Kernel methods for testing three-variable interactions

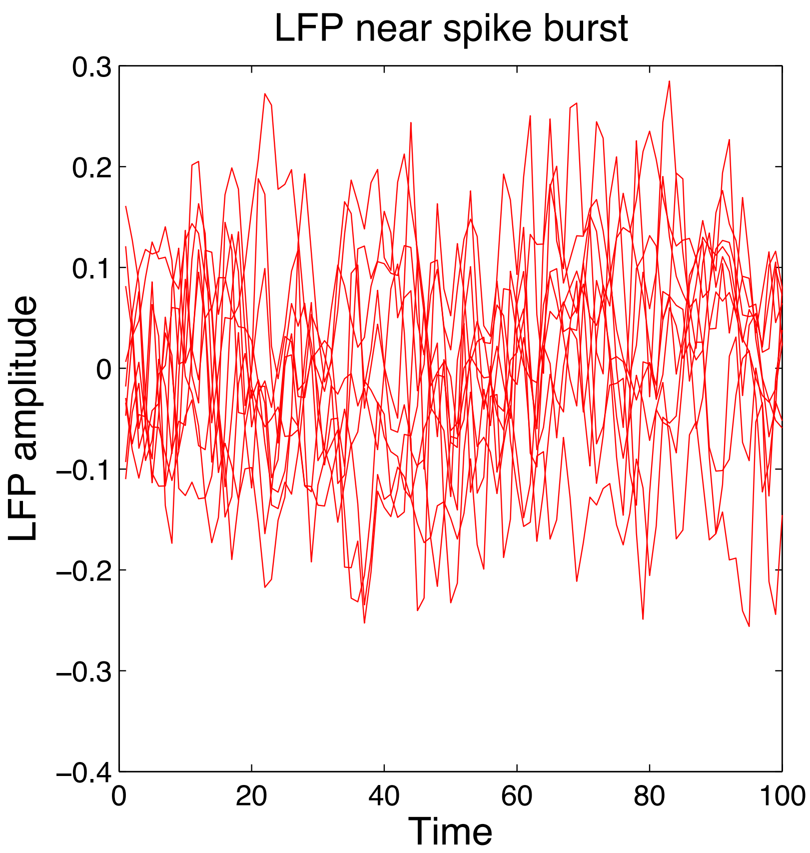
Arthur Gretton

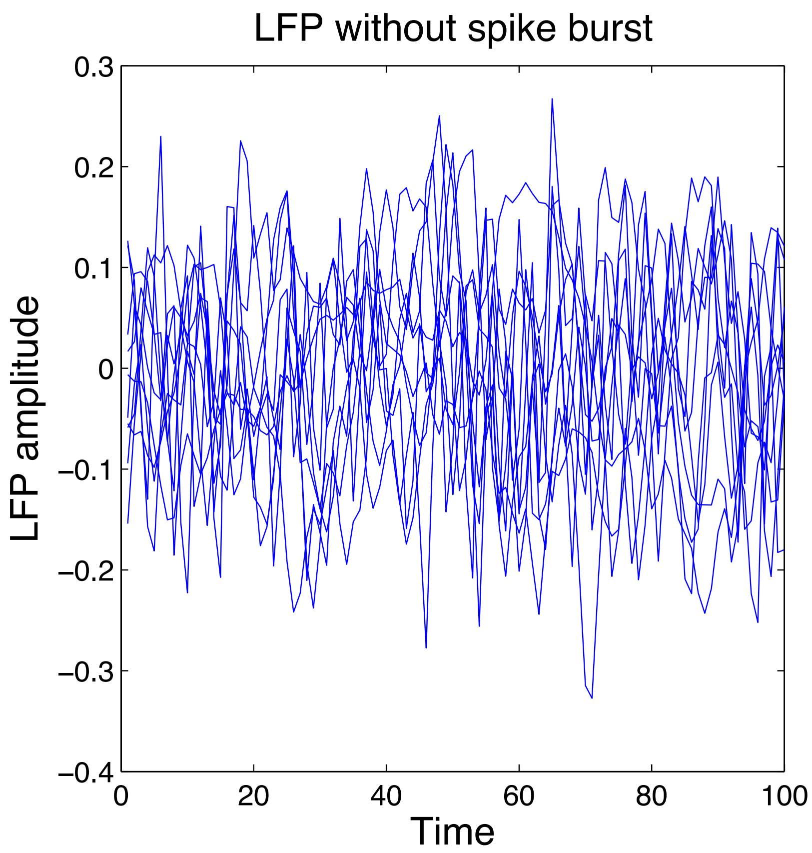
Gatsby Computational Neuroscience Unit

Tokyo, March 2014

Differences in brain signals

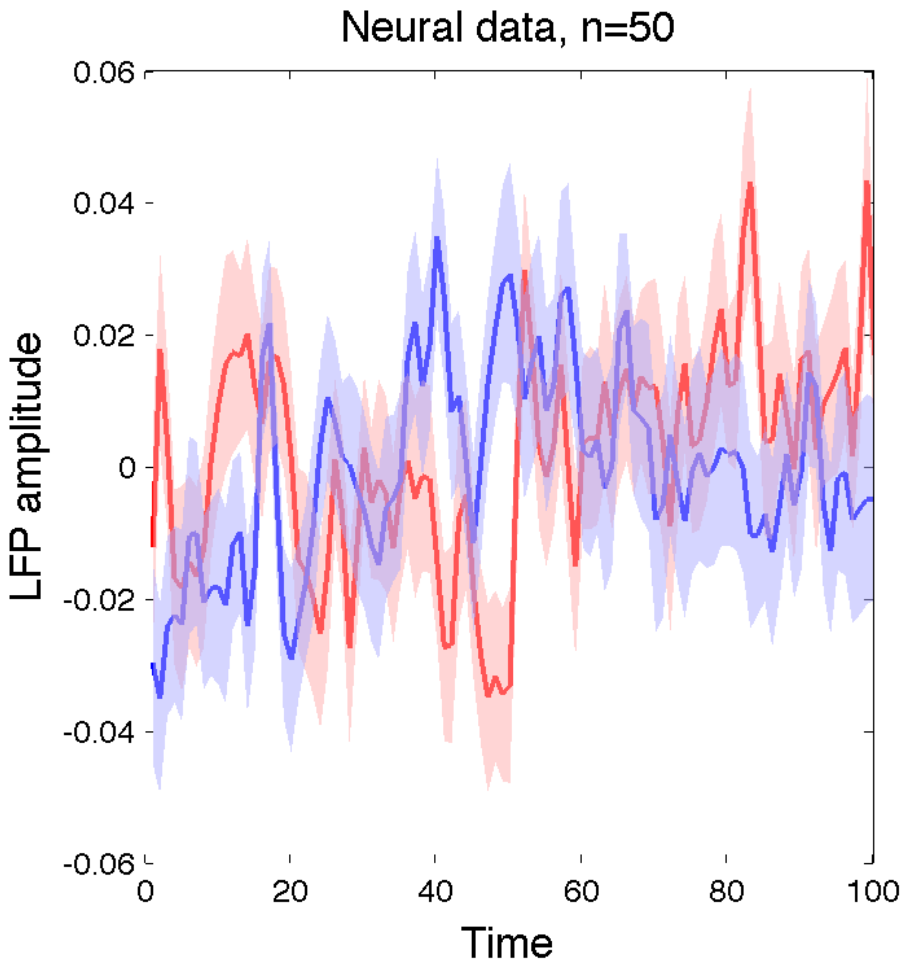
The problem: Do local field potential (LFP) signals change when measured near a spike burst?





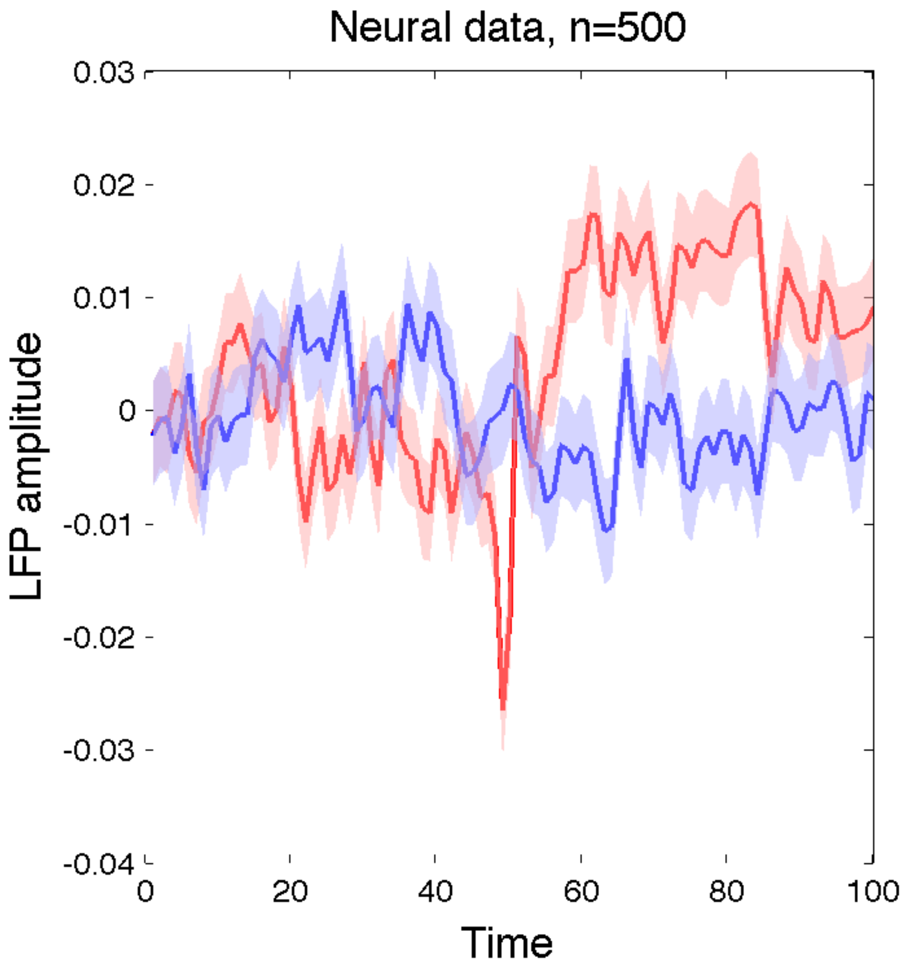
Differences in brain signals

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



Differences in brain signals

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



Detecting statistical dependence

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

Detecting statistical dependence

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

Detecting statistical dependence

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

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

Detecting statistical dependence

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

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

Detecting statistical dependence

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

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

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

…

$$\overset{?}{\Longleftrightarrow}$$

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

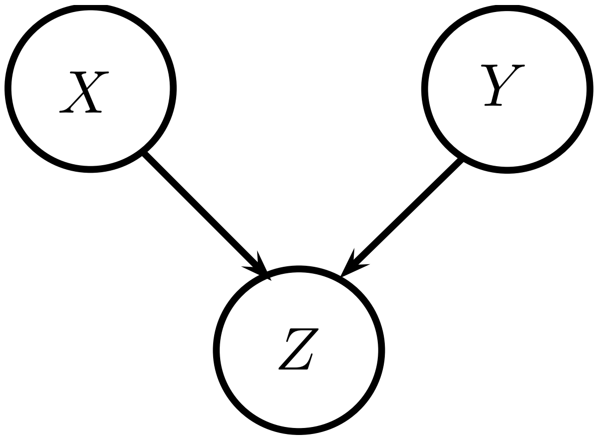
Y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement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

…

Are the French text extracts translations of the English ones?

Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?

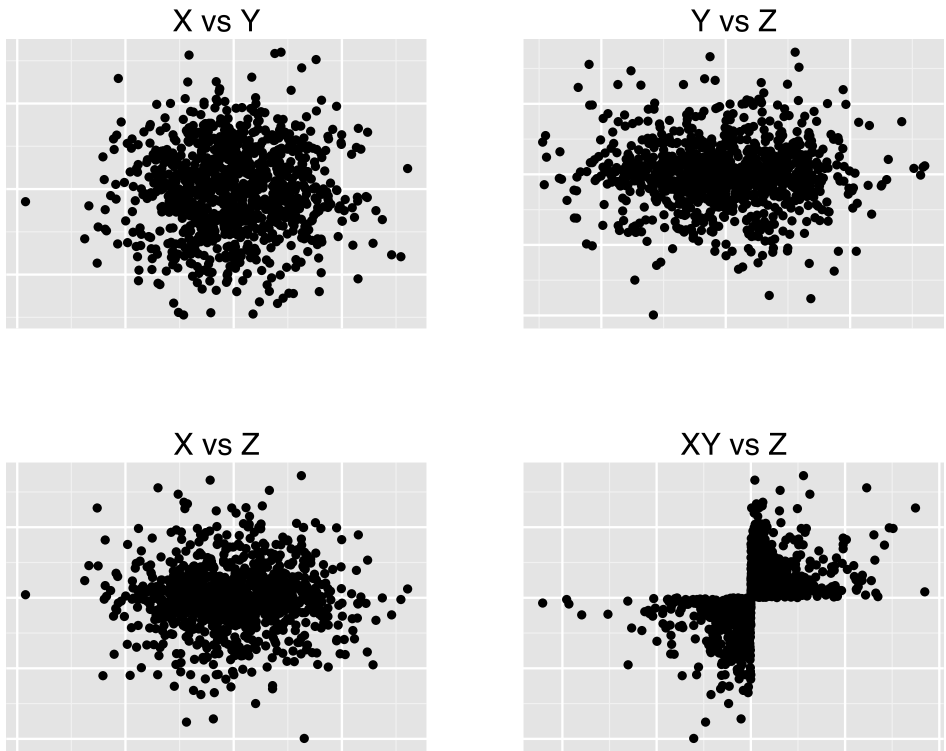


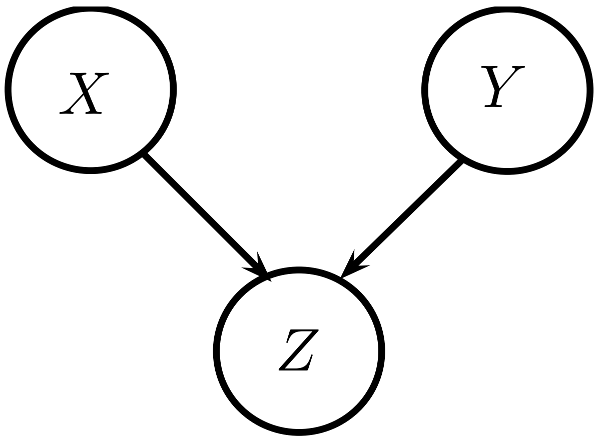
Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?

