Kernel Distribution Embeddings and Applications

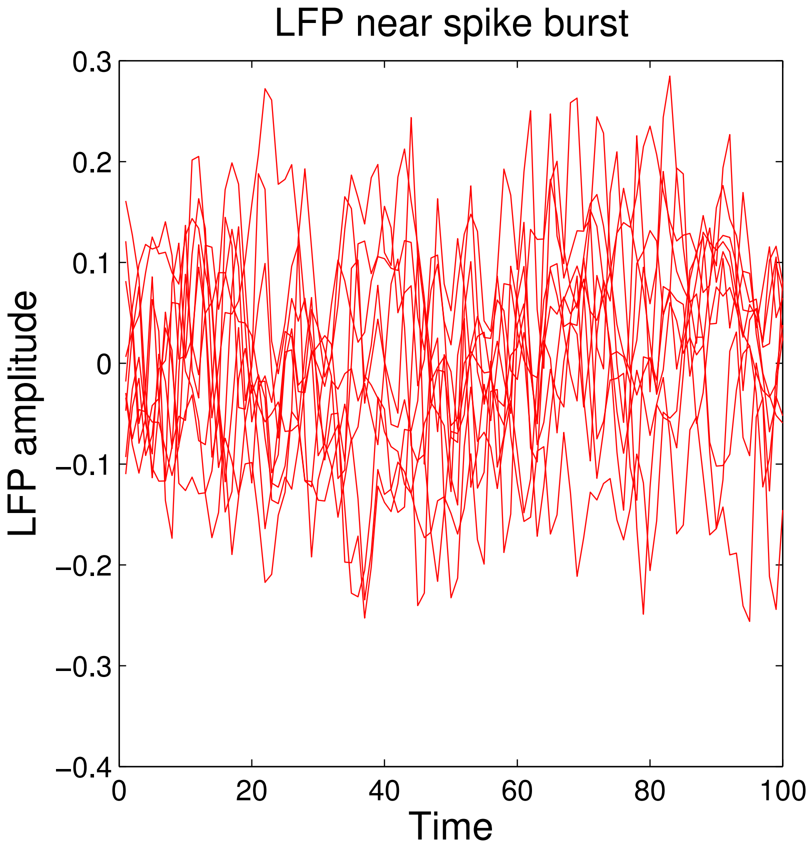
Kernel Methods in Machine Learning

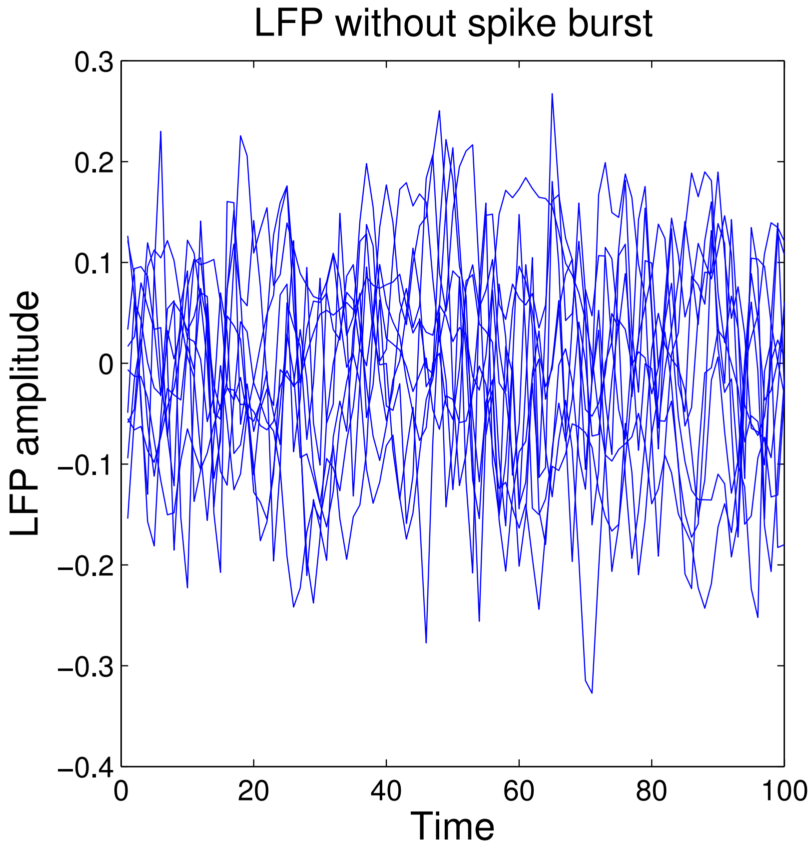
Arthur Gretton

Gatsby Computational Neuroscience Unit

Motivating question: differences in brain signals

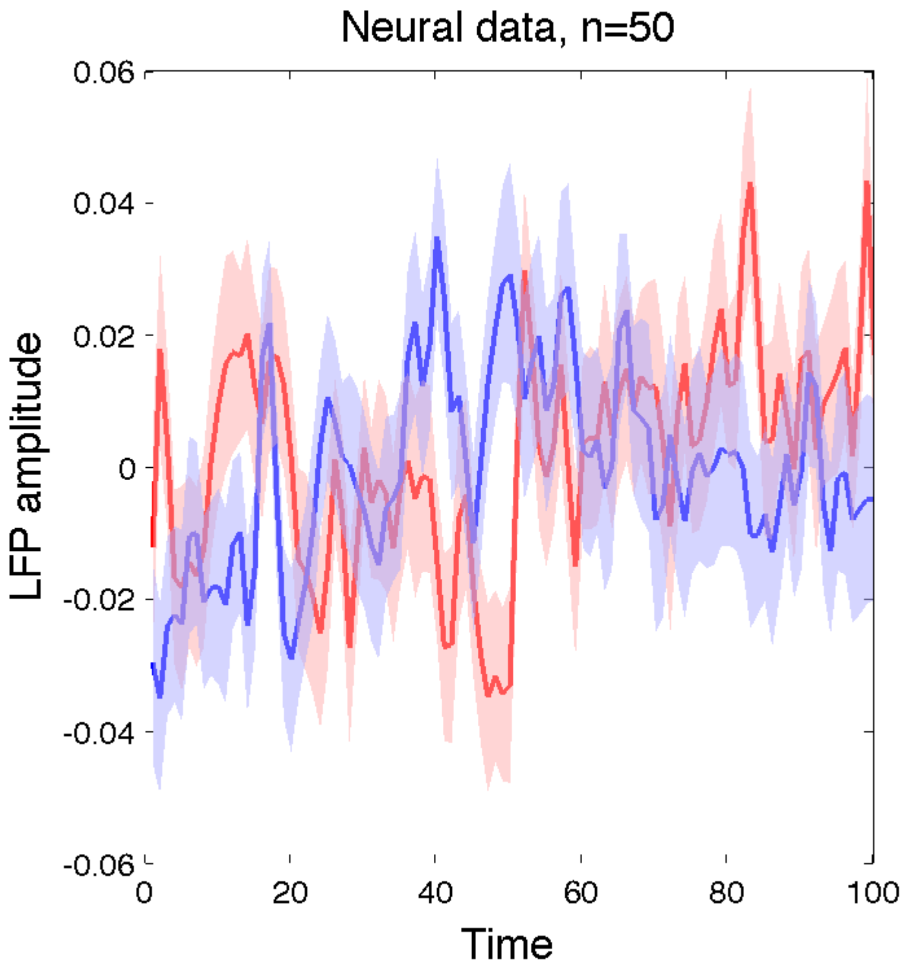
The problem: Do local field potential (LFP) signals change when measured near a spike burst?





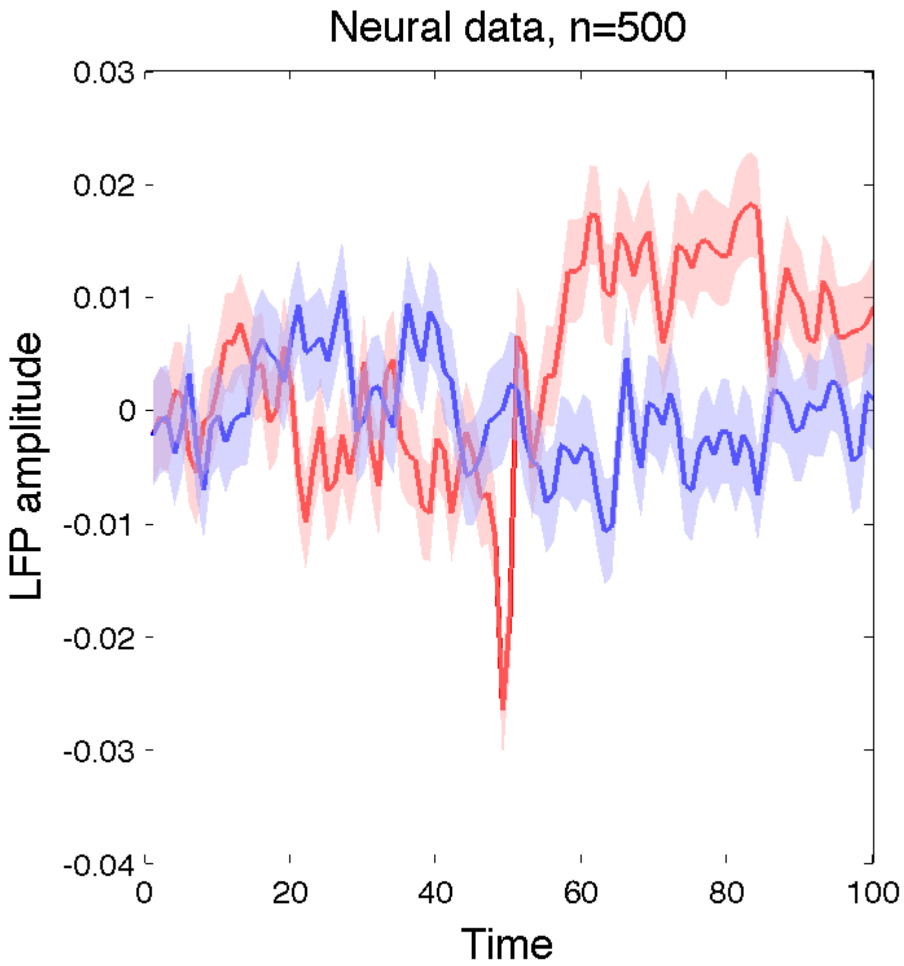
Motivating question: differences in brain signals

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

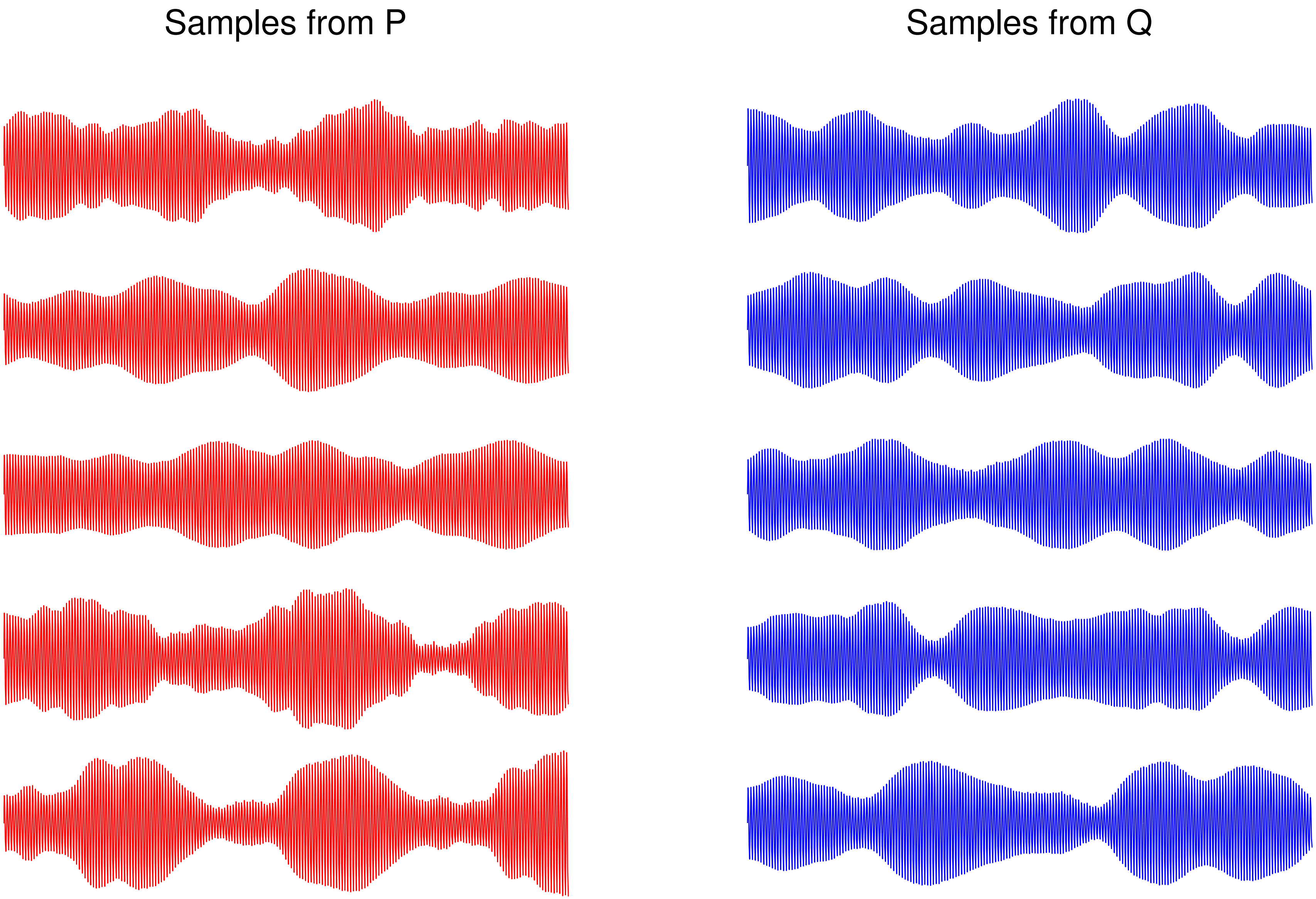


Motivating question: differences in brain signals

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



Motivating example: detect differences in AM signals



Case of discrete domains

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

Case of discrete domains

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

X_1: Now disturbing reports out of Newfoundland show that the fragile snow crab industry is in serious decline. First the west coast salmon, the east coast salmon and the cod, and now the snow crabs off Newfoundland.

X_2: To my pleasant surprise he responded that he had personally visited those wharves and that he had already announced money to fix them. What wharves did the minister visit in my riding and how much additional funding is he going to provide for Delaps Cove, Hampton, Port Lorne,

...

$$P_X \stackrel{?}{=} P_Y$$

Y_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.

Y_2: On the grain transportation system we have had the Estey report and the Kroeger report. We could go on and on. Recently programs have been announced over and over by the government such as money for the disaster in agriculture on the prairies and across Canada.

...

Are the pink extracts from the same distribution as the gray ones?

Another motivating question

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

Another motivating question

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

Another 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

Another 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

Another motivating question

- 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.

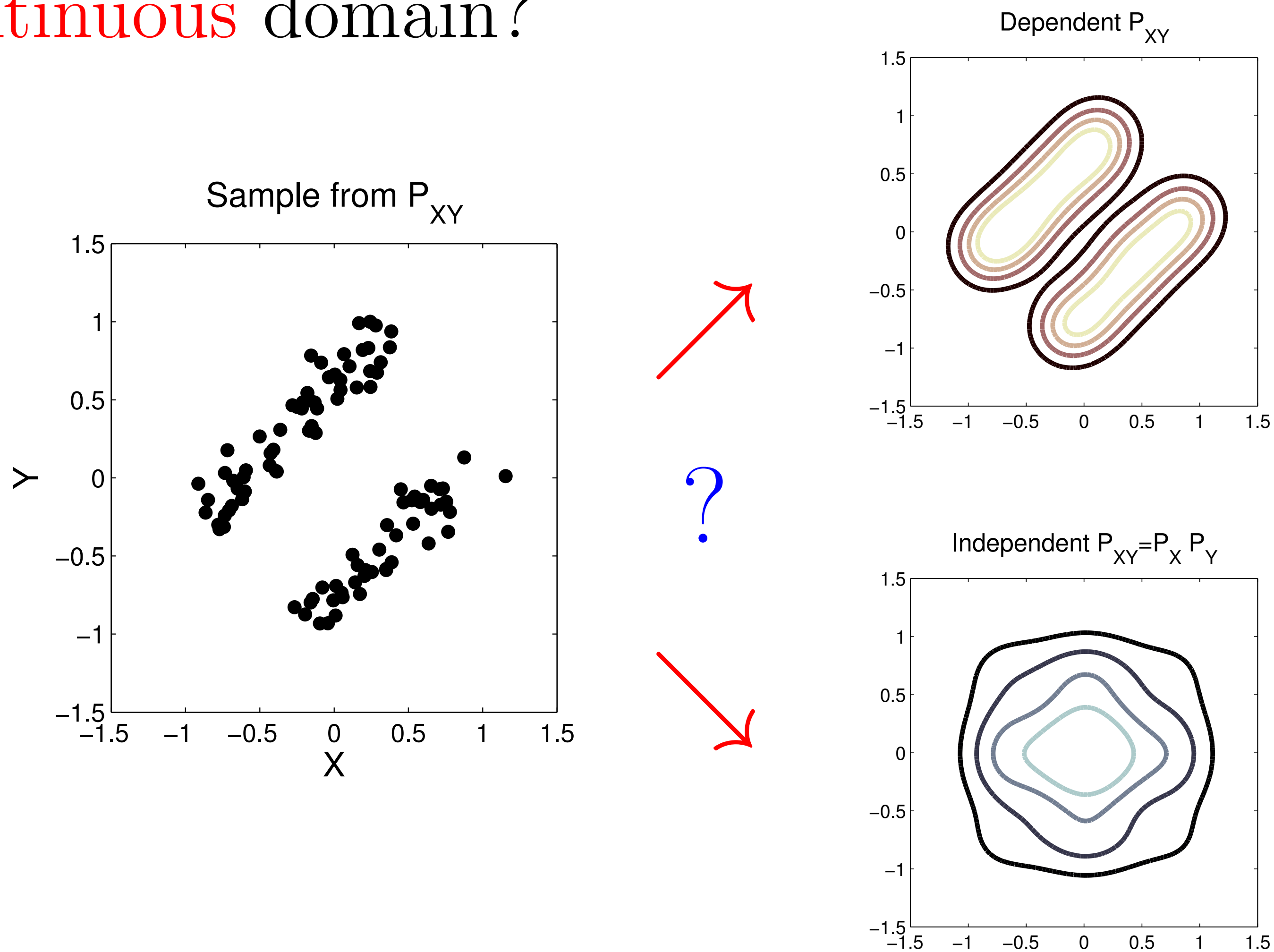
Y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement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

. . .

Are the French text extracts translations of the English ones?

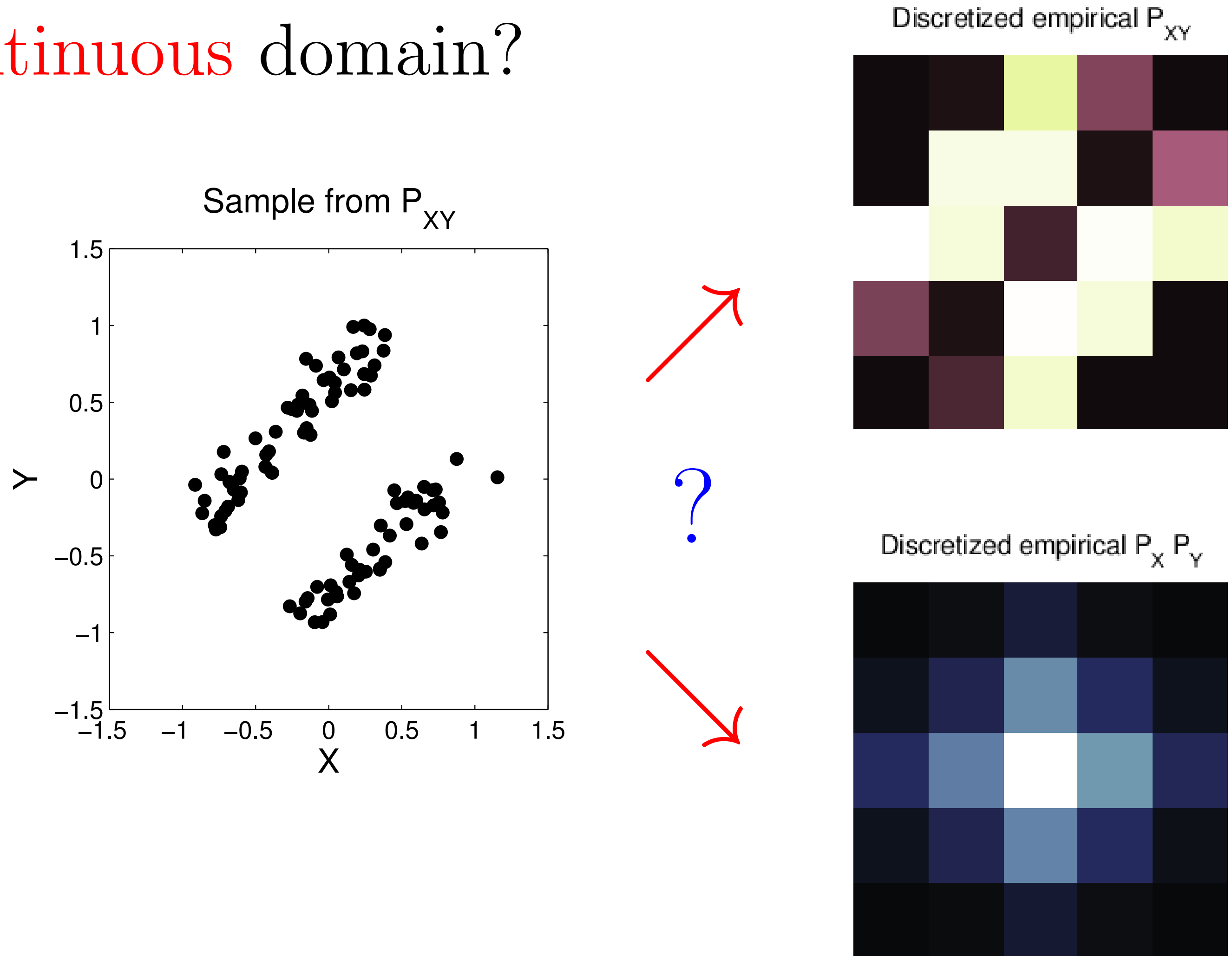
Another motivating question

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



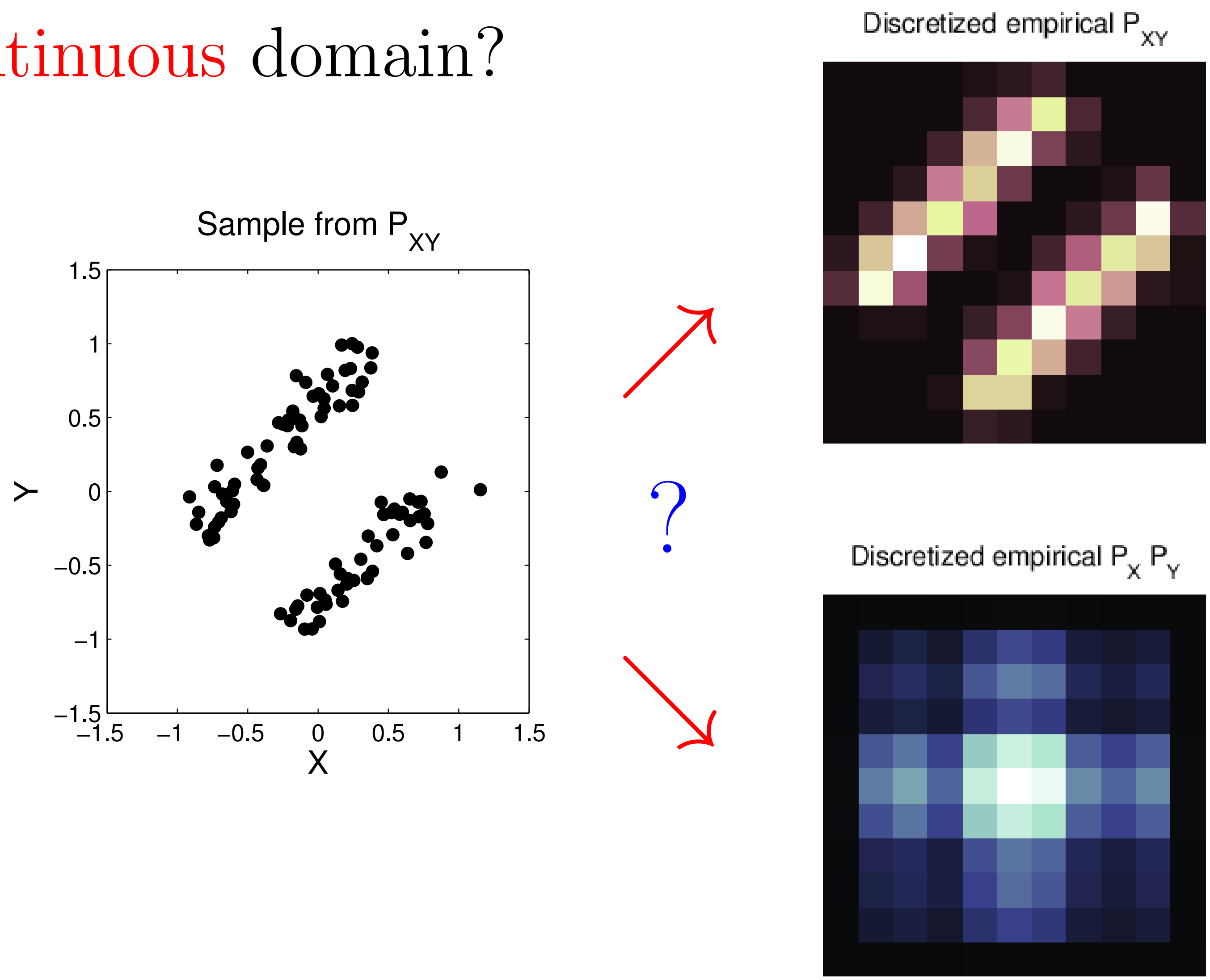
Another motivating question

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



Another motivating question

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



Another motivating question

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

Another motivating question

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

Can we represent and compare distributions in high dimensions?

Further motivating questions

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

Further motivating questions

- Compare distributions with high dimension/ low sample size/ "complex" structure
  - Microarray data (aggregation problem)
  - Neuroscience: naturalistic stimulus, complex response
  - Images and text on web (kernels on structured data)
- Discover structure in high dimensional data
  - Feature selection (microarrays, image and text,...)
  - Low dimensional visualization clustering, taxonomy fitting, max. variance unfolding,...
  - Blind source separation (e.g. ICA)

Outline

- Kernel metric on the space of probability measures
  - Function revealing differences in distributions
  - Distance between means in space of features (RKHS)
  - For which feature spaces are mappings unique?

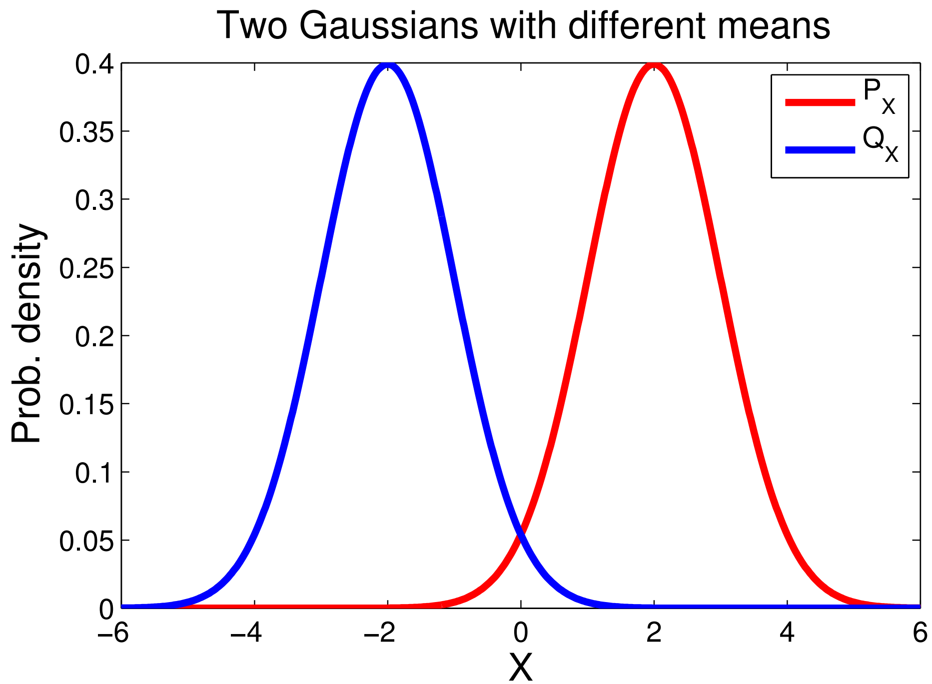
Outline

- Kernel metric on the space of probability measures
  - Function revealing differences in distributions
  - Distance between means in space of features (RKHS)
  - For which feature spaces are mappings unique?
- Dependence detection
  - Covariance and Correlation in feature space

Kernel distance between distributions

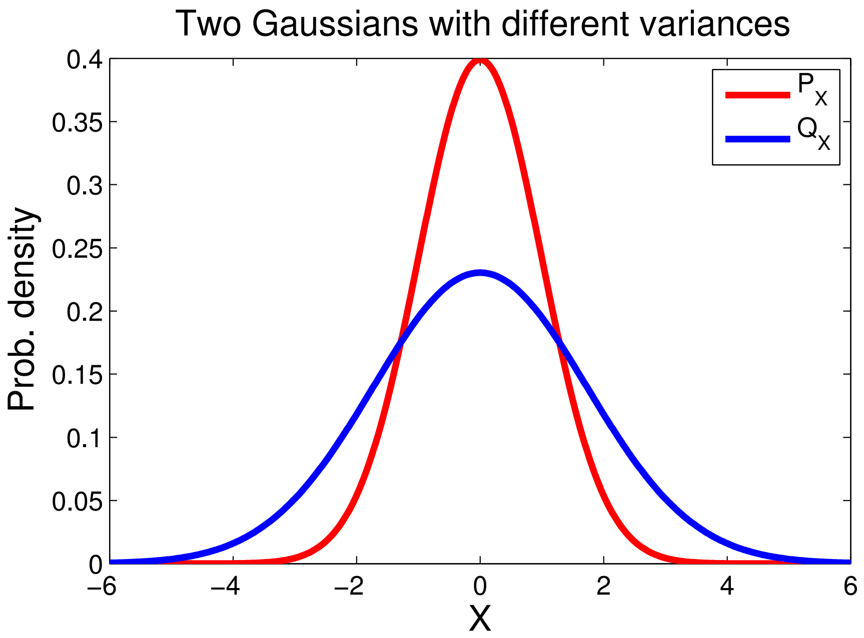
Feature mean difference

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



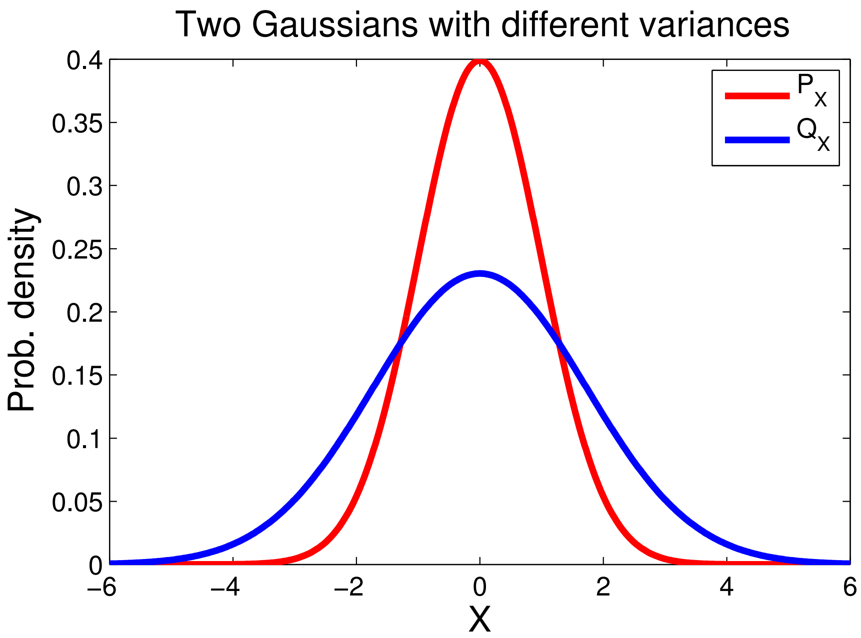
Feature mean difference

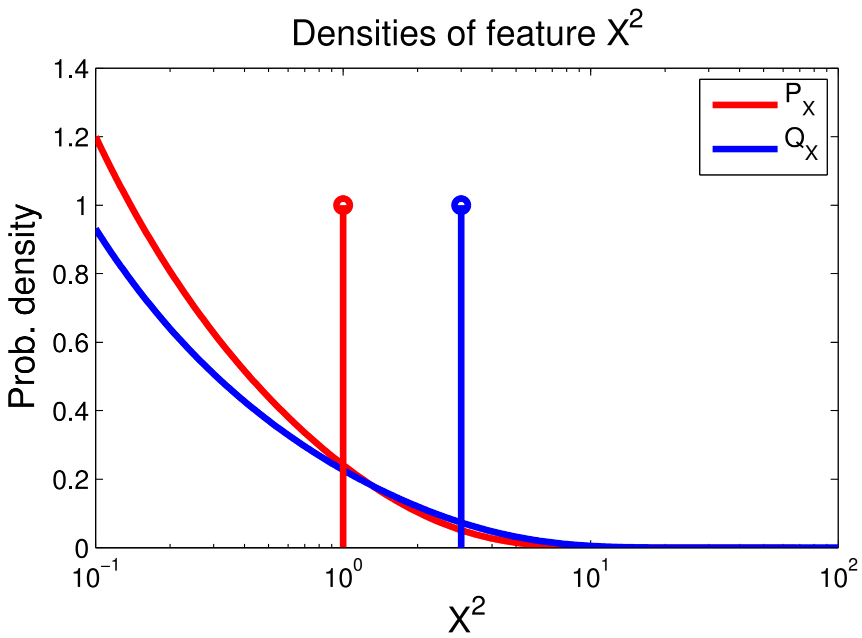
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$



Feature mean difference

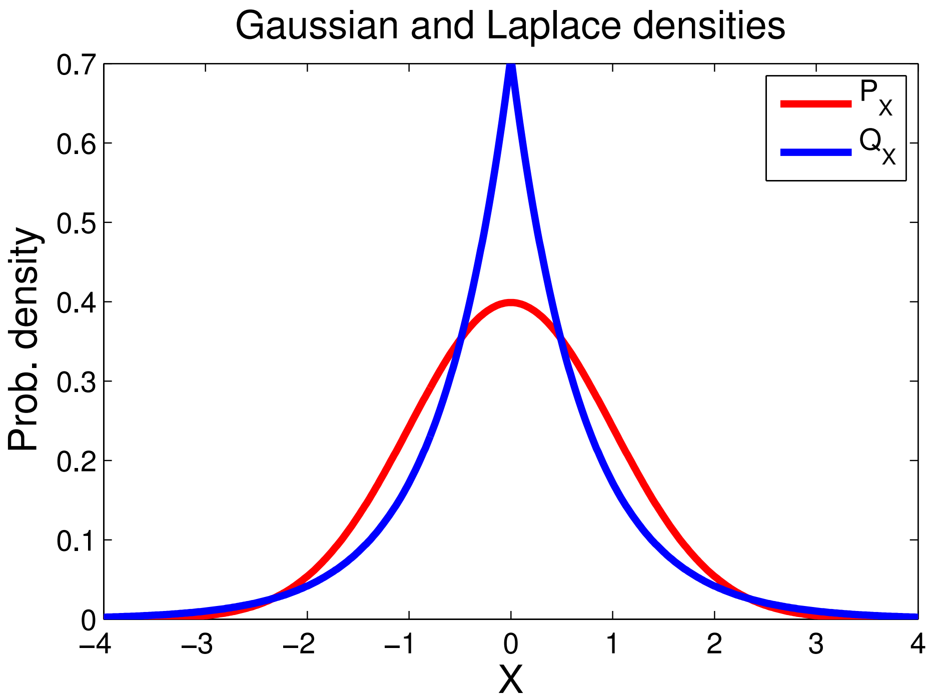
- Two Gaussians with same means, different variance
- Idea: look at difference in means of features of the RVs
- In Gaussian case: second order features of form $\varphi(x) = x^2$





Feature mean difference

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



Reminder: feature maps and the RKHS

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

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

Reminder: feature maps and the RKHS

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

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

- Inner product between feature maps:

$$\left\langle \varphi^{(p)}(x), \varphi^{(p)}(y) \right\rangle_{\mathcal{F}} = \langle x, y \rangle^2$$

$$\left\langle \varphi^{(g)}(x), \varphi^{(g)}(y) \right\rangle_{\mathcal{F}} = \exp\left(-\sigma^{-1} \|x - y\|^2 \right) = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x')$$

Reminder: feature maps and the RKHS

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

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

- Inner product between feature maps:

$$\left\langle \varphi^{(p)}(x), \varphi^{(p)}(y) \right\rangle_{\mathcal{F}} = \langle x, y \rangle^2 \qquad \left\langle \varphi^{(g)}(x), \varphi^{(g)}(y) \right\rangle_{\mathcal{F}} = \exp\left(-\sigma^{-1} \|x - y\|^2\right)$$

$$= \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x')$$

- 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

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: $\forall f \in \mathcal{F}$,

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

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

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

Feature embeddings of probabilities

The kernel trick:

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

The mean trick:

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

Empirical mean embedding:

$$\widehat{\mu}_{\mathsf{P}} = m^{-1} \sum_{i=1}^{m} \varphi_{x_i} \qquad x_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}$$

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

...but does this reasoning work in infinite dimensions?

Does the feature space mean exist?

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

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

Does the feature space mean exist?

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

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

We recall the concept of a bounded operator: a linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

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

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

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

Does the feature space mean exist?

