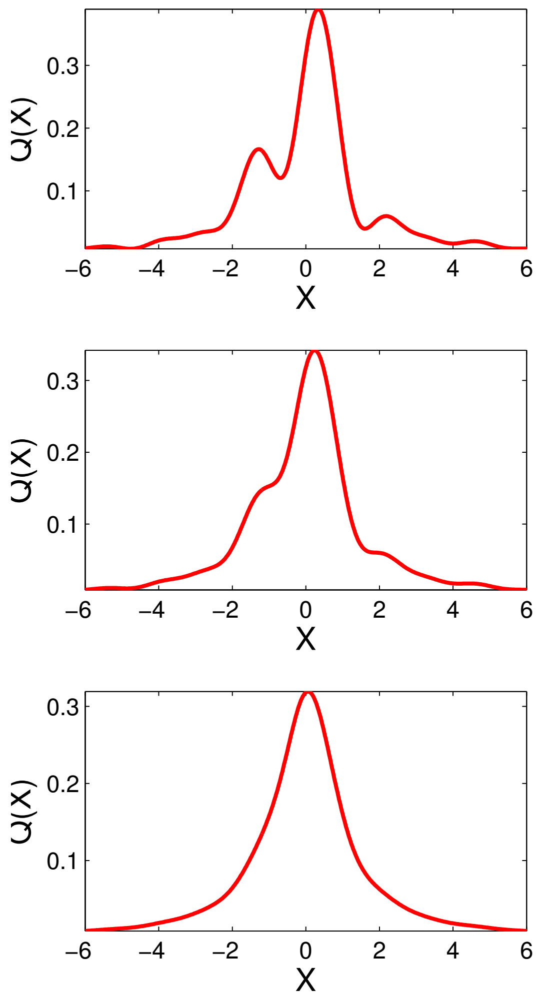
- As m increases, distinguish "closer" **P** and **Q** with same Type II error
- Example: $f_{\mathbf{P}}$ and $f_{\mathbf{Q}}$ probability densities, $f_{\mathbf{Q}} = f_{\mathbf{P}} + \delta g$, where $\delta \in \mathbb{R}$, g some *fixed* function such that $f_{\mathbf{Q}}$ is a valid density
  -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

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?

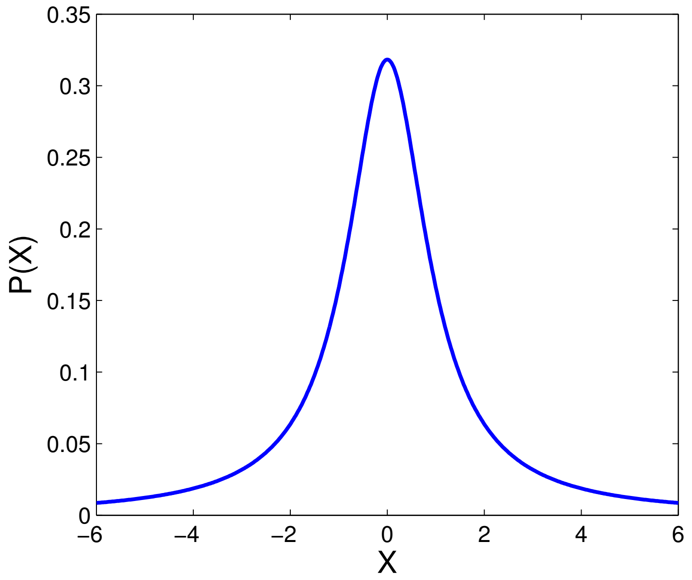
General characterization of local departures from $\mathcal{H}_0$:

- 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]

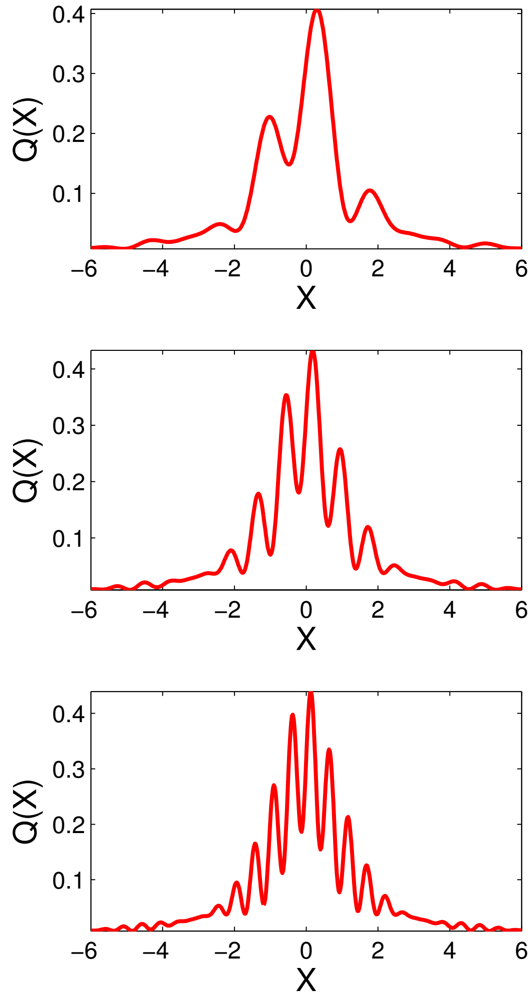
$$\|g_m\|_{\mathcal{F}} = cm^{-1/2}$$

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 this relate to Parzen density estimate? [Anderson et al., 1994]

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

Kernels vs kernels

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

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

- L_2 distance between Parzen density estimates:

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

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

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

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

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.

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

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems 15*, pages 513–520, Cambridge, MA, 2007. MIT Press.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22*, Red Hook, NY, 2009. Curran Associates Inc.

A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

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

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

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.

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.