Detecting a higher order interaction

- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- $X \perp\!\!\!\perp Y$, $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z$





- $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,
- $Z|\ X, Y \sim \text{sign}(XY) Exp(\frac{1}{\sqrt{2}})$

Faithfulness violated here

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 $MMD(\mathbf{P}_{xy}, \mathbf{P}_x\mathbf{P}_y)$

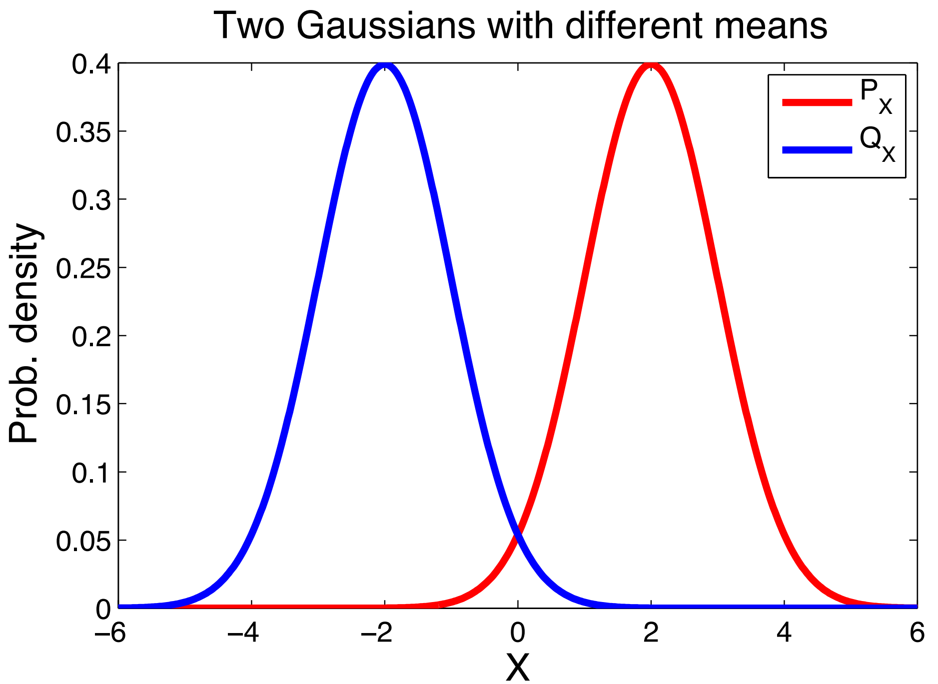
Overview

- 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 $MMD(\mathbf{P}_{xy}, \mathbf{P}_x\mathbf{P}_y)$
- Detecting three-way interactions
  - Parents with weak individual influence, strong combined influence
  - Avoid difficult problem of conditional dependence testing
  - Generalization of independence test