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

Proof:

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

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

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

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

μ_{P} is feature map of probability

Embedding of P to feature space

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

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

- What does prob. feature map look like?

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

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

Expectation of kernel!

- Empirical estimate:

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

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

Embedding of **P** to feature space

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

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

- What does prob. feature map look like?

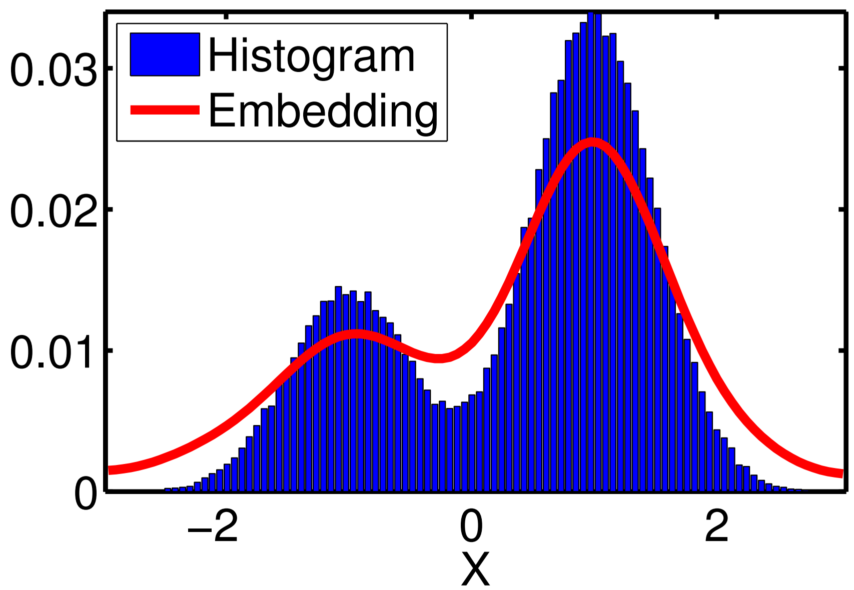
$$\mu_{\mathbf{P}}(x) = \langle \mu_{\mathbf{P}}(\cdot), \varphi(x) \rangle_{\mathcal{F}}$$

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

Expectation of kernel!

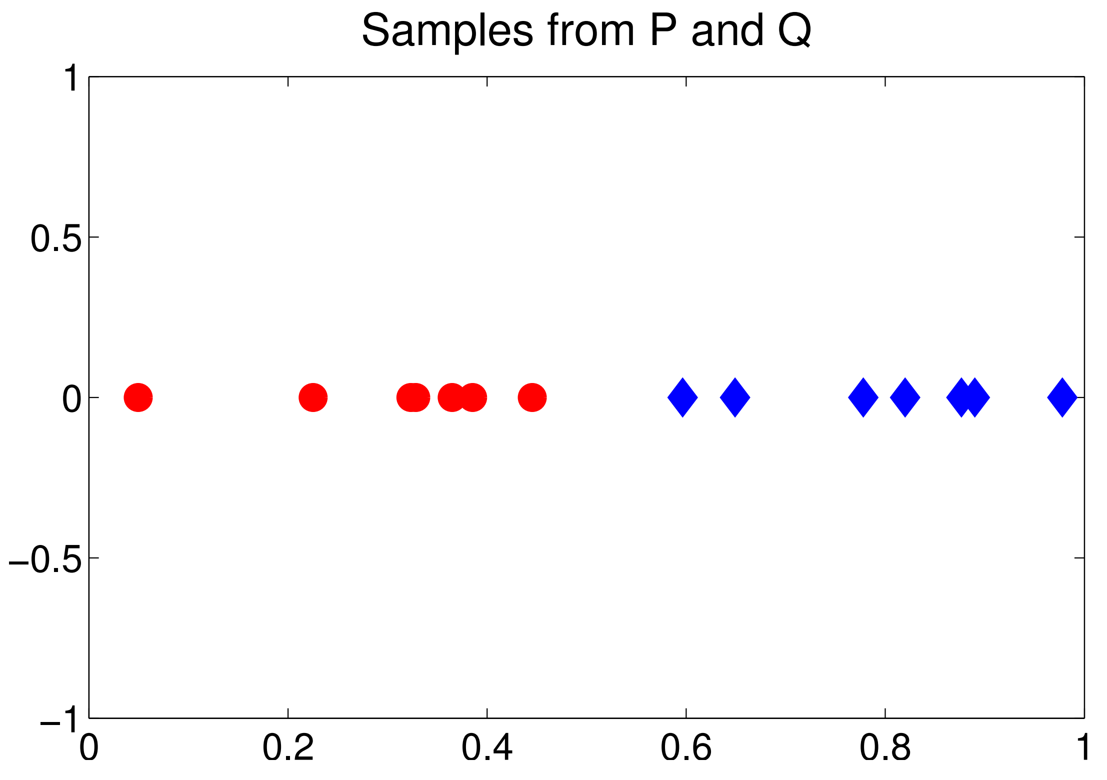
- Empirical estimate:

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



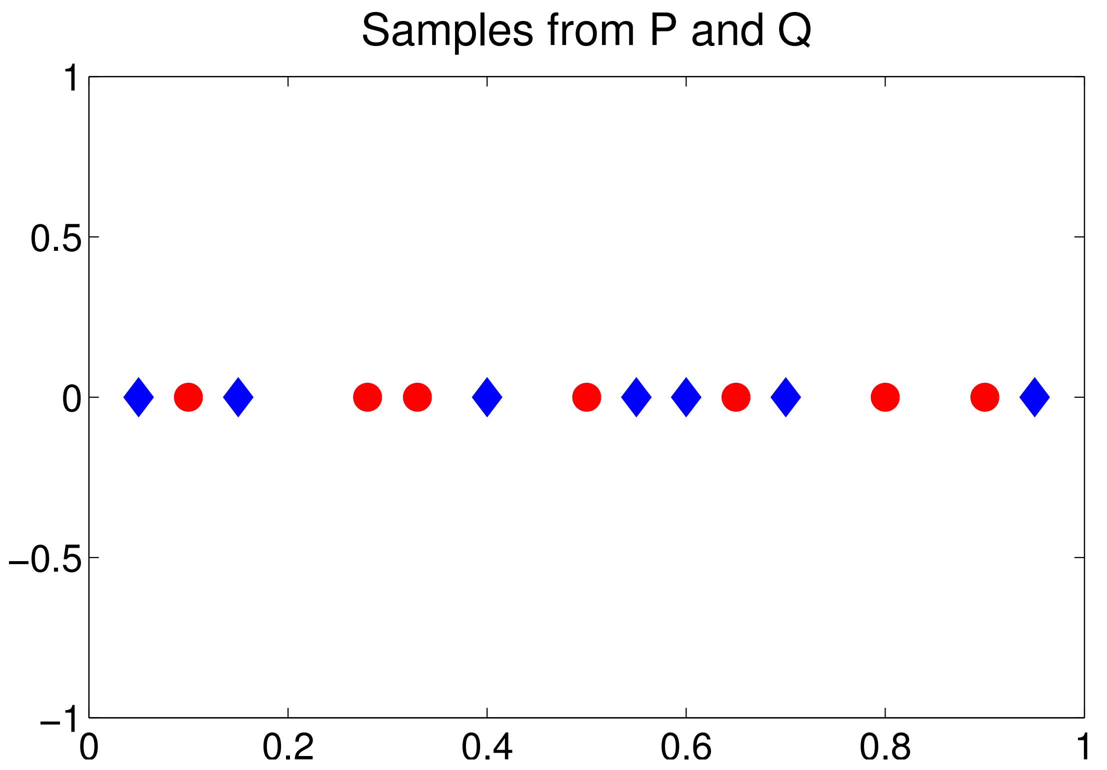
Function Showing Difference in Distributions

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



Function Showing Difference in Distributions

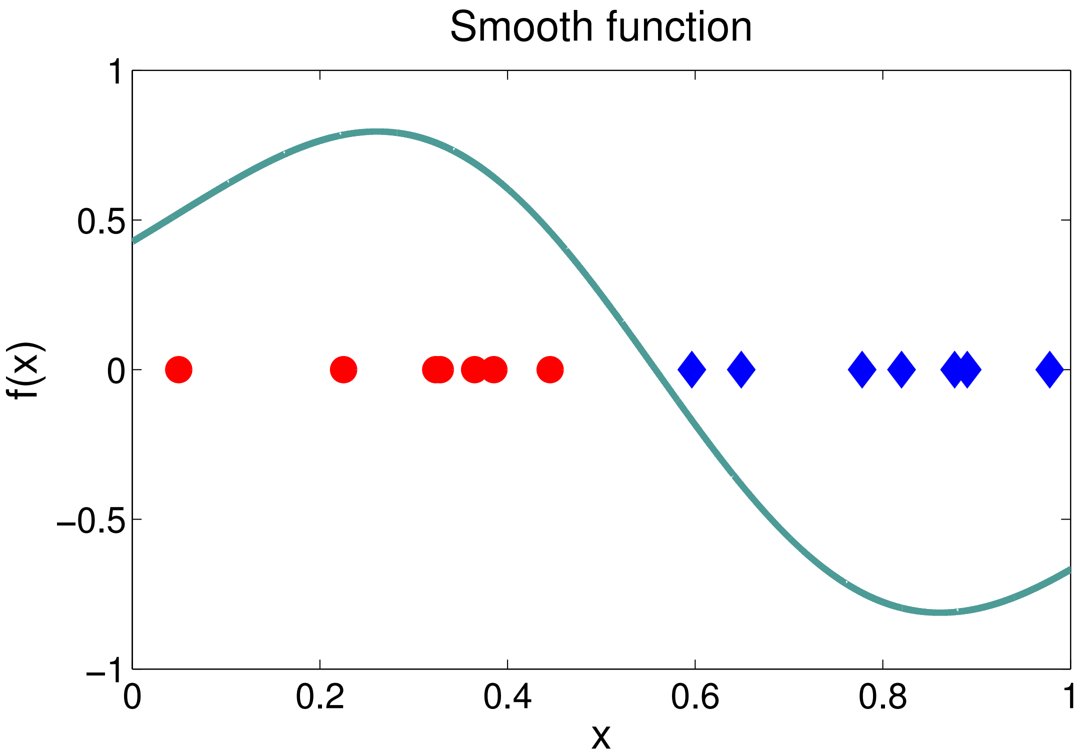
- Are **P** and **Q** different?



Function Showing Difference in Distributions

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

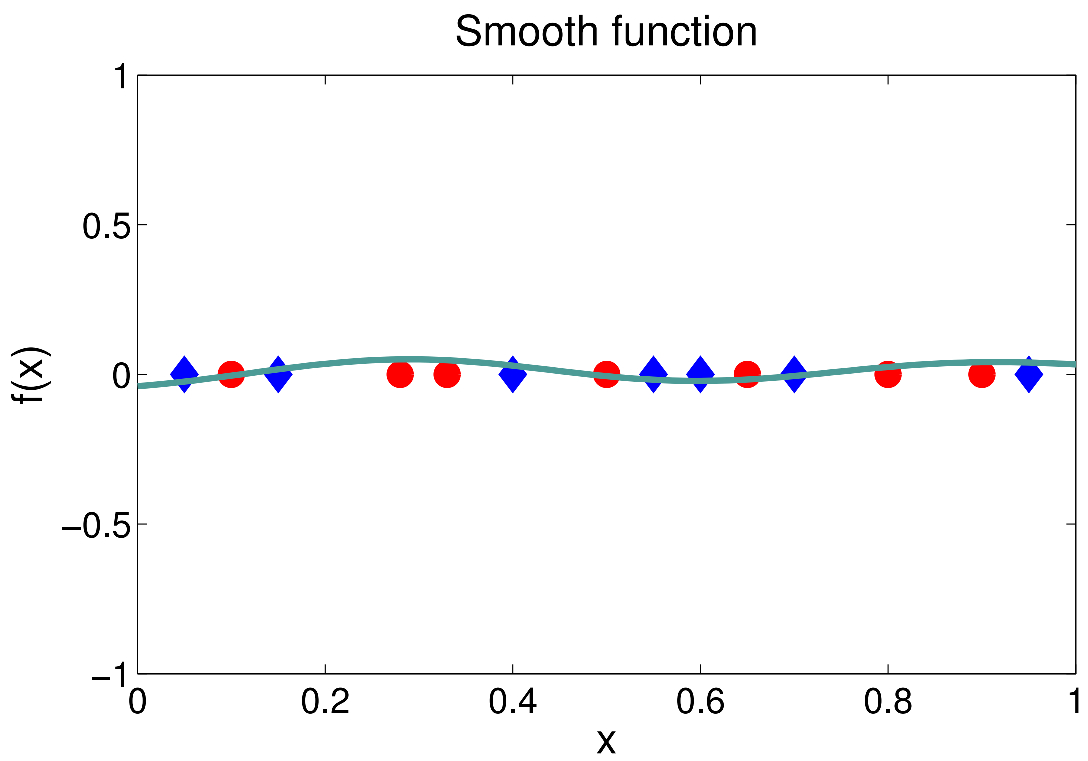
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathsf{y})\right].$$



Function Showing Difference in Distributions

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

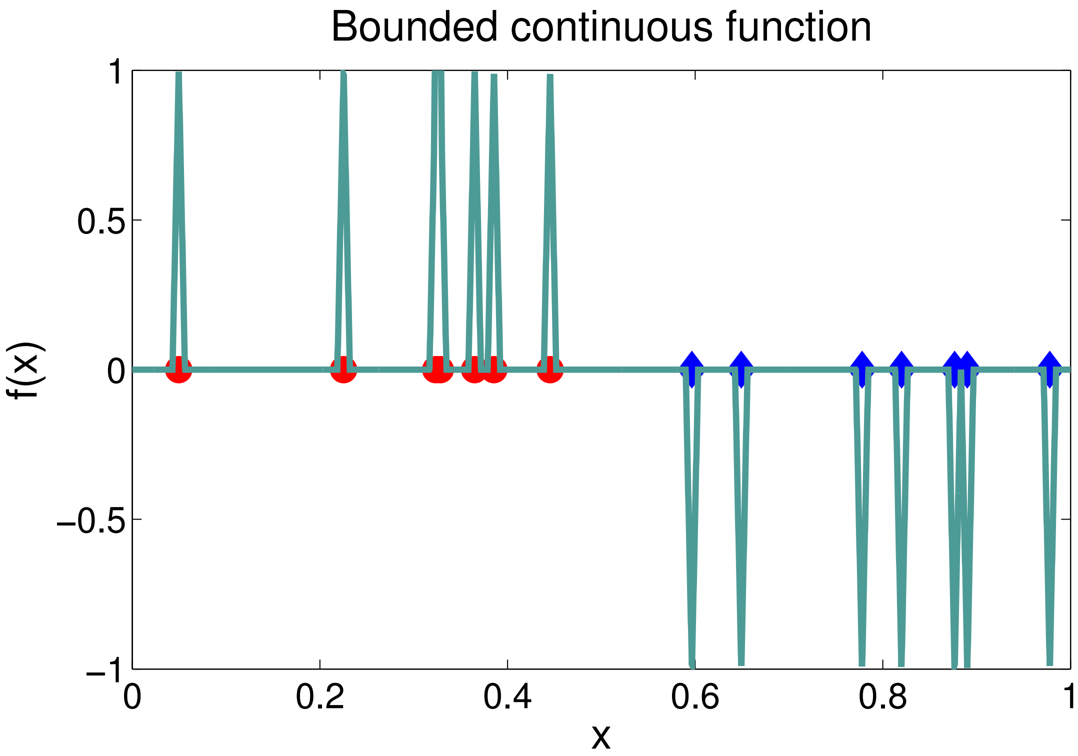
$$\mathrm{MMD}(\mathbf{P},\mathbf{Q};F) := \sup_{f\in F}\left[\mathbf{E}_{\mathbf{P}}\mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}}\mathbf{f}(\mathsf{y})\right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

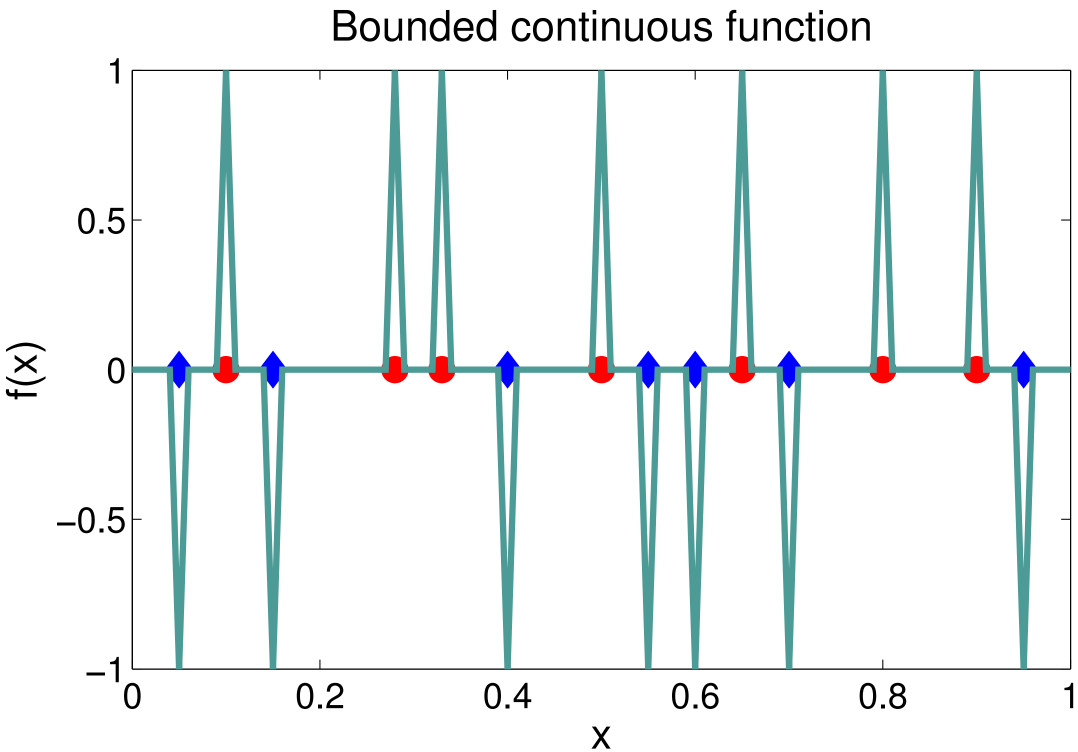
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \sup_{f \in F} \left[\mathbf{E}_{\mathbf{P}} \mathbf{f}(\mathsf{x}) - \mathbf{E}_{\mathbf{Q}} \mathbf{f}(\mathsf{y}) \right].$$



Function Showing Difference in Distributions

- What if the function is not smooth?

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

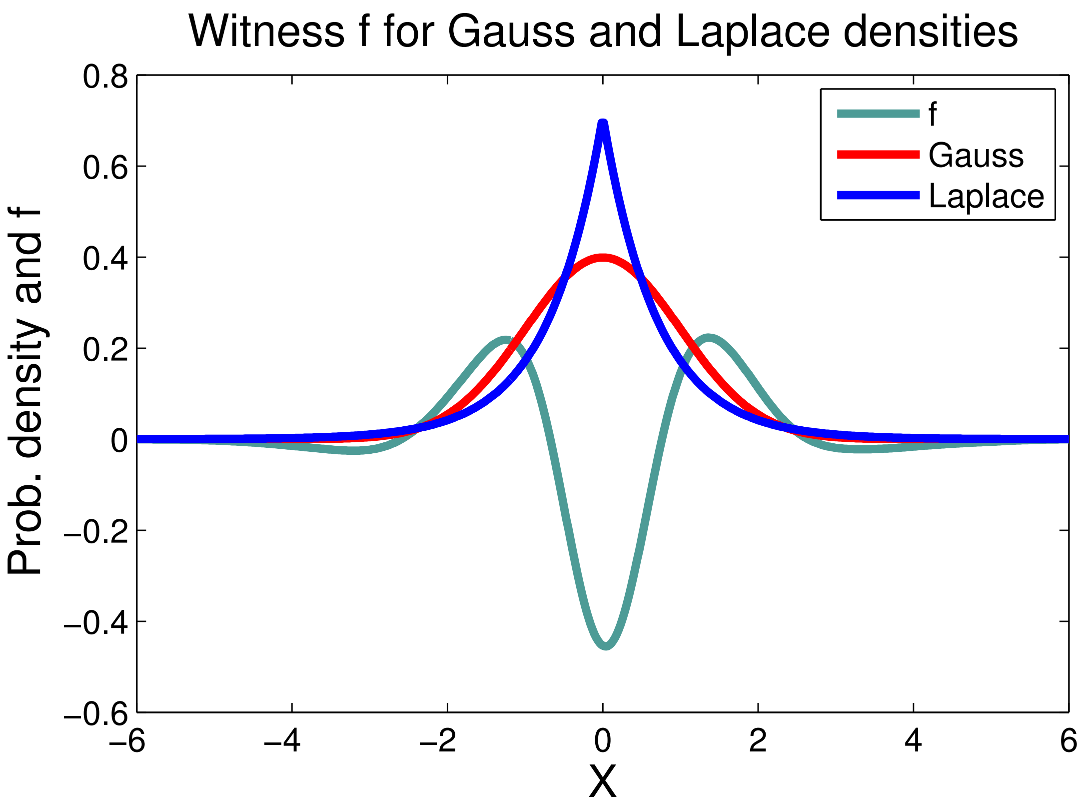


Function Showing Difference in Distributions

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

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

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



Function Showing Difference in Distributions

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

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

- Classical results: $\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) = 0$ iff $\mathbf{P} = \mathbf{Q}$, when
  - F =bounded continuous [Dudley, 2002]
  - F = bounded variation 1 (Kolmogorov metric) [Müller, 1997]
  - F = bounded Lipschitz (Earth mover's distances) [Dudley, 2002]

Function Showing Difference in Distributions

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

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

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

Function Showing Difference in Distributions

- Maximum mean discrepancy: smooth function for P vs Q

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

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

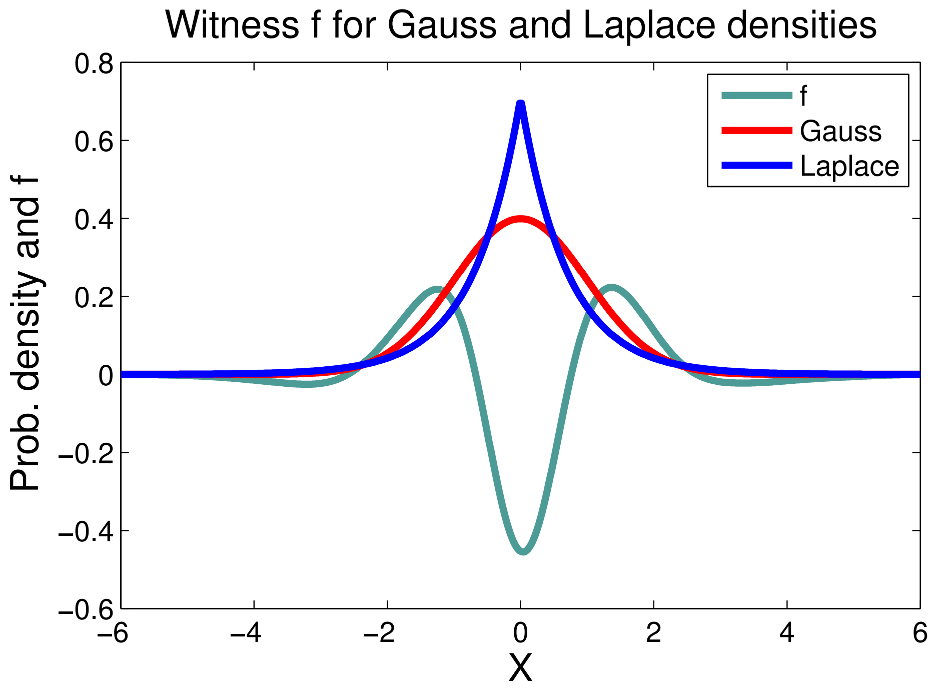
$\boxed{\text{How do smooth functions relate to feature maps?}}$

Function view vs feature mean view

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

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

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



Function view vs feature mean view

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

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

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

use

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

Function view vs feature mean view

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

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

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

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

use

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

Function view vs feature mean view

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

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

use

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

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

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

Empirical estimate of MMD

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

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

Empirical estimate of MMD

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

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

- Proof:

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

Empirical estimate of MMD

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

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

- Proof:

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

Empirical estimate of MMD

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

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

- Proof:

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

Empirical estimate of MMD

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

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

- Proof:

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

Empirical estimate of MMD

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

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

- Proof:

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

Empirical estimate of MMD

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

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

- Proof:

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

Empirical estimate of MMD

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

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

- Proof:

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

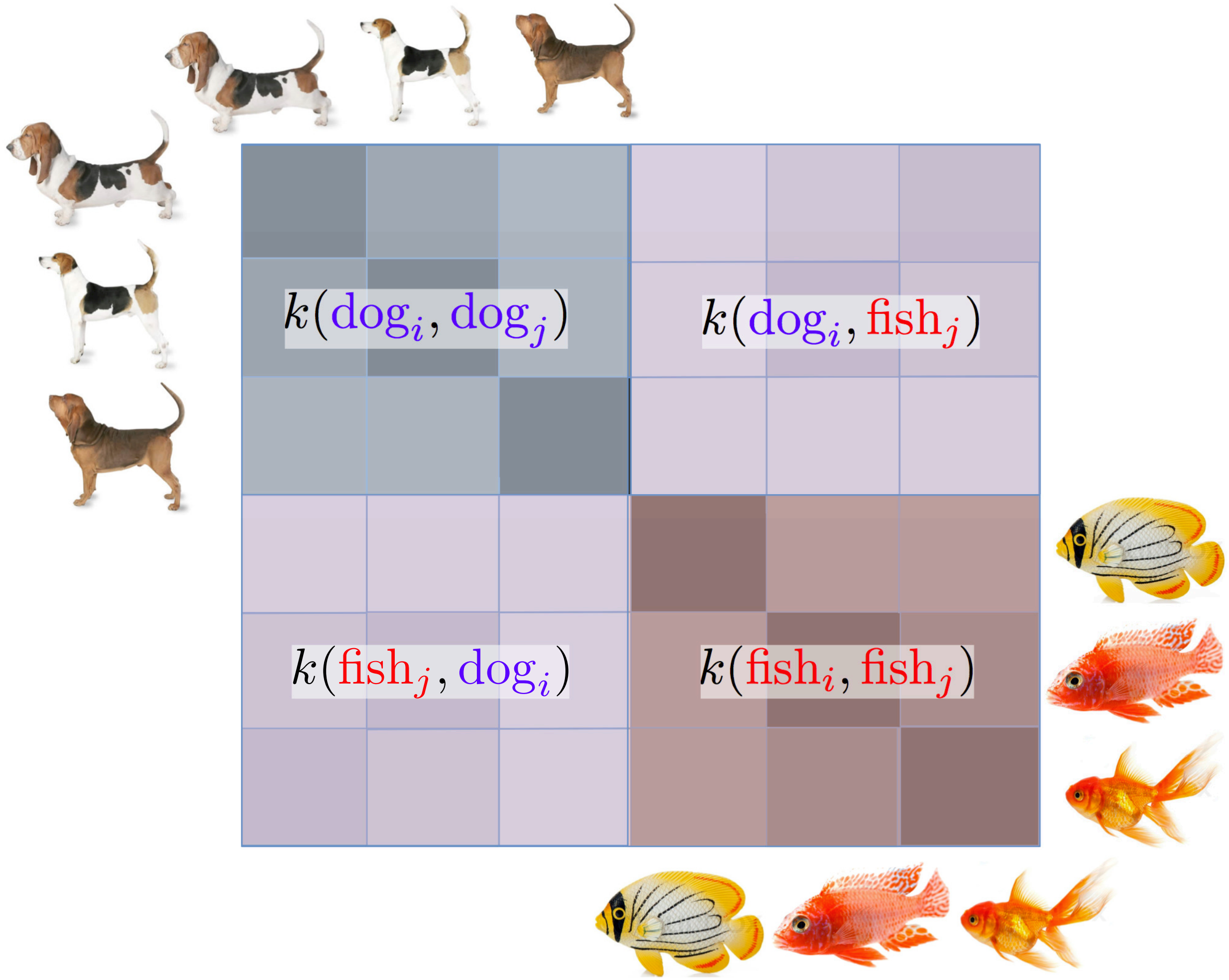
Then $\widehat{\mathbf{E}} k(\mathsf{x}, \mathsf{x}') = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j\neq i}^m k(x_i, x_j)$

The maximum mean discrepancy

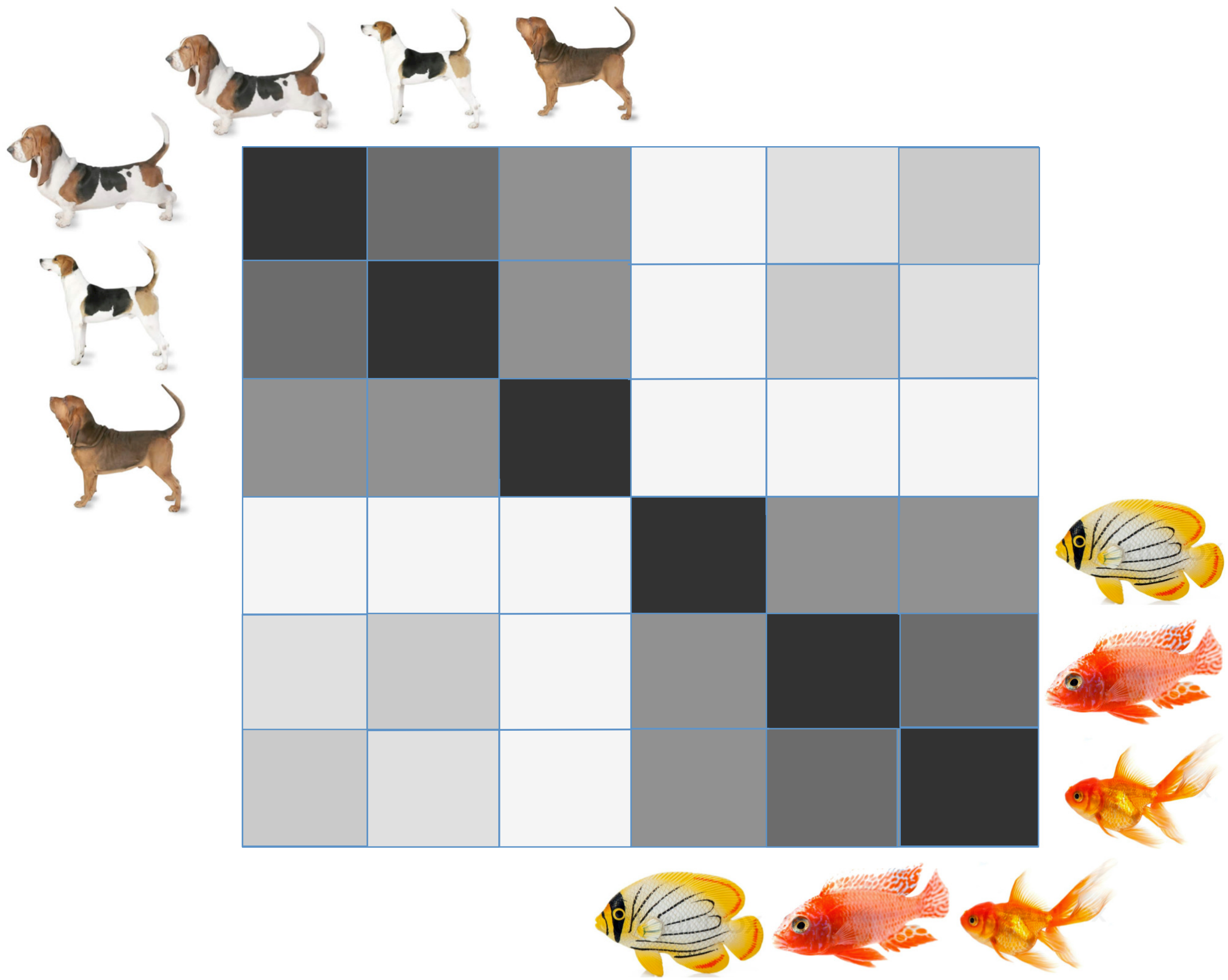
$$\sim P$$

$$\sim Q$$

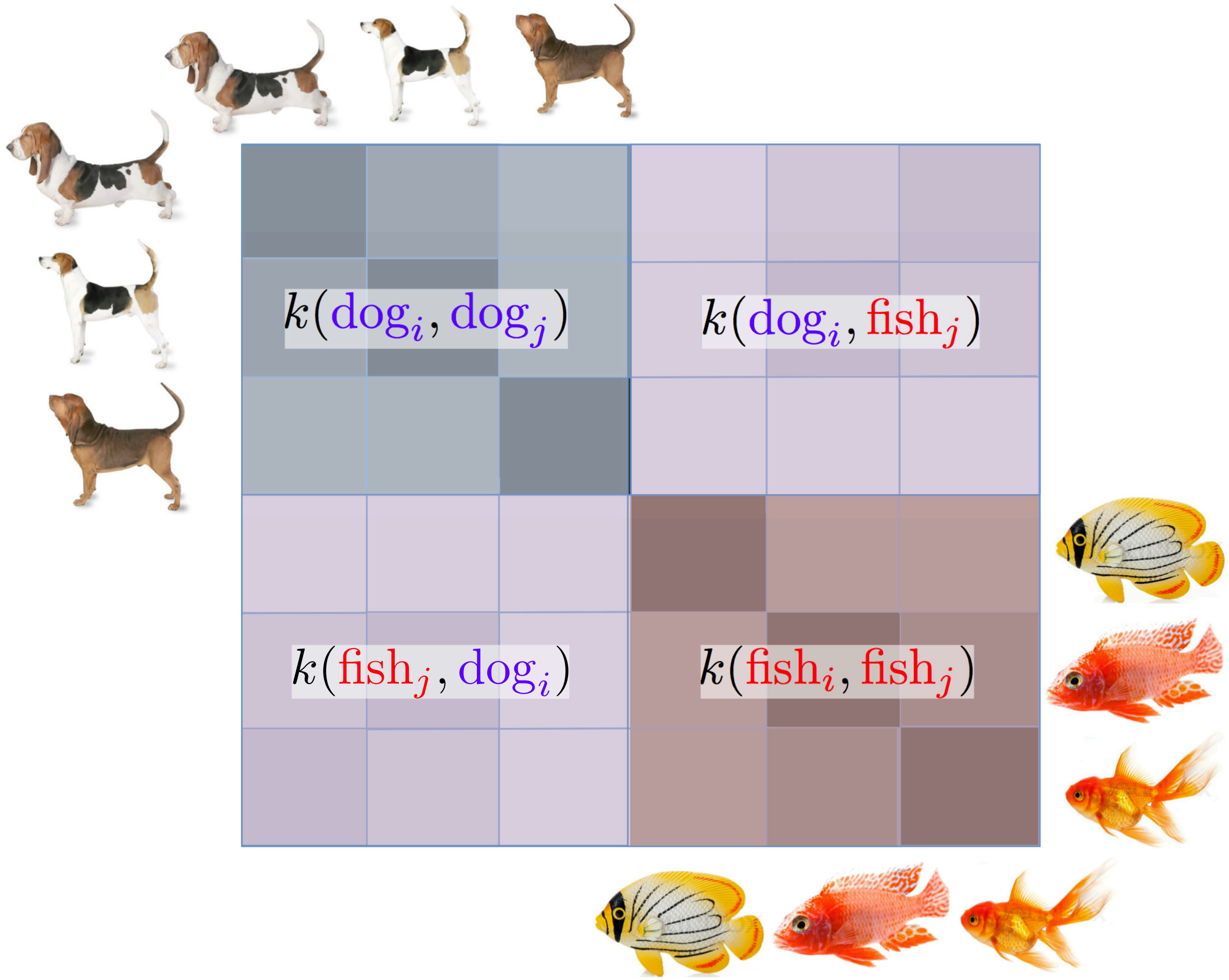
The maximum mean discrepancy



The maximum mean discrepancy



The maximum mean discrepancy



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

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

MMD for independence: HSIC

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

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

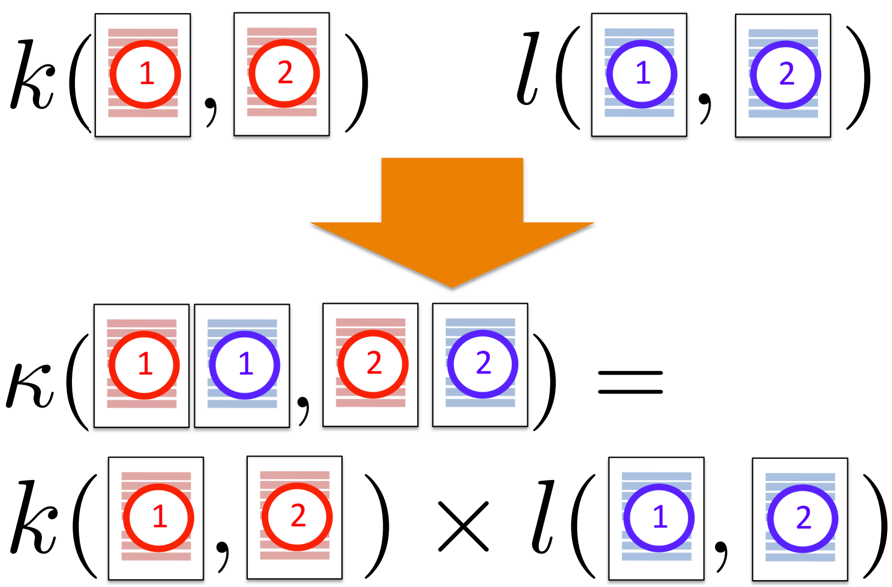
$$HSIC(\mathbf{P}_{XY}, \mathbf{P}_X\mathbf{P}_Y) := \left\| \mu_{\mathbf{P}_{XY}} - \mu_{\mathbf{P}_X\mathbf{P}_Y} \right\|^2$$

MMD for independence: HSIC

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

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

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



MMD for independence: HSIC

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

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

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

HSIC using expectations of kernels:

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

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

HSIC: empirical estimate and intuition

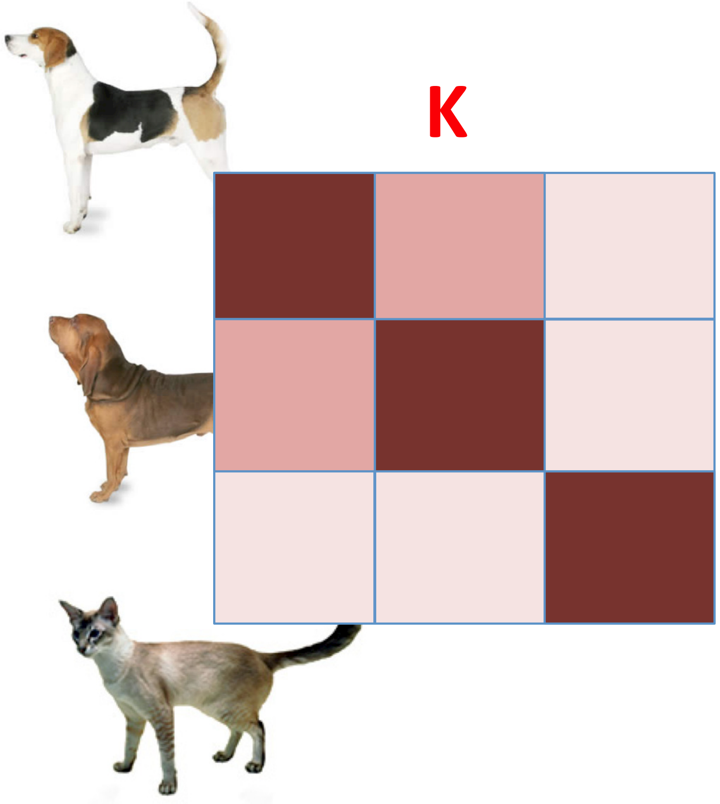
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

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

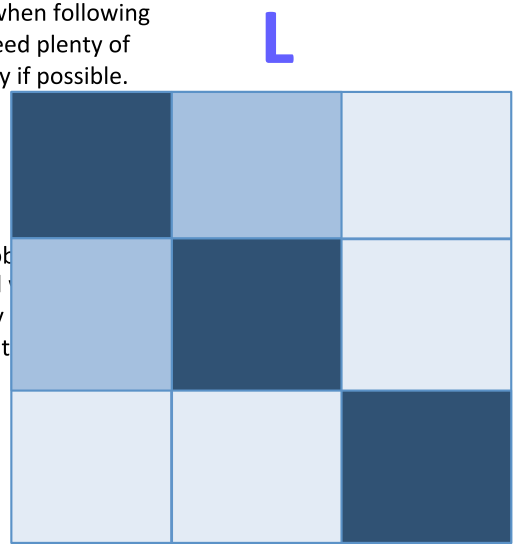
Known for their curiosity, intelligence, and excellent communication skills, the Javanese breed is perfect if you want a responsive, interactive pet, one that will blow in your ear and follow you everywhere.