Kernel distance between distributions

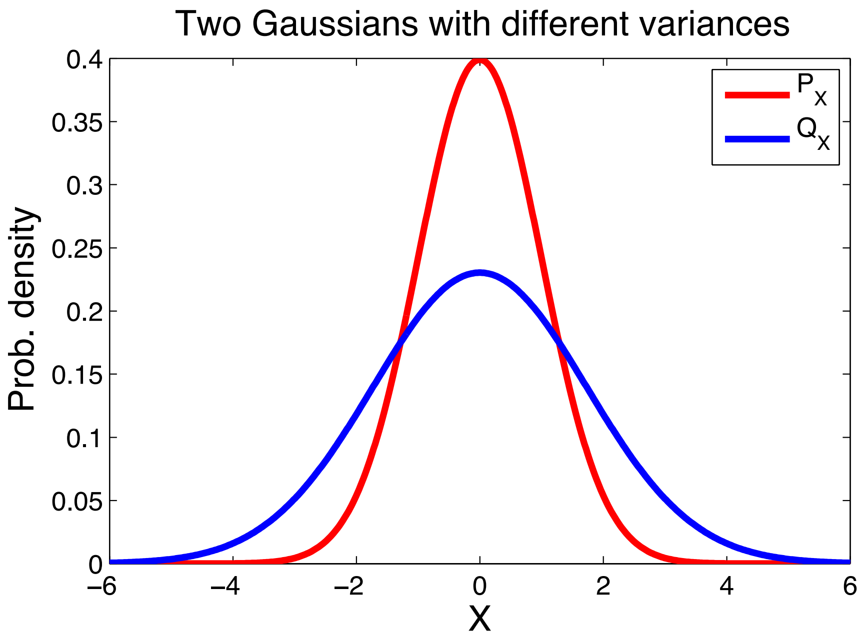
Feature mean difference

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



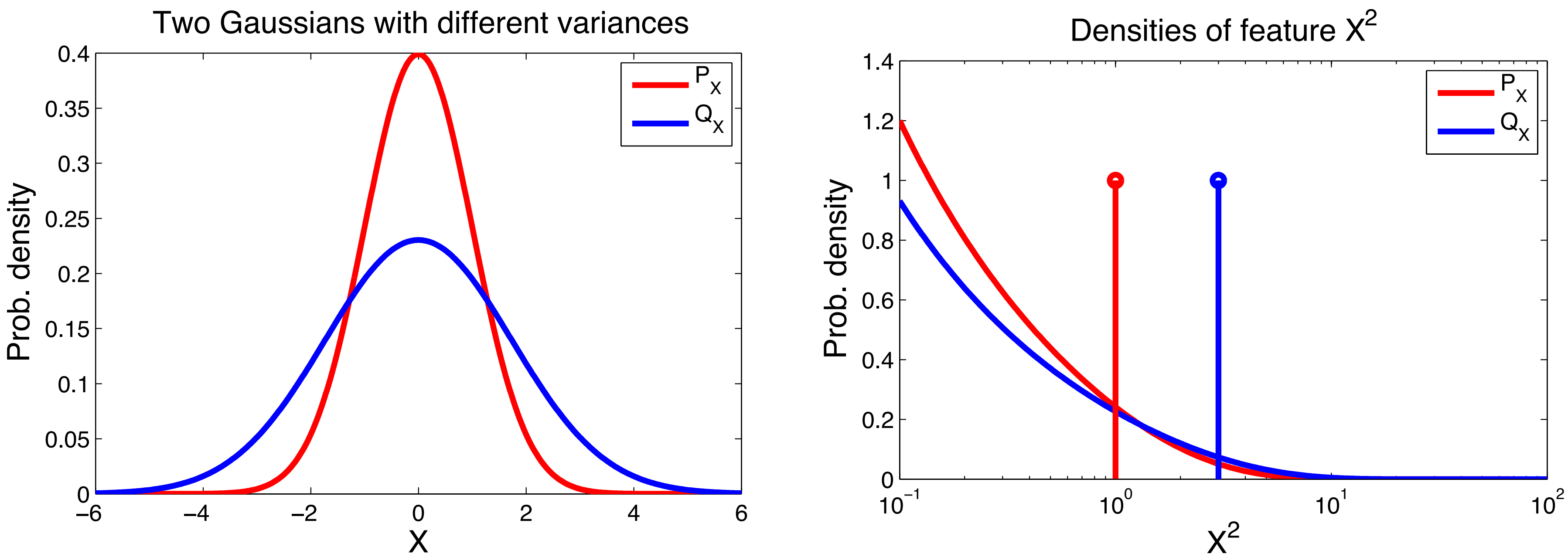
Feature mean difference

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



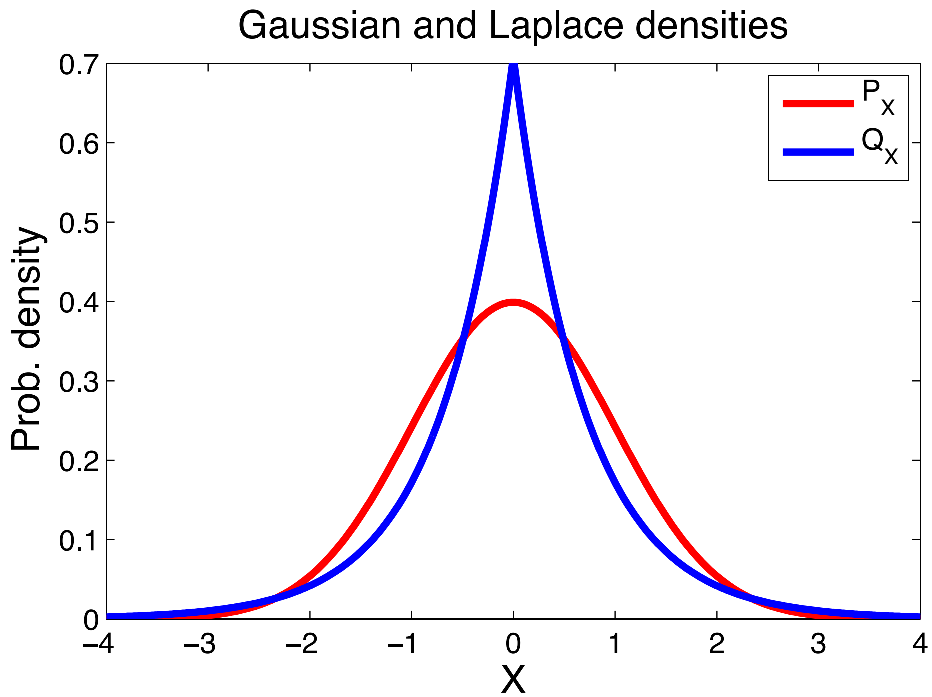
Feature mean difference

- 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 $\varphi_x = x^2$



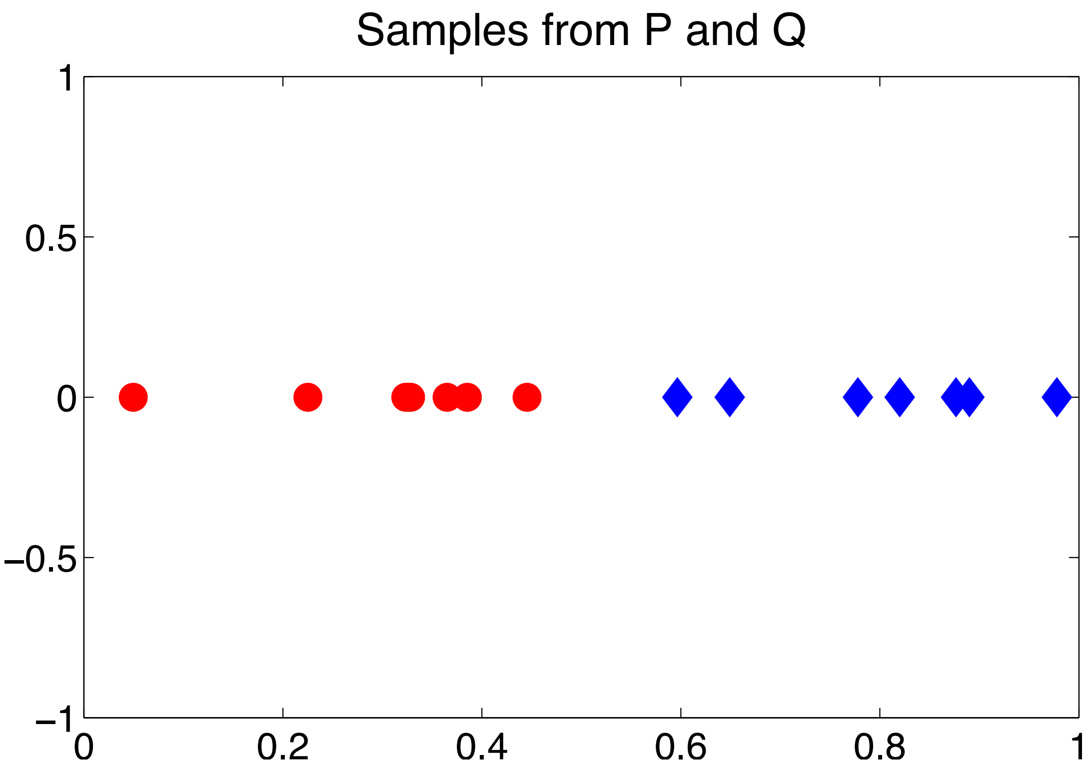
Feature mean difference

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



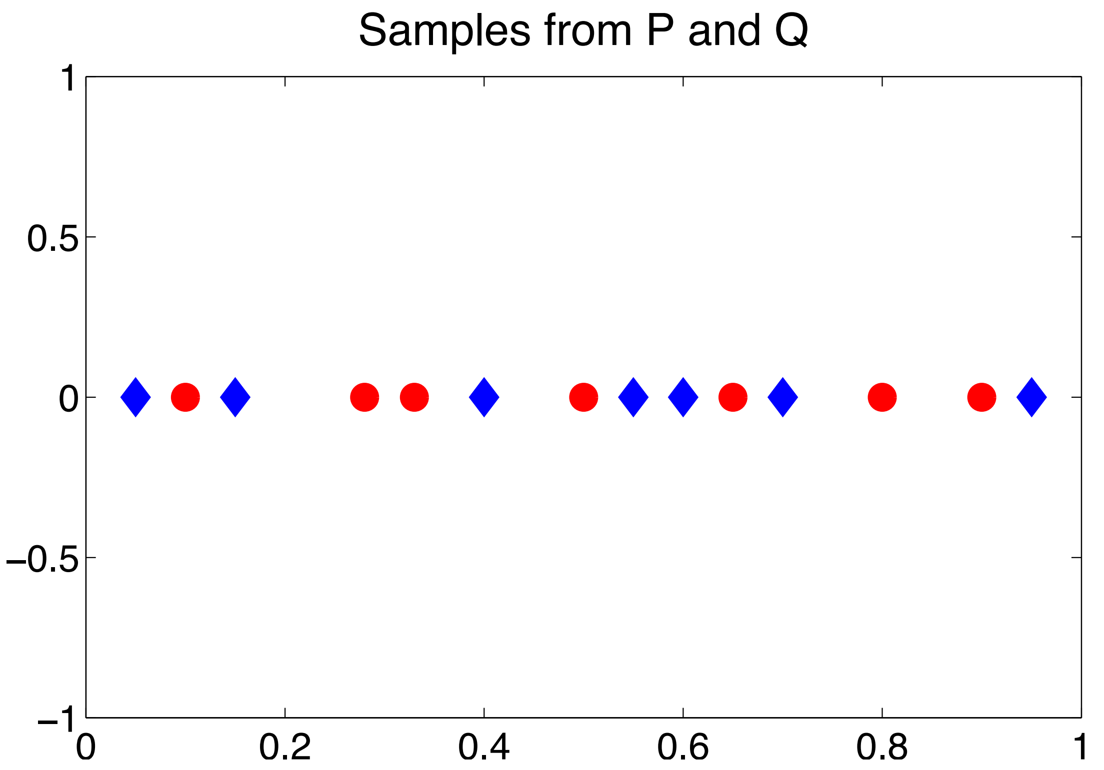
Function Showing Difference in Distributions

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



Function Showing Difference in Distributions

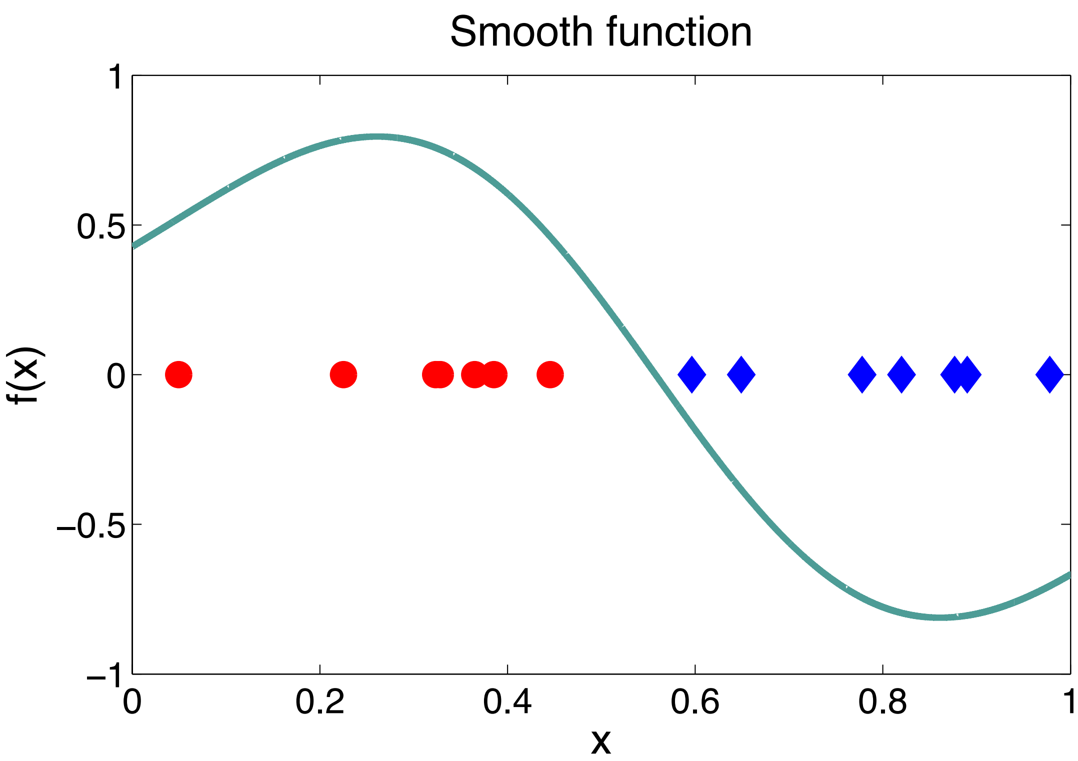
- Are **P** and **Q** different?



Function Showing Difference in Distributions

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

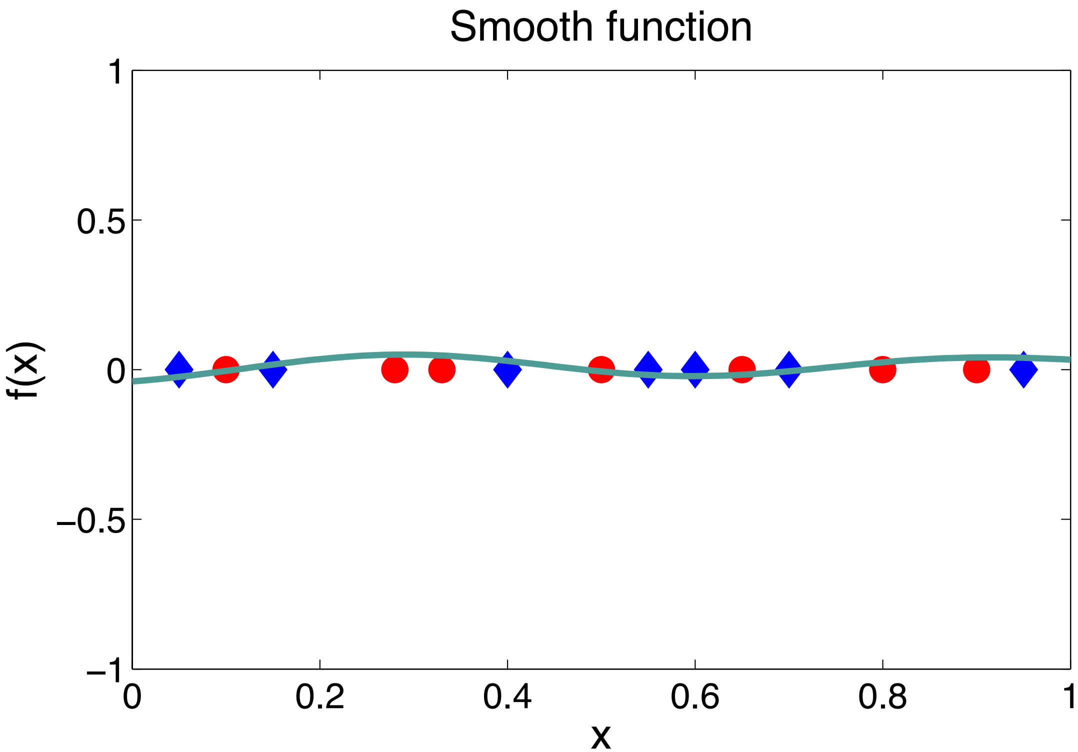
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathsf{y})\right].$$



Function Showing Difference in Distributions

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

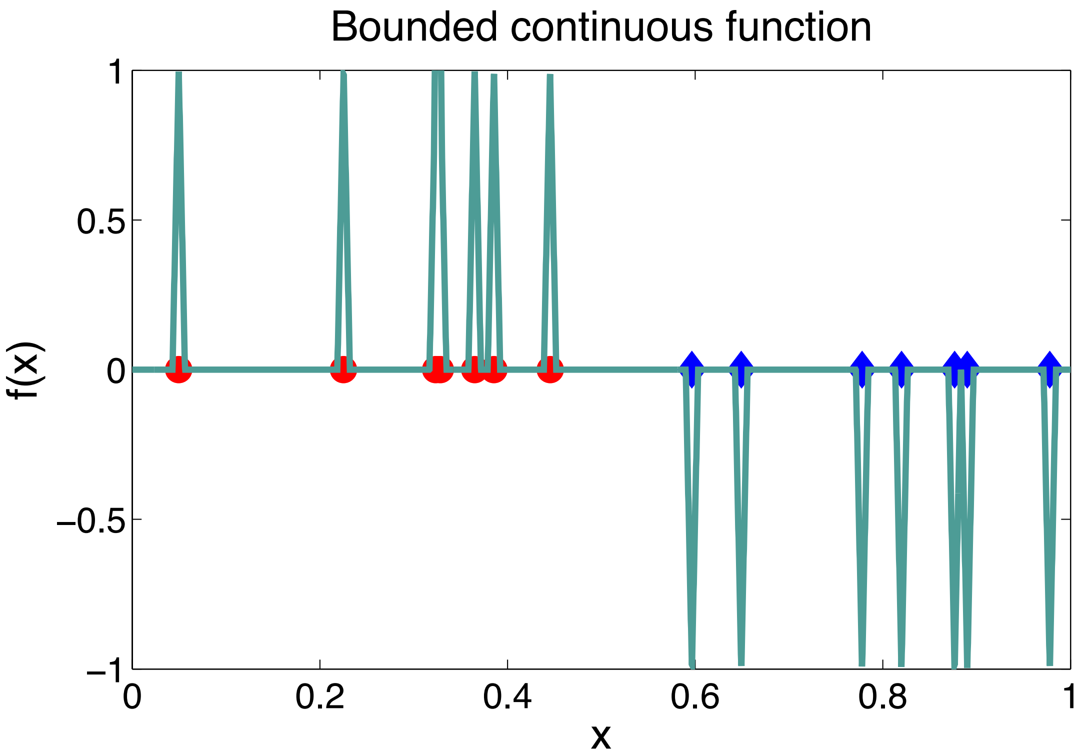
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathrm{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathrm{y})\right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

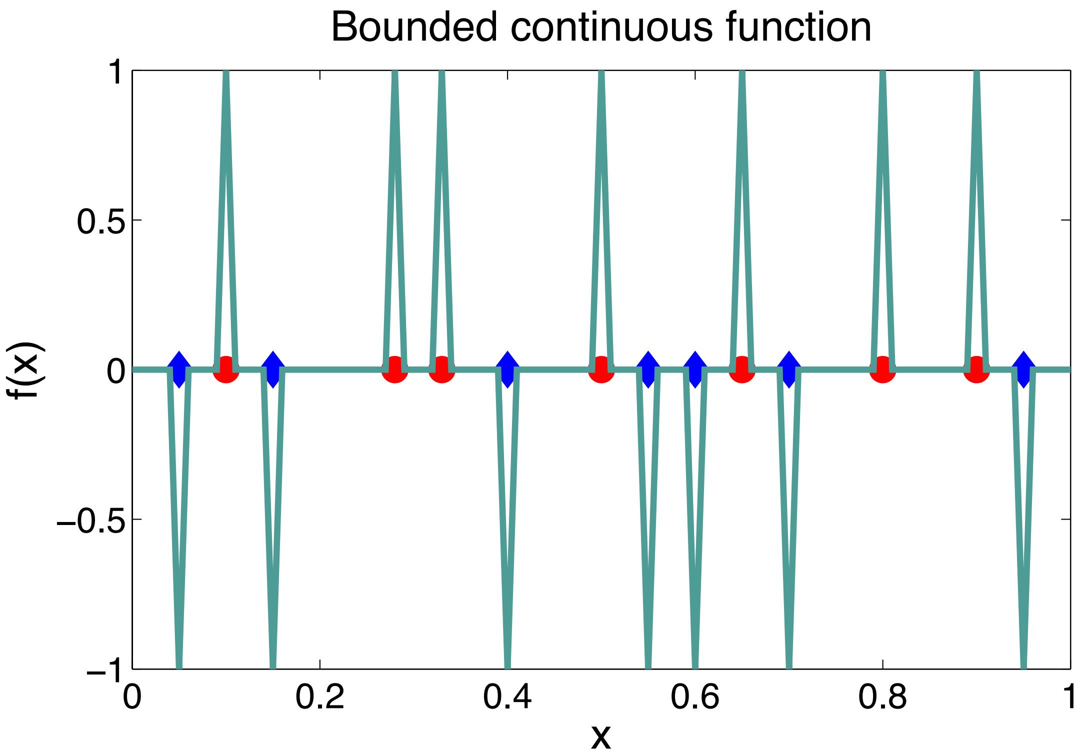
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