Text from dogtime.com and petfinder.com

HSIC: empirical estimate and intuition



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

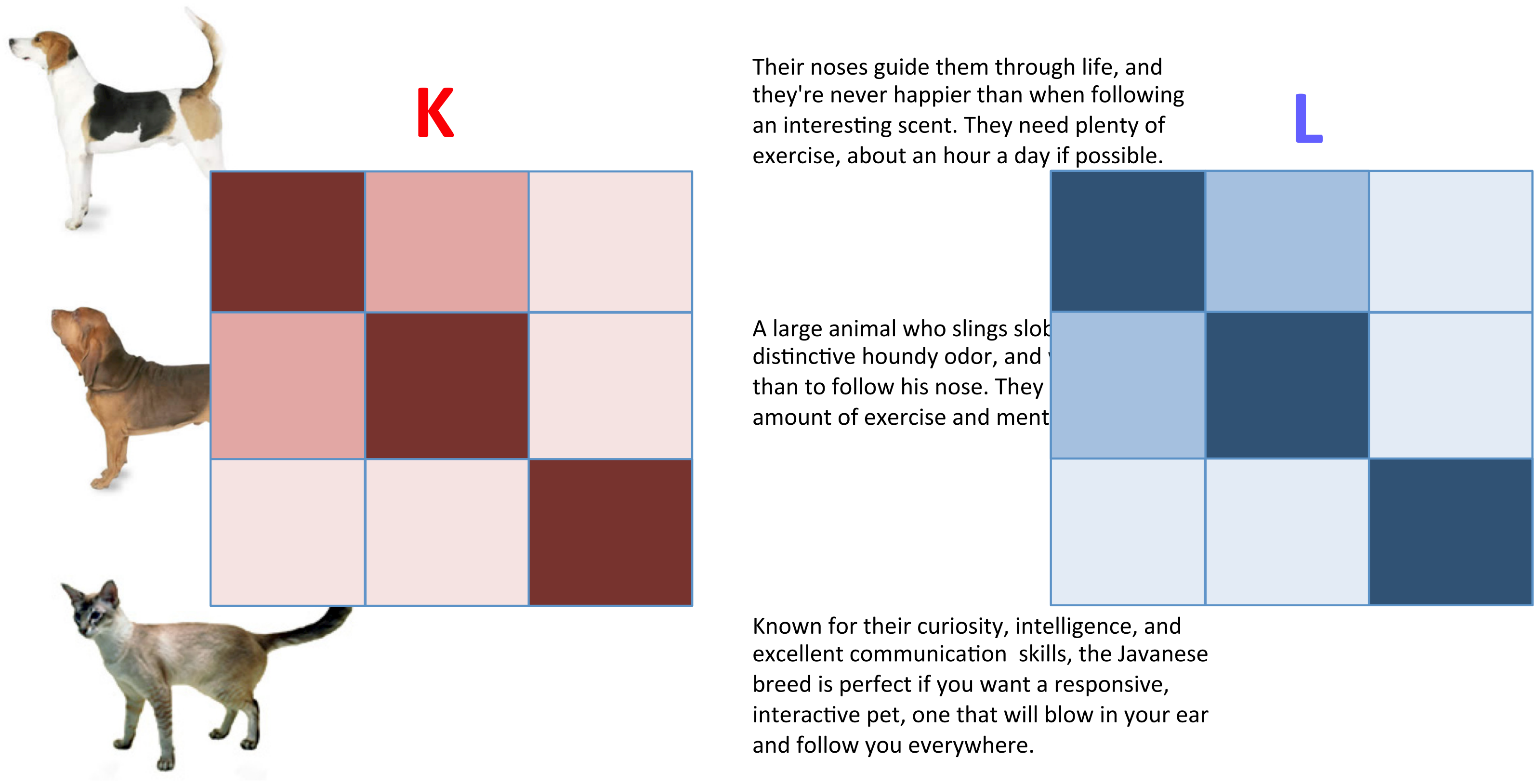


A large animal who slings slob
distinctive houndy odor, and
than to follow his nose. They
amount of exercise and ment

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

Text from dogtime.com and petfinder.com

HSIC: empirical estimate and intuition



Text from dogtime.com and petfinder.com

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

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

Characteristic kernels (Version 1: Via Universality)

Characteristic Kernels (via universality)

Characteristic kernels are those for which MMD is a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

Characteristic Kernels (via universality)

Characteristic kernels are those for which MMD is a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

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

Characteristic Kernels (via universality)

Characteristic kernels are those for which MMD is a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

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

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

Characteristic Kernels (via universality)

Characteristic kernels are those for which MMD is a metric (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]

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

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

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

Characteristic Kernels (via universality)

Proof:

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

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

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

Characteristic Kernels (via universality)

Proof:

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

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

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

We next make the expansion

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

The first and third terms satisfy

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

Characteristic Kernels (via universality)

Proof (continued):

Next, write

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

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

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

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

Characteristic kernels (Version 2: Via Fourier)

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

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

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

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

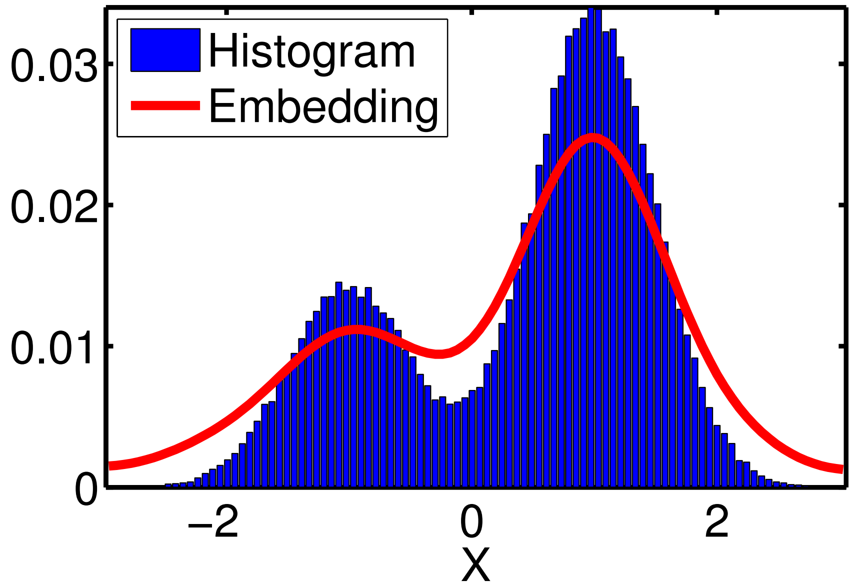
- What does prob. feature map look like?

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

Expectation of kernel!

- Maximum mean discrepancy

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

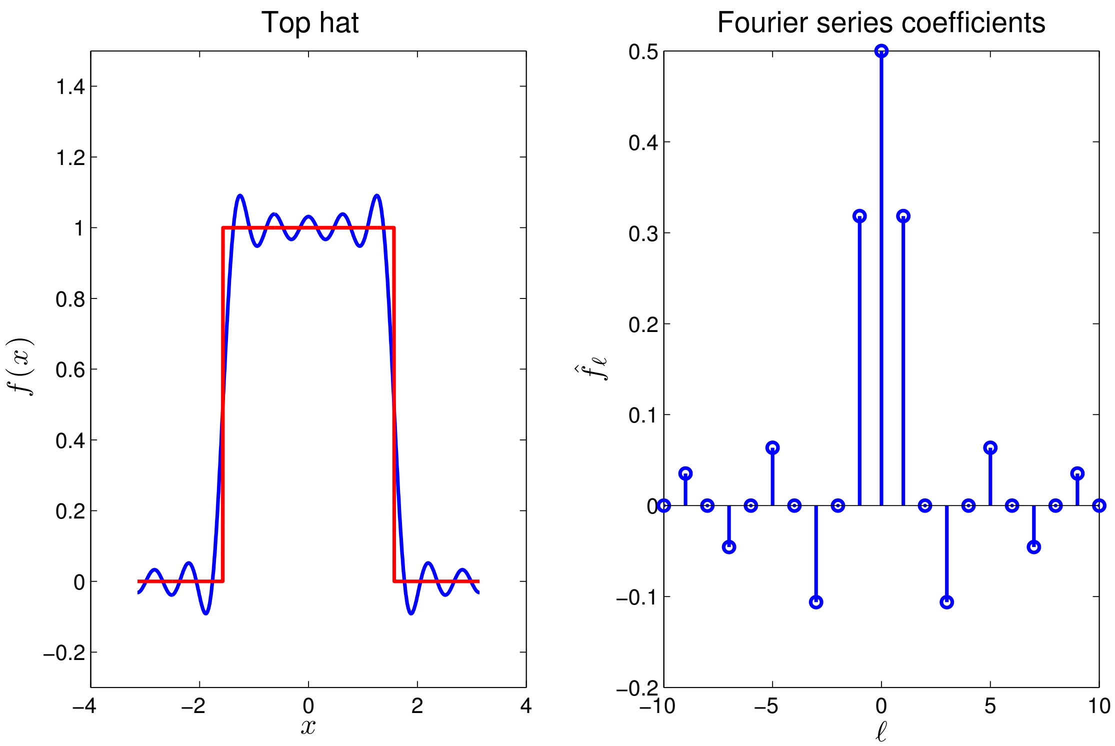


Characteristic Kernels (via Fourier)

Reminder: Fourier series

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

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



Characteristic Kernels (via Fourier)

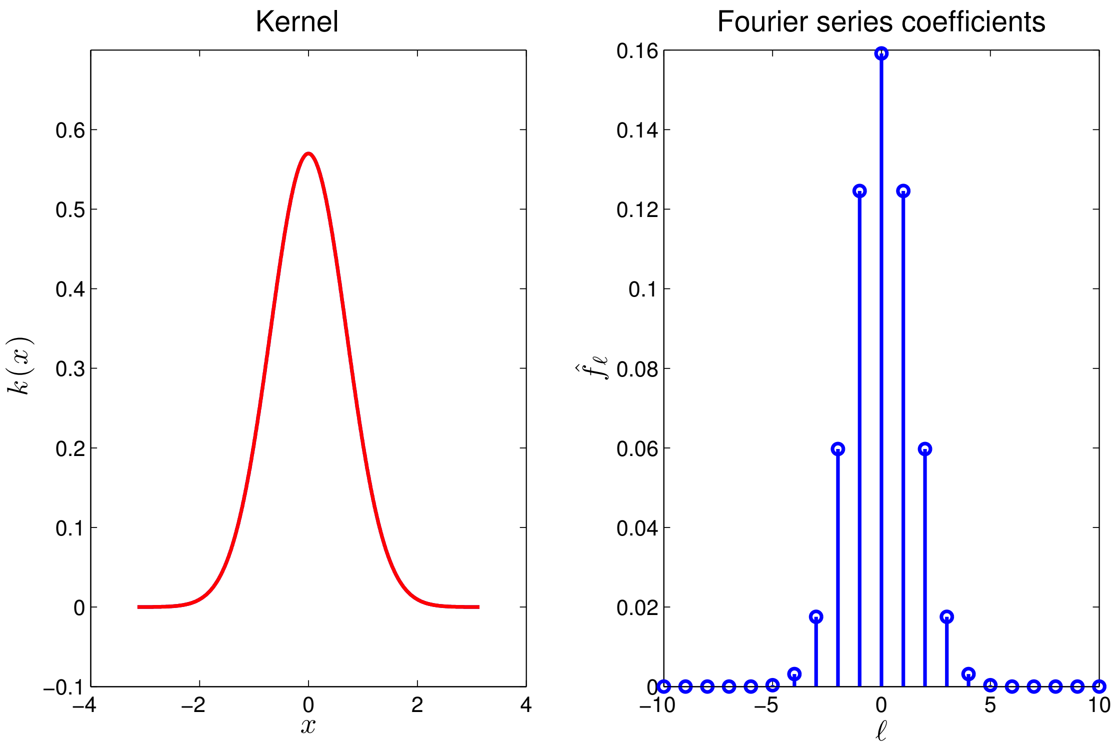
Reminder: Fourier series of kernel

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

E.g.,

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

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



Characteristic Kernels (via Fourier)

Maximum mean embedding via Fourier series:

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

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

Characteristic Kernels (via Fourier)

Maximum mean embedding via Fourier series:

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

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

- MMD can be written in terms of Fourier series:

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

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

A simpler Fourier expression for MMD

- Recall MMD expression:

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

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

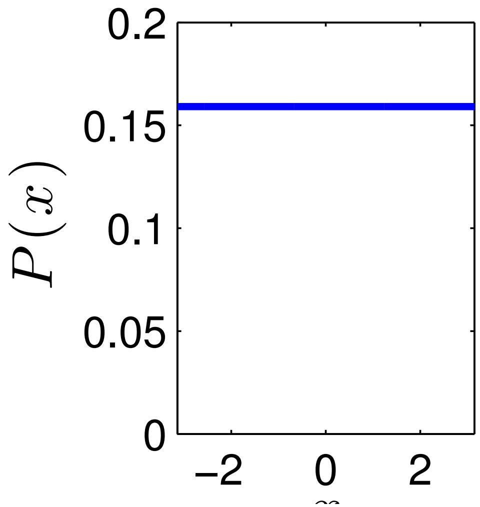
$$\|f\|_{\mathcal{F}}^2 = \langle f, f \rangle_{\mathcal{F}} = \sum_{l=-\infty}^{\infty} \frac{|\hat{f}_\ell|^2}{\hat{k}_\ell}.$$

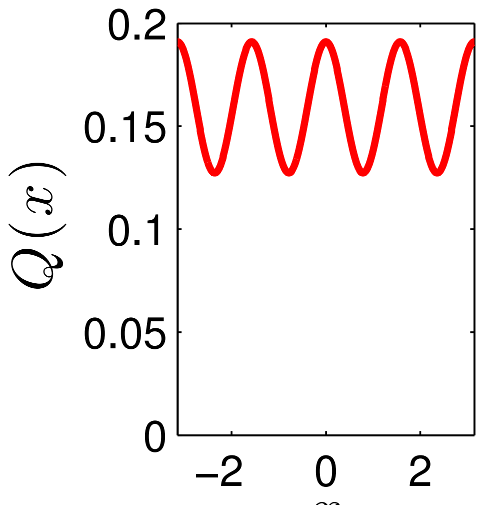
- Simple, interpretable expression for squared MMD:

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

Example

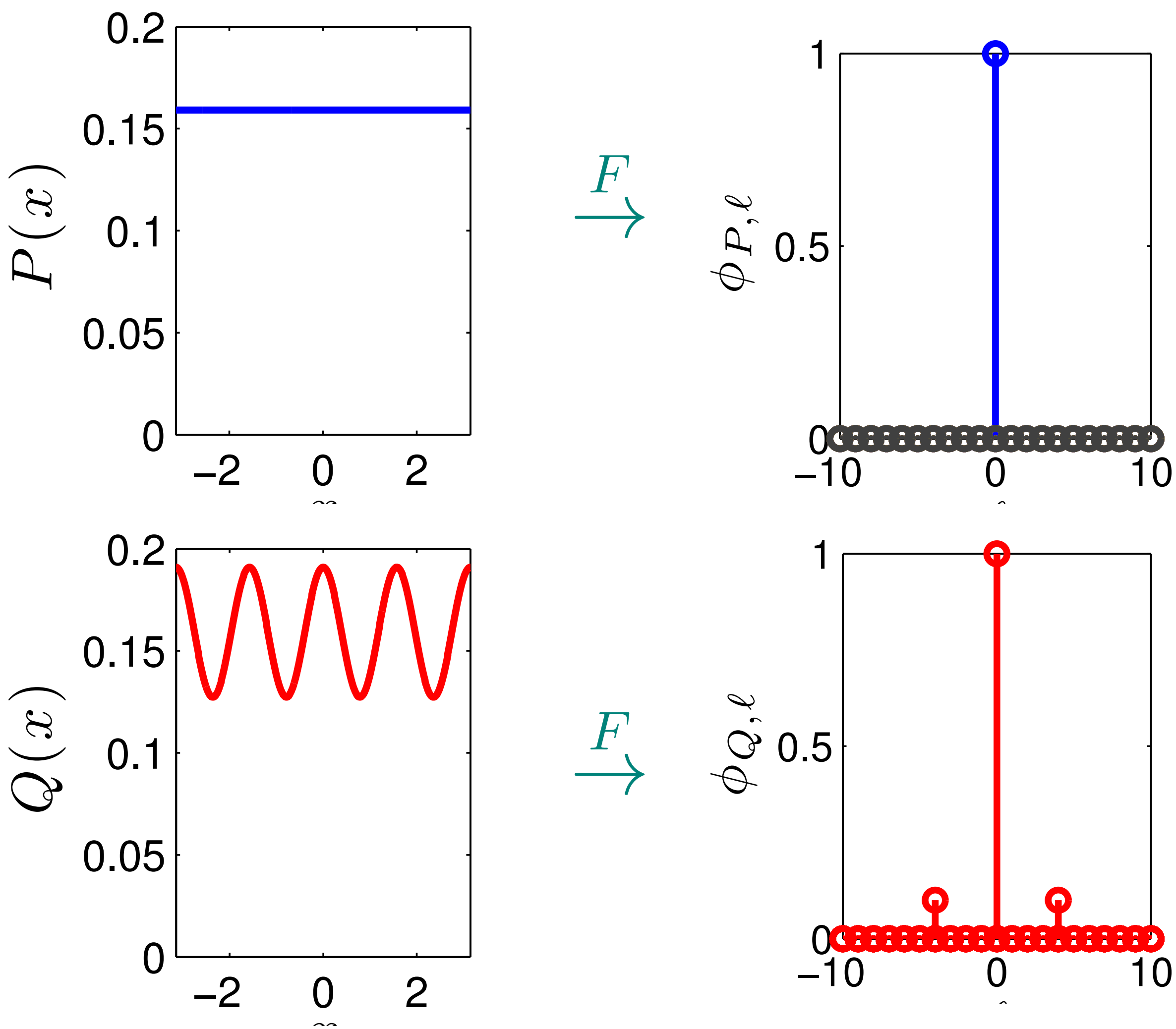
- Example: **P** differs from **Q** at one frequency





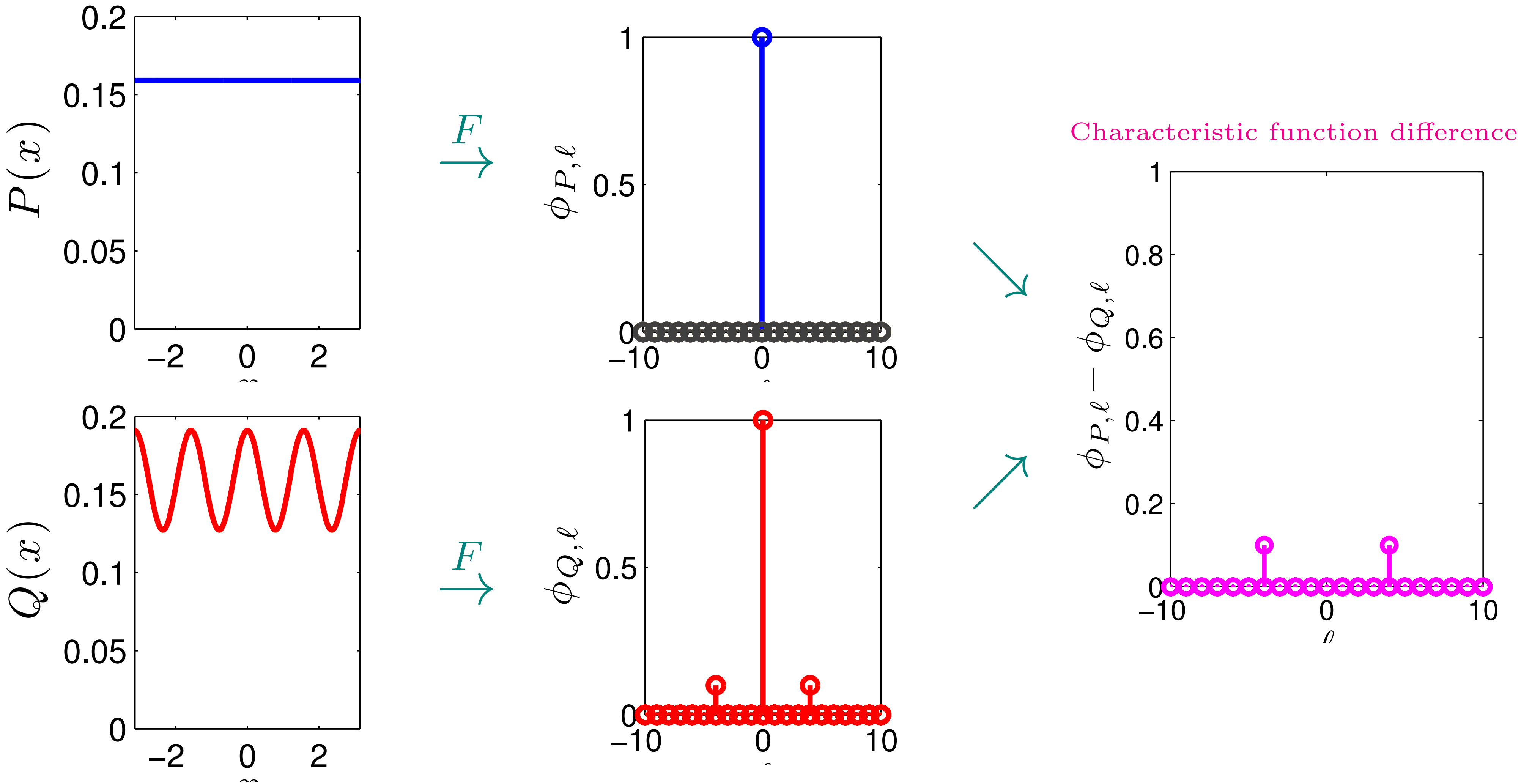
Characteristic Kernels (2)

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



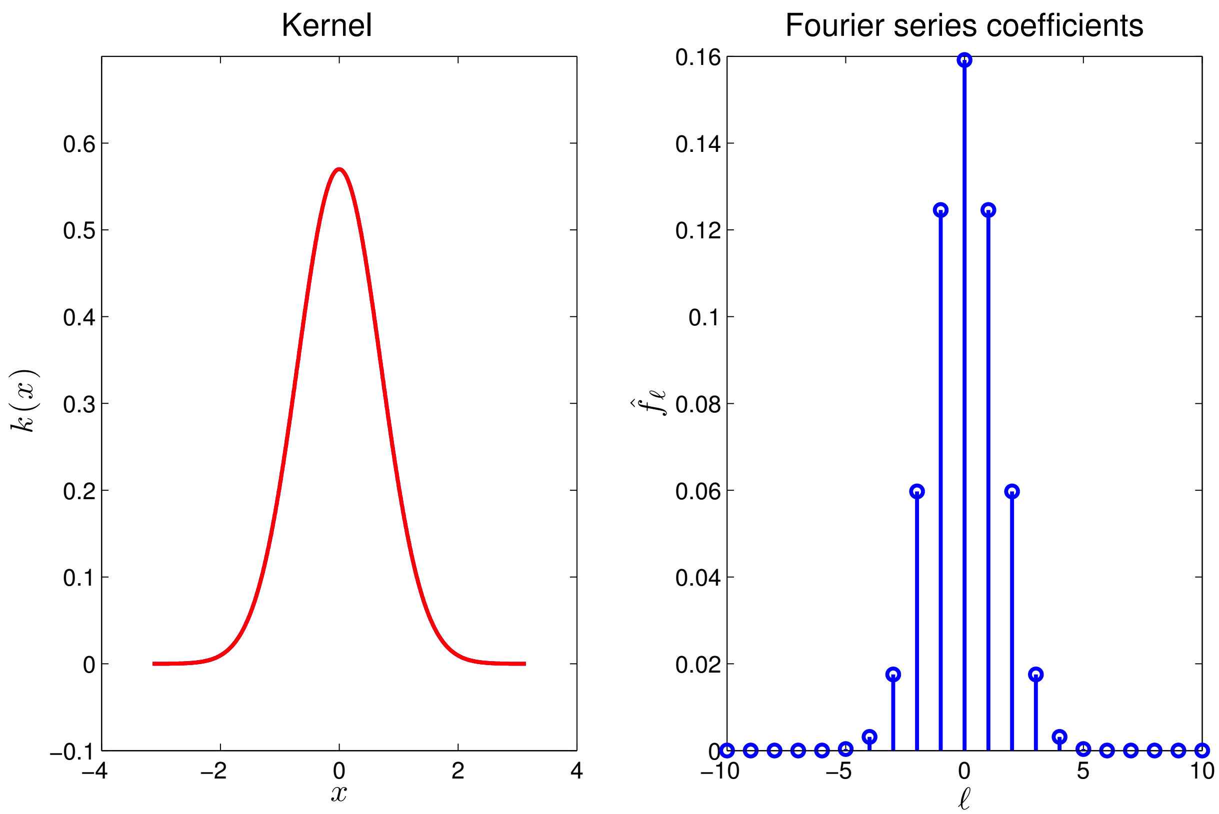
Characteristic Kernels (2)

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



Example

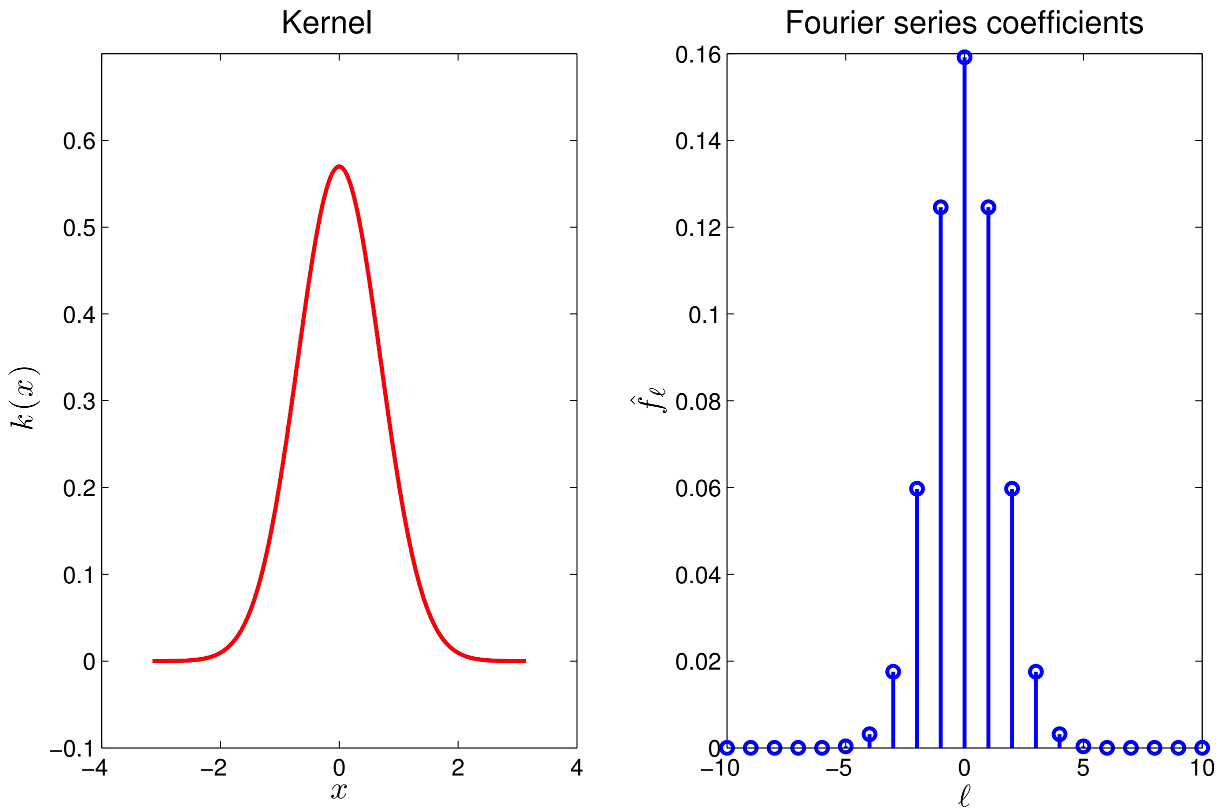
Is the Gaussian-spectrum kernel characteristic?



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

Example

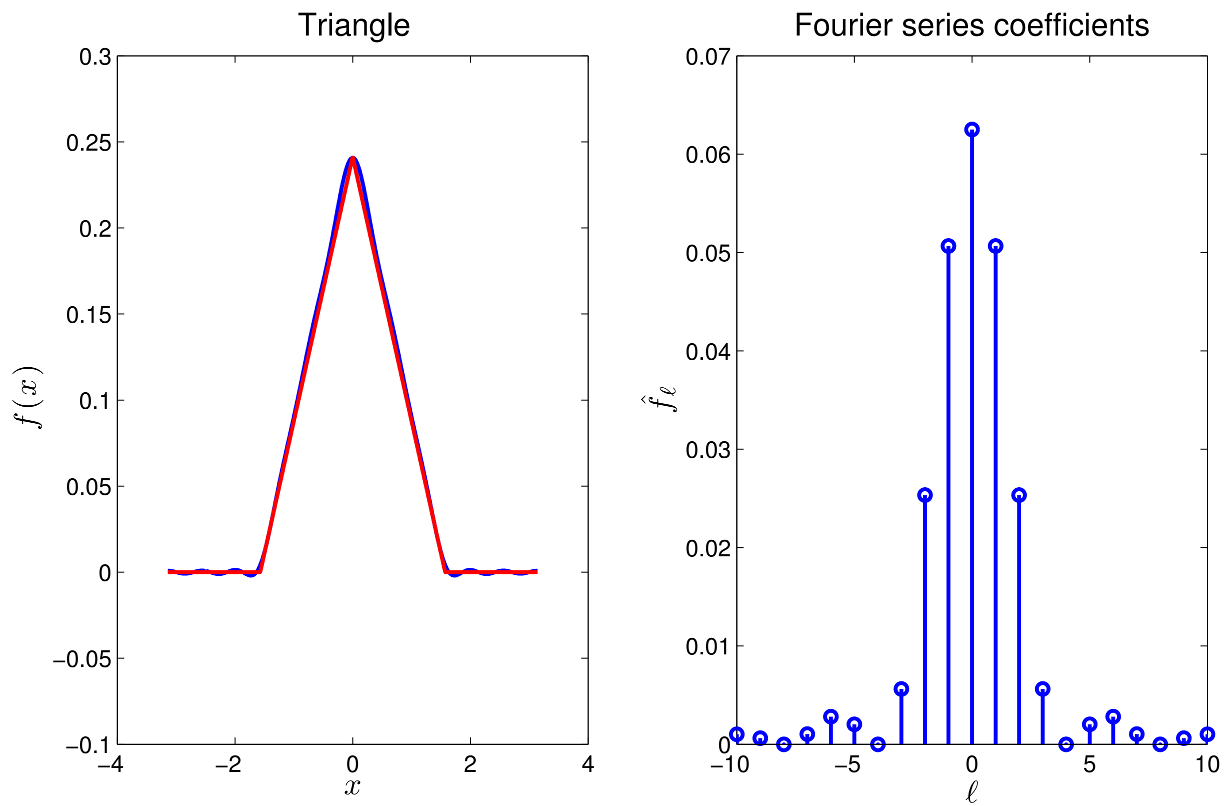
Is the Gaussian-spectrum kernel characteristic? YES



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

Example

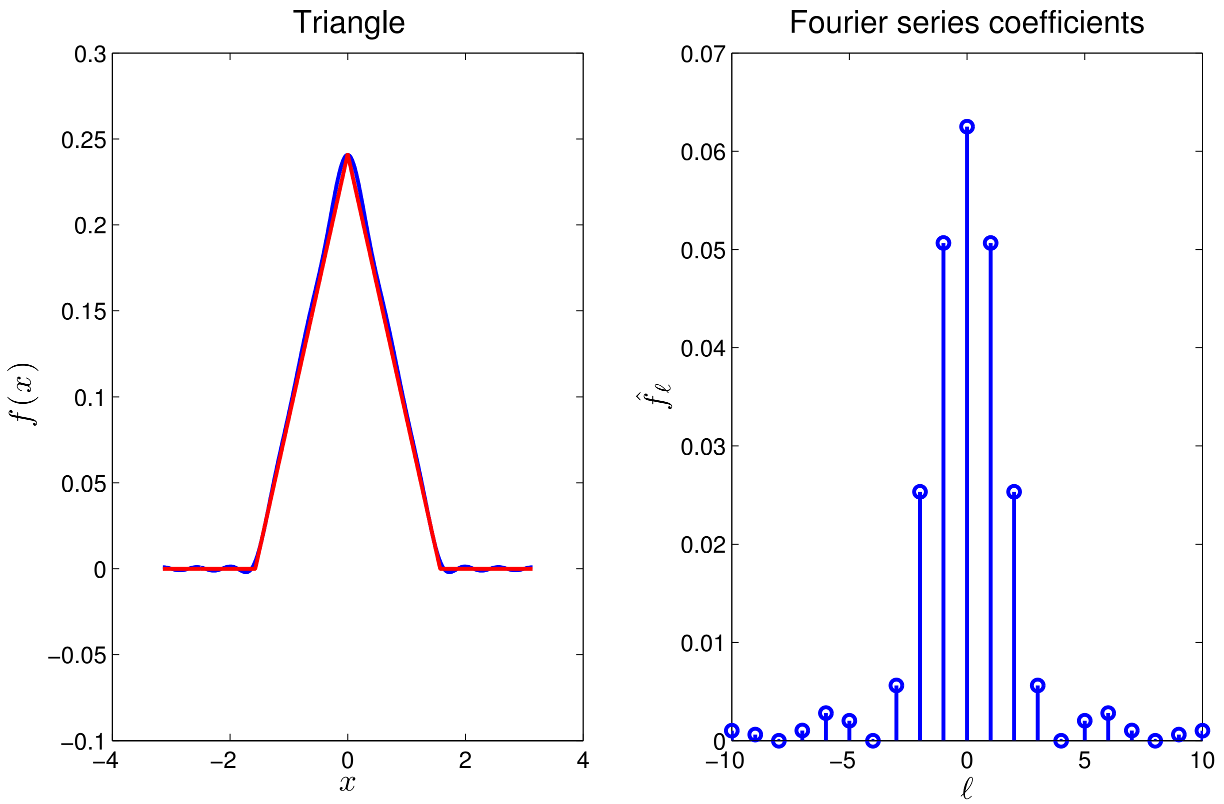
Is the triangle kernel characteristic?



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

Example

Is the triangle kernel characteristic? NO



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

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$? (not just $[\pi, \pi]$ periodic)

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$? (not just $[\pi, \pi]$ periodic)
- Characteristic function of $\mathbf{P}$ via Fourier transform

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

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$? (not just $[\pi, \pi]$ periodic)
- Characteristic function of **P** via Fourier transform

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

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

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

  - Λ finite non-negative Borel measure

Characteristic Kernels (via Fourier)

- Can we prove characteristic on $\mathbb{R}^d$? (not just $[\pi, \pi]$ periodic)
- Characteristic function of $\mathbf{P}$ via Fourier transform

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

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

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

  - Λ finite non-negative Borel measure

Characteristic Kernels (via Fourier)

Fourier representation of MMD:

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

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

Proof: Using Bochner's theorem (a) ...

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

Characteristic Kernels (via Fourier)

Fourier representation of MMD:

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

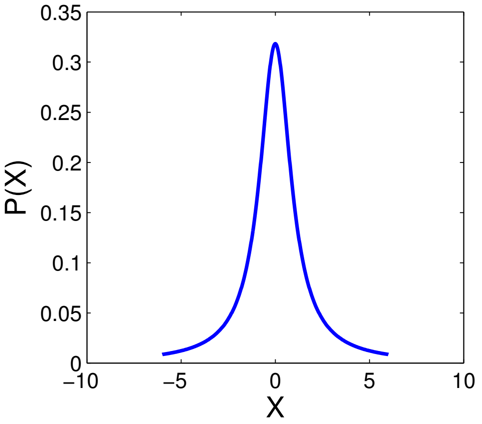
$\phi_{\mathbf{P}}$ characteristic function of $\mathbf{P}$

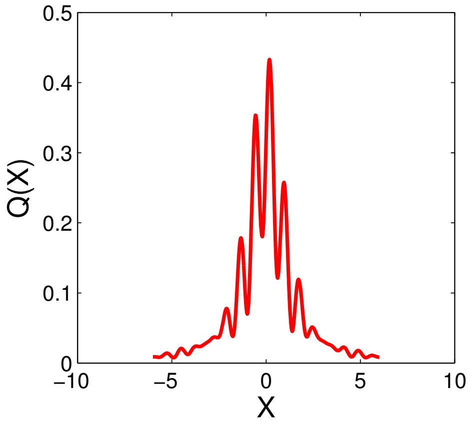
Proof: Using Bochner's theorem (a)... and Fubini's theorem (b)

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

Example

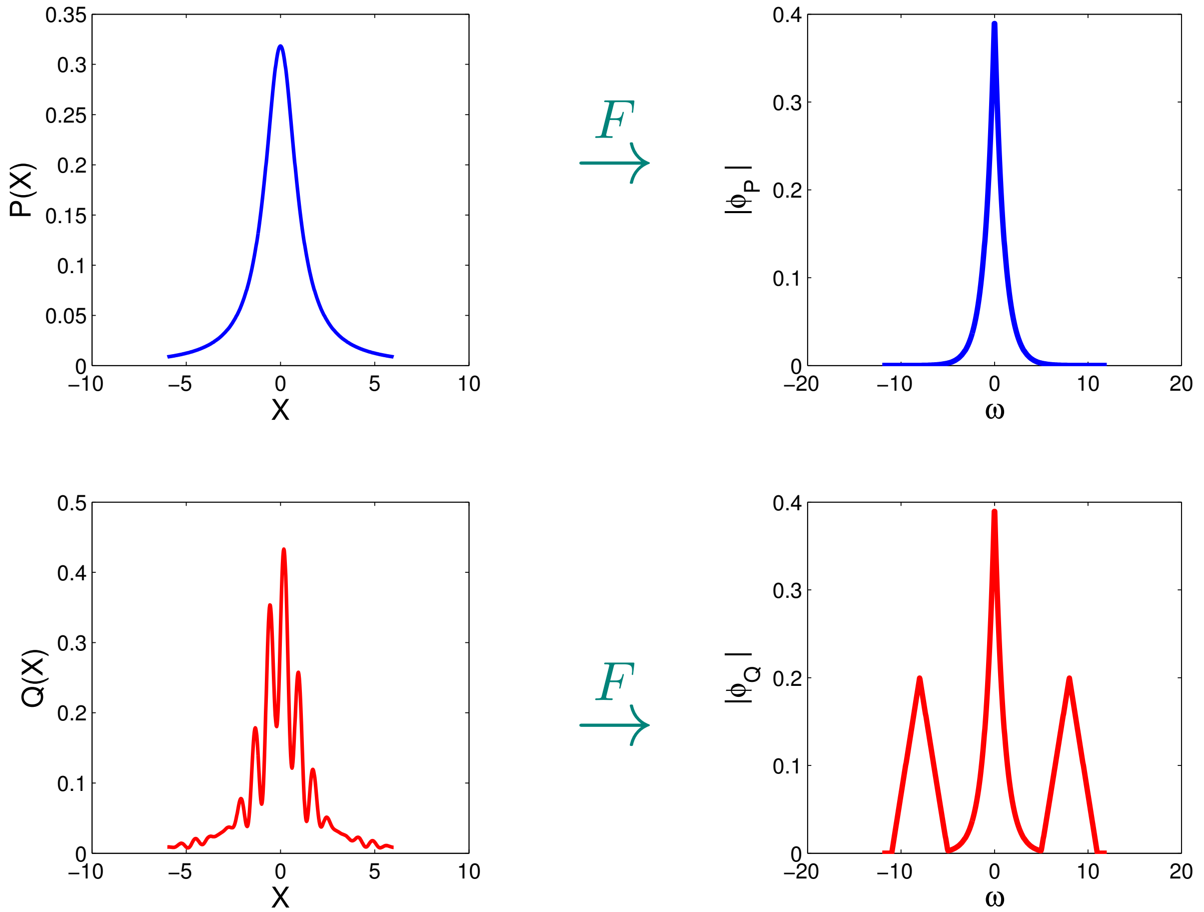
- Example: **P** differs from **Q** at (roughly) one frequency





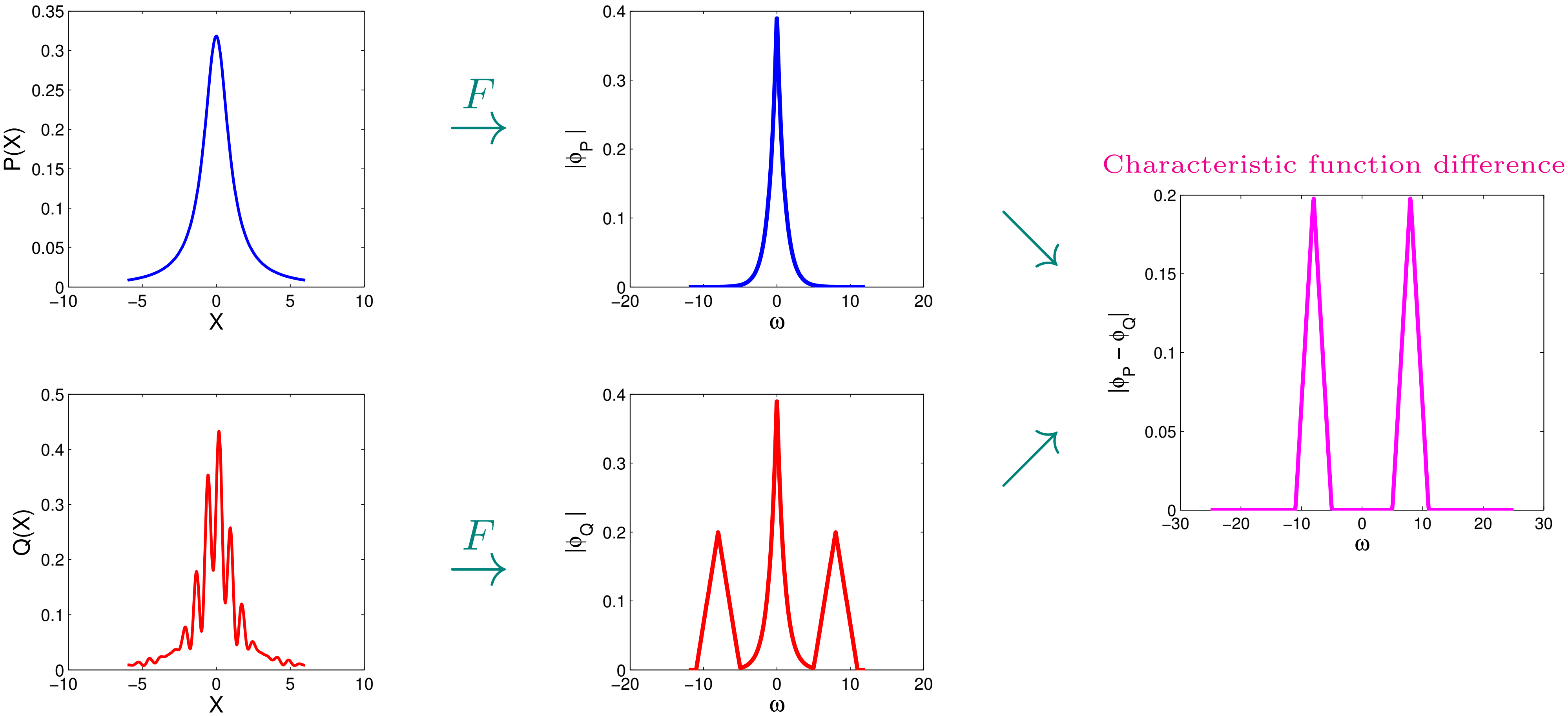
Example

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



Example

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

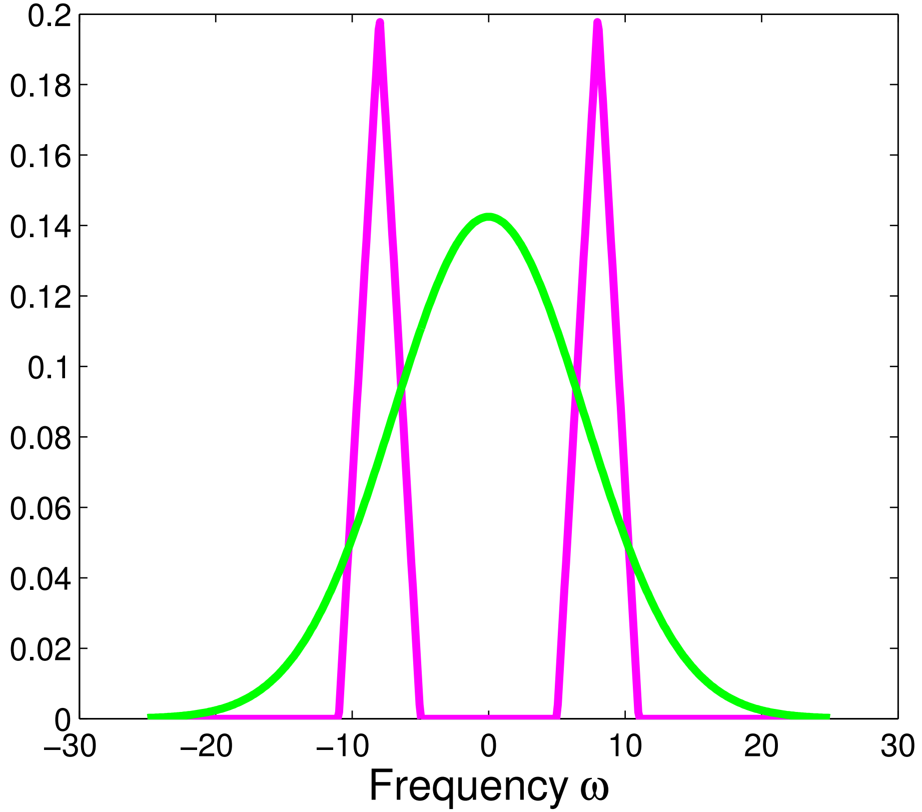


Example

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

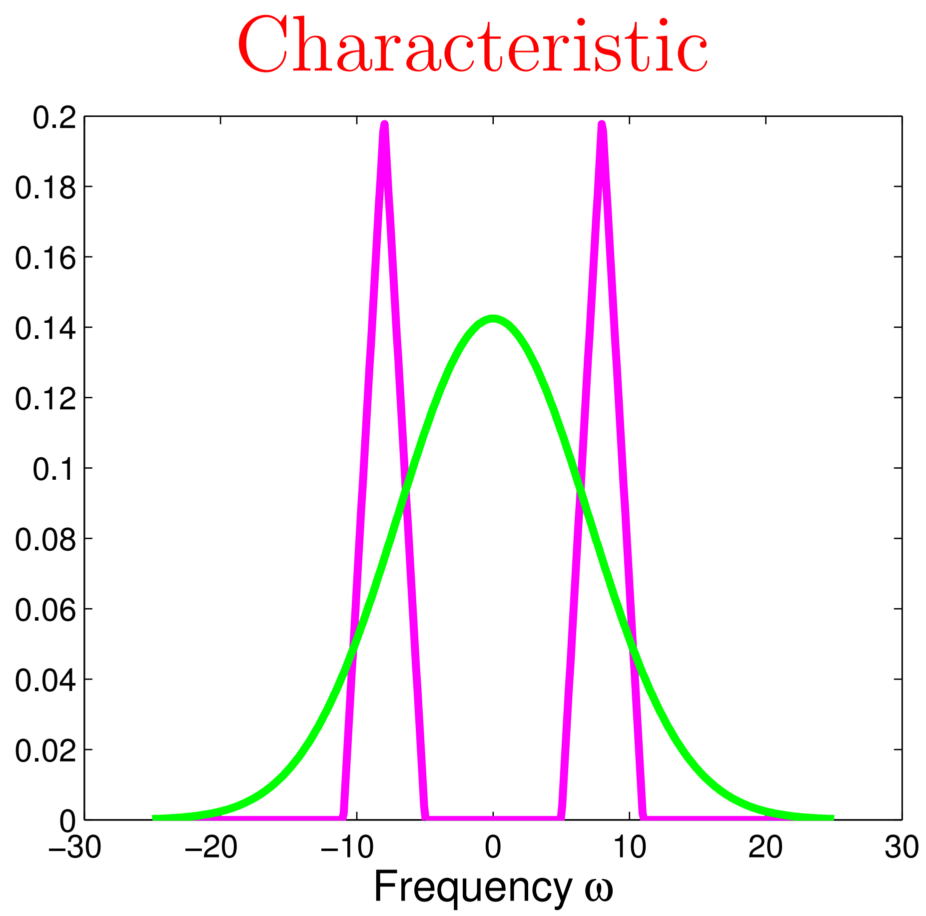
Gaussian kernel

Difference $|\phi_P - \phi_Q|$



Example

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

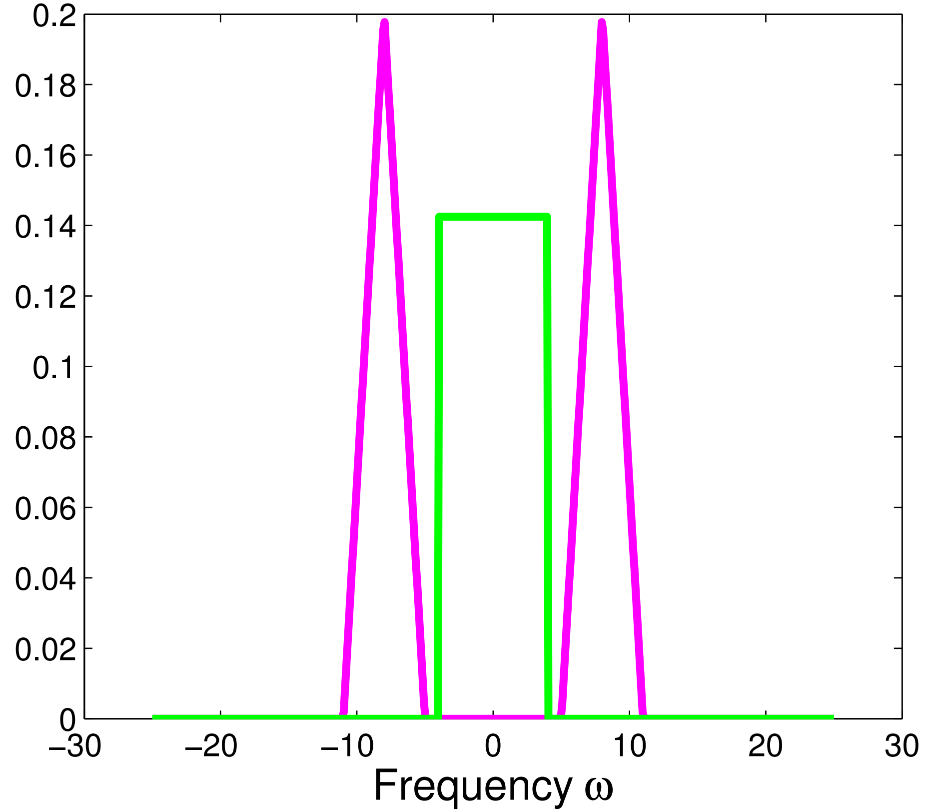


Example

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

Sinc kernel

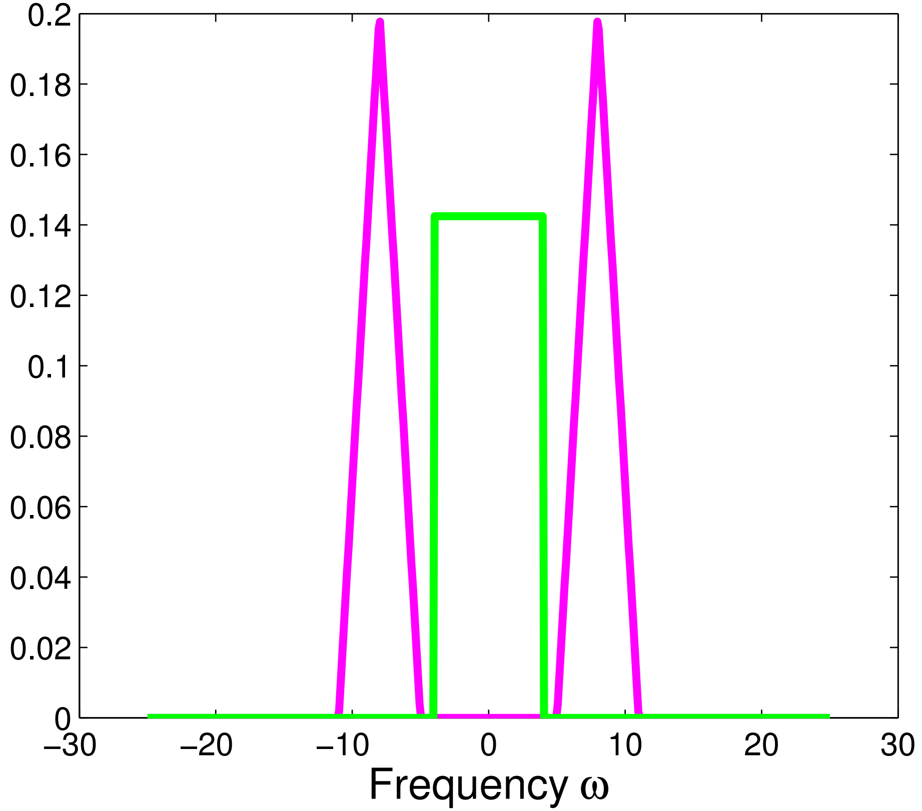
Difference $|\phi_P - \phi_Q|$



Example

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

NOT characteristic

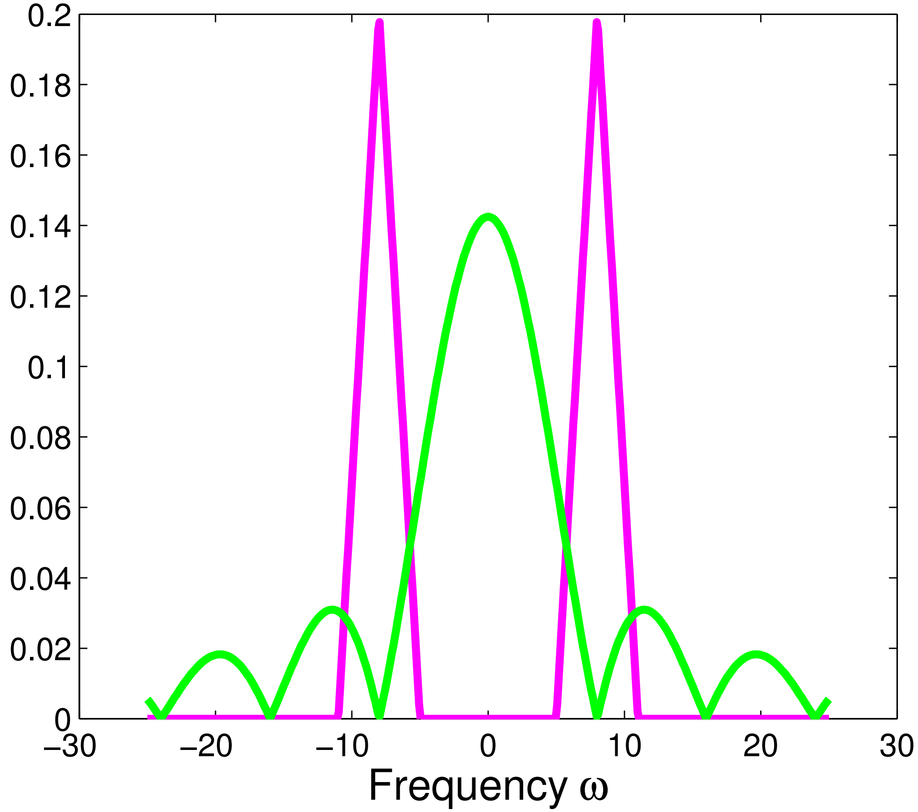


Example

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

Triangle (B-spline) kernel

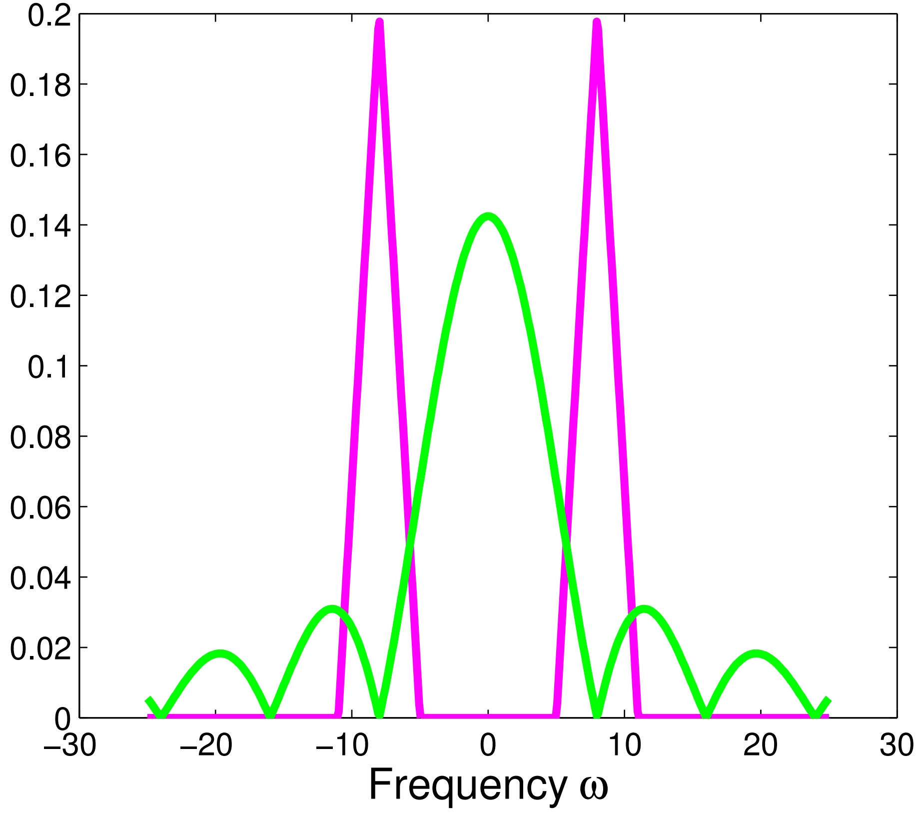
Difference $|\phi_P - \phi_Q|$



Example

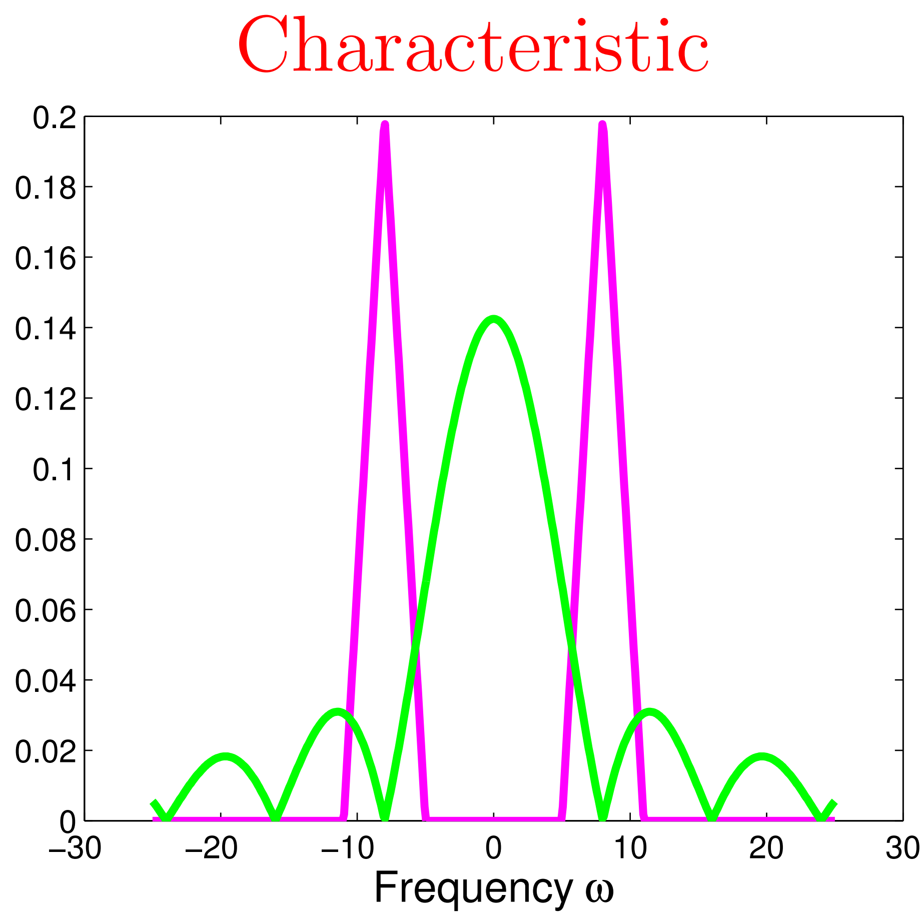
- Example: **P** differs from **Q** at (roughly) one frequency

???



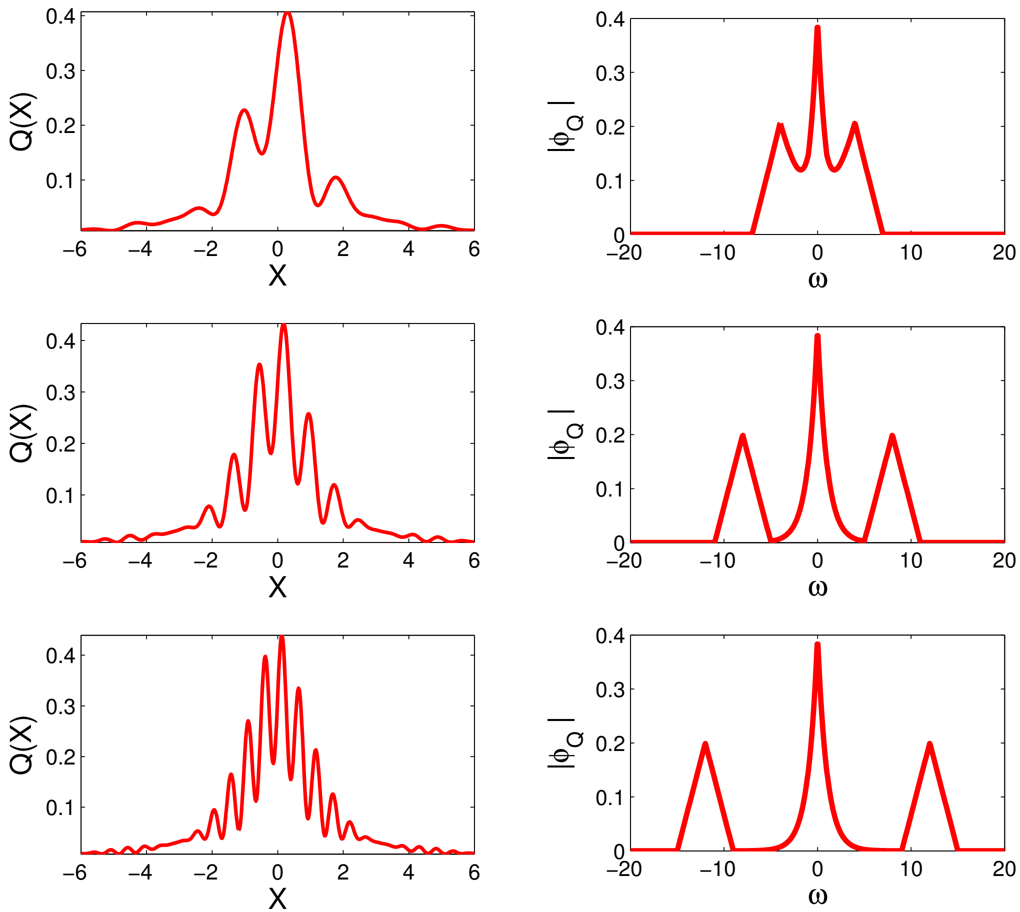
Example

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

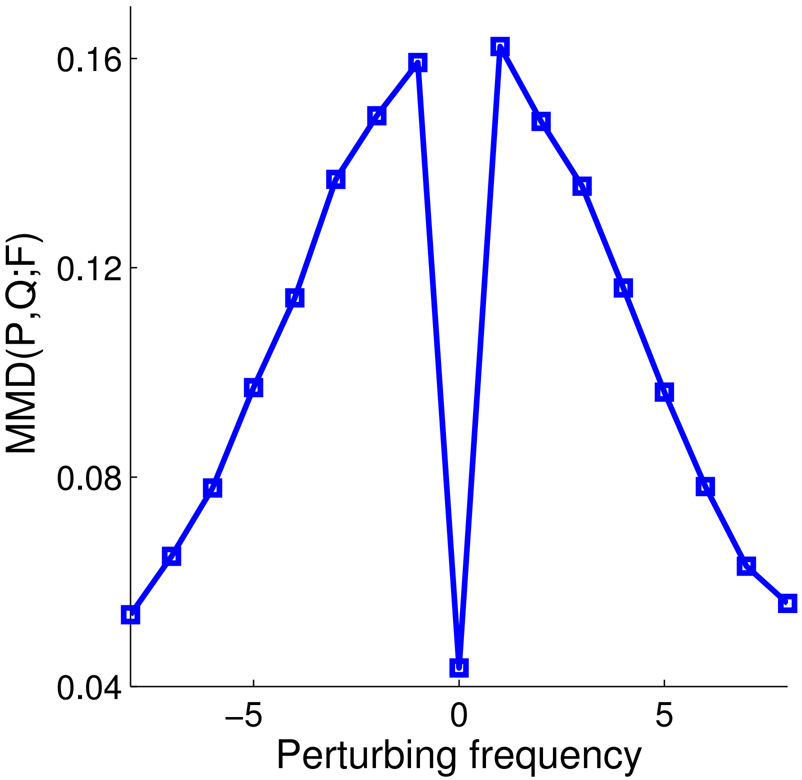


Choosing the kernel

- Gaussian kernel example

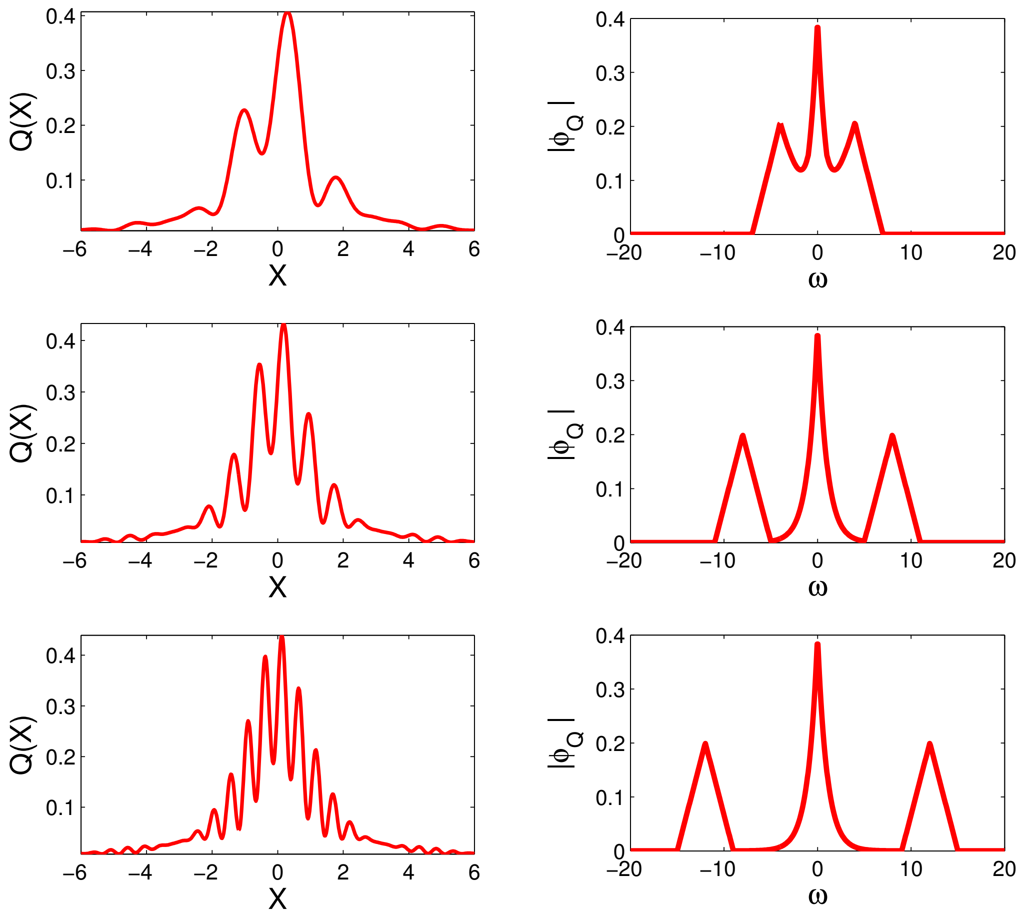


- MMD vs frequency of perturbation to **P**

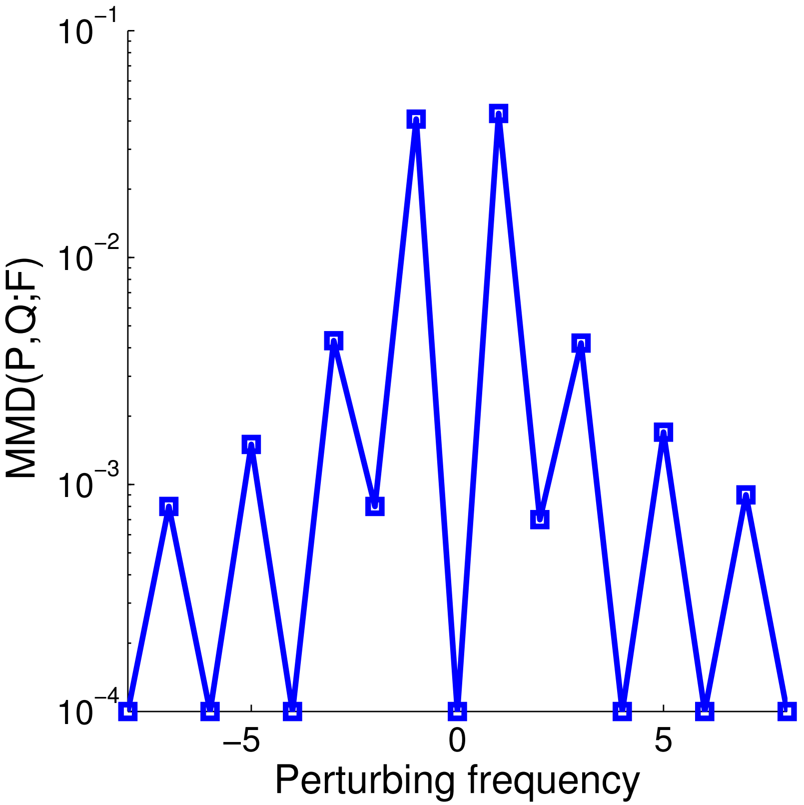


Choosing the kernel

- B-spline kernel example



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



Why does MMD decay with increasing perturbation freq.?

- Recall simple MMD expression, Fourier series case:

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

  and that $\hat{k}_\ell$ decays as ℓ grows.

- Fourier representation for more general case on $\mathbb{R}^d$:

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

  has similar behavior.