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

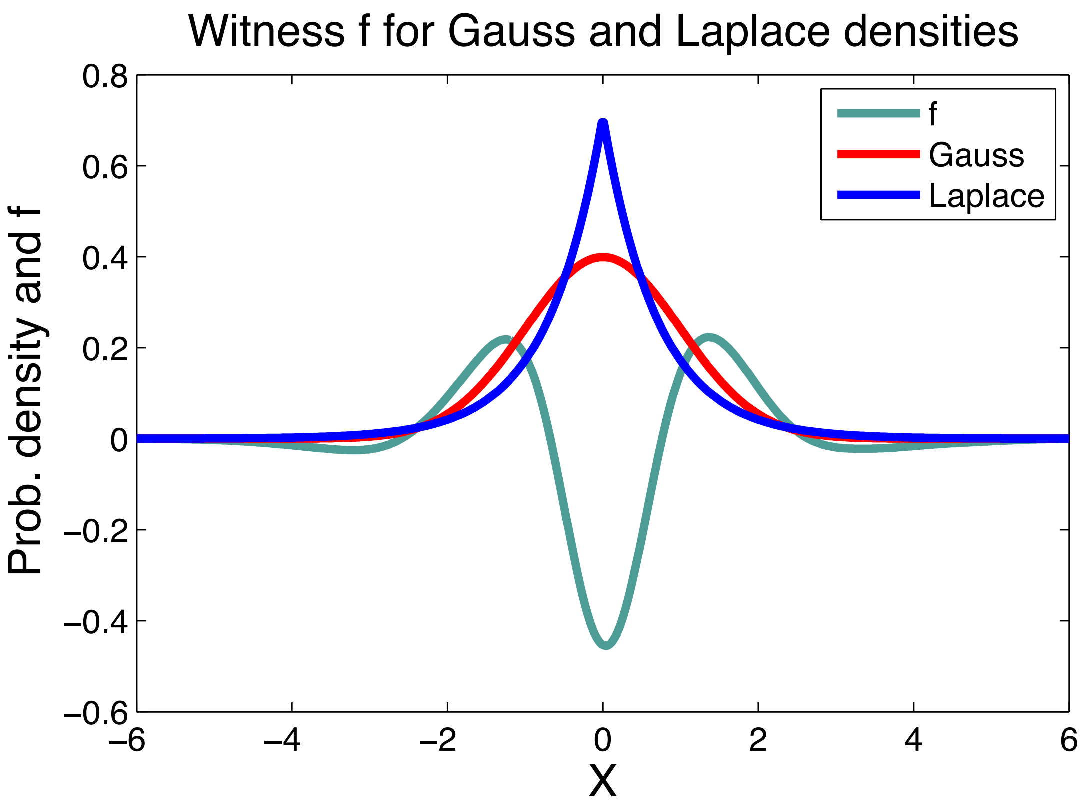


Function Showing Difference in Distributions

- 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

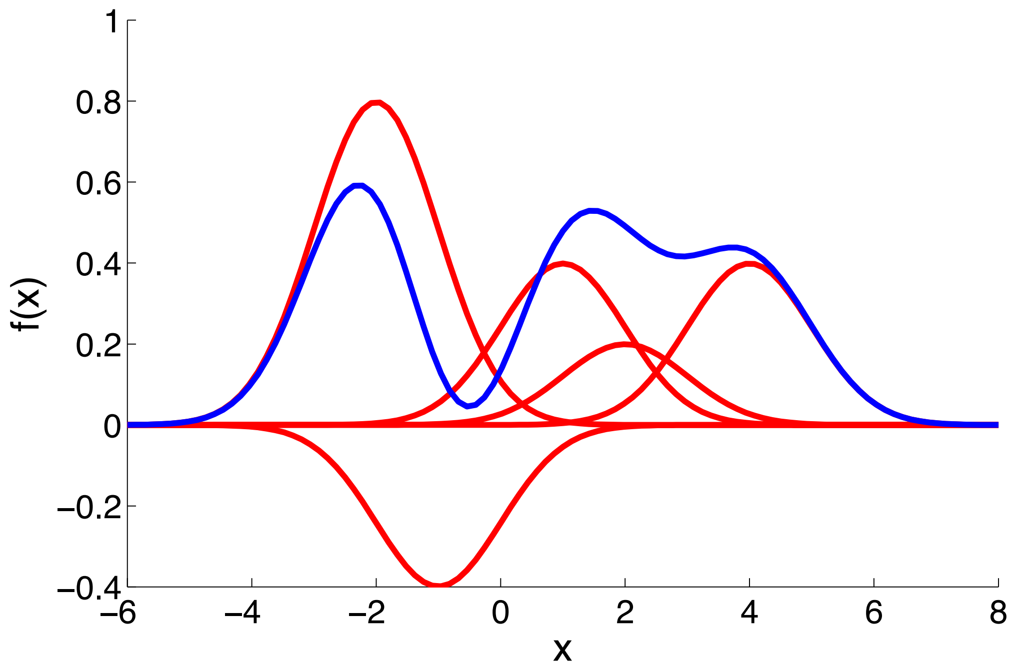
- Maximum mean discrepancy: smooth function for $\mathbf{P}$ vs $\mathbf{Q}$

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

- 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 $\mathcal{F}$ Sriperumbudur et al. (2010), Gretton et al. (2012), Sejdinovic et al. (2013)

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

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

The RKHS as feature map

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

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

- Inner product between feature maps:

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

The RKHS as feature map

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

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

- 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,

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

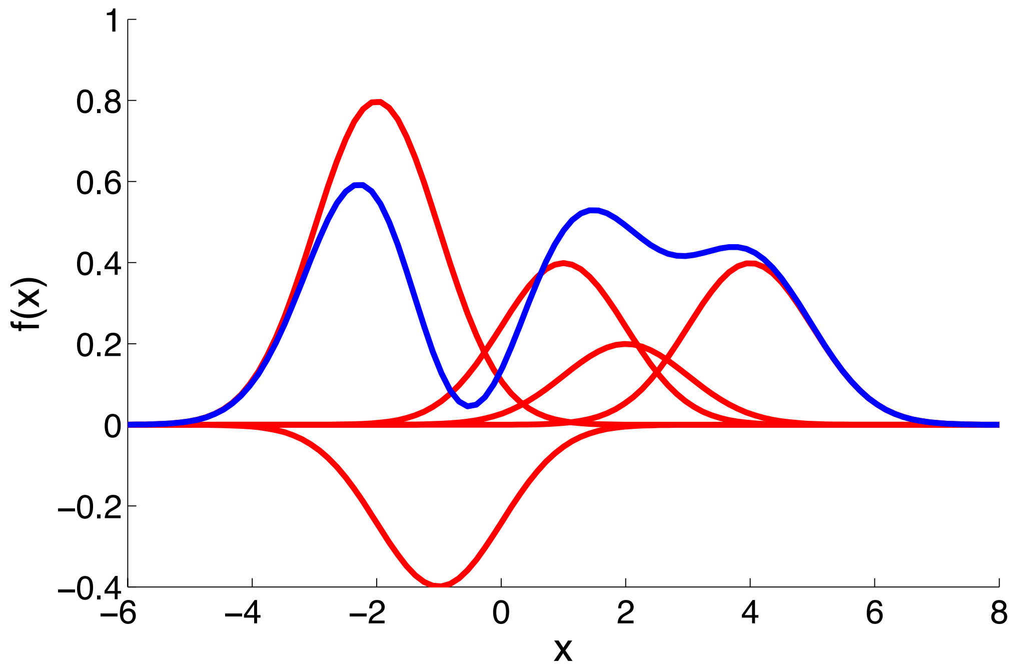
  for positive definite $k(x, y)$

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

The RKHS as feature map

- Function in RKHS:

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



Probabilities in feature space: the mean trick

The kernel trick

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

$$\varphi_x = \left[\dots \sqrt{\lambda_i} e_i(x) \dots \right] \in \ell_2$$

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

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

- The kernel trick: $\forall f \in \mathcal{F}$,

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

Probabilities in feature space: the mean trick

The kernel trick

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

$$\varphi_x = \left[\ldots \sqrt{\lambda_i} e_i(x) \ldots \right] \in \ell_2$$

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

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

- The kernel trick: $\forall f \in \mathcal{F}$,

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

The mean trick

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

$$\mu_{\mathsf{P}} = \left[\ldots \sqrt{\lambda_i} \mathbf{E}_{\mathsf{P}} \left[e_i(X) \right] \ldots \right] \in \ell_2$$

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

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

for $X \sim \mathsf{P}$ and $Y \sim \mathsf{Q}$.

- The mean trick:

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

Feature embeddings of probabilities

For all $f \in \mathcal{F}$,

The kernel trick:

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

The mean trick:

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

$\boxed{\mu_{\mathsf{P}} \text{ gives you expectations of all RKHS functions}}$

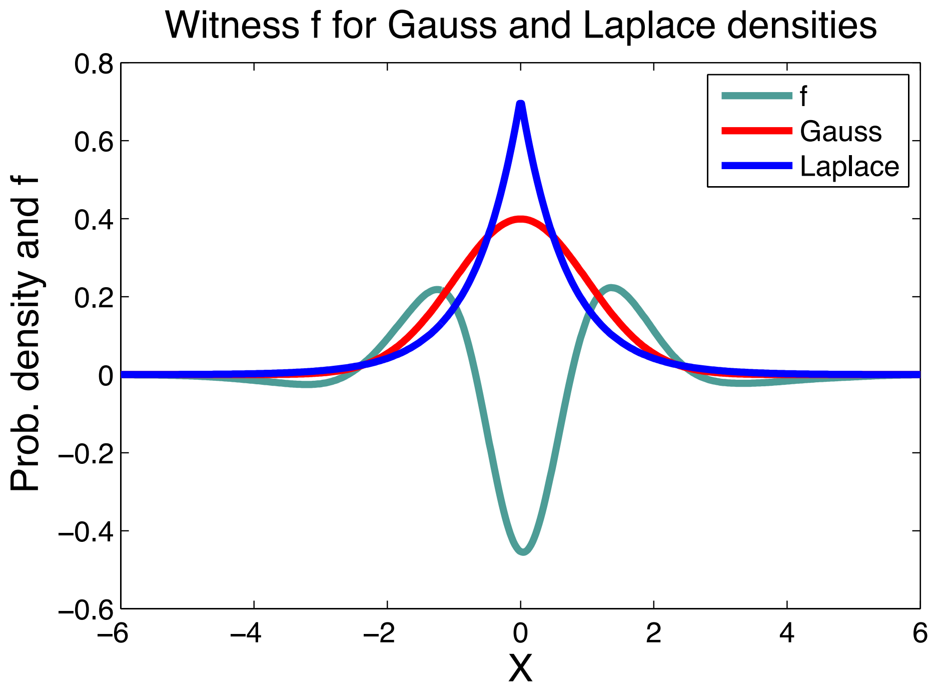
When k characteristic, then μ_{P} unique, e.g. Gauss, Laplace, ...

Function view vs feature mean view

- The (kernel) MMD:

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

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



Function view vs feature mean view

- The (kernel) MMD:

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

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

use

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

Function view vs feature mean view

- The (kernel) MMD:

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

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

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

use

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

Function view vs feature mean view

- The (kernel) MMD:

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

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

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

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

use

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

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

MMD in terms of kernels, empirical estimate

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

MMD in terms of kernels, empirical estimate

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

MMD in terms of kernels:

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

MMD in terms of kernels, empirical estimate

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

MMD in terms of kernels:

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

MMD in terms of kernels, empirical estimate

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

MMD in terms of kernels:

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

MMD in terms of kernels, empirical estimate

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

MMD in terms of kernels:

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

MMD in terms of kernels, empirical estimate

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

MMD in terms of kernels:

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

MMD in terms of kernels, empirical estimate

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

MMD in terms of kernels:

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

Empirical estimate: given i.i.d. $X := \{x_1, \ldots, x_m\}$

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

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

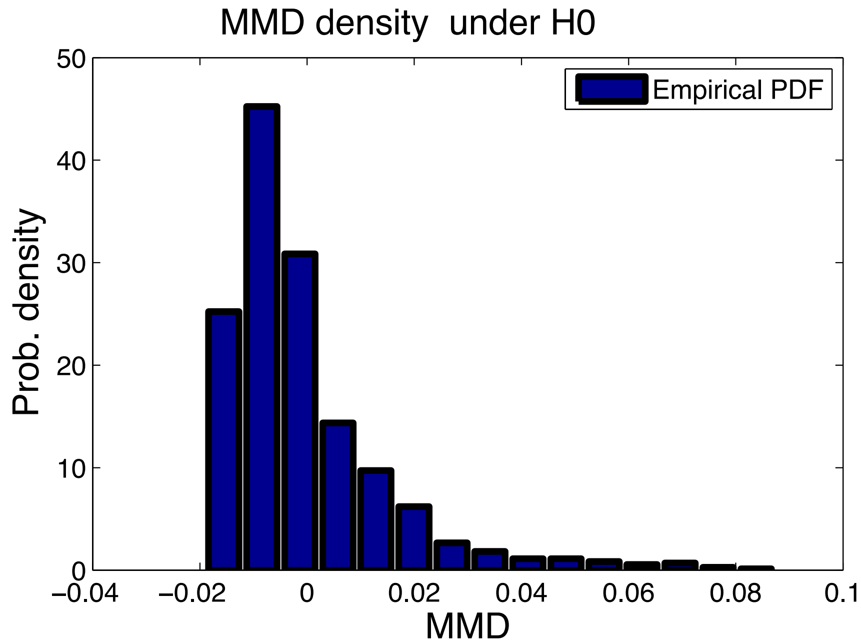
- Two hypotheses:
  - H_0: null hypothesis ($\mathsf{P} = \mathsf{Q}$)
  - H_1: alternative hypothesis ($\mathsf{P} \neq \mathsf{Q}$)
- Observe samples $\boldsymbol{x} := \{x_1, \ldots, x_m\}$ from P and $\boldsymbol{y}$ from Q
- If empirical $\widehat{\mathrm{MMD}}^2$ is
  - "far from zero": reject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