Summary: Characteristic Kernels

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

Summary: Characteristic Kernels

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

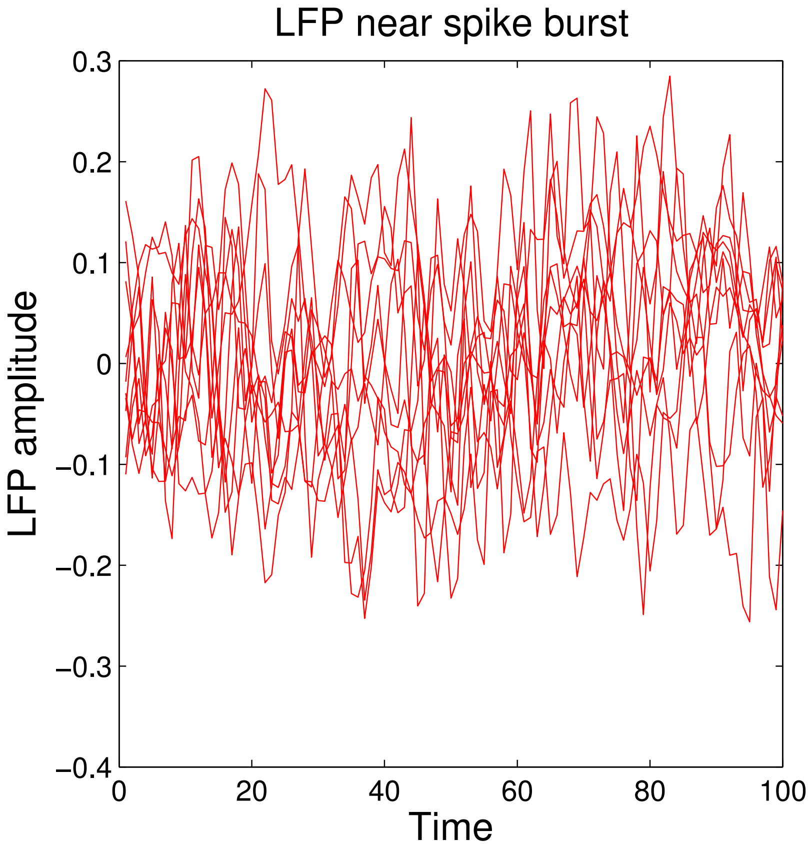
Summary: Characteristic Kernels

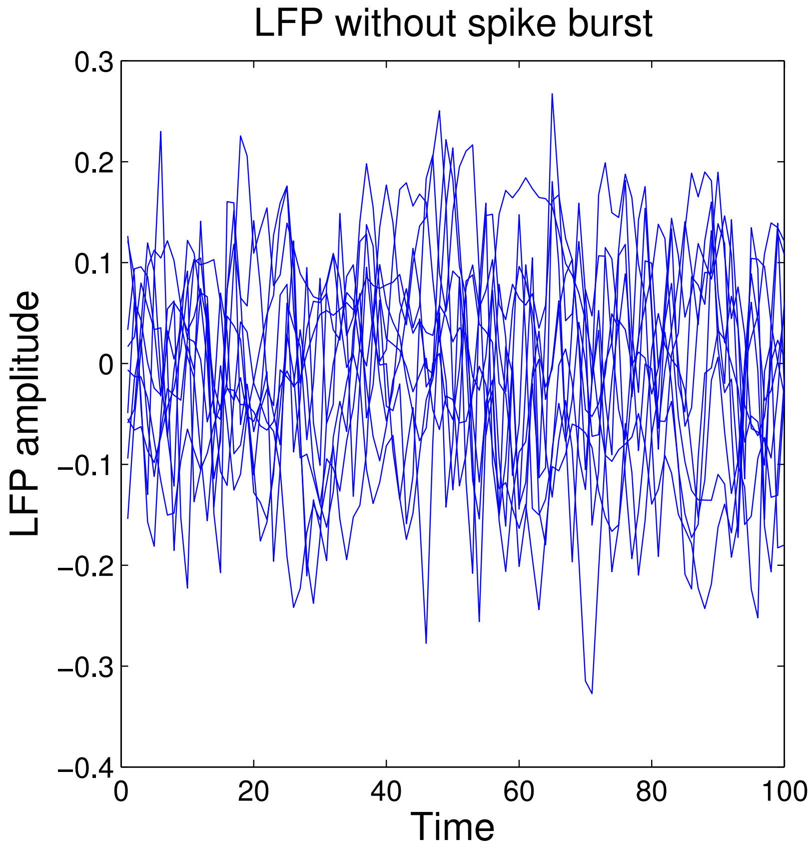
- Characteristic kernel: (MMD $= 0$ iff $\mathbf{P} = \mathbf{Q}$) [NIPS07b, COLT08]
- Main theorem: k characteristic for prob. measures on $\mathbb{R}^d$ if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ [COLT08, JMLR10]
  - Corollary: continuous, compactly supported k characteristic
- Similar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

Statistical hypothesis testing

Motivating question: differences in brain signals

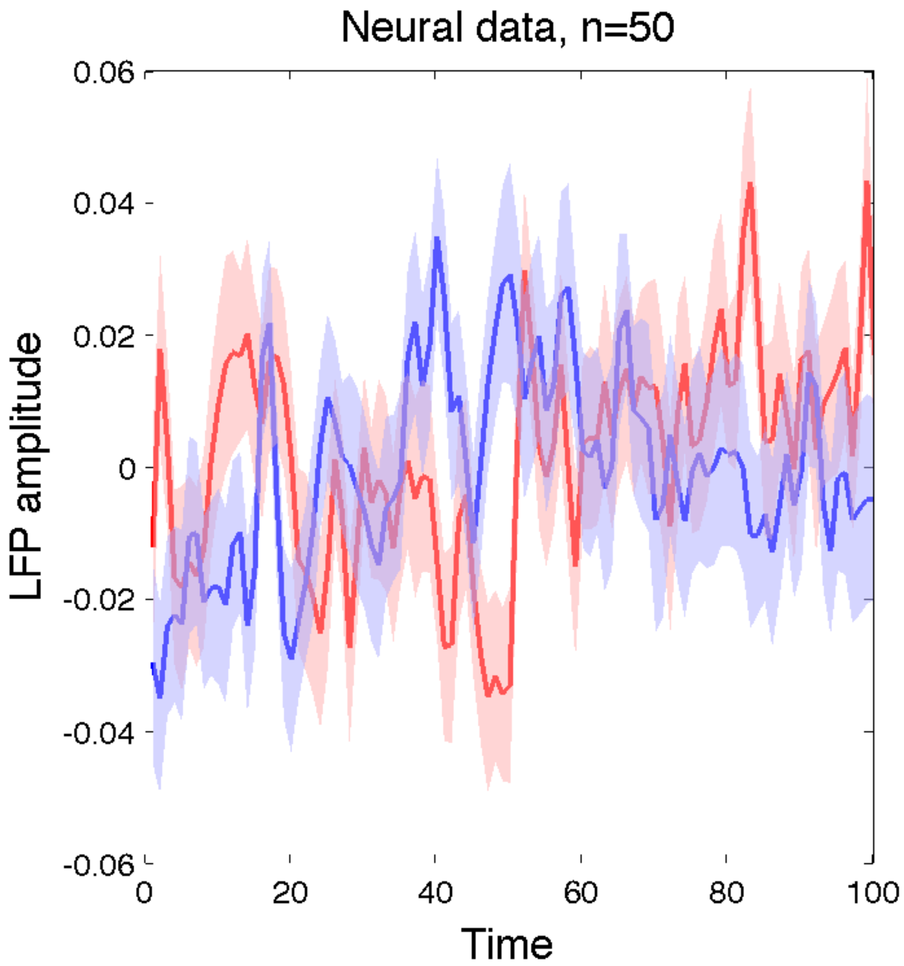
The problem: Do local field potential (LFP) signals change when measured near a spike burst?





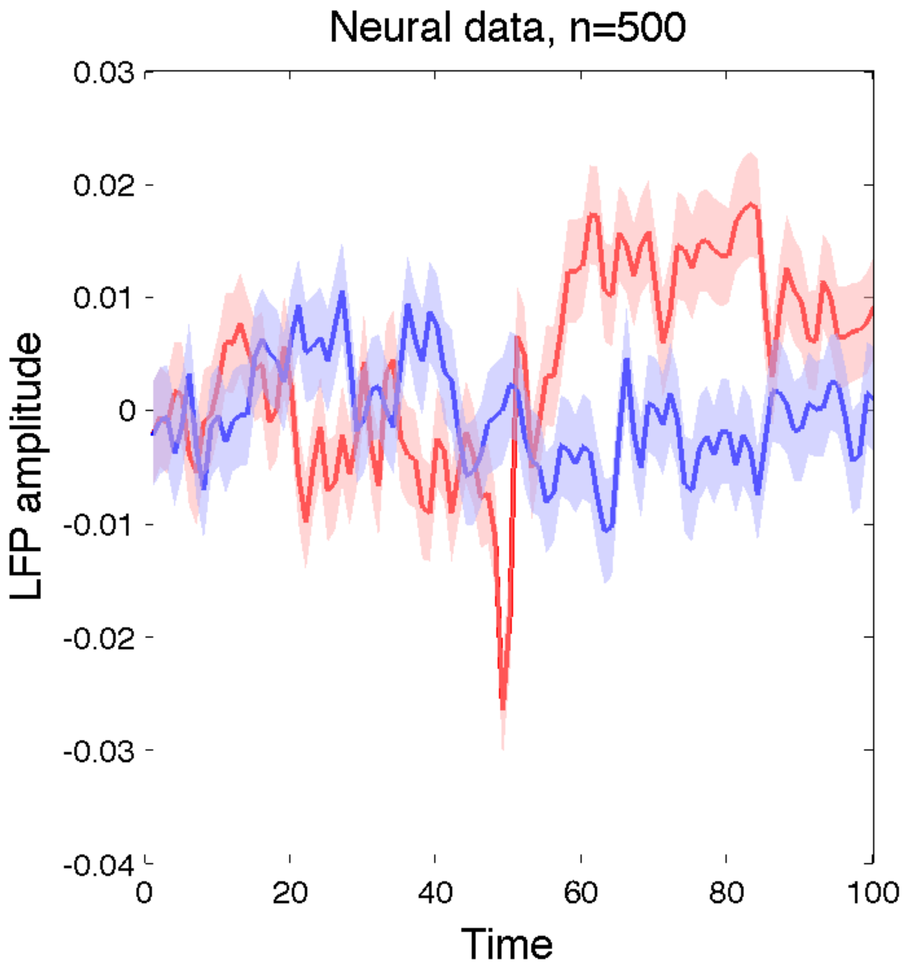
Motivating question: differences in brain signals

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



Motivating question: differences in brain signals

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



Statistical test using MMD (1)

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

Statistical test using MMD (1)

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

Statistical test using MMD (2)

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

Statistical test using MMD (2)

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

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

Statistical test using MMD (2)

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

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

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

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

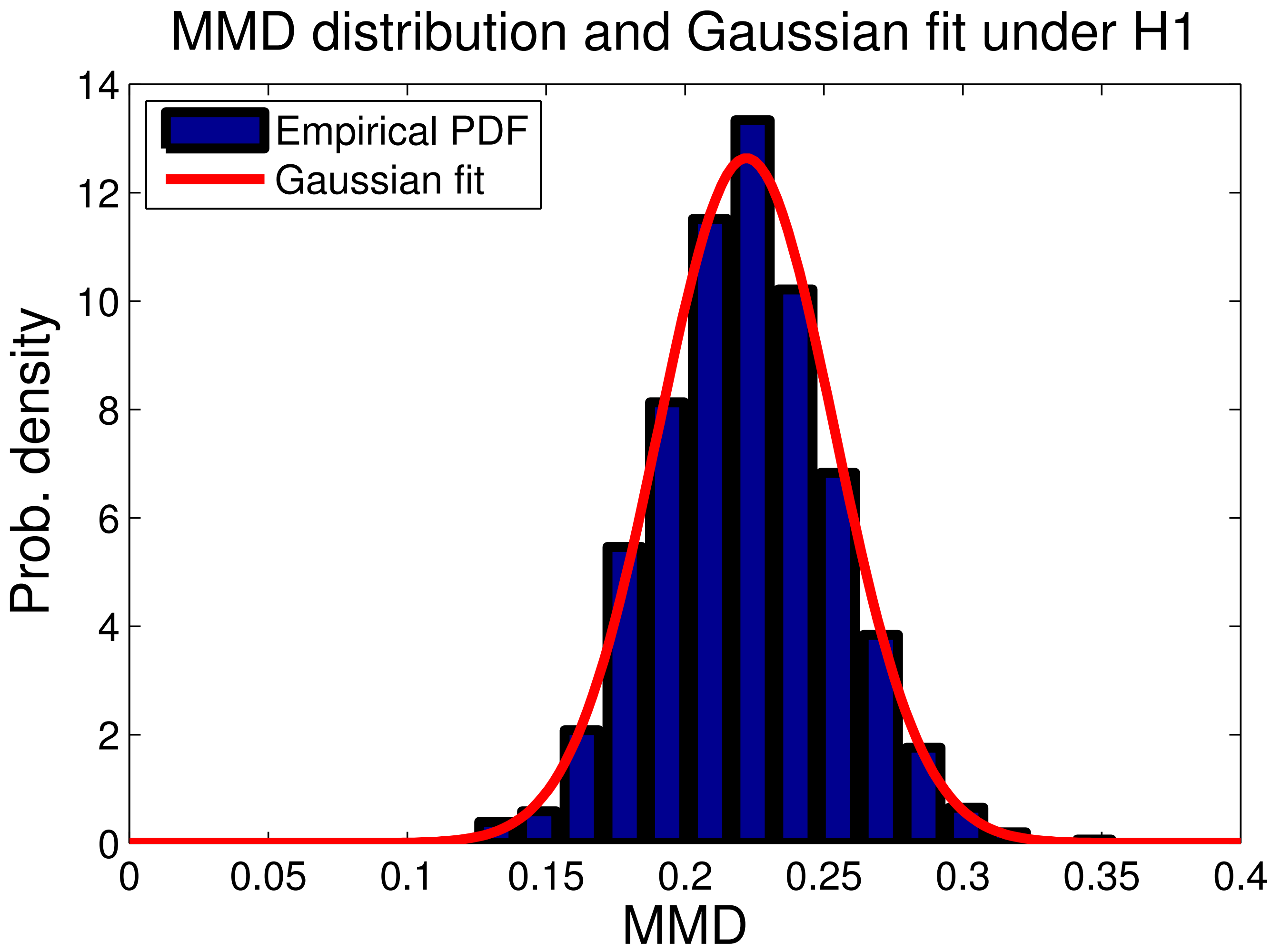
[Hoeffding, 1948, Serfling, 1980]

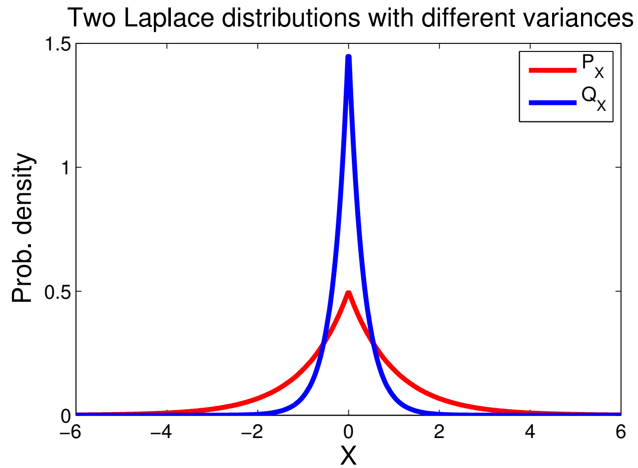
- Expression for the variance: $z_i := (x_i, y_i)$

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

Statistical test using MMD (3)

- Example: laplace distributions with different variance





Statistical test using MMD (4)

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

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

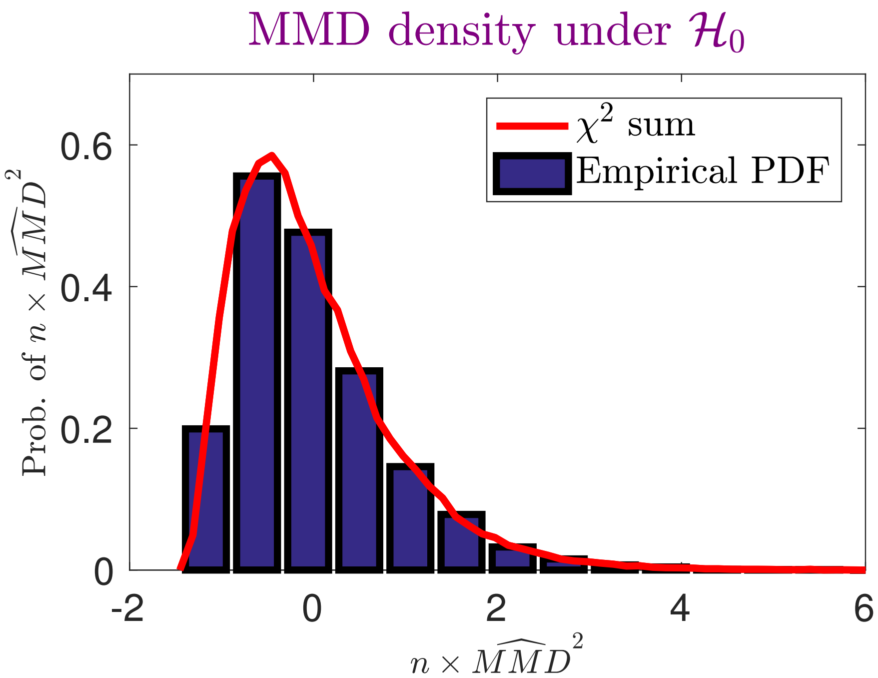
- where
  - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
  - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathsf{P}(x) = \lambda_i \psi_i(x')$

Statistical test using MMD (4)

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

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

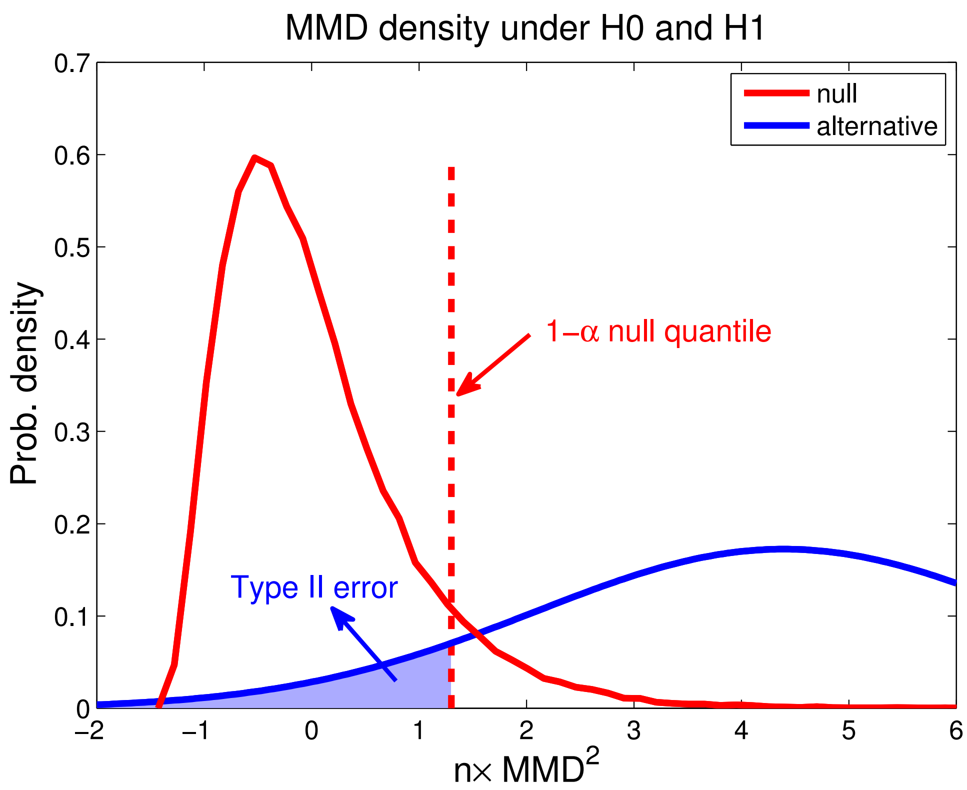
- where
  - $z_l \sim \mathcal{N}(0, 2)$ i.i.d
  - $\int_{\mathcal{X}} \underbrace{\tilde{k}(x, x')}_{\text{centred}} \psi_i(x) d\mathbf{P}(x) = \lambda_i \psi_i(x')$



Statistical test using MMD (5)

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

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



Statistical test using MMD (5)

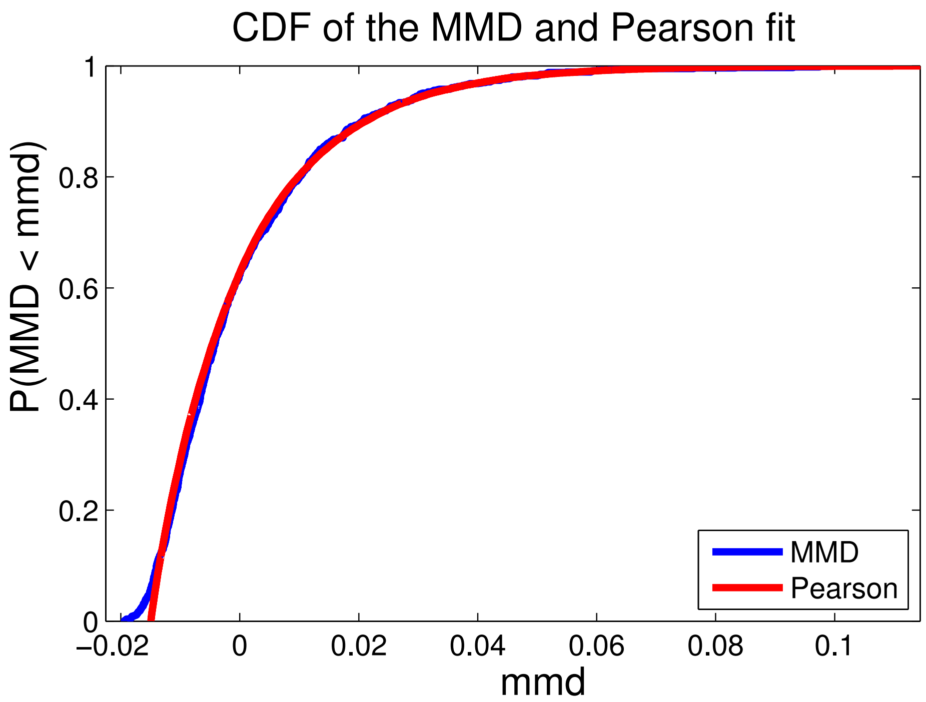
- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\text{MMD} > T) \leq 0.05$

Statistical test using MMD (5)

- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\mathrm{MMD} > T) \leq 0.05$
- Permutation for empirical CDF [Arcones and Giné, 1992]
- Pearson curves by matching first four moments [Johnson et al., 1994]
- Large deviation bounds [Hoeffding, 1963, McDiarmid, 1989]
- Consistent test using kernel eigenspectrum [NIPS09b]

Statistical test using MMD (5)

- Given $\mathbf{P} = \mathbf{Q}$, want threshold T such that $\mathbf{P}(\mathrm{MMD} > T) \leq 0.05$
- Permutation for empirical CDF [Arcones and Giné, 1992]
- Pearson curves by matching first four moments [Johnson et al., 1994]
- Large deviation bounds [Hoeffding, 1963, McDiarmid, 1989]
- Consistent test using kernel eigenspectrum [NIPS09b]



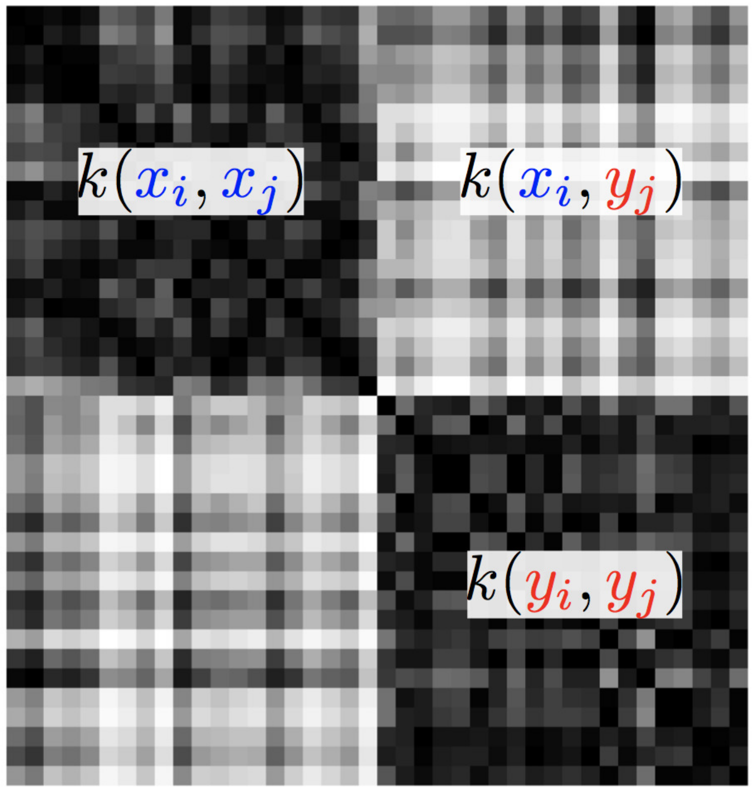
Approximate null distribution of $\widehat{MMD}$ via permutation

Original empirical MMD for dogs and fish:

$X = [$  $\dots]$

$Y = [$  $\dots]$

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



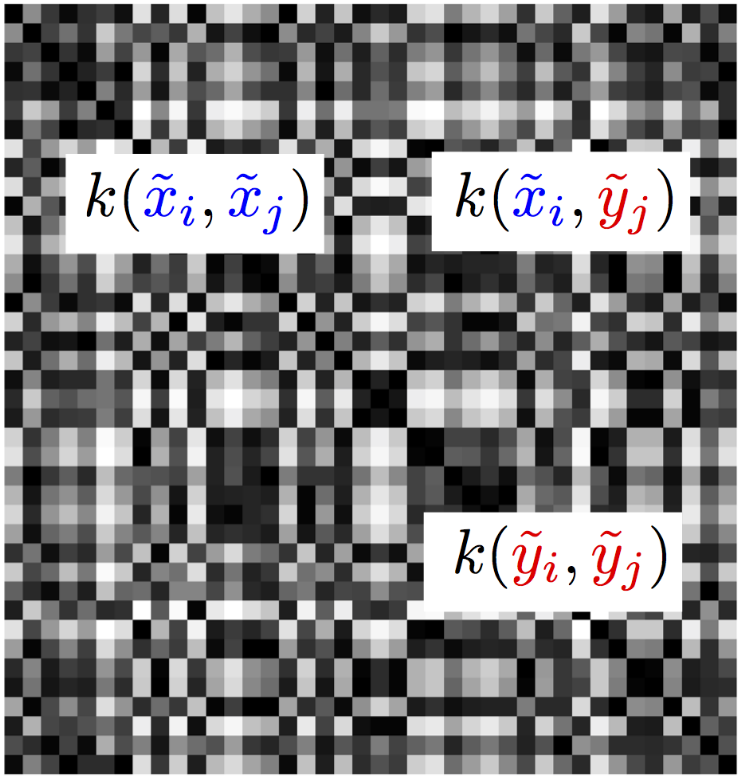
Approximate null distribution of $\widehat{MMD}$ via permutation

Permuted dog and fish samples (**merdogs**):

$\widetilde{X} = [$  $\dots]$

$\widetilde{Y} = [$  $\dots]$

$$\widehat{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{x}_i, \tilde{x}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\tilde{y}_i, \tilde{y}_j) - \frac{2}{n^2} \sum_{i,j} k(\tilde{x}_i, \tilde{y}_j)$$

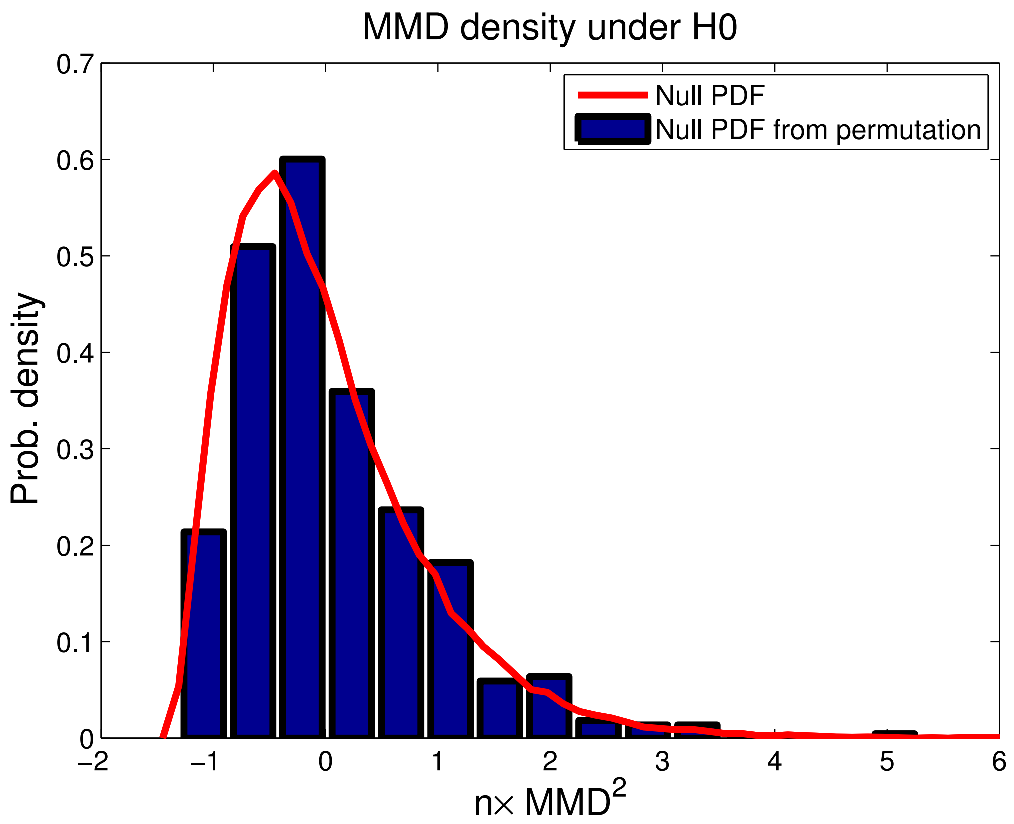


Permutation simulates

$P = Q$

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

- Null distribution estimated from 500 permutations
- $P = Q = \mathcal{N}(0, 1)$



Consistent test w/o bootstrap (not examinable)

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

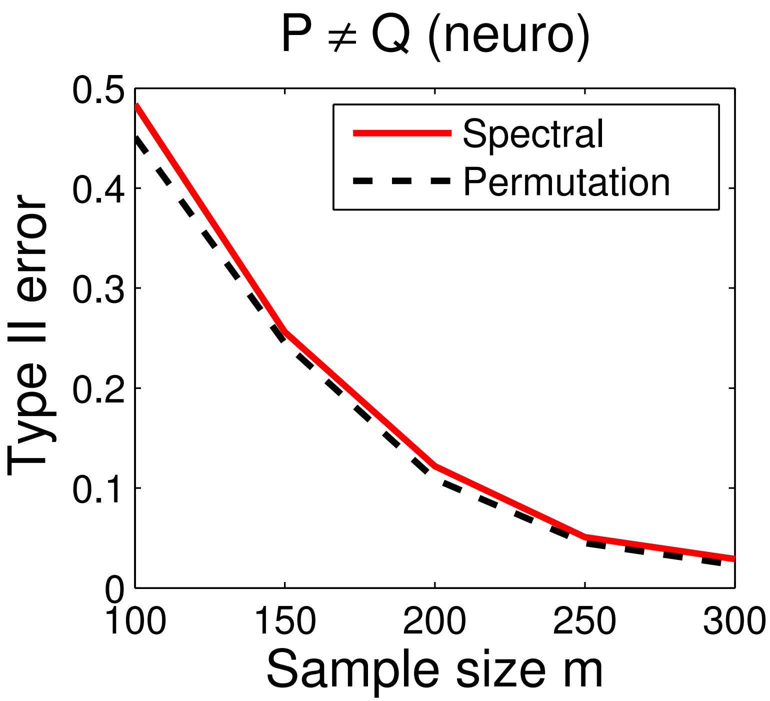
$$\mathrm{MMD}(\mathbf{P}, \mathbf{Q}; F) := \|\mu_{\mathbf{P}} - \mu_{\mathbf{Q}}\|^2_{\mathcal{F}}$$

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

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

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

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



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

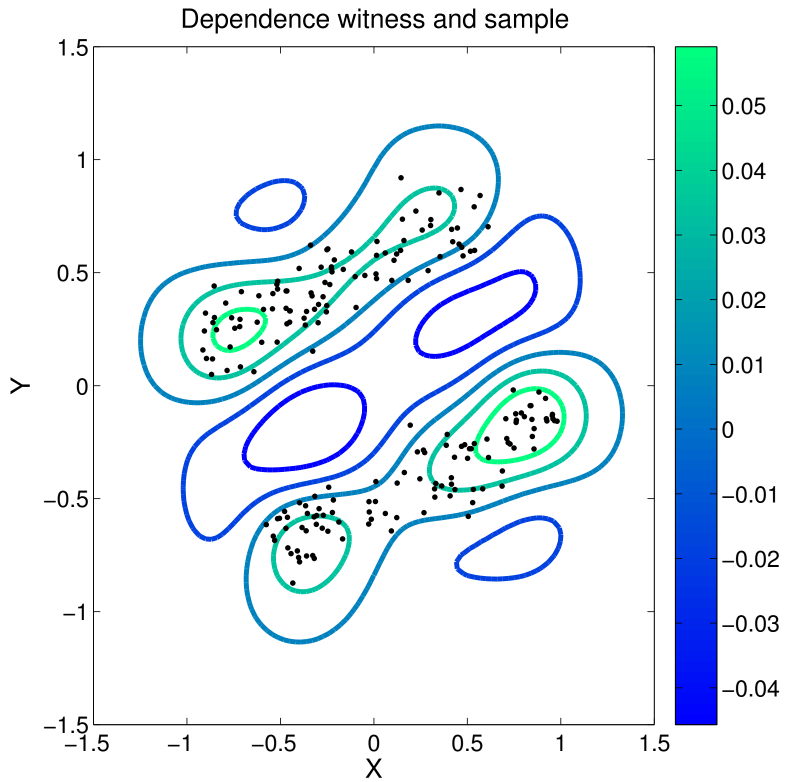
Kernel dependence measures

Reminder: MMD can be used as a dependence measure

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

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

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



Kernels on image-caption pairs

Kernel k on images with feature space $\mathcal{F}$,

$k($  ,  $)$

Kernel l on captions with feature space $\mathcal{G}$,

$l($ A large animal who slings slobber, ... , A responsive, interactive pet ... $)$

Kernels on image-caption pairs

Kernel k on images with feature space $\mathcal{F}$,

$k($  ,  $)$

Kernel l on captions with feature space $\mathcal{G}$,

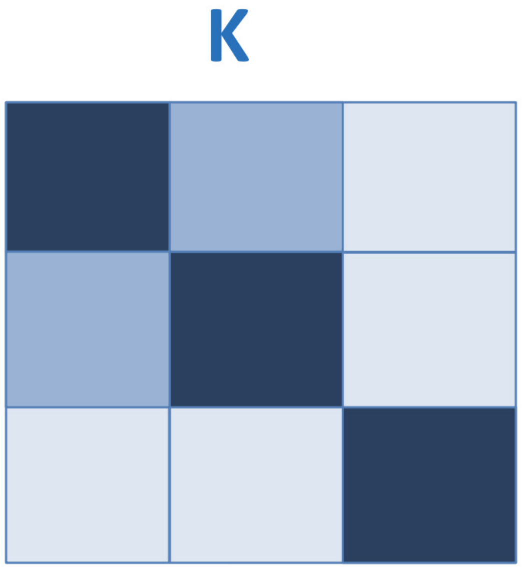
$l($ A large animal who slings slobber, ... , A responsive, interactive pet ... $)$

Kernel κ on image-text *pairs*: are images **and** captions similar?

$\kappa($  A large animal who slings slobber, ... ,  A responsive, interactive pet, ... $)$

$=k($  ,  $)\times l($ A large animal who slings slobber, ... , A responsive, interactive pet, ... $)$

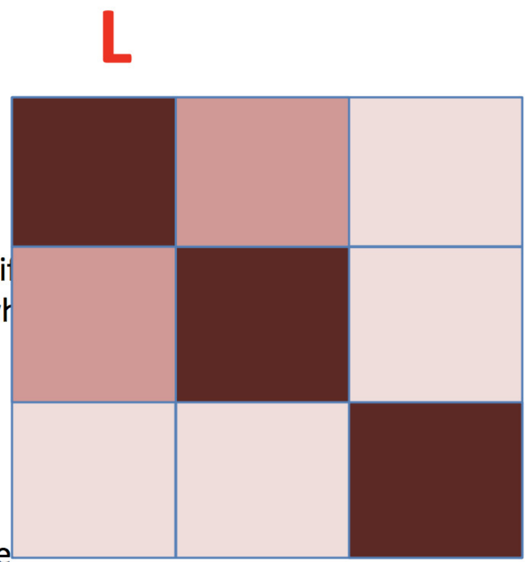
HSIC: empirical estimate and intuition



A large animal who slings slobber, exudes a distinctive houndy odor, ...

Their noses guide them through li
and they're never happier than wh
following an interesting scent.

A responsive, interactive pet, one that will blow in your ear and follow you everywhere.



Text from dogtime.com and petfinder.com

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

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

MMD as a dependence measure

Two questions:

- Why the product kernel? Many ways to combine kernels - why not eg a sum?
- Is there a more interpretable way of defining this dependence measure?

Covariance to reveal dependence

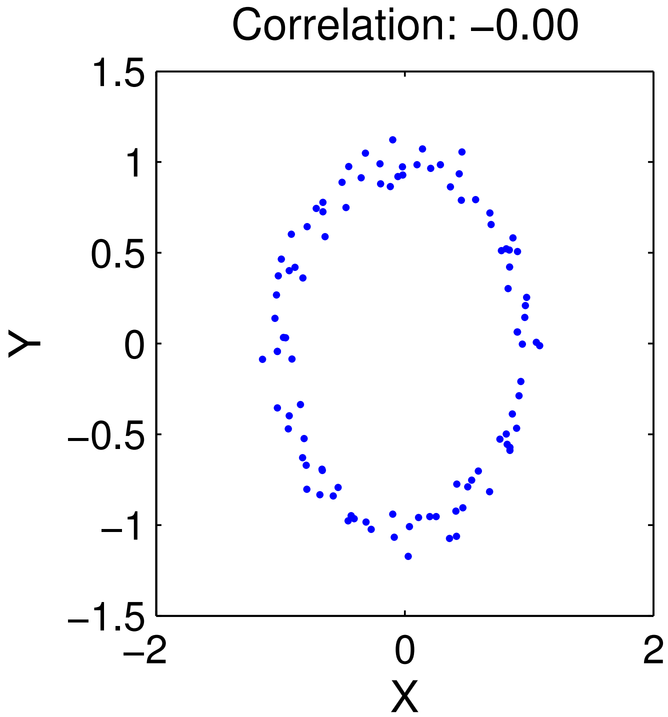
A more intuitive idea: maximize covariance of smooth mappings:

$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$

Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

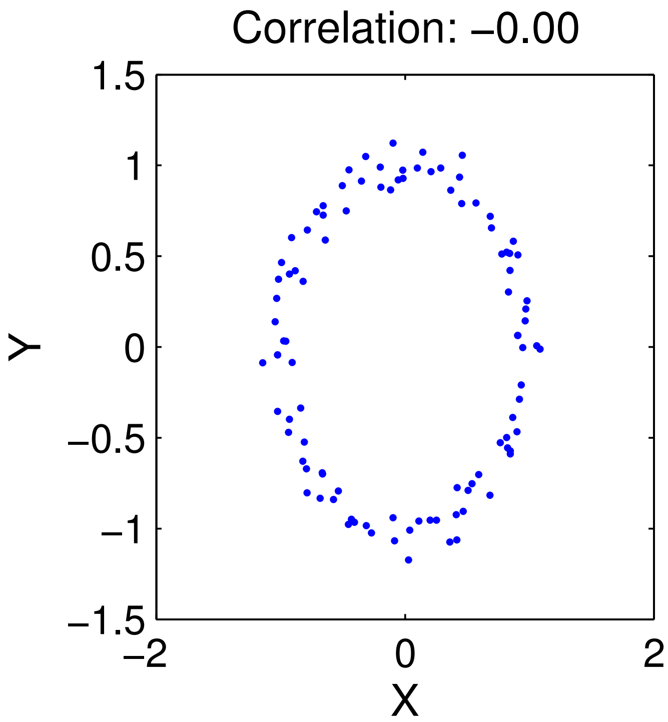
$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$

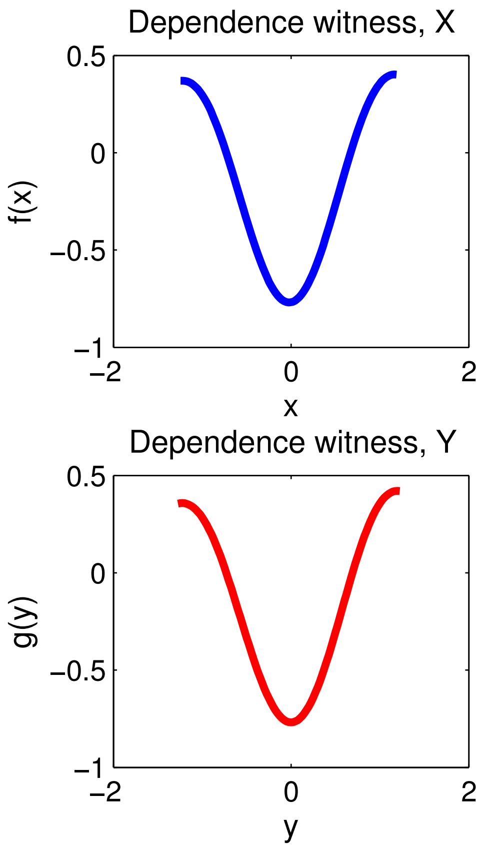


Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

$$\mathrm{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})] \right)$$

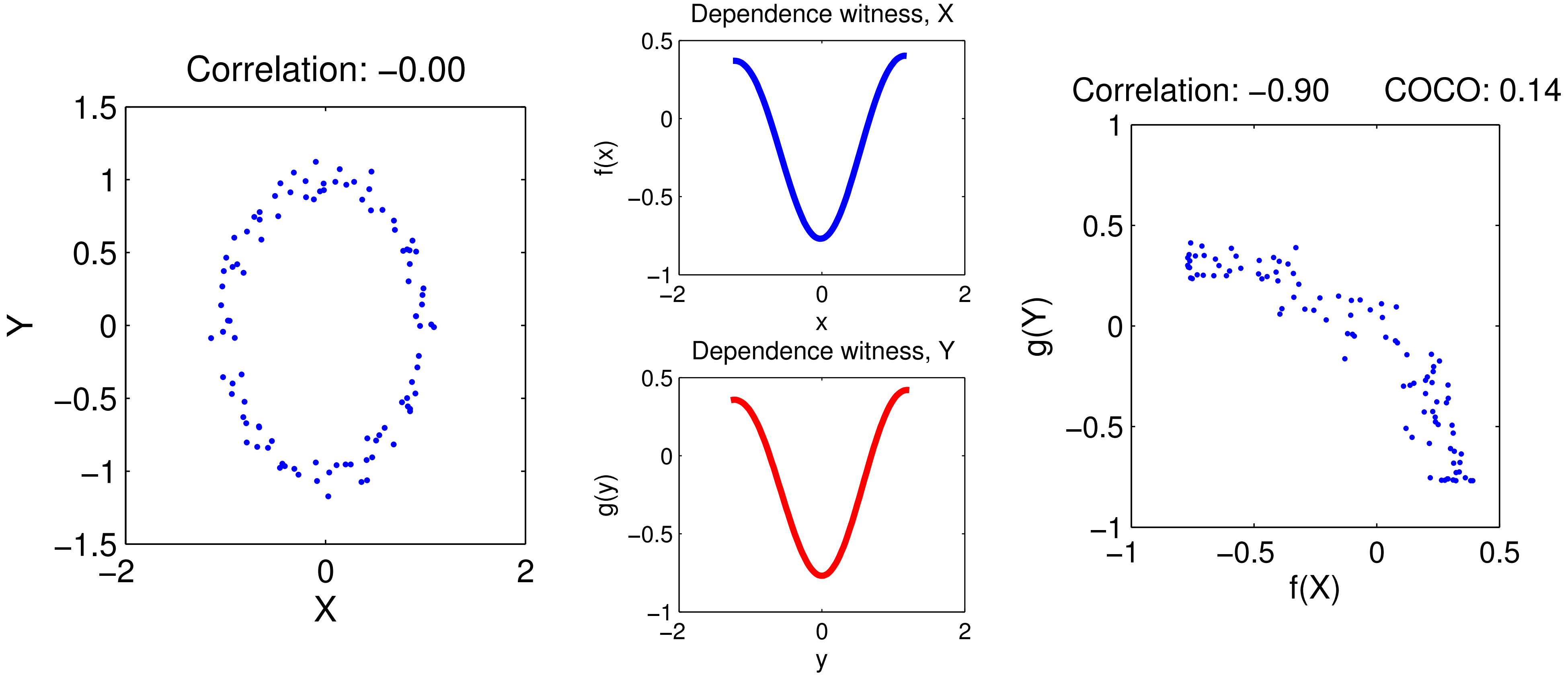




Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

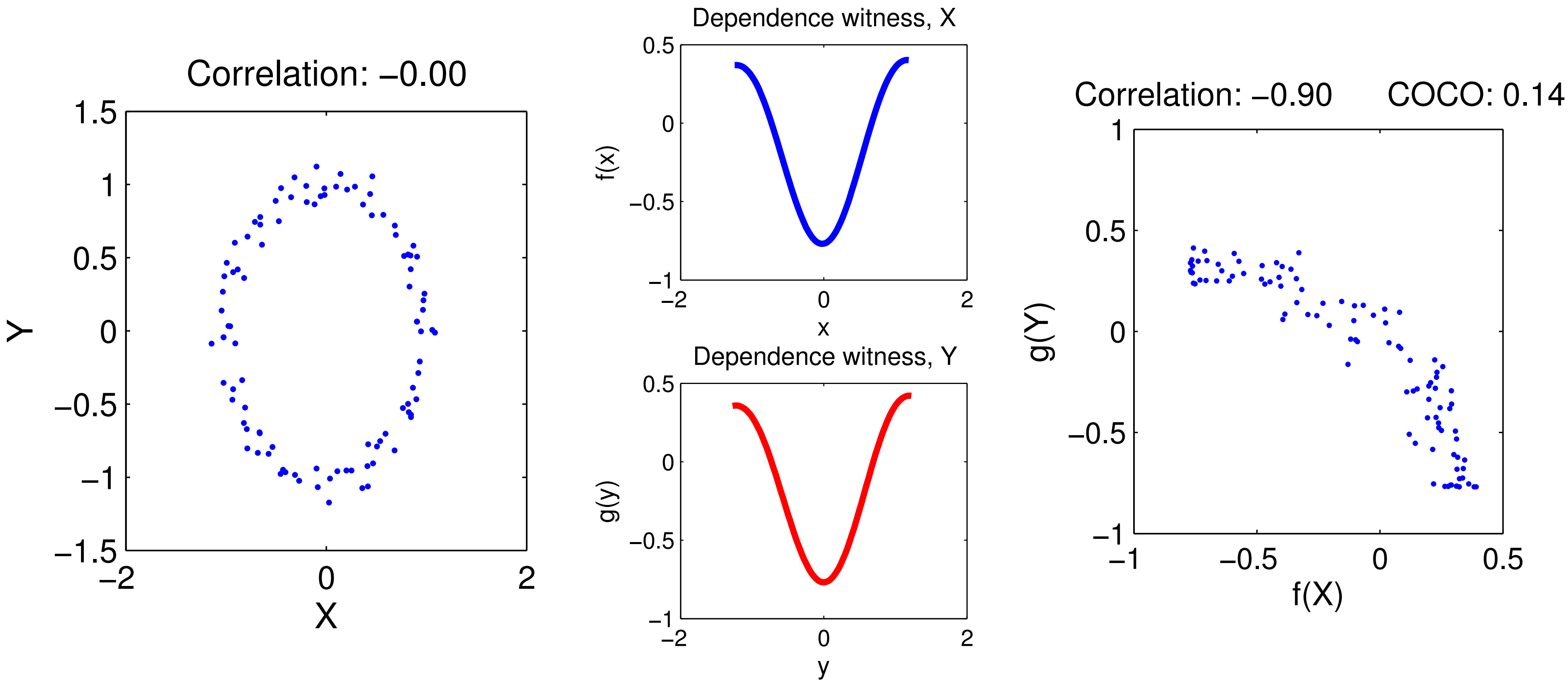
$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)]\right)$$



Covariance to reveal dependence

A more intuitive idea: maximize covariance of smooth mappings:

$$\mathrm{COCO}(\mathbf{P}; \mathcal{F}, \mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1, \|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})] \right)$$



How do we define covariance in infinite feature space?

Covariance to reveal dependence

How do we do this in RKHS? Let's first look at finite linear case.

We have two random vectors $\mathsf{x} \in \mathbb{R}^d$, $\mathsf{y} \in \mathbb{R}^{d'}$. Are they linearly dependent?

Covariance to reveal dependence

How do we do this in RKHS? Let's first look at finite linear case.

We have two random vectors $\mathsf{x} \in \mathbb{R}^d$, $\mathsf{y} \in \mathbb{R}^{d'}$. Are they linearly dependent?

Compute their covariance matrix: (ignore centering)

$$C_{xy} = \mathbf{E}\left(\mathsf{x}\mathsf{y}^\top\right)$$

...but this is a $d \times d'$ matrix! How to get a single "summary" number?

Covariance to reveal dependence

How do we do this in RKHS? Let's first look at finite linear case.

We have two random vectors $\mathsf{x} \in \mathbb{R}^d$, $\mathsf{y} \in \mathbb{R}^{d'}$. Are they linearly dependent?

Compute their covariance matrix: (ignore centering)

$$C_{xy} = \mathbf{E}\left(\mathsf{x}\mathsf{y}^\top\right)$$

...but this is a $d \times d'$ matrix! How to get a single "summary" number?

Solve for vectors $f \in \mathbb{R}^d$, $g \in \mathbb{R}^{d'}$

$$\begin{aligned}
\underset{\|f\|=1,\|g\|=1}{\operatorname{argmax}} f^\top C_{xy} g &= \underset{\|f\|=1,\|g\|=1}{\operatorname{argmax}} \mathbf{E}_{\mathsf{x}\mathsf{y}}\left[\left(f^\top \mathsf{x}\right)\left(g^\top \mathsf{y}\right)\right] \\
&= \underset{\|f\|=1,\|g\|=1}{\operatorname{argmax}} \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] = \underset{\|f\|=1,\|g\|=1}{\operatorname{argmax}} \operatorname{cov}\left(f(\mathsf{x})g(\mathsf{y})\right)
\end{aligned}$$

(maximum singular value) of C_{xy}.

Challenges in defining feature space covariance

Given features $\phi(x) \in \mathcal{F}$ and $\psi(y) \in \mathcal{G}$:

Challenge 1: Can we define a feature space analog to $x\, y^\top$?
YES:

- Given $f \in \mathbb{R}^d$, $g \in \mathbb{R}^{d'}$, $h \in \mathbb{R}^{d'}$, define matrix $f\, g^\top$ such that $(f\, g^\top)h = f(g^\top h)$.
- Given $f \in \mathcal{F}$, $g \in \mathcal{G}$, $h \in \mathcal{G}$, define tensor product operator $f \otimes g$ such that $(f \otimes g)h = f\langle g, h\rangle_{\mathcal{G}}$.
- Now just set $f := \phi(x)$, $g = \psi(y)$, to get $x\, y^\top \to \phi(x) \otimes \psi(y)$

Challenges in defining feature space covariance

Given features $\phi(x) \in \mathcal{F}$ and $\psi(y) \in \mathcal{G}$:

Challenge 2: Does a covariance "matrix" (operator) in feature space exist?
I.e. is there some $C_{XY} : \mathcal{G} \to \mathcal{F}$ such that

$$\langle f, C_{XY} g \rangle_{\mathcal{F}} = \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] = \operatorname{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)$$

Does "something" exist $\to$ Riesz theorem. Can we write (finite dimensional) covariance as a dot product?

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

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

Challenges in defining feature space covariance

Given features $\phi(x) \in \mathcal{F}$ and $\psi(y) \in \mathcal{G}$:

Challenge 2: Does a covariance "matrix" (operator) in feature space exist?
I.e. is there some $C_{XY} : \mathcal{G} \to \mathcal{F}$ such that

$$\langle f, C_{XY} g \rangle_{\mathcal{F}} = \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] = \operatorname{cov}(f(\mathsf{x}), g(\mathsf{y}))$$

Hints:

- In the finite dimensional case, and given basis vectors $g_j \in \mathbb{R}^{d'}$, $C_{XY} \in \mathbb{R}^{d \times d'}$ is in a vector space, with inner product

$$\langle C_{XY}, A \rangle_{\mathrm{HS}} = \operatorname{trace}(C_{XY}{}^{\top} A) = \sum_{j \in J} (C_{XY} g_j)^{\top} (A g_j),$$

- In particular

$$\langle C_{XY}, f\, g^{\top} \rangle_{\mathrm{HS}} = \operatorname{trace}(C_{XY}{}^{\top} (f\, g)^{\top}) = f^{\top} C_{XY} g = \mathbf{E}_{\mathsf{xy}} [f(\mathsf{x}) g(\mathsf{y})]$$

Challenges in defining feature space covariance

Given features $\phi(x) \in \mathcal{F}$ and $\psi(y) \in \mathcal{G}$:

Challenge 2 (reformulated via the hints): does there exist $C_{XY} : \mathcal{G} \to \mathcal{F}$ in a Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ such that:

$$\langle C_{XY}, A \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}$$

and in particular,

$$\langle C_{XY}, f \otimes g \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{xy}} \left[f(\mathsf{x}) g(\mathsf{y}) \right]$$

The Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

- $\mathcal{F}$ and $\mathcal{G}$ separable Hilbert spaces.
- $(g_j)_{j \in J}$ orthonormal basis for $\mathcal{G}$.
- Index set J either finite or countably infinite.

$$\langle g_i, g_j \rangle_{\mathcal{G}} := \begin{cases} 1 & i = j, \\ 0 & i \neq j \end{cases}$$

The Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

- $\mathcal{F}$ and $\mathcal{G}$ separable Hilbert spaces.
- $(g_j)_{j\in J}$ orthonormal basis for $\mathcal{G}$.
- Index set J either finite or countably infinite.
- Linear operators $L : \mathcal{G} \to \mathcal{F}$ and $M : \mathcal{G} \to \mathcal{F}$.
- Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$, with inner product

$$\langle L, M \rangle_{\mathrm{HS}} = \sum_{j\in J} \langle L g_j, M g_j \rangle_{\mathcal{F}} , \tag{2}$$

(independent of orthonormal basis)

The Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

- $\mathcal{F}$ and $\mathcal{G}$ separable Hilbert spaces.
- $(g_j)_{j \in J}$ orthonormal basis for $\mathcal{G}$.
- Index set J either finite or countably infinite.
- Linear operators $L : \mathcal{G} \to \mathcal{F}$ and $M : \mathcal{G} \to \mathcal{F}$.
- Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$, with inner product

$$\langle L, M \rangle_{\mathrm{HS}} = \sum_{j \in J} \langle L g_j, M g_j \rangle_{\mathcal{F}} \,, \tag{3}$$

(independent of orthonormal basis)

- Hilbert-Schmidt norm of the operators L:

$$\|L\|_{\mathrm{HS}}^2 = \sum_{j \in J} \|L g_j\|_{\mathcal{F}}^2$$

L is Hilbert-Schmidt when this norm is finite.

The tensor product $a \otimes b$ is in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, we earlier defined the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$ (generalize finite case $a\, b^\top$)

$$(a \otimes b) g \;\mapsto\; \langle g, b \rangle_{\mathcal{G}}\, a.$$

Is $a \otimes b \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$?

The tensor product $a \otimes b$ is in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given $a \in \mathcal{F}$ and $b \in \mathcal{G}$, we earlier defined the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$ (generalize finite case $a\, b^\top$)

$$(a \otimes b)g \;\mapsto\; \langle g, b \rangle_{\mathcal{G}}\, a.$$

Is $a \otimes b \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$?

$$\begin{aligned}
\|a \otimes b\|^2_{\mathrm{HS}} &= \sum_{j \in J} \|(a \otimes b)g_j\|^2_{\mathcal{F}} \\
&= \sum_{j \in J} \left\|a \, \langle b, g_j \rangle_{\mathcal{G}}\right\|^2_{\mathcal{F}} \\
&= \|a\|^2_{\mathcal{F}} \sum_{j \in J} \left|\langle b, g_j \rangle_{\mathcal{G}}\right|^2 \\
&= \|a\|^2_{\mathcal{F}} \|b\|^2_{\mathcal{G}},
\end{aligned} \tag{5}$$

where we use Parseval's identity. Thus, the operator is Hilbert-Schmidt.

Inner product of $a \otimes b$ with $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given a Hilbert-Schmidt operator $L : \mathcal{G} \to \mathcal{F}$,

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \tag{6}$$

Special case:

$$\langle u \otimes v, a \otimes b \rangle_{\mathrm{HS}} = \langle u, a \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}} .$$

Proof: Use expansion

$$b = \sum_{j \in J} \langle b, g_j \rangle_{\mathcal{G}} \, g_j$$

Inner product of $a \otimes b$ with $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$

Given a Hilbert-Schmidt operator $L : \mathcal{G} \to \mathcal{F}$,

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \tag{7}$$

Special case:

$$\langle u \otimes v, a \otimes b \rangle_{\mathrm{HS}} = \langle u, a \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}} .$$

Proof: Use expansion

$$b = \sum_{j \in J} \langle b, g_j \rangle_{\mathcal{G}} \, g_j$$

Then

$$\begin{aligned} \langle a, Lb \rangle &= \left\langle a, L \left(\sum_j \langle b, g_j \rangle_{\mathcal{G}} g_j \right) \right\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, g_j \rangle_{\mathcal{G}} \, \langle a, L g_j \rangle_{\mathcal{F}} \end{aligned}$$

Inner product of $a \otimes b$ with $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$

Proof (continued)

$$\begin{aligned}
\langle a \otimes b, L \rangle_{\mathrm{HS}} &:= \sum_j \langle L g_j, (a \otimes b) g_j \rangle_{\mathcal{F}} \\
&= \sum_j \langle b, g_j \rangle_{\mathcal{G}} \langle L g_j, a \rangle_{\mathcal{F}} .
\end{aligned}$$

Covariance operator in RKHS

Given RKHS $\mathcal{F}$ with feature map $\phi(x)$ and kernel $k(x, x')$, RKHS $\mathcal{G}$ with feature map $\psi(x)$ and kernel $l(y, y')$.

Challenge 2 (reminder): does there exist $C_{XY} : \mathcal{G} \to \mathcal{F}$ in some Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ such that:

$$\langle C_{XY}, A \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}$$

and in particular,

$$\langle C_{XY}, f \otimes g \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{xy}} [f(\mathsf{x}) g(\mathsf{y})] = \mathrm{cov} [f(\mathsf{x}) g(\mathsf{y})]$$

(ignoring centering)

Covariance operator in RKHS

Given RKHS $\mathcal{F}$ with feature map $\phi(x)$ and kernel $k(x, x')$, RKHS $\mathcal{G}$ with feature map $\psi(x)$ and kernel $l(y, y')$.

Challenge 2 (reminder): does there exist $C_{XY} : \mathcal{G} \to \mathcal{F}$ in some Hilbert space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ such that:

$$\langle C_{XY}, A \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}$$

and in particular,

$$\langle C_{XY}, f \otimes g \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{xy}} \left[f(\mathsf{x}) g(\mathsf{y}) \right] = \mathrm{cov} \left[f(\mathsf{x}) g(\mathsf{y}) \right]$$

(ignoring centering)

- Define $\phi(\mathsf{x}) \otimes \psi(\mathsf{y})$ a random variable in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$
- The covariance operator, written C_{XY}, is the unique element satisfying

$$\langle C_{XY}, A \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}} \tag{9}$$

Covariance operator in RKHS

Proof: Use Riesz representer theorem. The operator

$$\begin{aligned} T_{\mathsf{xy}} : \mathrm{HS}(\mathcal{G},\mathcal{F}) &\to \mathbb{R} \\ A &\mapsto \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}} \end{aligned}$$

is bounded when $\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right) < \infty$, since

$$\begin{aligned} \left| \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}} \right| &\leq \mathbf{E}_{\mathsf{x},\mathsf{y}} \left| \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}} \right| \\ &\leq \|A\|_{\mathrm{HS}} \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right). \end{aligned}$$

(first Jensen, then Cauchy-Schwarz). Thus covariance operator exists by Riesz.

I.e. there exists C_{XY} such that

$$\langle C_{XY}, A \rangle_{HS} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}$$

Covariance operator in RKHS

Proof: Use Riesz representer theorem. The operator

$$\begin{aligned} T_{\mathsf{xy}} : \mathrm{HS}(\mathcal{G}, \mathcal{F}) &\to \mathbb{R} \\ A &\mapsto \mathbf{E}_{\mathsf{x,y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}} \end{aligned}$$

is bounded when $\mathbf{E}_{\mathsf{x,y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right) < \infty$, since

$$\begin{aligned} \left|\mathbf{E}_{\mathsf{x,y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}\right| &\leq \mathbf{E}_{\mathsf{x,y}} \left|\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}\right| \\ &\leq \|A\|_{\mathrm{HS}} \mathbf{E}_{\mathsf{x,y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right). \end{aligned}$$

(first Jensen, then Cauchy-Schwarz). Thus covariance operator exists by Riesz.

Simpler condition:

$$\begin{aligned} \mathbf{E}_{\mathsf{x,y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right) &= \mathbf{E}_{\mathsf{x,y}} \left(\|\phi(\mathsf{x})\|_{\mathcal{F}} \|\psi(\mathsf{y})\|_{\mathcal{G}} \right) \\ &= \mathbf{E}_{\mathsf{x,y}} \left(\sqrt{k(\mathsf{x},\mathsf{x}) l(\mathsf{y},\mathsf{y})} \right) < \infty. \end{aligned}$$

Covariance operator in RKHS

Now just prove the special case,

$$\langle C_{XY}, f \otimes g \rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{xy}} \left[f(\mathsf{x}) g(\mathsf{y}) \right]$$

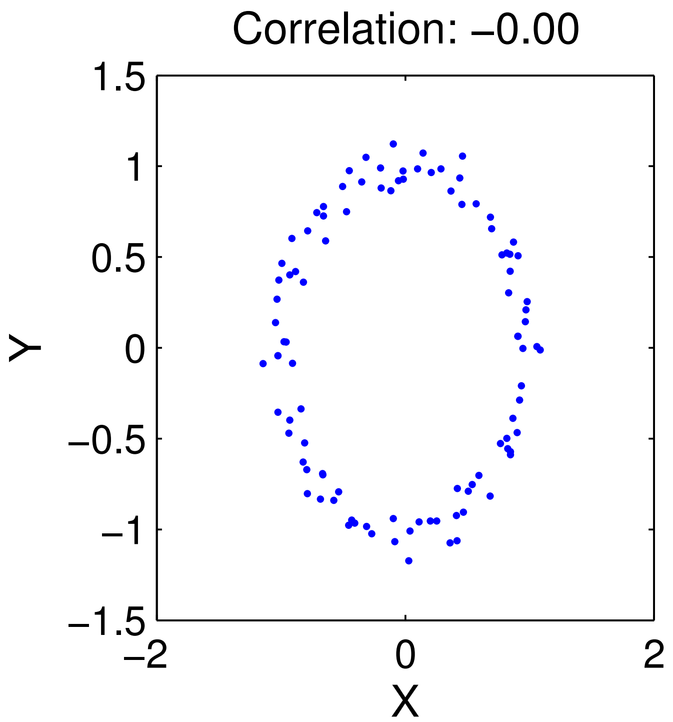
Proof:

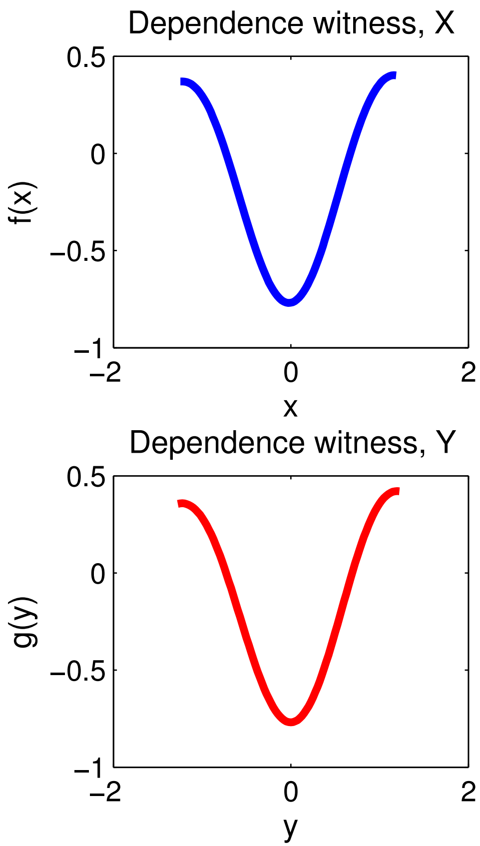
$$\begin{aligned}
\langle f, C_{XY} g \rangle_{\mathcal{F}} &= \langle C_{XY}, f \otimes g \rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), f \otimes g \rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{xy}} \left[\langle f, \phi(\mathsf{x}) \rangle_{\mathcal{F}} \langle g, \psi(\mathsf{y}) \rangle_{\mathcal{F}} \right] \\
&= \mathbf{E}_{\mathsf{xy}} \left[f(\mathsf{x}) g(\mathsf{y}) \right] \\
&= \mathrm{cov}(f, g).
\end{aligned}$$

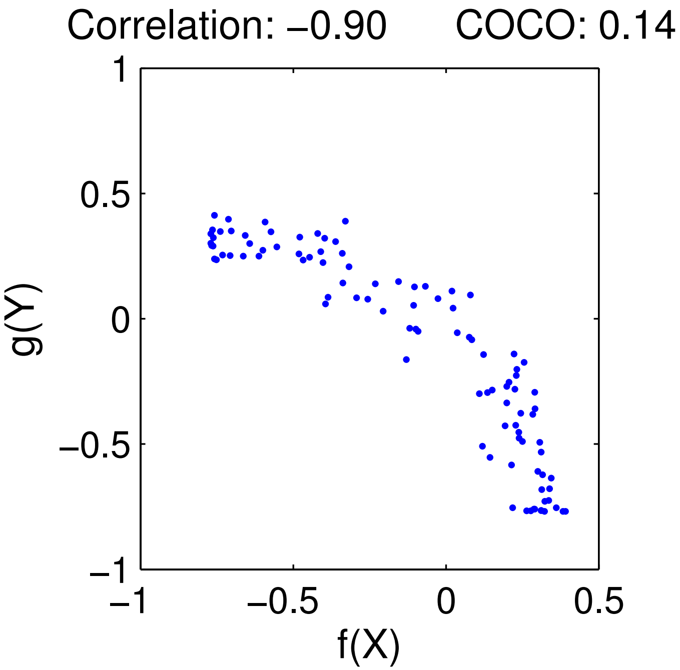
Thus, we proved C_{XY} exists and behaves as expected.

REMINDER: functions revealing dependence

$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$

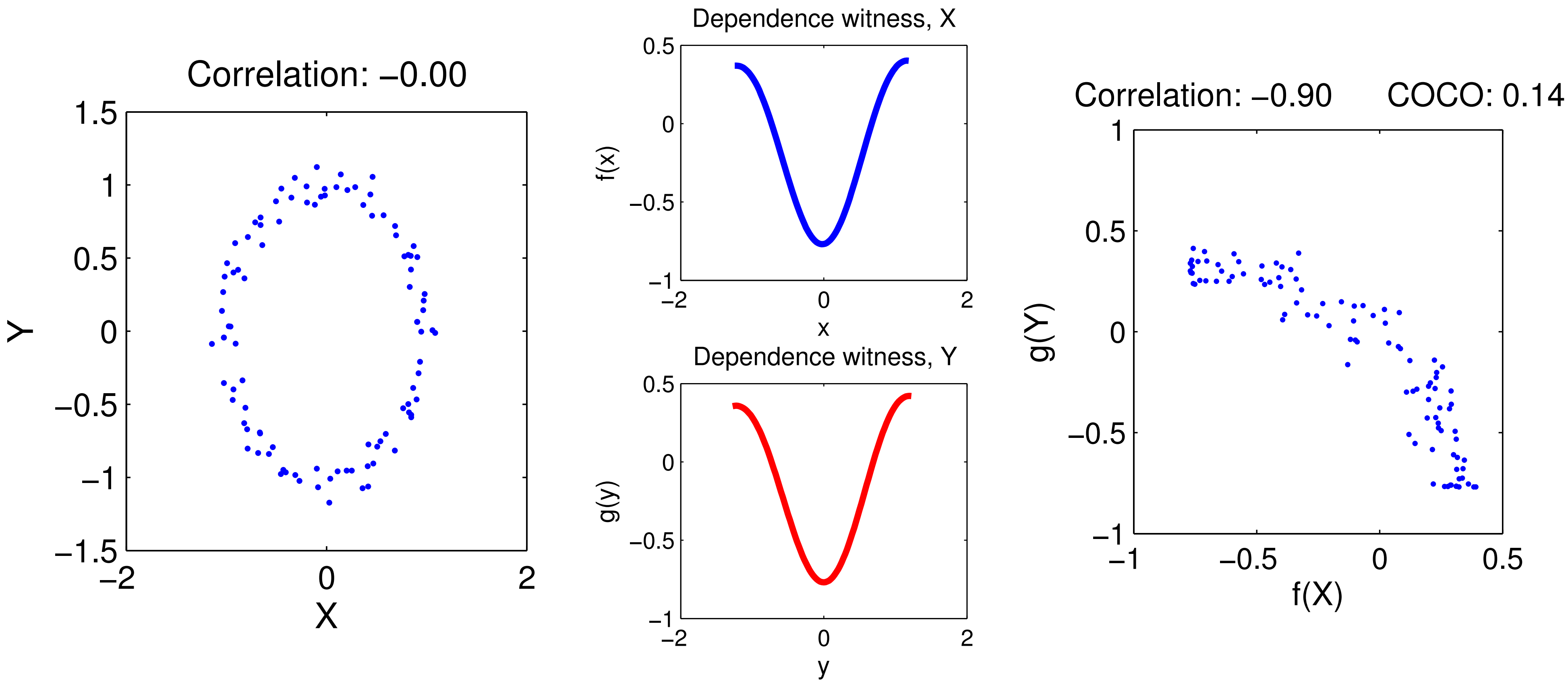






REMINDER: functions revealing dependence

$$\mathrm{COCO}(\mathbf{P};\mathcal{F},\mathcal{G}) := \sup_{\|f\|_{\mathcal{F}}=1,\|g\|_{\mathcal{G}}=1} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})] - \mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})]\right)$$



How do we compute this from finite data?

Empirical covariance operator

The empirical covariance given $\boldsymbol{z} := (x_i, y_i)_{i=1}^n$ (now include centring)

$$\widehat{C}_{XY} := \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \otimes \hat{\mu}_y,$$

where $\hat{\mu}_x := \frac{1}{n}\sum_{i=1}^n \phi(x_i)$. More concisely,

$$\widehat{C}_{XY} = \frac{1}{n} XHY^\top,$$

where $H = I_n - n^{-1}\mathbf{1}_n$, and $\mathbf{1}_n$ is an $n \times n$ matrix of ones, and

$$X = \begin{bmatrix} \phi(x_1) & \ldots & \phi(x_n) \end{bmatrix} \qquad Y = \begin{bmatrix} \psi(y_1) & \ldots & \psi(y_n) \end{bmatrix}.$$

Define the kernel matrices

$$K_{ij} = \left(X^\top X\right)_{ij} = k(x_i, x_j) \qquad L_{ij} = l(y_i, y_j),$$

Functions revealing dependence

Optimization problem:

$$\begin{aligned} \mathrm{COCO}(\boldsymbol{z};\mathcal{F},\mathcal{G}) := \max \quad & \left\langle f, \widehat{C}_{XY} g \right\rangle_{\mathcal{F}} \\ \text{subject to} \quad & \|f\|_{\mathcal{F}} = 1 && (10) \\ & \|g\|_{\mathcal{G}} = 1 && (11) \end{aligned}$$

Assume

$$f = \sum_{i=1}^{n} \alpha_i \left[\phi(x_i) - \hat{\mu}_x\right] = XH\alpha \qquad g = \sum_{j=1}^{n} \beta_i \left[\psi(y_i) - \hat{\mu}_y\right] = YH\beta,$$

The associated Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \langle f, \widehat{C}_{XY} g \rangle_{\mathcal{F}} - \frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) - \frac{\gamma}{2}\left(\|g\|_{\mathcal{F}}^2 - 1\right),$$

Functions revealing dependence

We now write this in terms of α and β:

$$\begin{aligned} f^\top \widehat{C}_{XY} g &= \frac{1}{n}\alpha^\top H X^\top \left(X H Y^\top \right) Y H \beta \\ &= \frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta, \end{aligned}$$

where we note that $H = HH$. Similarly

$$\|f\|_{\mathcal{F}}^2 = \alpha^\top H X X^\top H \alpha = \alpha^\top \widetilde{K}\alpha.$$

Substituting these into the Lagrangian,

$$\mathcal{L}(\alpha,\beta,\lambda,\gamma) = \frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta - \frac{\lambda}{2}\left(\alpha^\top \widetilde{K}\alpha - 1\right) - \frac{\gamma}{2}\left(\beta^\top \widetilde{L}\beta - 1\right).$$

Kernel matrices between centred variables,

$$\widetilde{K} = HKH \qquad \widetilde{L} = HLH$$

Functions revealing dependence

Maximize wrt the primal variables α, β.

Differentiating wrt α and β and setting to zero,

$$\frac{1}{n}\widetilde{K}\widetilde{L}\beta - \lambda\widetilde{K}\alpha = 0 \tag{12}$$

$$\frac{1}{n}\widetilde{L}\widetilde{K}\alpha - \gamma\widetilde{L}\beta = 0 \tag{13}$$

Multiply the first equation by $\alpha^\top$, and the second by $\beta^\top$,

$$\begin{aligned} \frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta &= \lambda\alpha^\top\widetilde{K}\alpha \\ \frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha &= \gamma\beta^\top\widetilde{L}\beta \end{aligned}$$

Functions revealing dependence

Subtracting first expression from the second,

$$\lambda \alpha^\top \widetilde{K} \alpha = \gamma \beta^\top \widetilde{L} \beta.$$

Recall the constraints $\alpha^\top \widetilde{K} \alpha = 1$ and $\beta^\top \widetilde{L} \beta = 1$. Thus $\lambda = \gamma$.