Statistical test using MMD

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

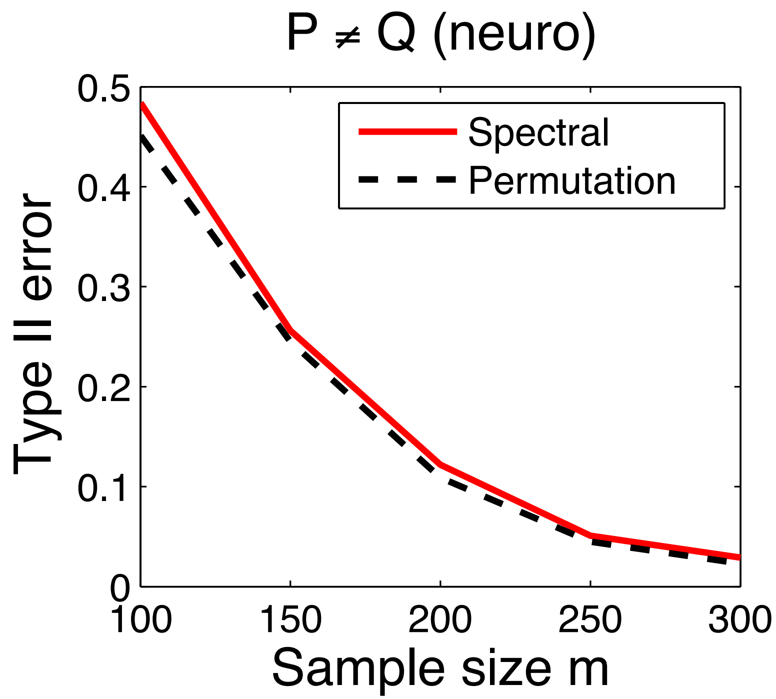
$$m\widehat{\mathrm{MMD}}^2 \sim \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2\right]$$

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

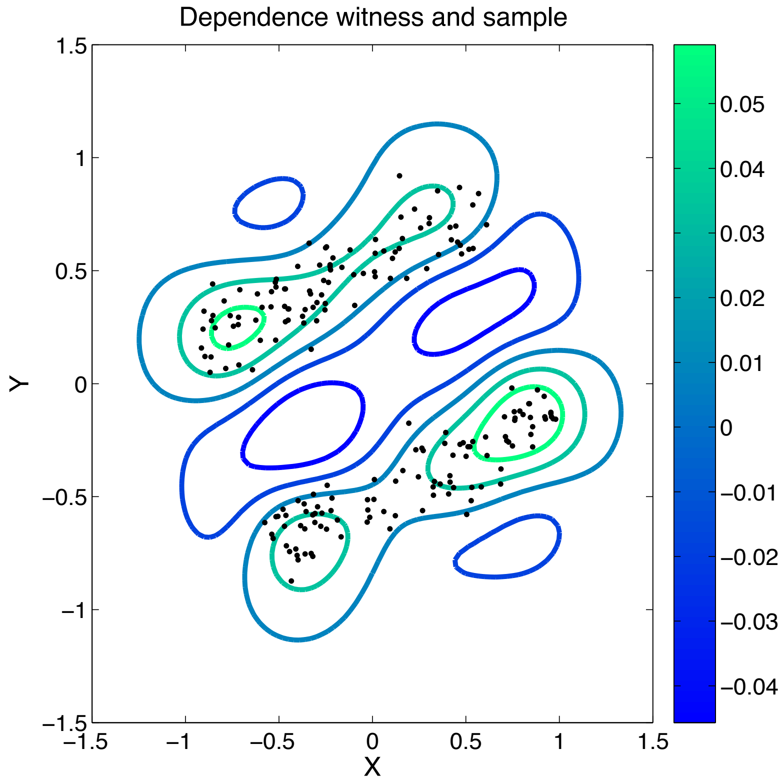


MMD for independence

- Dependence measure: Gretton et al. (2008)

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

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

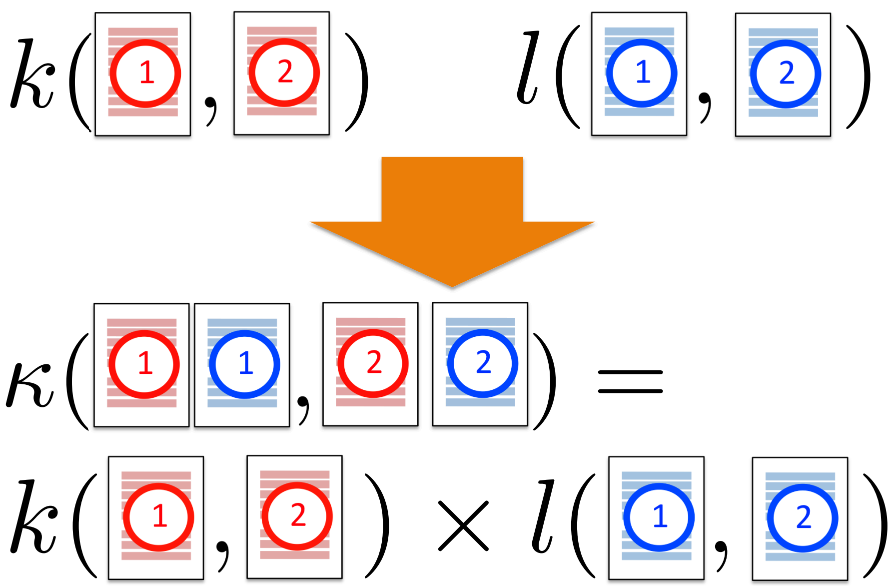


MMD for independence

- Dependence measure: Gretton et al. (2008)

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

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



Experiment: dependence testing for translation

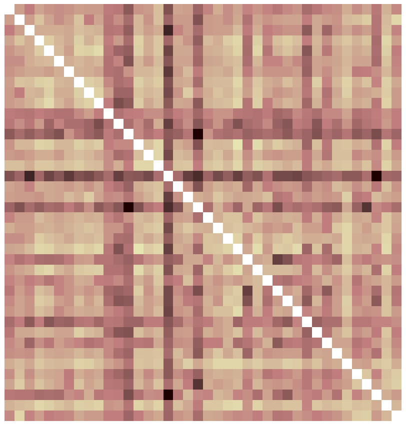
- Translation example: [NIPS07b] Canadian Hansard (agriculture)
- 5-line extracts, k-spectrum kernel, $k = 10$, repetitions=300, sample size 10
- Empirical $MMD(\mathbf{P}_{XY}, \mathbf{P}_X \mathbf{P}_Y)$:

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

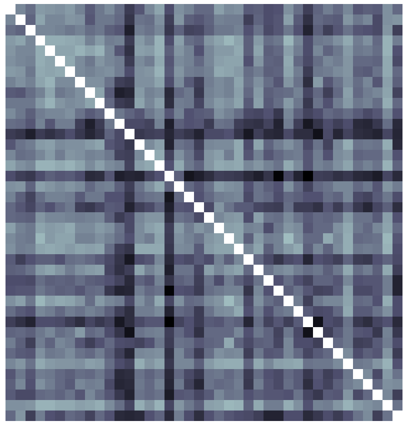
... no doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development...

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

$\Downarrow$ $\Downarrow$



$\Rightarrow$MMD$\Leftarrow$

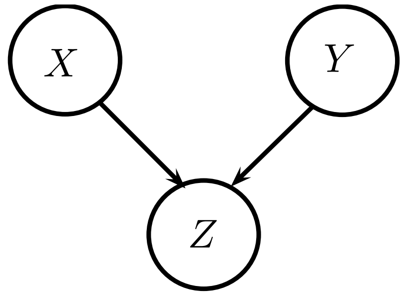


K L

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

Lancaster (3-way) Interactions

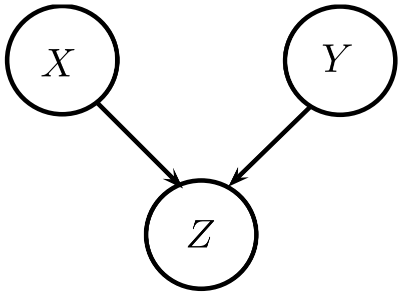
V-structure Discovery



Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

- CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ (Zhang et al 2011) or

V-structure Discovery

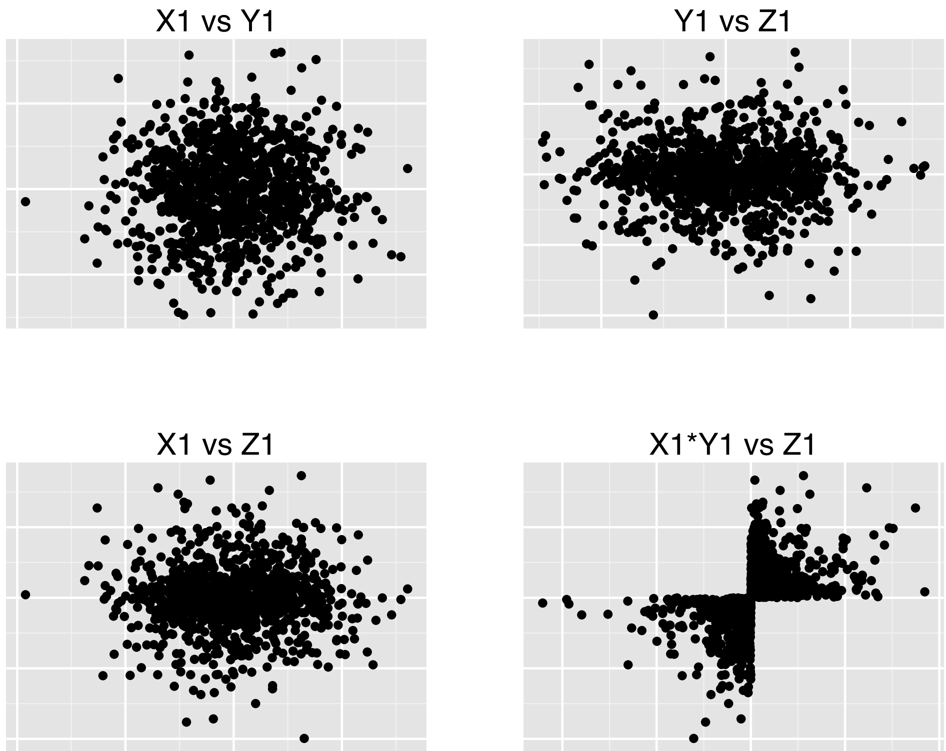


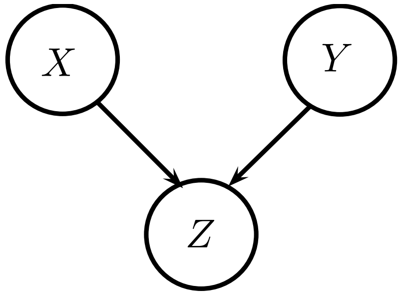
Assume $X \perp\!\!\!\perp Y$ has been established. V-structure can then be detected by:

- CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ (Zhang et al 2011) or
- Factorisation test: $\mathbf{H_0} : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)
  - compute p-values for each of the marginal tests for $(Y, Z) \perp\!\!\!\perp X$, $(X, Z) \perp\!\!\!\perp Y$, or $(X, Y) \perp\!\!\!\perp Z$
  - apply Holm-Bonferroni (**HB**) sequentially rejective correction (Holm 1979)

V-structure Discovery (2)

- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- $X \perp\!\!\!\perp Y$, $Y \perp\!\!\!\perp Z$, $X \perp\!\!\!\perp Z$

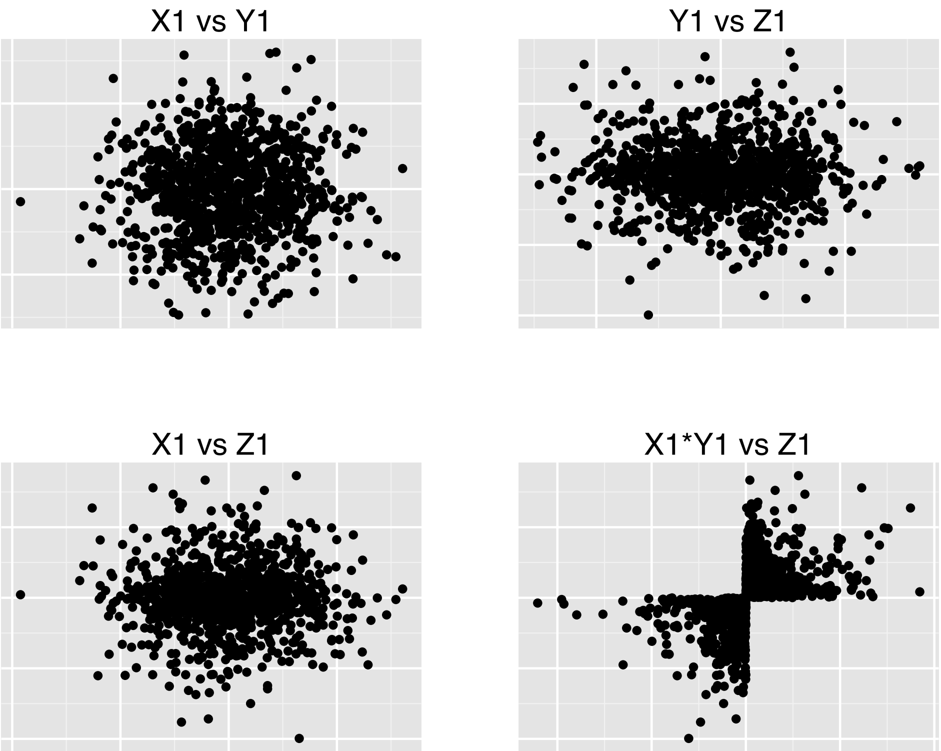


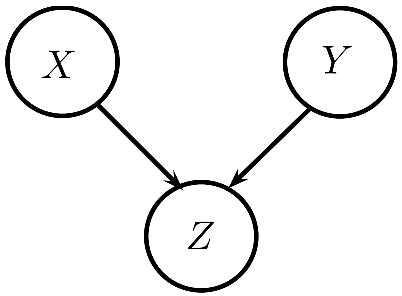


- $X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$,
- $Z_1 | \ X_1, Y_1 \sim \text{sign}(X_1 Y_1) Exp(\frac{1}{\sqrt{2}})$

V-structure Discovery (2)

- How to detect V-structures with pairwise weak (or nonexistent) dependence?
- $X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$





- $X_1, Y_1 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,
- $Z_1 | \; X_1, Y_1 \sim \text{sign}(X_1 Y_1) Exp(\frac{1}{\sqrt{2}})$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$
- (Note: violates faithfulness)

V-structure Discovery (3)

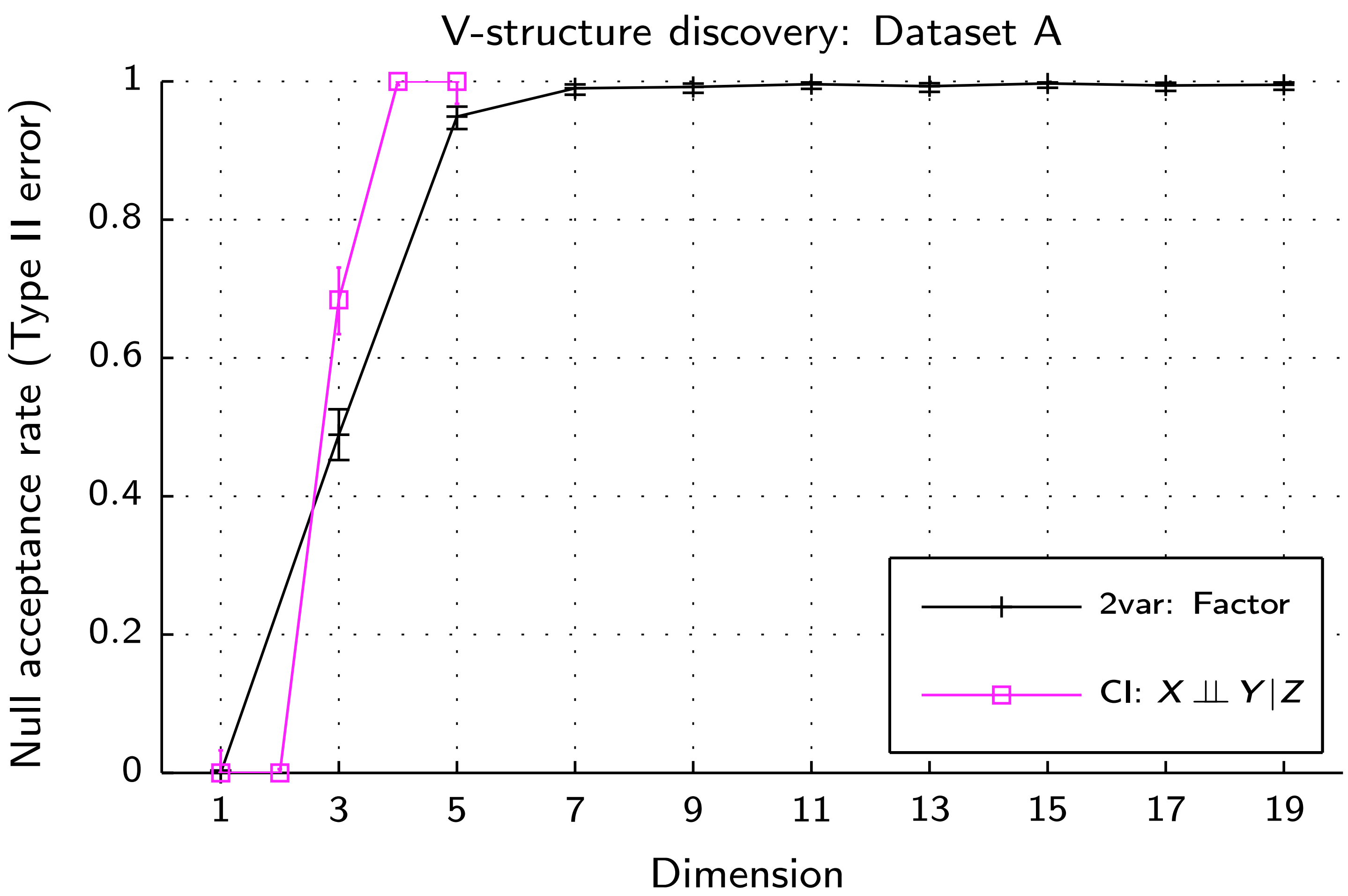


Figure 1: CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), and a factorisation test with a **HB** correction, $n = 500$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2 :$ $\qquad \Delta_L P = P_{XY} - P_X P_Y$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

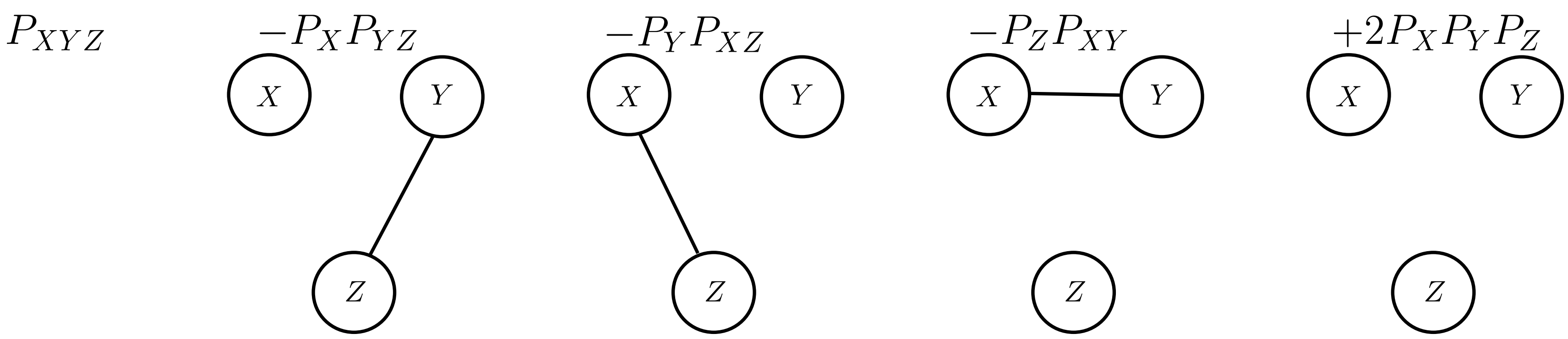
- $D = 2:$ $\qquad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2P_X P_Y P_Z$

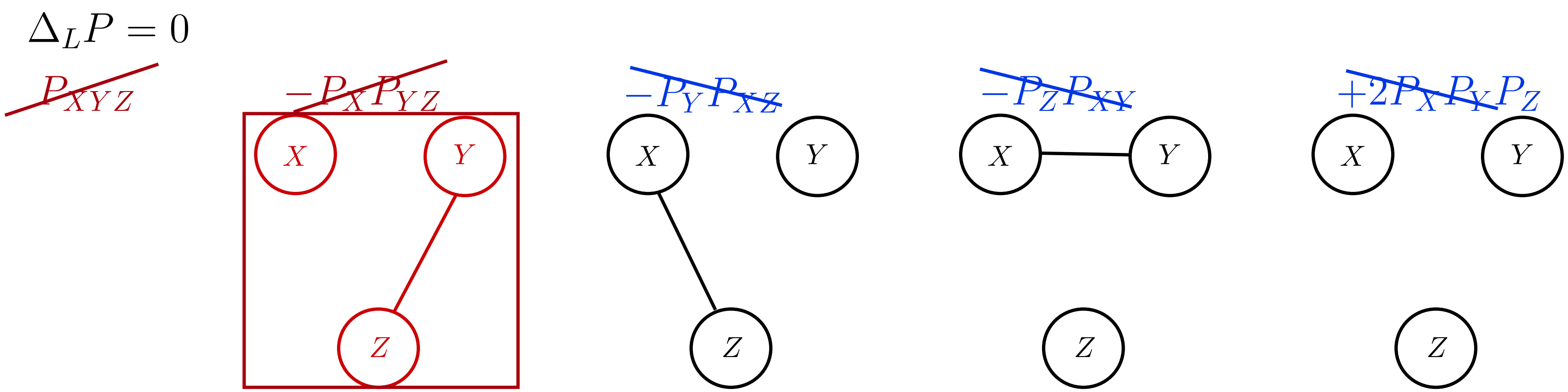
$\Delta_L P =$



Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$



Case of $P_X \perp\!\!\!\perp P_{YZ}$

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2:$ $\quad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3:$ $\quad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$

$$(X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X \ \Rightarrow \ \Delta_L P = 0.$$

...so what might be missed?

Lancaster Interaction Measure

[Bahadur (1961); Lancaster (1969)] **Interaction measure** of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised in a non-trivial way as a product of its (possibly multivariate) marginal distributions.

- $D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$
- $D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$

$$\Delta_L P = 0 \not\Rightarrow (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$$

Example:

$P(0,0,0) = 0.2$	$P(0,0,1) = 0.1$	$P(1,0,0) = 0.1$	$P(1,0,1) = 0.1$
$P(0,1,0) = 0.1$	$P(0,1,1) = 0.1$	$P(1,1,0) = 0.1$	$P(1,1,1) = 0.2$

A Test using Lancaster Measure

- Test statistic is empirical estimate of $\|\mu_\kappa\left(\Delta_L P\right)\|^2_{\mathcal{H}_\kappa}$, where $\kappa = k \otimes l \otimes m$:

$$\|\mu_\kappa(P_{XYZ} - P_{XY}P_Z - \cdots)\|^2_{\mathcal{H}_\kappa} =$$

$$\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ}\rangle_{\mathcal{H}_\kappa} - 2\,\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY}P_Z\rangle_{\mathcal{H}_\kappa} \cdots$$

Inner Product Estimators

$\nu \backslash \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(\mathbf{K} \circ \mathbf{L} \circ \mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{L})\,\mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{M})\,\mathbf{L})_{++}$	$((\mathbf{M} \circ \mathbf{L})\,\mathbf{K})_{++}$	$tr(\mathbf{K}_+ \circ \mathbf{L}_+ \circ \mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K} \circ \mathbf{L})_{++}\,\mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K} \circ \mathbf{M})_{++}\,\mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L} \circ \mathbf{M})_{++}\,\mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_XP_YP_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

Table 1: V-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$

Inner Product Estimators

$\nu\backslash\nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(\mathbf{K}\circ\mathbf{L}\circ\mathbf{M})_{++}$	$((\mathbf{K}\circ\mathbf{L})\,\mathbf{M})_{++}$	$((\mathbf{K}\circ\mathbf{M})\,\mathbf{L})_{++}$	$((\mathbf{M}\circ\mathbf{L})\,\mathbf{K})_{++}$	$tr(\mathbf{K}_+\circ\mathbf{L}_+\circ\mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K}\circ\mathbf{L})_{++}\,\mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K}\circ\mathbf{M})_{++}\,\mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L}\circ\mathbf{M})_{++}\,\mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_XP_YP_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

Table 2: V-statistic estimators of $\langle\mu_\kappa\nu, \mu_\kappa\nu'\rangle_{\mathcal{H}_\kappa}$

$$\boxed{\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa} = \frac{1}{n^2}\left(H\mathbf{K}H\circ H\mathbf{L}H\circ H\mathbf{M}H\right)_{++}.}$$