We must maximize the following expression relating α, β:

$$\begin{bmatrix} 0 & \frac{1}{n} \widetilde{K} \widetilde{L} \\ \frac{1}{n} \widetilde{L} \widetilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

At solution (given eq. (12) on previous slide, and $\alpha^\top \widetilde{K} \alpha = 1$),

$$\gamma^* = \frac{1}{n} \beta^\top \widetilde{L} \widetilde{K} \alpha = \mathrm{COCO}(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$$

Covariance to reveal dependence

- Empirical $\mathrm{COCO}(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$ largest eigenvalue of

$$\begin{bmatrix} 0 & \frac{1}{n}\widetilde{K}\widetilde{L} \\ \frac{1}{n}\widetilde{L}\widetilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

- $\widetilde{K}$ and $\widetilde{L}$ are matrices of inner products between centred observations in respective feature spaces:

$$\widetilde{K} = HKH \qquad \text{where} \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

Covariance to reveal dependence

- Empirical $\mathrm{COCO}(\boldsymbol{z};\mathcal{F},\mathcal{G})$ largest eigenvalue of

$$\begin{bmatrix} 0 & \frac{1}{n}\widetilde{K}\widetilde{L} \\ \frac{1}{n}\widetilde{L}\widetilde{K} & 0 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

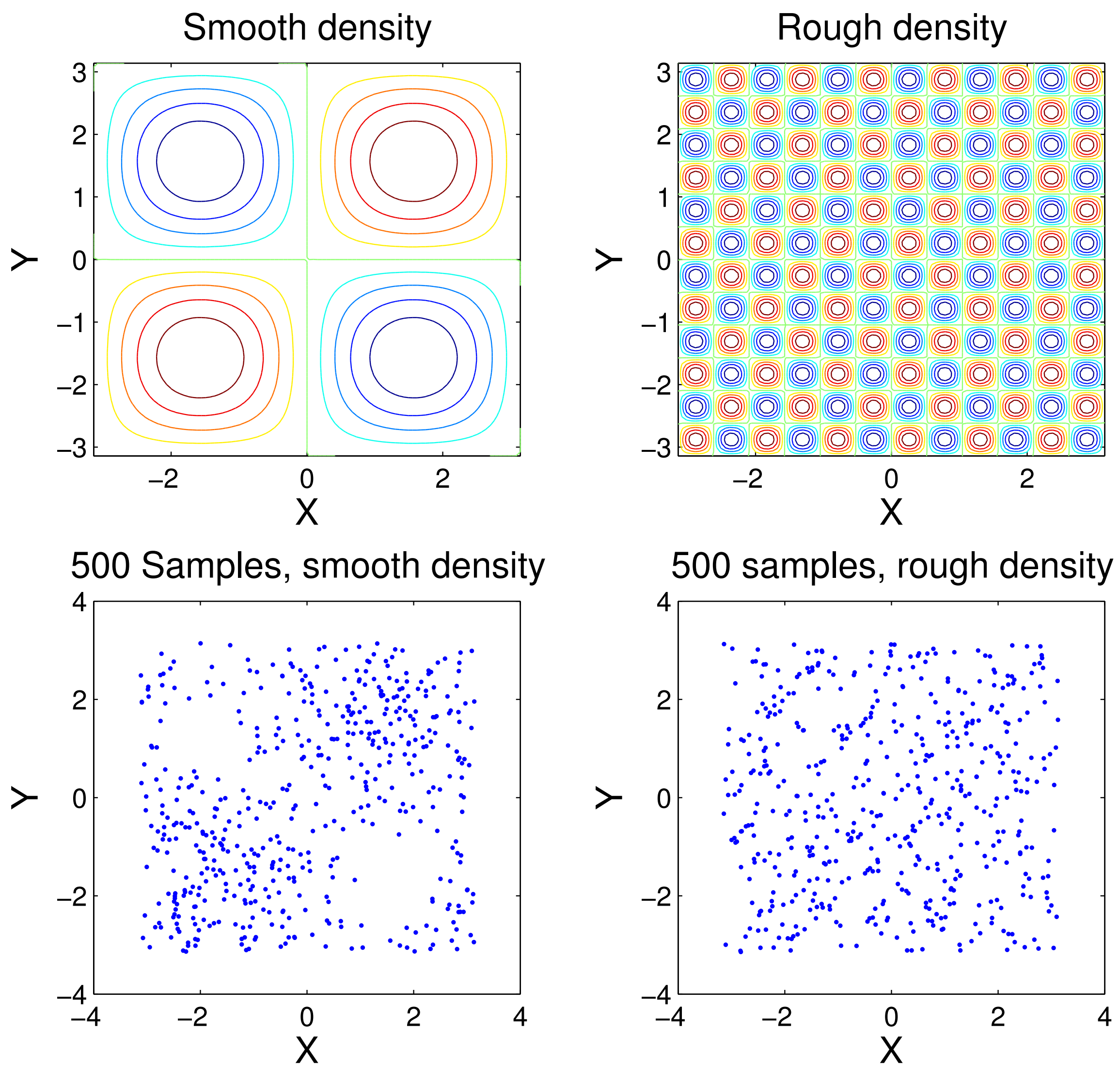
- $\widetilde{K}$ and $\widetilde{L}$ are matrices of inner products between centred observations in respective feature spaces:

$$\widetilde{K} = HKH \qquad \text{where} \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

- Mapping function for x:

$$f(x) = \sum_{i=1}^{n} \alpha_i \left(k(x_i, x) - \frac{1}{n}\sum_{j=1}^{n} k(x_j, x) \right)$$

Hard-to-detect dependence

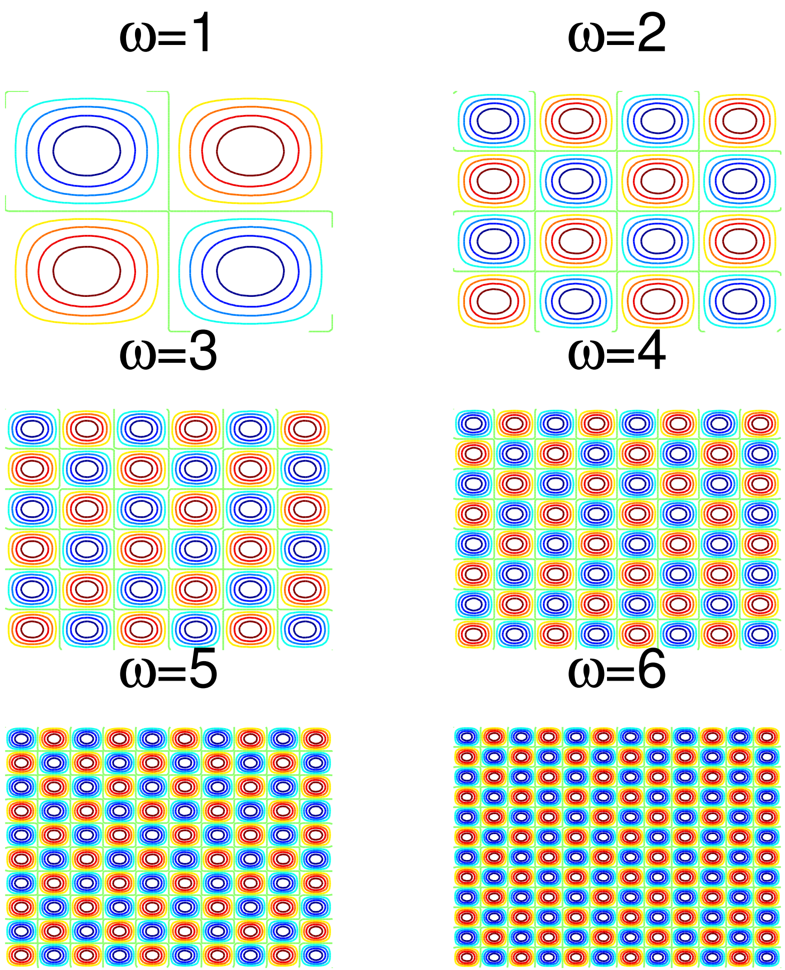


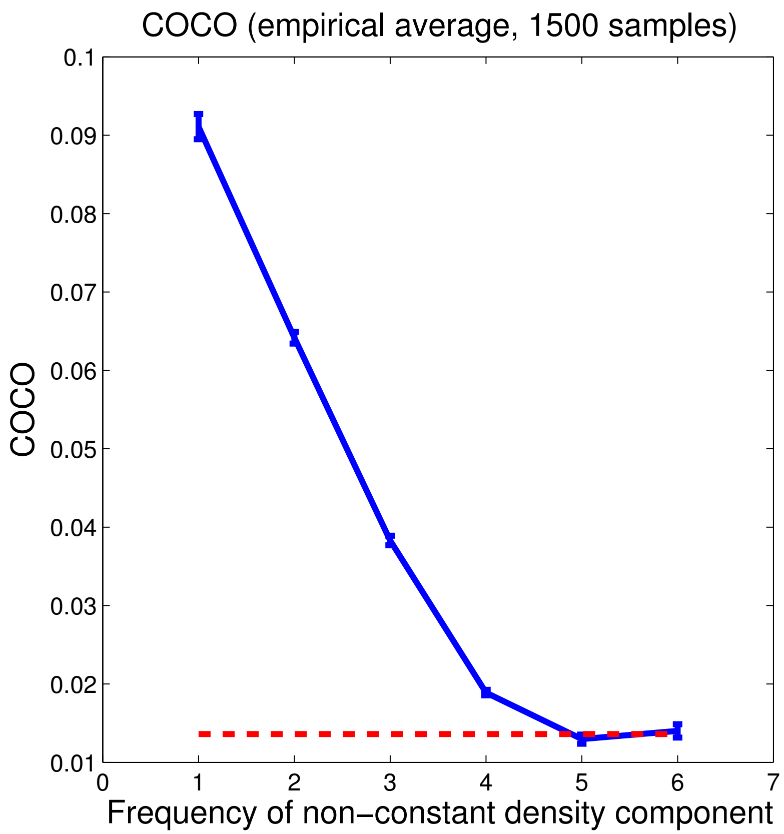
Density takes the form:

$$\mathbf{P}_{x,y} \propto 1 + \sin(\omega x)\sin(\omega y)$$

Hard-to-detect dependence

- Example: sinusoids of increasing frequency





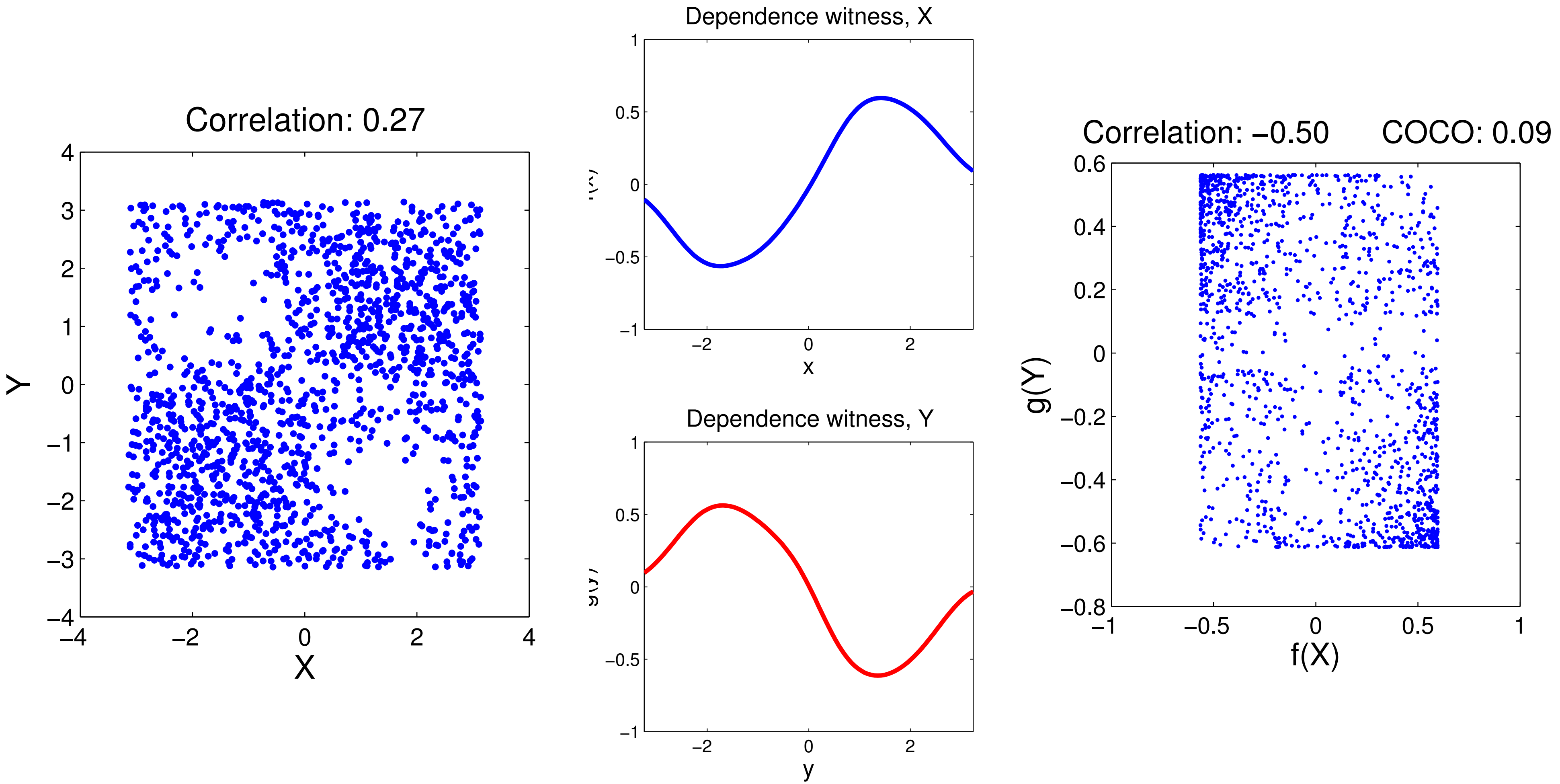
Hard-to-detect dependence

Why does COCO decay when dependence encoded at higher frequencies?

Hard-to-detect dependence

Why does COCO decay when dependence encoded at higher frequencies?

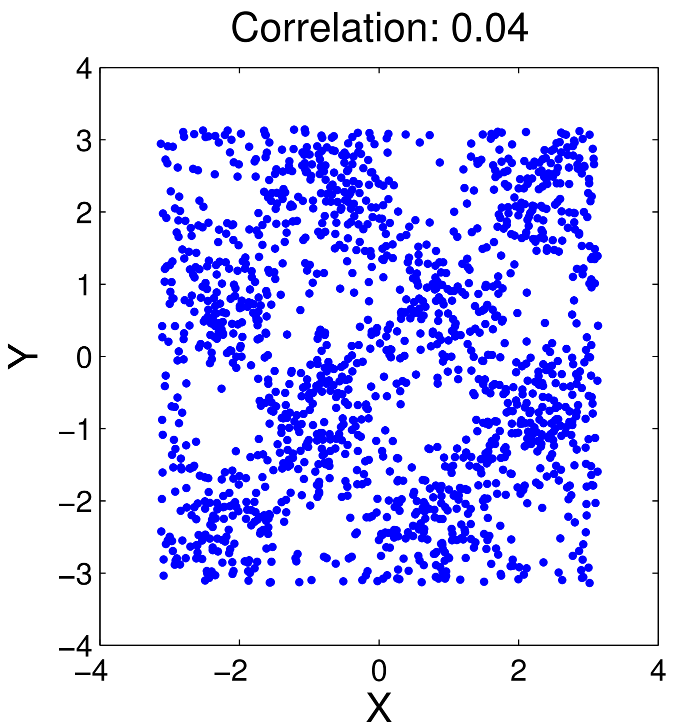
Case of $\omega = 1$

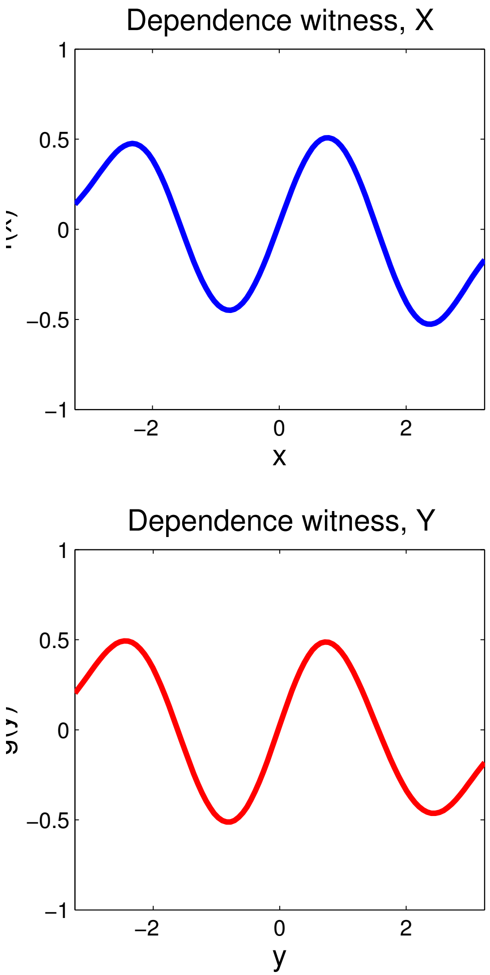


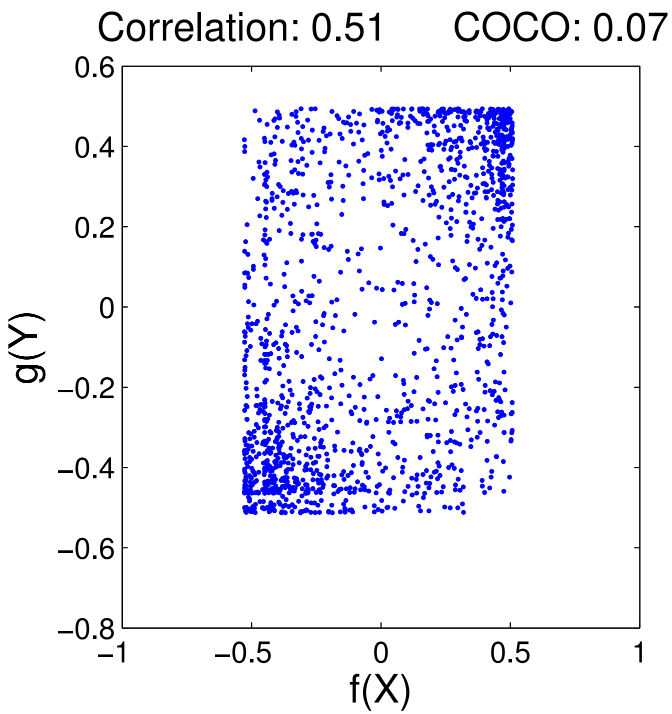
Hard-to-detect dependence

Why does COCO decay when dependence encoded at higher frequencies?

Case of $\omega = 2$



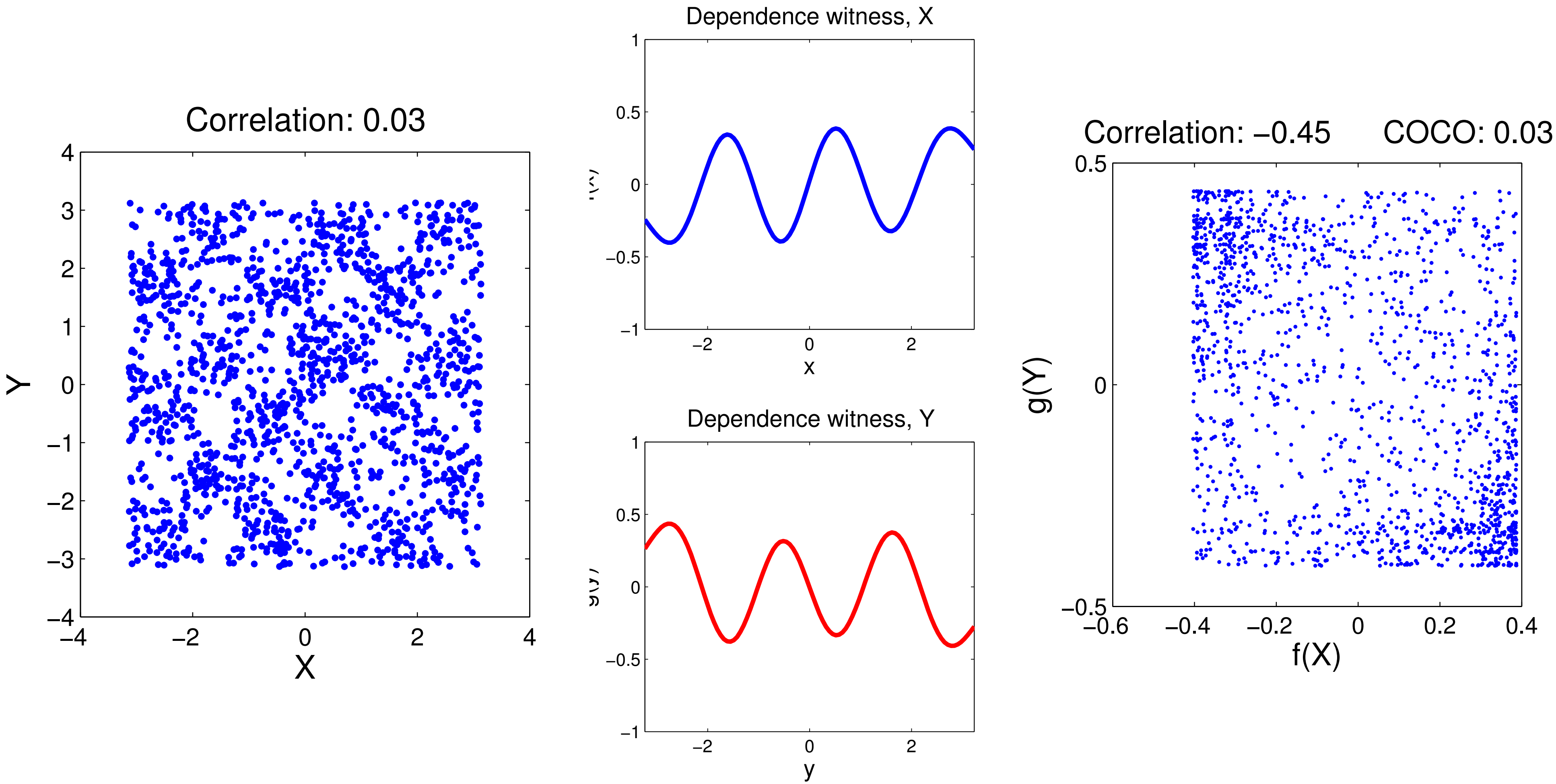




Hard-to-detect dependence

Why does COCO decay when dependence encoded at higher frequencies?

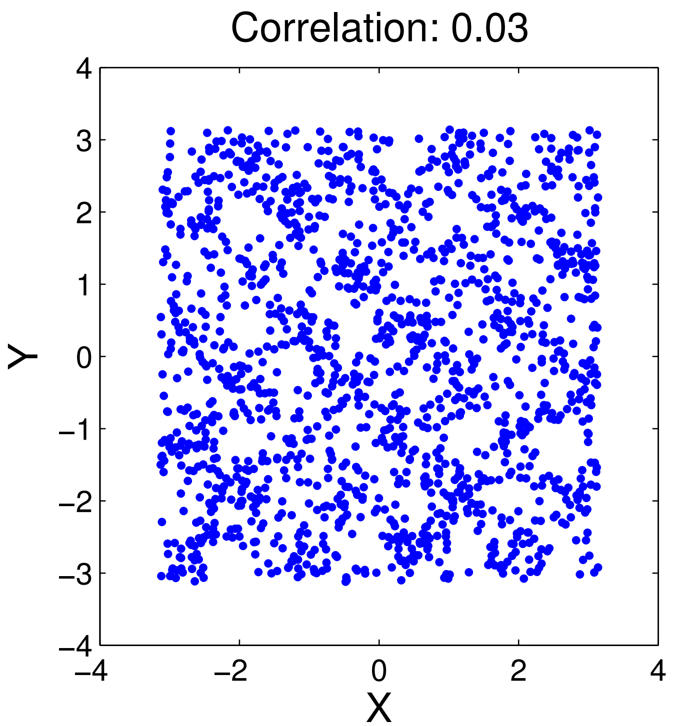
Case of $\omega = 3$

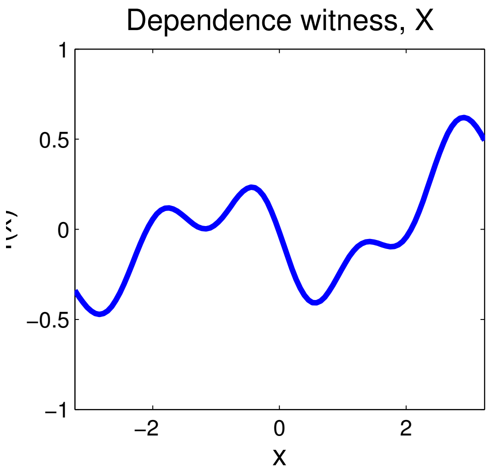


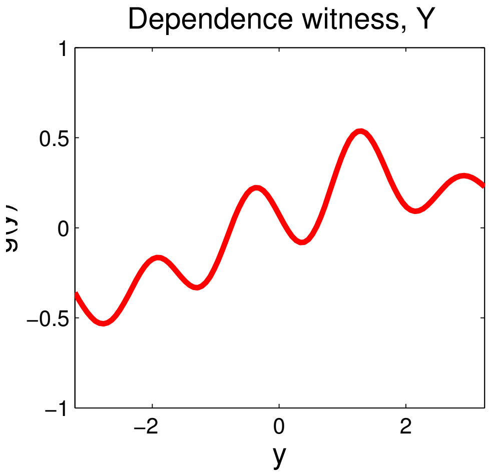
Hard-to-detect dependence

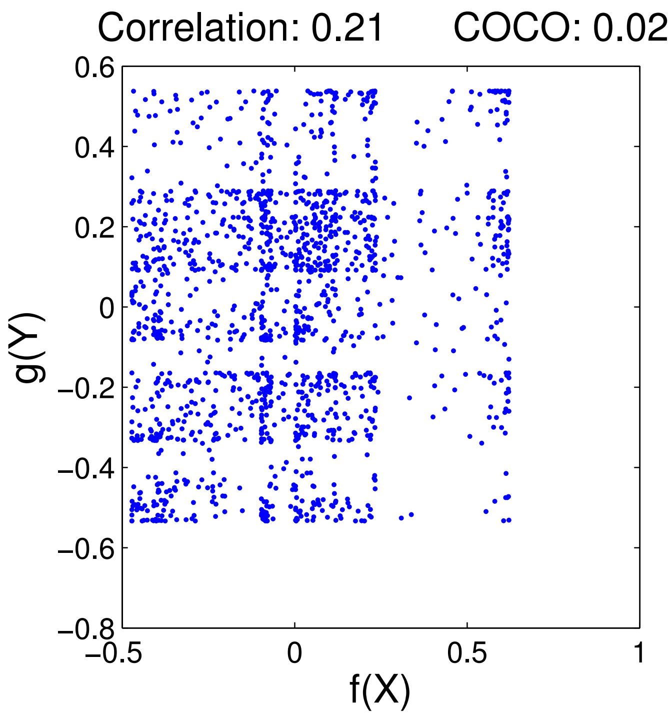
Why does COCO decay when dependence encoded at higher frequencies?

Case of $\omega = 4$





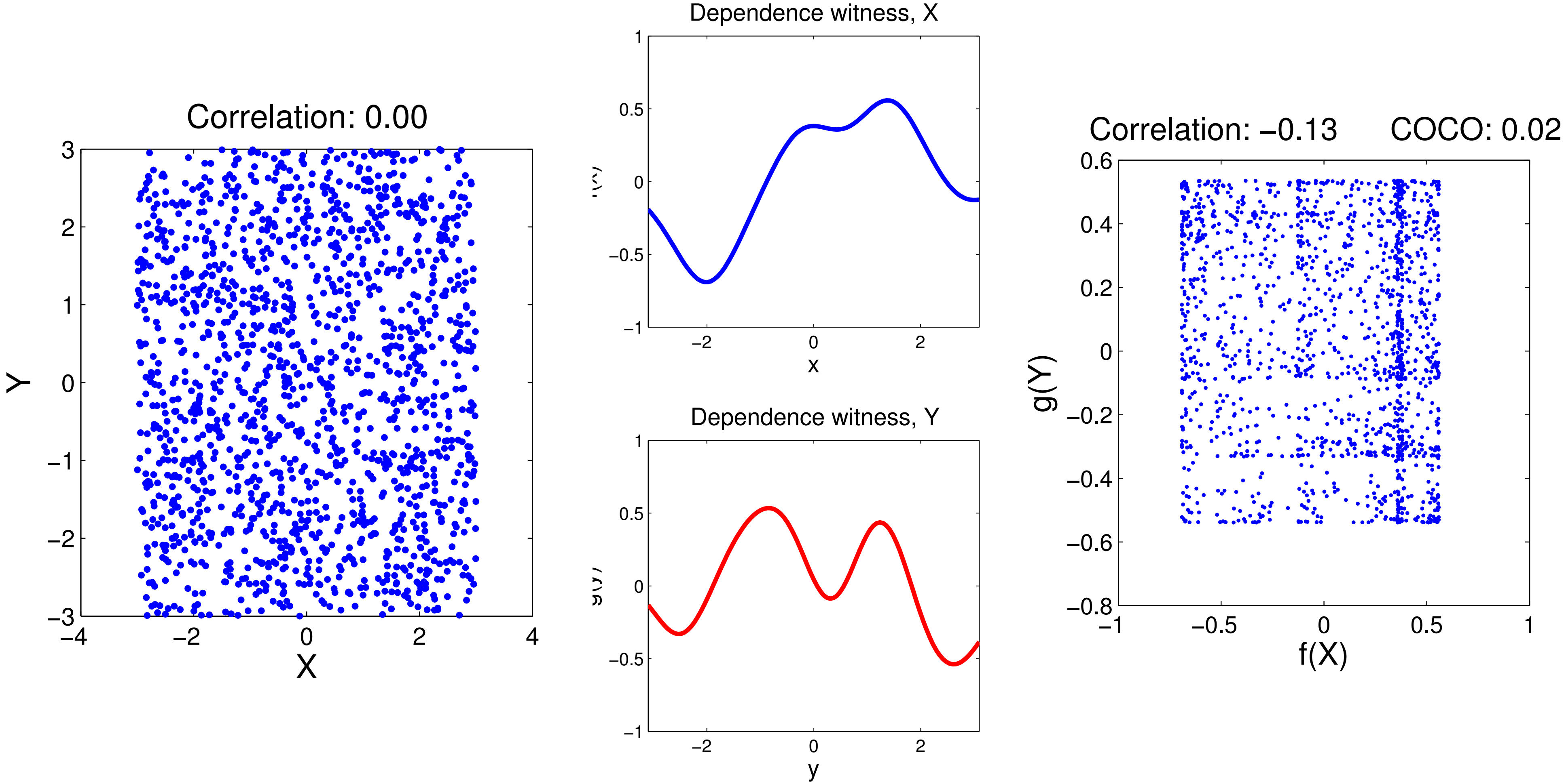




Hard-to-detect dependence

Why does COCO decay when dependence encoded at higher frequencies?

Case of $\omega =??$

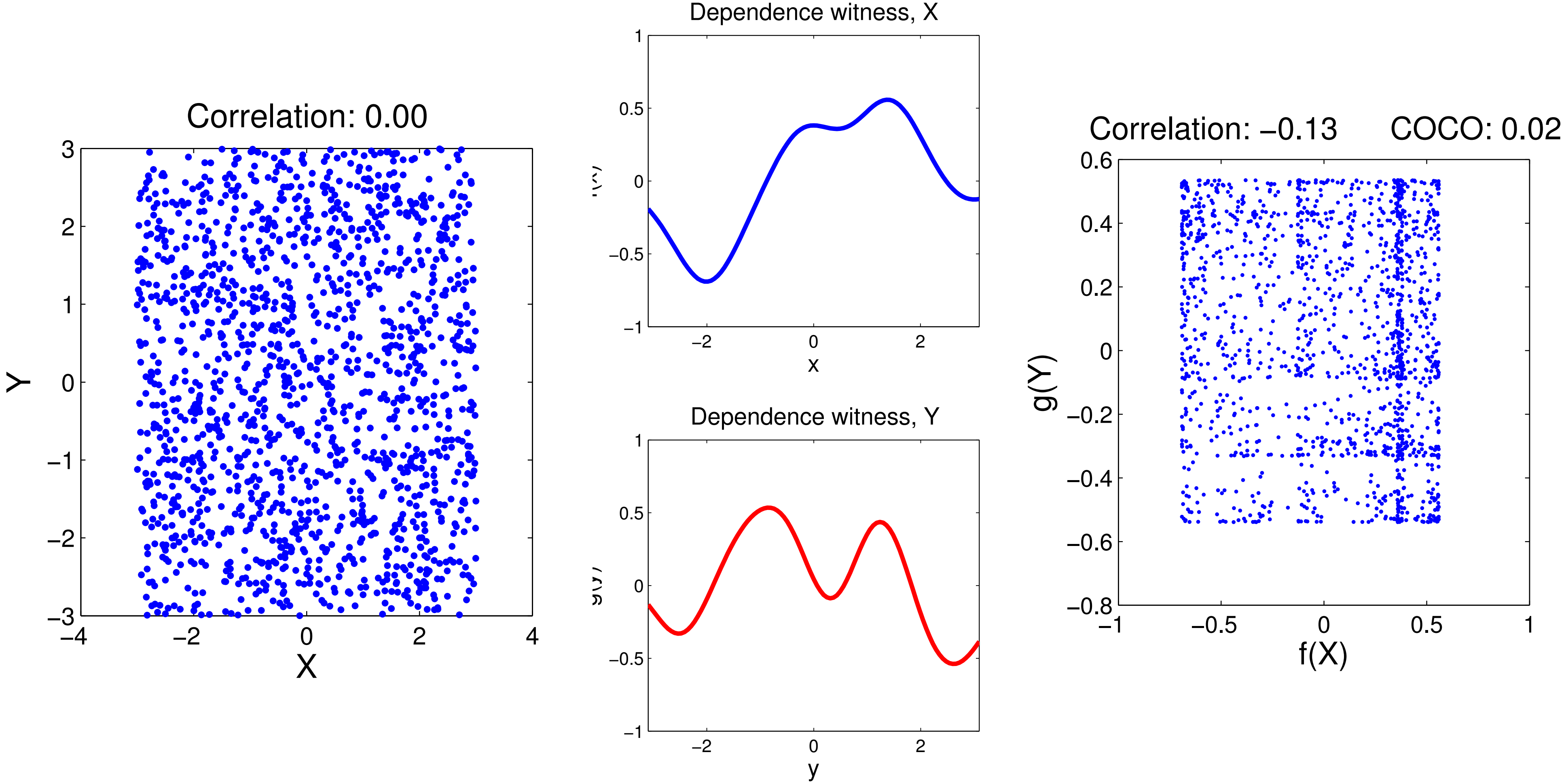


Hard-to-detect dependence

Why does COCO decay when dependence encoded at higher frequencies?

Case of uniform noise!

This bias will decrease with increasing sample size.



Hard-to-detect dependence

Why does COCO decay when dependence encoded at higher frequencies?

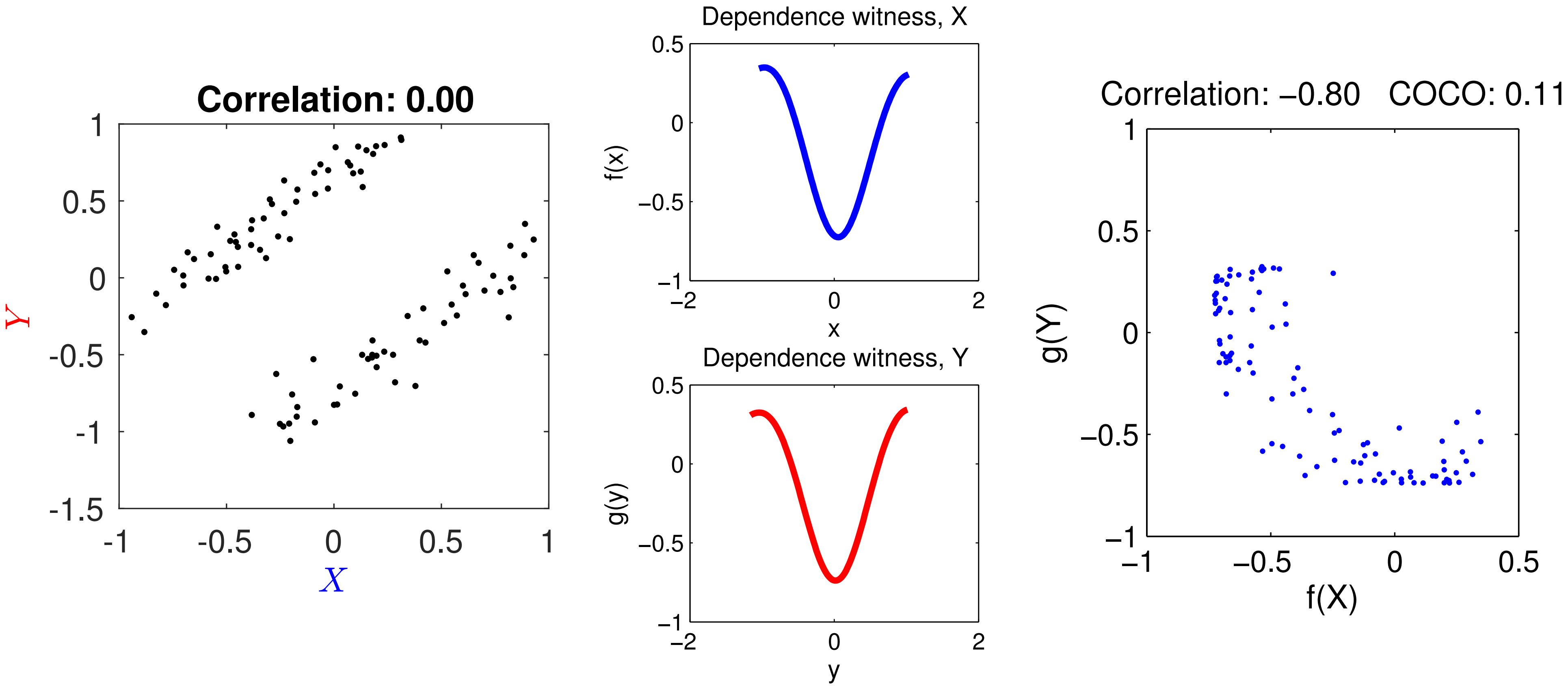
- As dependence is encoded at higher frequencies, the smooth mappings f, g achieve lower linear dependence.
- Even for independent variables, COCO will **not** be zero at finite sample sizes, since some mild linear dependence will be induced by f, g (bias)
- This bias will decrease with increasing sample size.

More functions revealing dependence

- Can we do better than COCO?