Empirical joint central moment in the feature space

Example A: factorisation tests

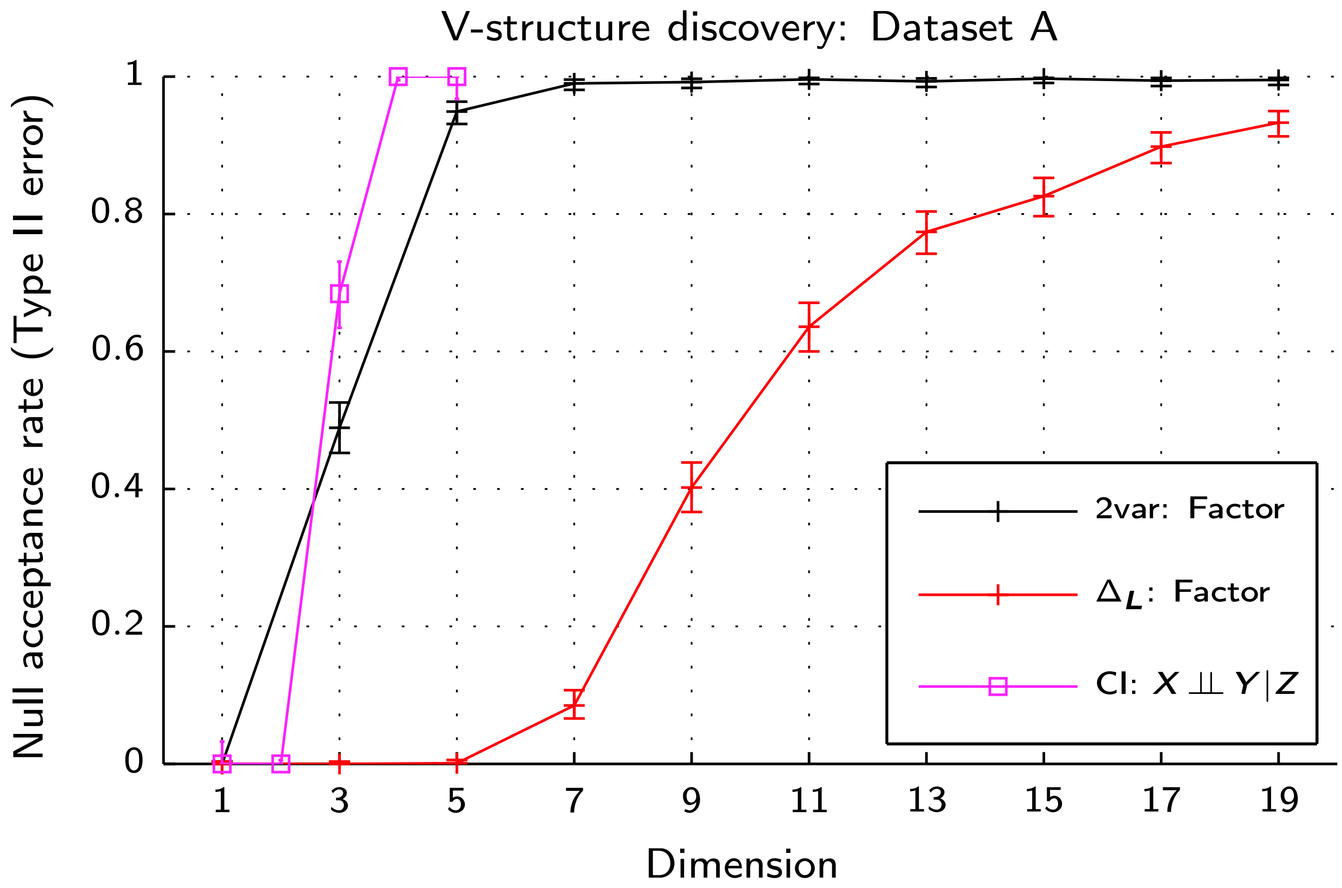


Figure 2: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Example B: Joint dependence can be easier to detect

- $X_1, Y_1 \stackrel{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $Z_1 = \begin{cases} X_1^2 + \epsilon, & w.p.\ 1/3, \\ Y_1^2 + \epsilon, & w.p.\ 1/3, \\ X_1 Y_1 + \epsilon, & w.p.\ 1/3, \end{cases}$ where $\epsilon \sim \mathcal{N}(0, 0.1^2)$.
- $X_{2:p}, Y_{2:p}, Z_{2:p} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$
- dependence of Z on pair (X, Y) is stronger than on X and Y individually
- Satisfies faithfulness

Example B: factorisation tests

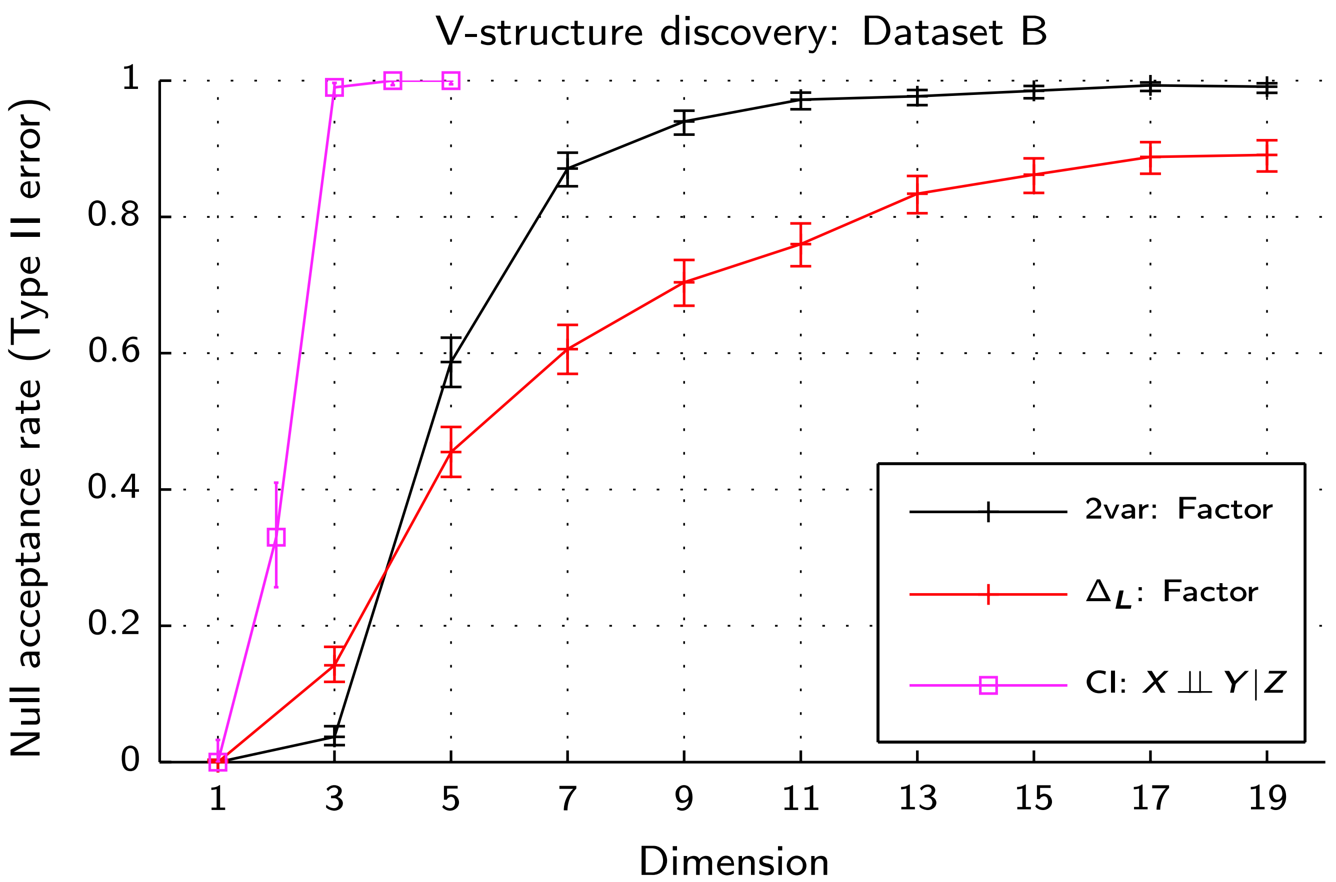


Figure 3: Factorisation hypothesis: Lancaster statistic vs. a two-variable based test (both with **HB** correction); Test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al (2011), $n = 500$

Interaction for $D \geq 4$

- Interaction measure valid for all D (Streitberg, 1990):

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} \left(|\pi| - 1\right)! J_\pi P$$

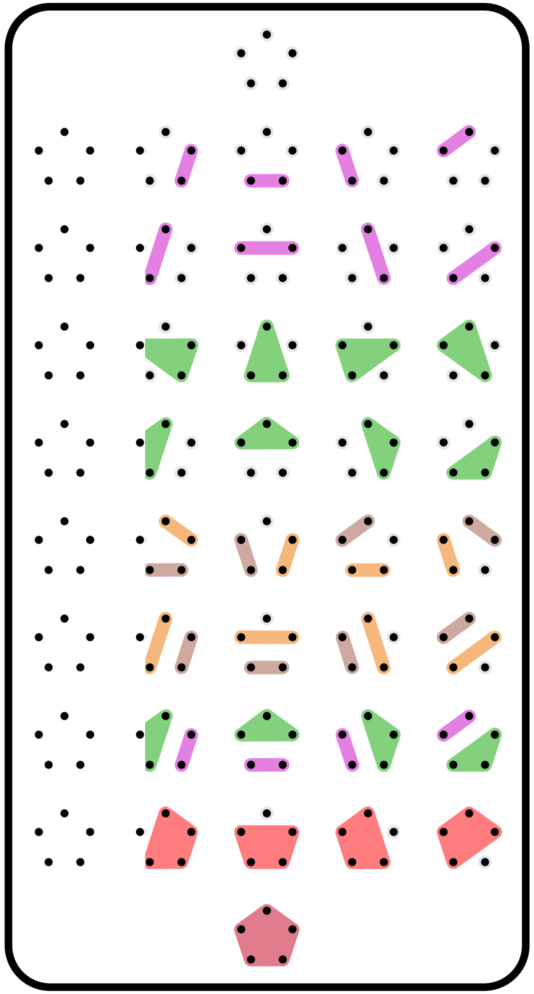
  – For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

Interaction for $D \geq 4$

- Interaction measure valid for all D (Streitberg, 1990):

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_\pi P$$

  – For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

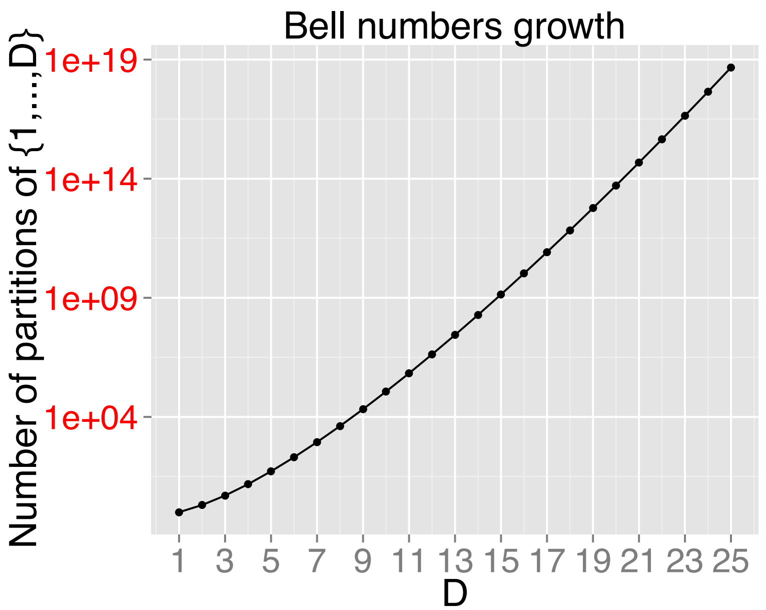


Interaction for $D \geq 4$

- Interaction measure valid for all D (Streitberg, 1990):

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_\pi P$$

  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.



joint central moments (Lancaster interaction)

vs.

joint cumulants (Streitberg interaction)

Total independence test

- Total independence test:
  $\mathbf{H_0} : P_{XYZ} = P_X P_Y P_Z$ vs. $\mathbf{H}_1 : P_{XYZ} \neq P_X P_Y P_Z$

Total independence test

- Total independence test:
  $\mathbf{H_0} : P_{XYZ} = P_X P_Y P_Z$ vs. $\mathbf{H}_1 : P_{XYZ} \neq P_X P_Y P_Z$

- For $(X_1, \ldots, X_D) \sim P_{\mathbf{X}}$, and $\kappa = \bigotimes_{i=1}^{D} k^{(i)}$:

$$\left\| \mu_\kappa \underbrace{\left(\hat{P}_{\mathbf{X}} - \prod_{i=1}^{D} \hat{P}_{X_i} \right)}_{\Delta_{tot} \hat{P}} \right\|_{\mathcal{H}_\kappa}^2 = \frac{1}{n^2} \sum_{a=1}^{n} \sum_{b=1}^{n} \prod_{i=1}^{D} K_{ab}^{(i)} - \frac{2}{n^{D+1}} \sum_{a=1}^{n} \prod_{i=1}^{D} \sum_{b=1}^{n} K_{ab}^{(i)} + \frac{1}{n^{2D}} \prod_{i=1}^{D} \sum_{a=1}^{n} \sum_{b=1}^{n} K_{ab}^{(i)}.$$

- Coincides with the test proposed by Kankainen (1995) using empirical characteristic functions: similar relationship to that between dCov and HSIC (DS et al, 2013)

Example B: total independence tests

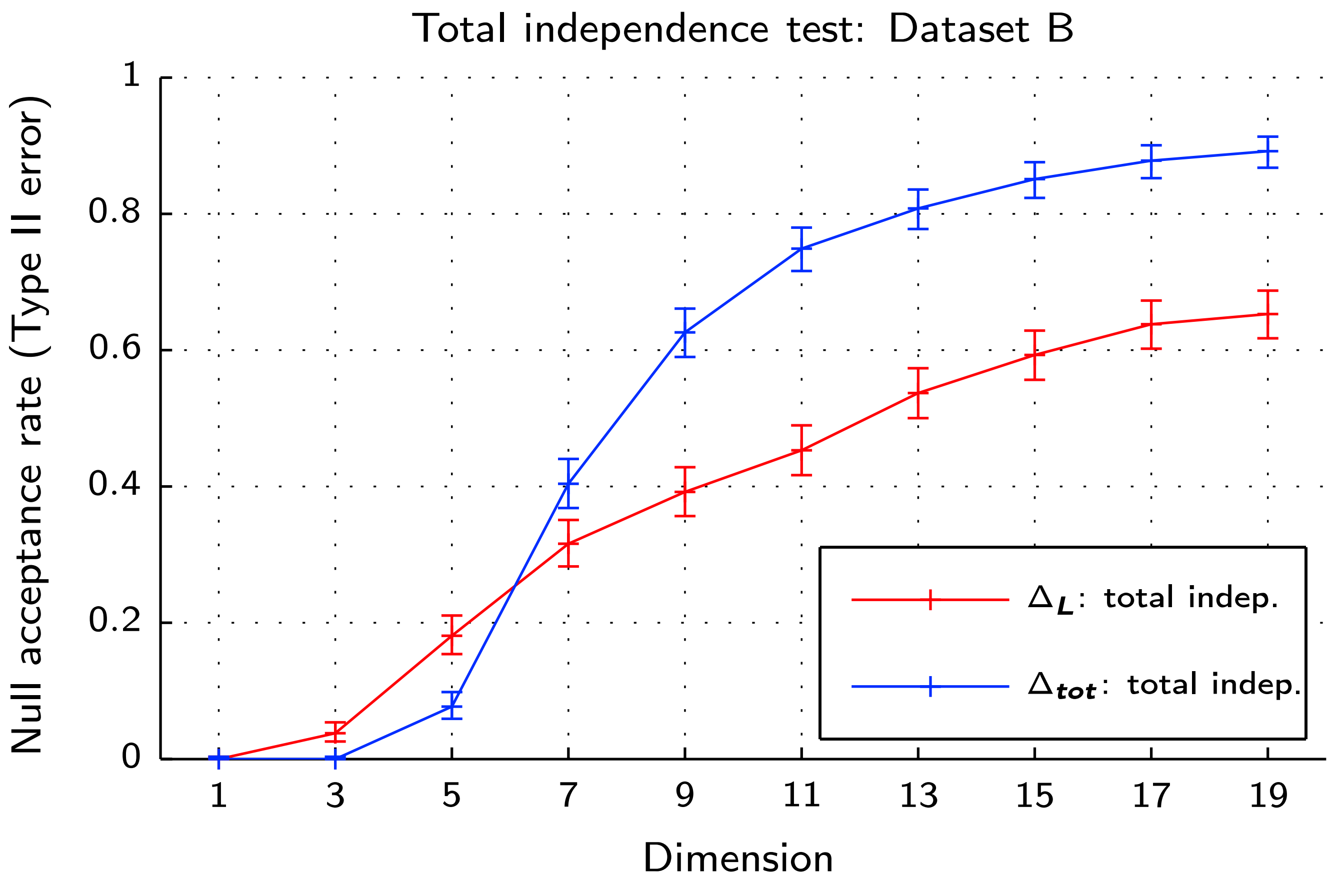


Figure 4: Total independence: $\Delta_{tot}\hat{P}$ vs. $\Delta_L\hat{P}$, $n = 500$

Conclusion

- 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 $MMD(\mathbf{P}_{xy}, \mathbf{P}_x\mathbf{P}_y)$
- Detecting three-way interactions
  - Parents with weak individual influence, strong combined influence
  - Avoid difficult problem of conditional dependence testing
  - Generalization of independence test

Co-authors

- Wicher Bergsma
- Karsten Borgwardt
- Kenji Fukumizu
- Dino Sejdinovic
- Bharath Sriperumbudur
- Bernhard Schoelkopf
- Alex Smola

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

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

Three-variable interaction tests:

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

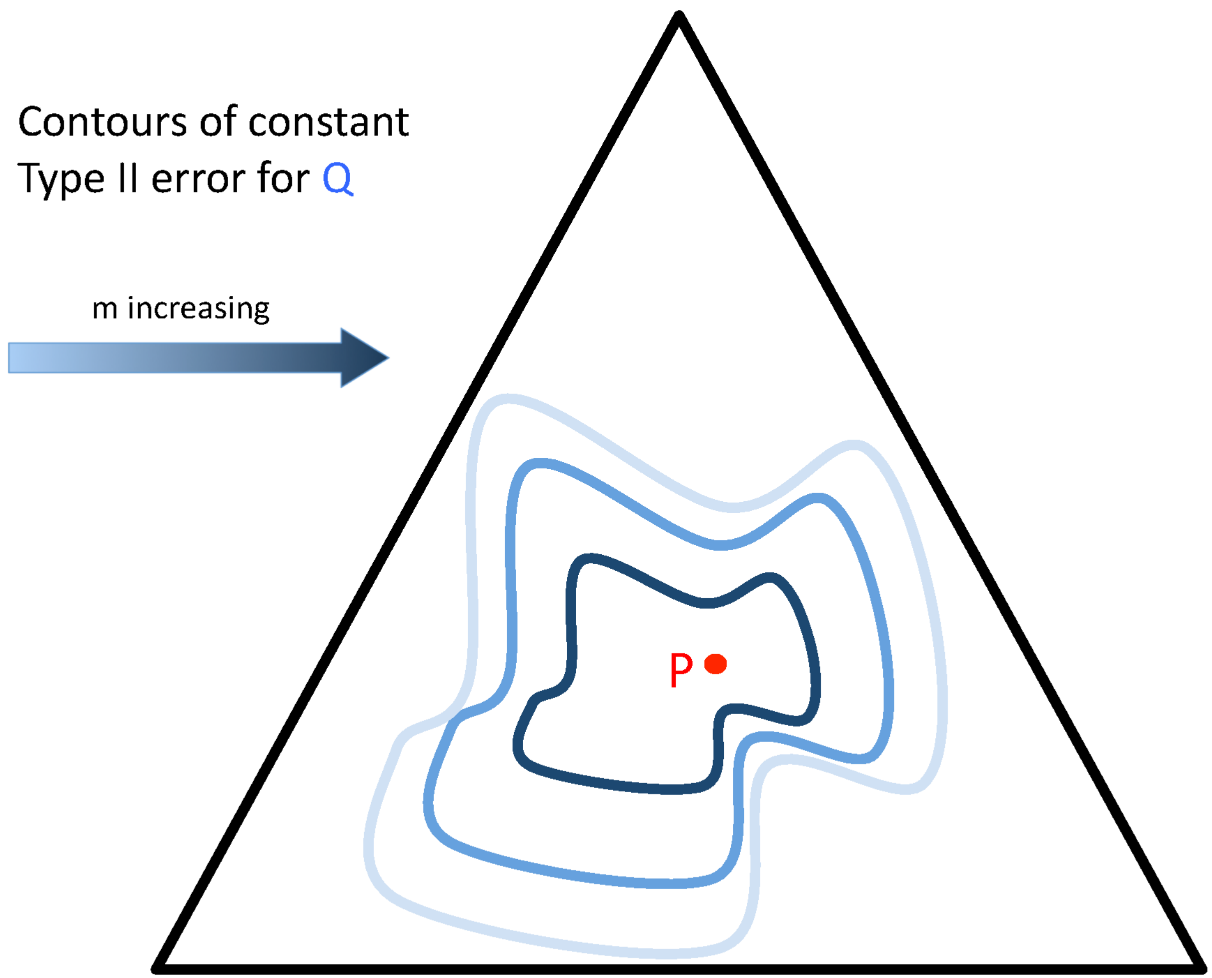
Local departures from the null

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



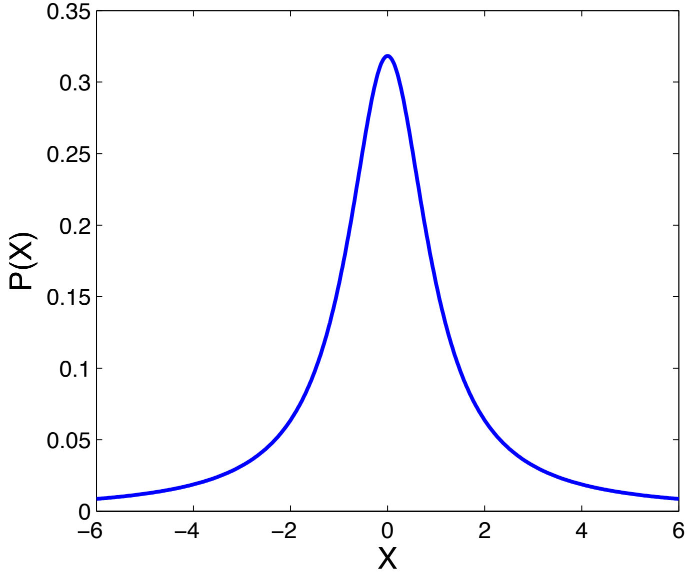
Local departures from the null

What is a hard testing problem?

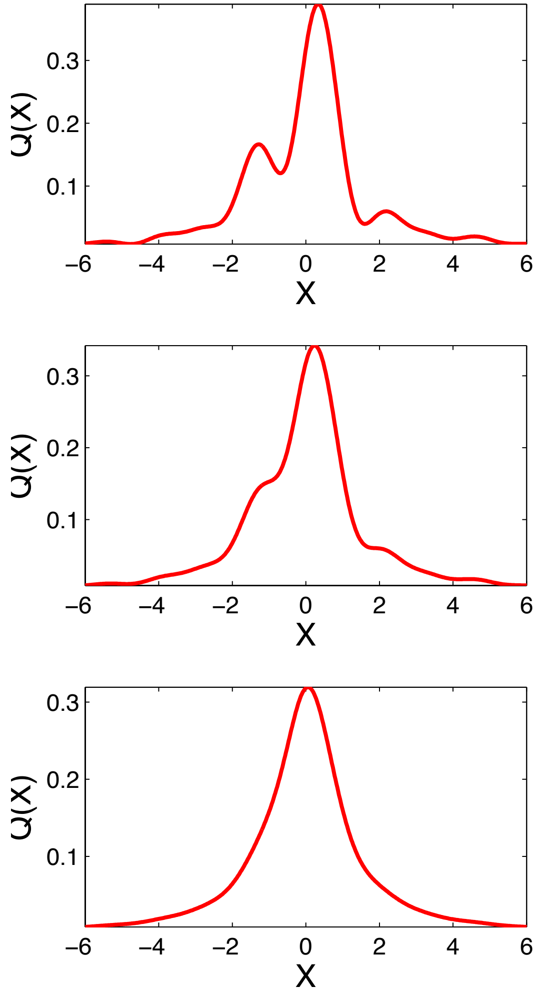
- As m increases, distinguish "closer" **P** and **Q** with same Type II error
- Example: $f_{\mathbf{P}}$ and $f_{\mathbf{Q}}$ probability densities, $f_{\mathbf{Q}} = f_{\mathbf{P}} + \delta g$, where $\delta \in \mathbb{R}$, g some *fixed* function such that $f_{\mathbf{Q}}$ is a valid density
  -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

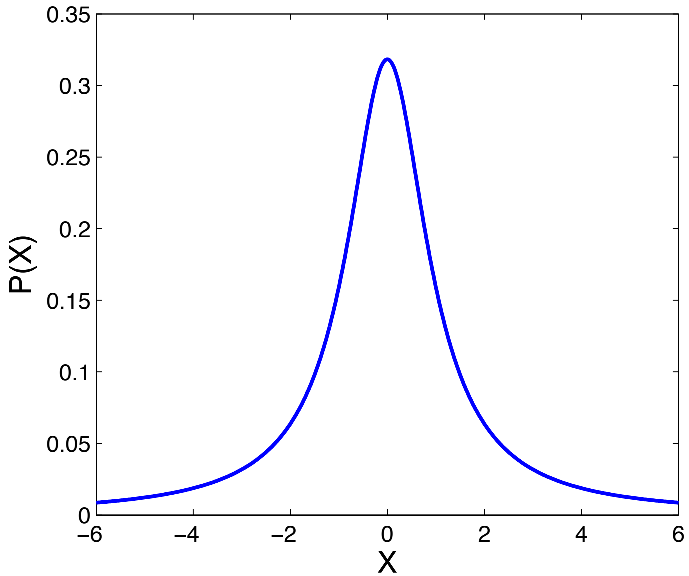
General characterization of local departures from $\mathcal{H}_0$:

- 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]

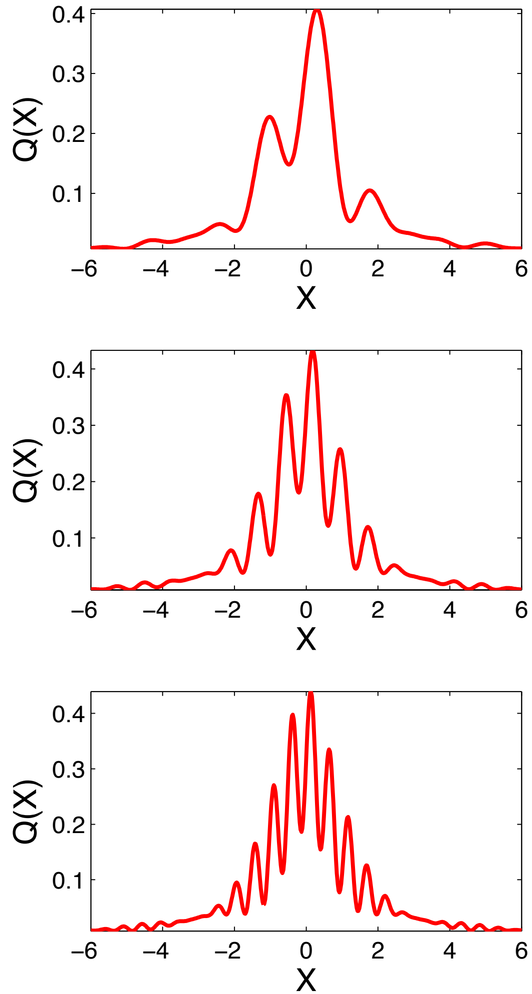
$$\|g_m\|_{\mathcal{F}} = cm^{-1/2}$$

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



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.

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

A. Ihler and D. McAllester. Particle belief propagation. In *AISTATS*, pages 256–263, 2009.

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.

Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. 3-d depth reconstruction from a single still image. *International Journal on Computer Vision*, 76 (1):53–69, 2007.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Proc. Intl. Conference on Artificial Intelligence and Statistics*, volume 10 of *JMLR workshop and conference proceedings*, 2011.

E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In *AISTATS*, pages 781–788, 2010.