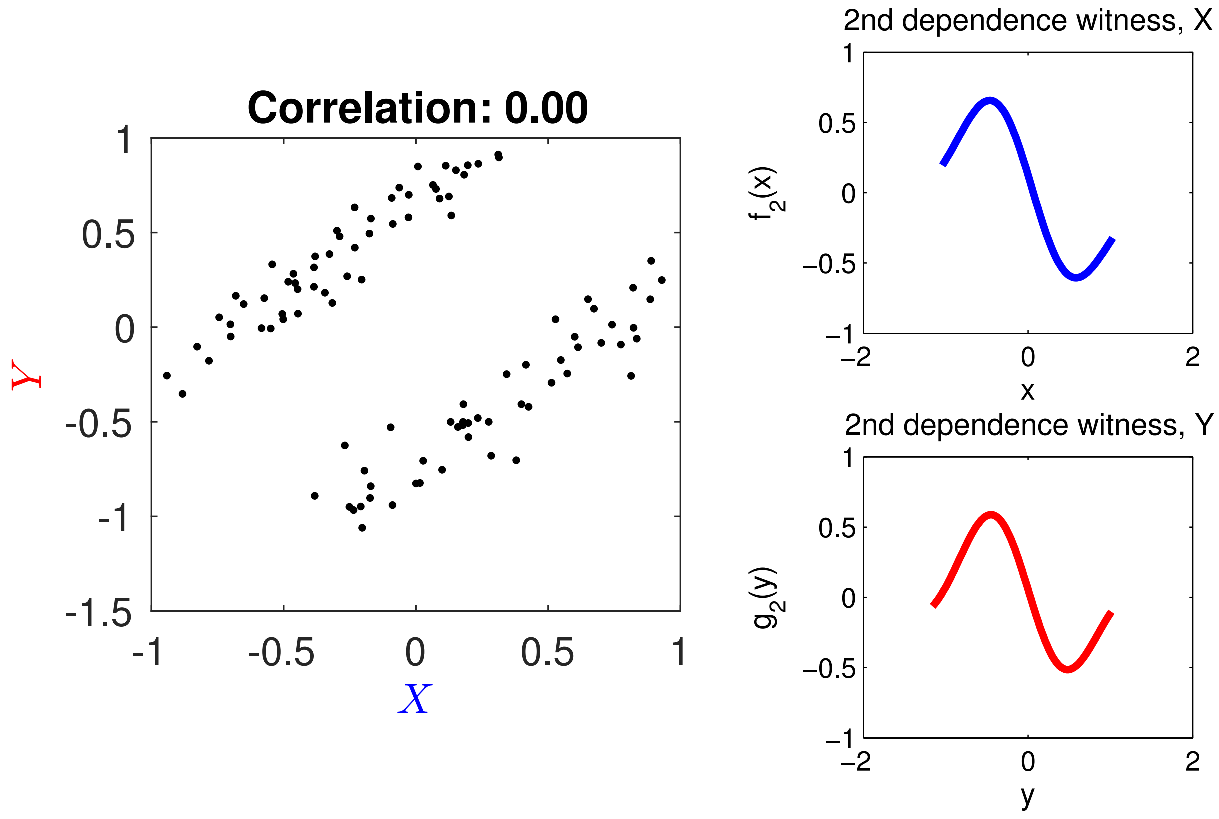
More functions revealing dependence

- Can we do better than COCO?
- A second example with zero correlation



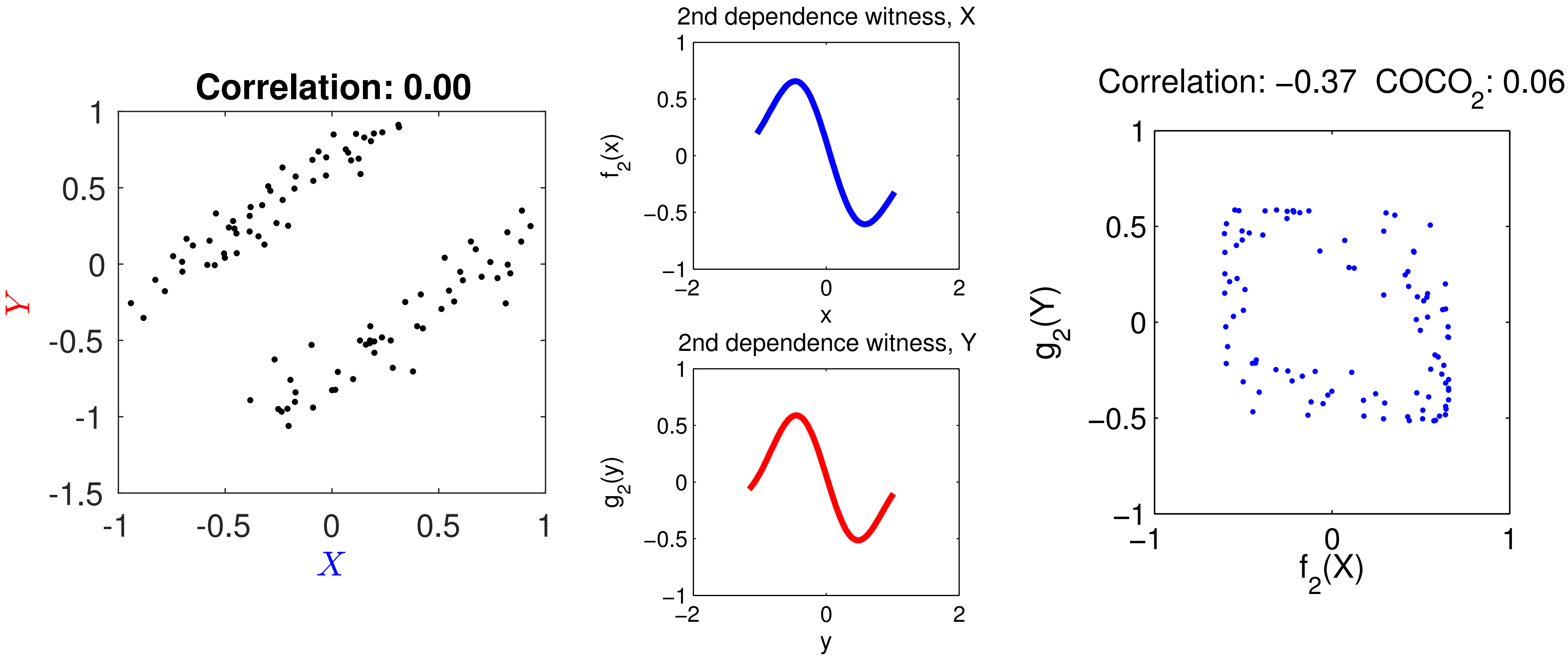
More functions revealing dependence

- Can we do better than COCO?
- A second example with zero correlation



More functions revealing dependence

- Can we do better than COCO?
- A second example with zero correlation



Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \mathrm{COCO}_i(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10] :

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

Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \mathrm{COCO}_i(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10] :

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

- In limit of infinite samples:

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{P}; F, G) &:= \|\widetilde{C}_{XY} - \mu_X \otimes \mu_Y\|_{\mathrm{HS}}^2 \\
&= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\mathrm{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\mathrm{HS}} \\
&\quad - 2 \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} [k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}')] + \mathbf{E}_{\mathsf{x},\mathsf{x}'} [k(\mathsf{x}, \mathsf{x}')] \mathbf{E}_{\mathsf{y},\mathsf{y}'} [l(\mathsf{y}, \mathsf{y}')] \\
&\quad - 2\mathbf{E}_{\mathsf{x},\mathsf{y}} \left[\mathbf{E}_{\mathsf{x}'} [k(\mathsf{x}, \mathsf{x}')] \mathbf{E}_{\mathsf{y}'} [l(\mathsf{y}, \mathsf{y}')] \right]
\end{aligned}$$

- $\widetilde{C}_{XY}$ uncentered covariance, x' indep. copy of x, y' indep. copy of y

Hilbert-Schmidt Independence Criterion

- Given $\gamma_i := \mathrm{COCO}_i(\boldsymbol{z}; \mathcal{F}, \mathcal{G})$, define Hilbert-Schmidt Independence Criterion (HSIC) [ALT05, NIPS07a, JMLR10] :

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

- In limit of infinite samples:

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{P}; F, G) &:= \|\widetilde{C}_{XY} - \mu_X \otimes \mu_Y\|_{\mathrm{HS}}^2 \\
&= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\mathrm{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\mathrm{HS}} \\
&\quad - 2 \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} [k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}')] + \mathbf{E}_{\mathsf{x},\mathsf{x}'} [k(\mathsf{x}, \mathsf{x}')] \mathbf{E}_{\mathsf{y},\mathsf{y}'} [l(\mathsf{y}, \mathsf{y}')] \\
&\quad - 2 \mathbf{E}_{\mathsf{x},\mathsf{y}} \left[\mathbf{E}_{\mathsf{x}'} [k(\mathsf{x}, \mathsf{x}')] \mathbf{E}_{\mathsf{y}'} [l(\mathsf{y}, \mathsf{y}')] \right]
\end{aligned}$$

- NOTE: HSIC is identical to $MMD^2(\mathbf{P}_{XY}, \mathbf{P}_X \mathbf{P}_Y)$ (exercise!)

Hilbert-Schmidt Independence Criterion

Proof: Recall:

$$\langle L, a \otimes b \rangle_{\mathrm{HS}} = \langle a, Lb \rangle_{\mathcal{F}} \qquad \left\langle \widetilde{C}_{XY}, A \right\rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}$$

and

$$[a \otimes b]c = \langle b, c \rangle a$$

Assume uncentred covariance. Applying covariance operator definition twice,

$$\begin{aligned}
\|\widetilde{C}_{XY}\|^2_{\mathrm{HS}} &= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \widetilde{C}_{XY} \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \left\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \phi(\mathsf{x}') \otimes \psi(\mathsf{y}') \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \left\langle \phi(\mathsf{x}), [\phi(\mathsf{x}') \otimes \psi(\mathsf{y}')] \psi(\mathsf{y}) \right\rangle_{\mathcal{F}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \left[\left\langle \phi(\mathsf{x}), \phi(\mathsf{x}') \right\rangle_{\mathcal{F}} \left\langle \psi(\mathsf{y}'), \psi(\mathsf{y}) \right\rangle_{\mathcal{G}} \right] \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \left[k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}'). \right]
\end{aligned}$$

Estimates of HSIC

Unbiased estimate: define $\widehat{A}$ as the empirical estimator of $\|\widetilde{C}_{XY}\|^2_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}}\mathbf{E}_{\mathsf{x}',\mathsf{y}'}\left[k(\mathsf{x},\mathsf{x}')l(\mathsf{y},\mathsf{y}').\right]$,

$$\widehat{A} := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, x_j) l(y_i, y_j)$$

Estimates of HSIC

Unbiased estimate: define $\widehat{A}$ as the empirical estimator of $\|\widetilde{C}_{XY}\|^2_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}}\mathbf{E}_{\mathsf{x}',\mathsf{y}'}\left[k(\mathsf{x},\mathsf{x}')l(\mathsf{y},\mathsf{y}').\right]$,

$$\widehat{A} := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j)l(y_i,y_j)$$

Alternative: plug in empirical covariance operator (uncentered),

$$\check{C}_{XY} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\otimes\psi(y_i),$$

Biased estimate:

$$\widehat{A}_b = \left\|\check{C}_{XY}\right\|^2 = \left\langle \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\otimes\psi(y_i), \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\otimes\psi(y_i)\right\rangle_{\mathrm{HS}}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i,x_j)l(y_i,y_j) = \frac{1}{n^2}\mathrm{tr}(KL),$$

How large is the bias?

Difference is:

$$\begin{aligned}\widehat{A}_b - \widehat{A} &= \frac{1}{n^2}\sum_{i,j=1}^{n} k_{ij}l_{ij} - \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k_{ij}l_{ij} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} k_{ii}l_{ii} + \left(\frac{1}{n^2} - \frac{1}{n(n-1)}\right)\left(\sum_{i\neq j}^{n} k_{ij}l_{ij}\right) \\ &= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} k_{ii}l_{ii} - \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k_{ij}l_{ij}\right),\end{aligned}$$

where $k_{ij} = k(x_i, x_j)$.

thus the *expectation* of this difference (i.e., the bias) is of $O(n^{-1})$.

Remaining terms covered in lecture notes.

Distribution of HSIC at independence

- (Biased) empirical HSIC a v-statistic

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

  - Statistical testing: How do we find when this is larger enough that the null hypothesis $\mathbf{P} = \mathbf{P}_x\mathbf{P}_y$ is unlikely?
  - Formally: given $\mathbf{P} = \mathbf{P}_x\mathbf{P}_y$, what is the threshold T such that $\mathbf{P}(\mathrm{HSIC} > T) < \alpha$ for small α?

Distribution of HSIC at independence

- (Biased) empirical HSIC a v-statistic

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

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

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

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

Distribution of HSIC at independence

- (Biased) empirical HSIC a v-statistic

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

- Associated U-statistic degenerate when $\mathbf{P} = \mathbf{P}_{\mathsf{x}}\mathbf{P}_{\mathsf{y}}$ [Serfling, 1980]:

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

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

- First two moments [NIPS07b]

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

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

Statistical testing with HSIC

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

Statistical testing with HSIC

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

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

where

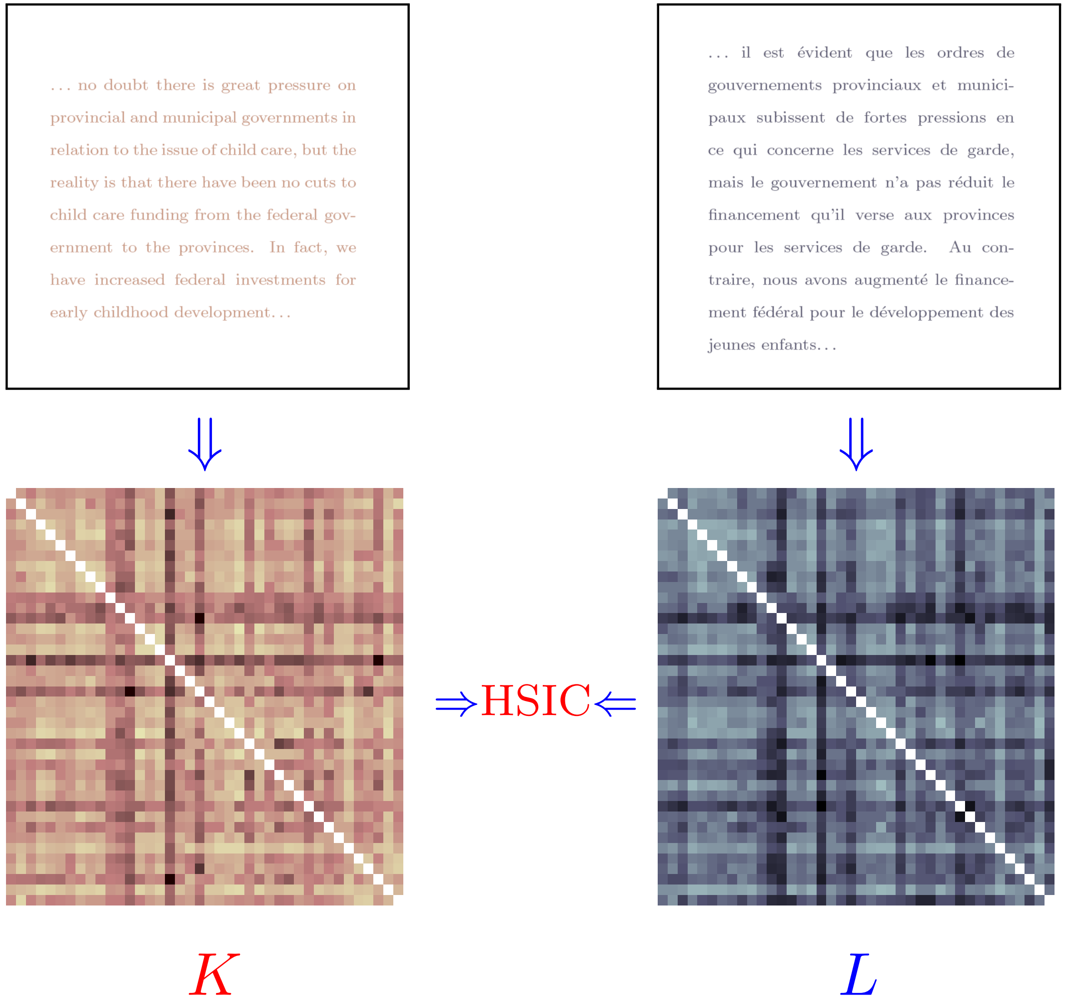
$$\alpha = \frac{(\mathbf{E}(\text{HSIC}_b))^2}{\text{var}(\text{HSIC}_b)}, \quad \beta = \frac{\text{var}(\text{HSIC}_b)}{n\mathbf{E}(\text{HSIC}_b)}.$$

Experiment: dependence testing for translation

- (Biased) empirical HSIC:

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

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



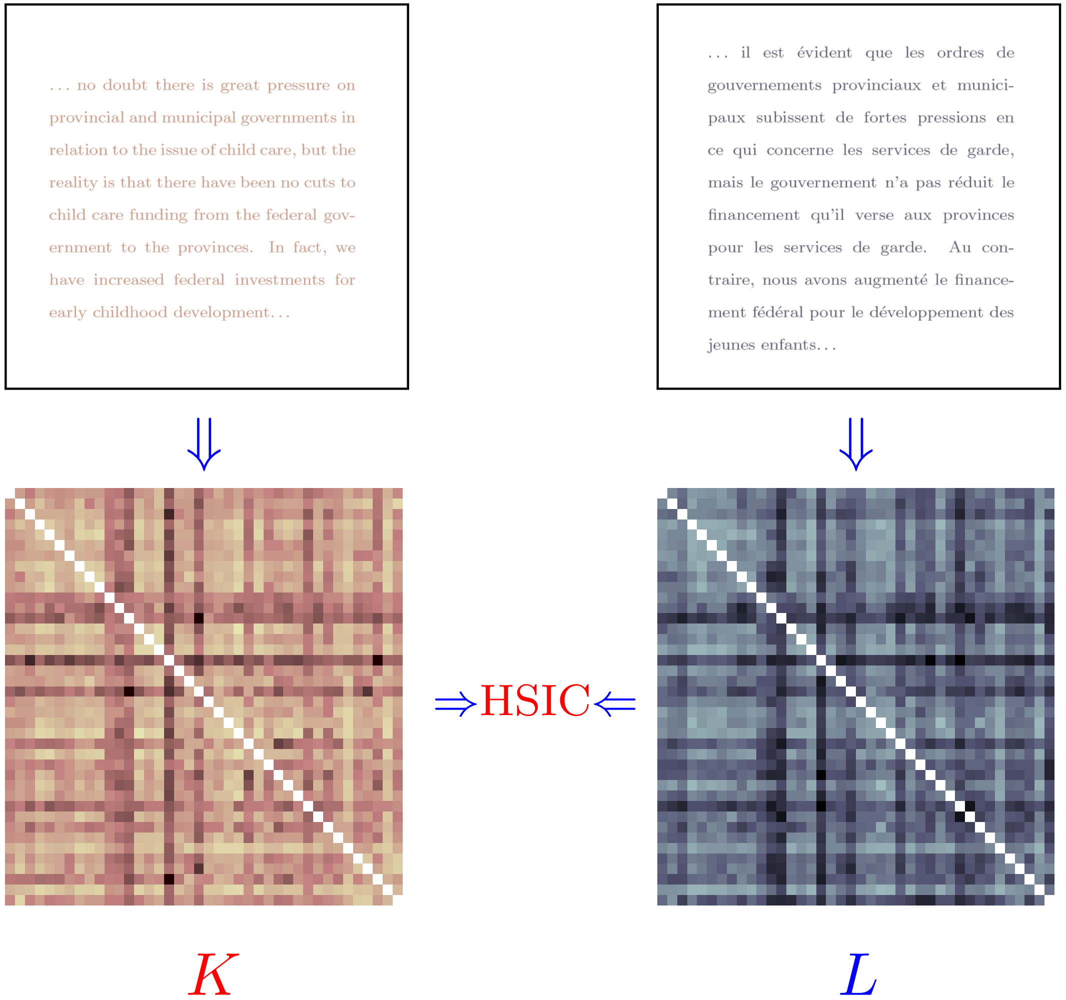
- k-spectrum kernel: average Type II error 0 ($\alpha = 0.05$)

Experiment: dependence testing for translation

- (Biased) empirical HSIC:

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

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



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

Application of HSIC: Feature Selection

HSIC for Microarray feature selection

- Select genes from microarray data for classification
- Different methods choose features optimising different criteria

HSIC for Microarray feature selection

- Select genes from microarray data for classification
- Different methods choose features optimising different criteria
- Several criteria special cases of HSIC: [ICML07a,ISMB07]
  - Pearson's correlation (normalise by standard deviation) [van't Veer et al., 2002, Ein-Dor et al., 2006]
  - Mean difference and variants [Bedo et al., 2006, Hastie et al., 2001]
  - Shrunken centroid [Tibshirani et al., 2002, 2003]
  - (Kernel) ridge regression [Li and Yang, 2005]

HSIC for Microarray feature selection

- Select genes from microarray data for classification
- Different methods choose features optimising different criteria
- Several criteria special cases of HSIC: [ICML07a,ISMB07]
  - Pearson's correlation (normalise by standard deviation) [van't Veer et al., 2002, Ein-Dor et al., 2006]
  - Mean difference and variants [Bedo et al., 2006, Hastie et al., 2001]
  - Shrunken centroid [Tibshirani et al., 2002, 2003]
  - (Kernel) ridge regression [Li and Yang, 2005]
- When are nonlinear feature maps justified?

Feature selection: BAHSIC (1)

- Backwards elimination of irrelevant features to maximise dependence (HSIC). Why backwards?

Feature selection: BAHSIC (1)

- Backwards elimination of irrelevant features to maximise dependence (HSIC). Why backwards?

Input: The full set of features $\mathcal{S}$

Output: An ordered set of features $\mathcal{S}^\dagger$

1: $\mathcal{S}^\dagger \leftarrow \varnothing$
2: **repeat**
3: Adapt kernel parameter σ_0
4: Remove **individual** features to maximize HSIC,
$\mathcal{I} \leftarrow \arg\max_{\mathcal{I}} \sum_{j \in \mathcal{I}} \mathrm{HSIC}(\sigma_0, \mathcal{S} \setminus \{j\}), \quad \mathcal{I} \subset \mathcal{S}$
5: $\mathcal{S} \leftarrow \mathcal{S} \setminus \mathcal{I}$
6: $\mathcal{S}^\dagger \leftarrow (\mathcal{S}^\dagger, \mathcal{I})$
7: **until** $\mathcal{S} = \varnothing$

- Application: feature selection in microarrays [ICML07a,ISMB07, JMLR12]

Relation of HSIC to mean difference

- (Biased) empirical HSIC: $\text{HSIC}(X, Y) := \mathsf{Tr}(KHLH)$

Relation of HSIC to mean difference

- (Biased) empirical HSIC: $\text{HSIC}(X,Y) := \mathsf{Tr}(KHLH)$
- HSIC equivalent to difference in means
  - Linear input kernel $K_\ell = x[\ell]\,(x[\ell])^\top$, $K = \sum_\ell K_\ell$ (single feature, HSIC is sum of all feature scores)
  - Linear output kernel, $1/n_+$ for one class, $-1/n_-$ for the other
  - Warning: for nonlinear kernel, features can interact.

$$\mathsf{Tr}(K_\ell HLH) = \left(\frac{1}{n_+} \sum_{i=1}^{n_+} x_i[\ell] - \frac{1}{n_-} \sum_{i=n_++1}^{n} x_i[\ell] \right)^2$$

Relation of HSIC to mean difference

- (Biased) empirical HSIC: $\text{HSIC}(X,Y) := \mathsf{Tr}(KHLH)$
- HSIC equivalent to difference in means
  - Linear input kernel $K_\ell = x[\ell]\,(x[\ell])^\top$, $K = \sum_\ell K_\ell$ (single feature, HSIC is sum of all feature scores)
  - Linear output kernel, $1/n_+$ for one class, $-1/n_-$ for the other
  - Warning: for nonlinear kernel, features can interact.

$$\mathsf{Tr}(K_\ell HLH) = \left(\frac{1}{n_+}\sum_{i=1}^{n_+} x_i[\ell] - \frac{1}{n_-}\sum_{i=n_++1}^{n} x_i[\ell]\right)^2$$

- HSIC equivalent to shrunken centroid
  - Linear kernels, $Y = \begin{pmatrix} \frac{\mathbf{1}_{n_+}}{n_+} - \frac{\mathbf{1}_{n_+}}{n} & -\frac{\mathbf{1}_{n_+}}{n} \\ -\frac{\mathbf{1}_{n_-}}{n} & \frac{\mathbf{1}_{n_-}}{n_-} - \frac{\mathbf{1}_{n_-}}{n} \end{pmatrix}_{n\times 2}$.

$$\mathsf{Tr}(K_\ell HLH) = (\bar{x}_+[\ell] - \bar{x}[\ell])^2 + (\bar{x}_-[\ell] - \bar{x}[\ell])^2$$

Relation of HSIC to ridge regression

- Objective: given $y = [y_1 \dots y_n]^\top$, minimise

$$R = \|y - Vw\|^2 + \lambda \|w\|^2$$

where

$$V = \begin{pmatrix} k(x_1, \cdot) \\ \vdots \\ k(x_n, \cdot) \end{pmatrix} \quad \text{and} \quad w := \sum_i \alpha_i k(x_i, \cdot)$$

- Solution is:

$$R^* = y^\top y - y^\top (K + \lambda I)^{-1} K y$$

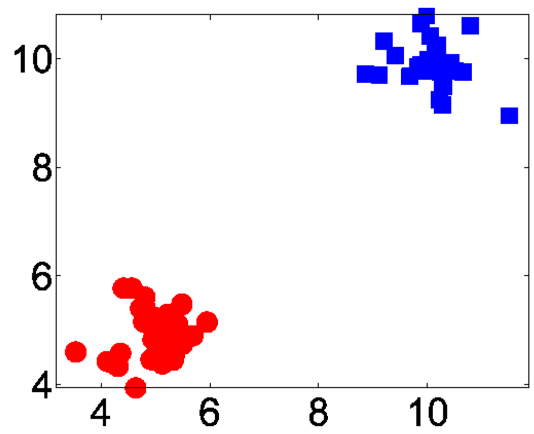
- Features that minimise R^* $\Leftrightarrow$ maximise HSIC with kernel

$$\mathfrak{K} = (K + \lambda I)^{-1} K$$

(but take care with centering: either $\sum_i y_i = 0$ or $K = HKH$)

Linear vs nonlinear kernel: idea

- For microarray data (esp. 2 class), difference in means with linear kernel usually works best.



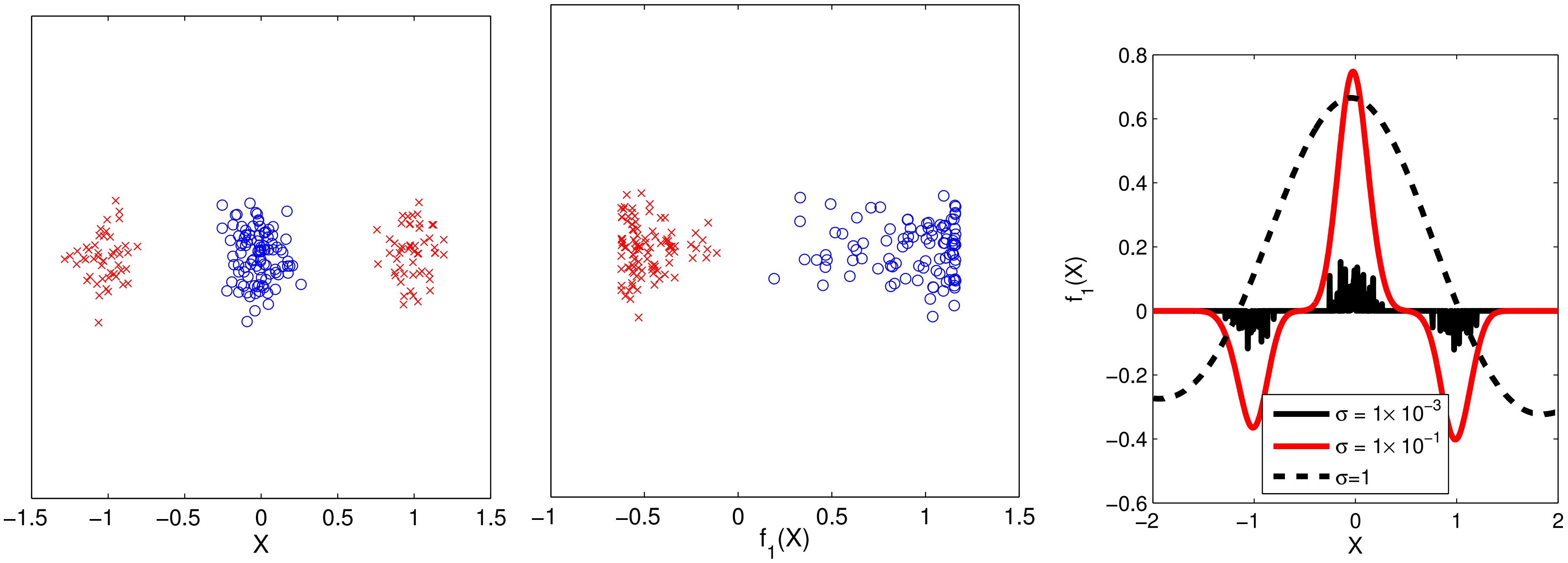
Linear vs nonlinear kernel: idea

- For microarray data (esp. 2 class), difference in means with linear kernel usually works best.
- Exceptions:
  - Nonlinear dependence between features and labels (e.g class with multiple subclasses)
  - Multiple classes, different features serve different purposes

$$L = Y^\top Y = \begin{bmatrix} n_1^{-2} & 0 & \dots & 0 \\ 0 & n_2^{-2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_d^{-2} \end{bmatrix}$$

Linear vs nonlinear kernel: application 1

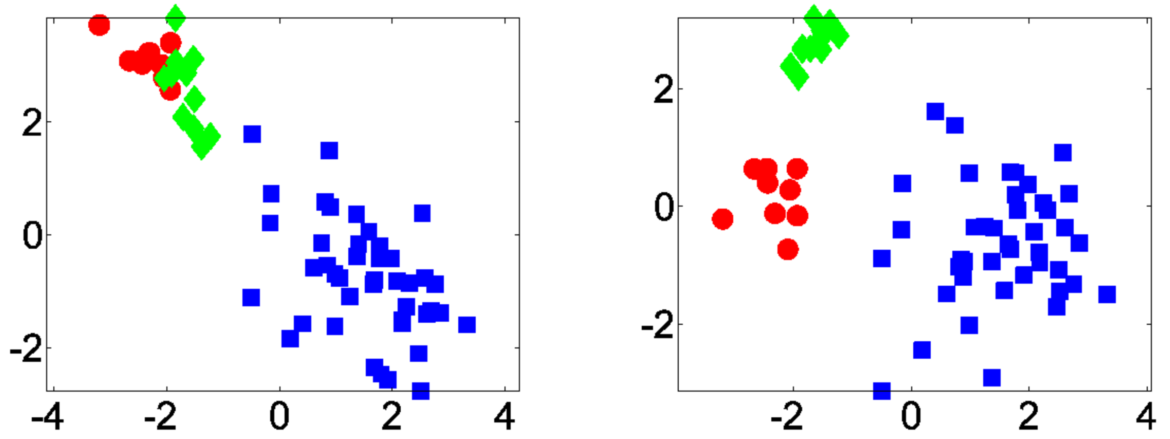
- Two classes, nonlinear relation
- Plot of maximum singular function $f_1(x)$ on $\mathcal{X}$ (as for COCO)



Linear vs nonlinear kernel: application 2

- Three cancer subtypes (diffuse large B-cell lymphoma and leukemia, follicular lymphoma, and chronic lymphocytic leukemia)

Linear Nonlinear

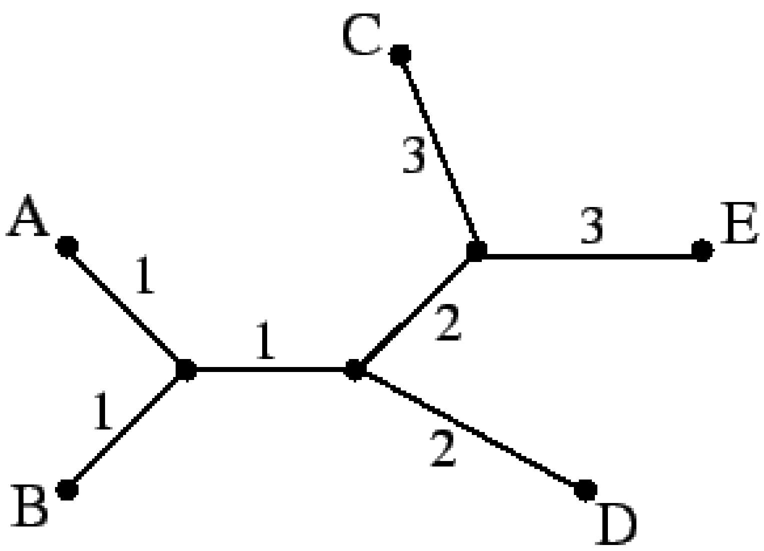


Application 2: Taxonomy Discovery

Overview: HSIC-based taxonomy discovery

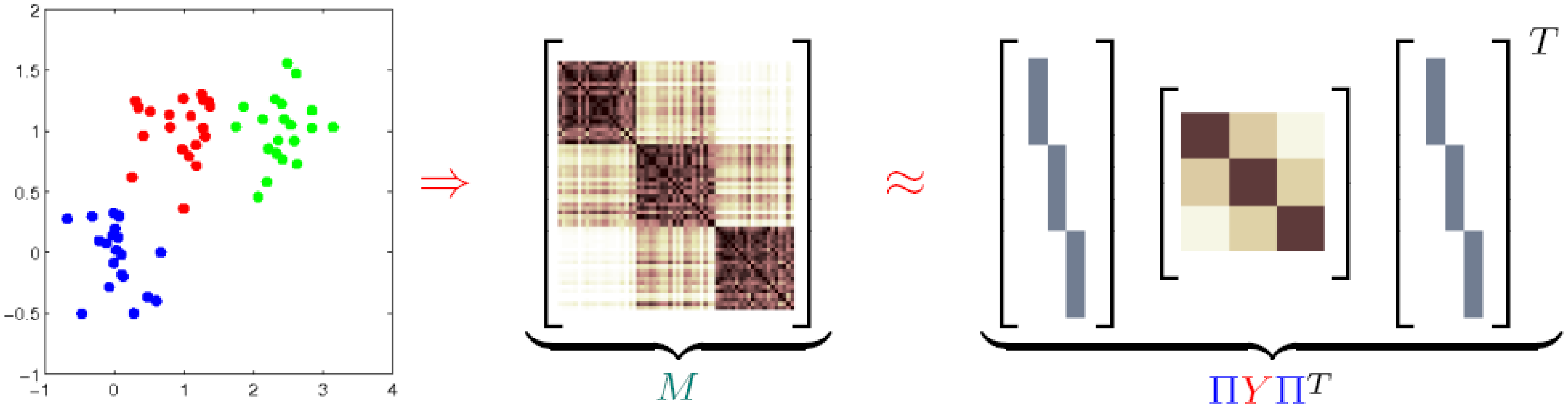
- Simultaneous clustering and taxonomy fitting
  $\rightarrow$ Numerical Taxonomy Clustering [NIPS08b]
- Maximise dependence (HSIC) between data and clusters

	A	B	C	D	E
A	0	2	7	4	7
B		0	7	4	7
C			0	7	6
D				0	7
E					0



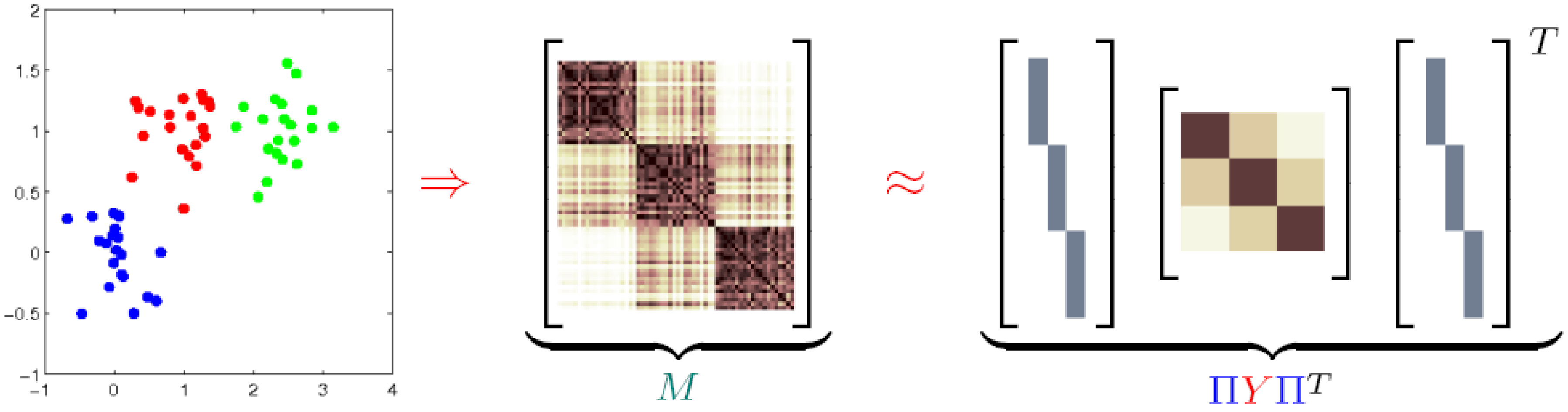
Dependence Maximization

Idea:



Dependence Maximization

Idea:



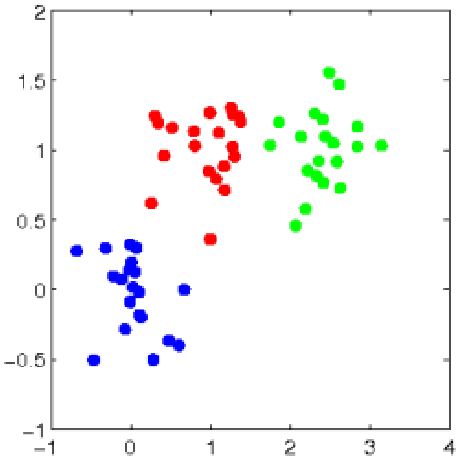
Objective:

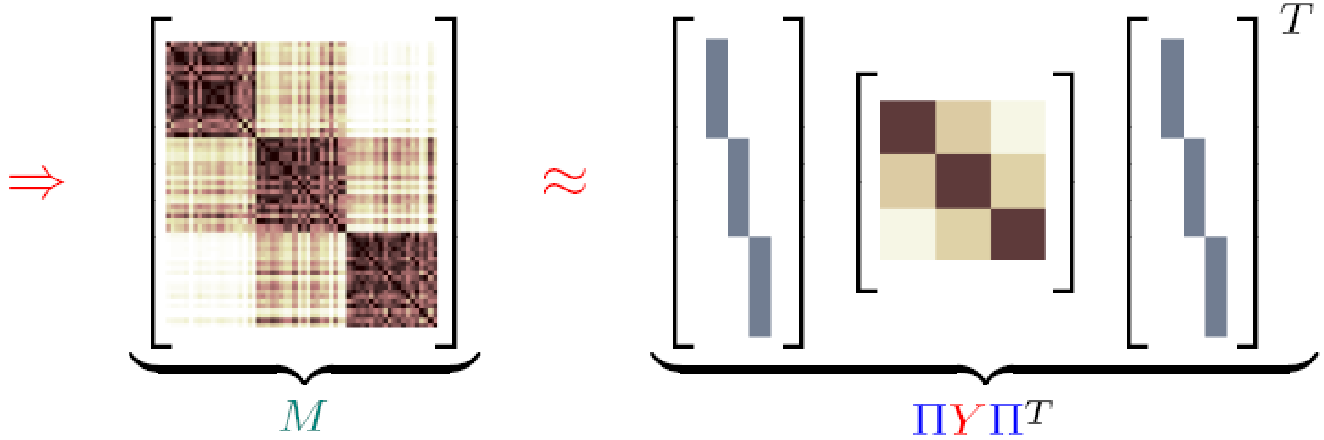
$$\max_{Y,\Pi} \frac{\operatorname{Tr}\left[M H \Pi Y \Pi^T H\right]}{\|H \Pi Y \Pi^T H\|_{\mathrm{HS}}}.$$

- Data kernel matrix: M
- Π is $n \times k$ cluster assignment matrix, $\Pi 1 = 1$, $\Pi_{i,j} \in \{0,1\}$.
- $Y \succeq \mathbf{0}$ Gram matrix between clusters

Dependence Maximization

Idea:

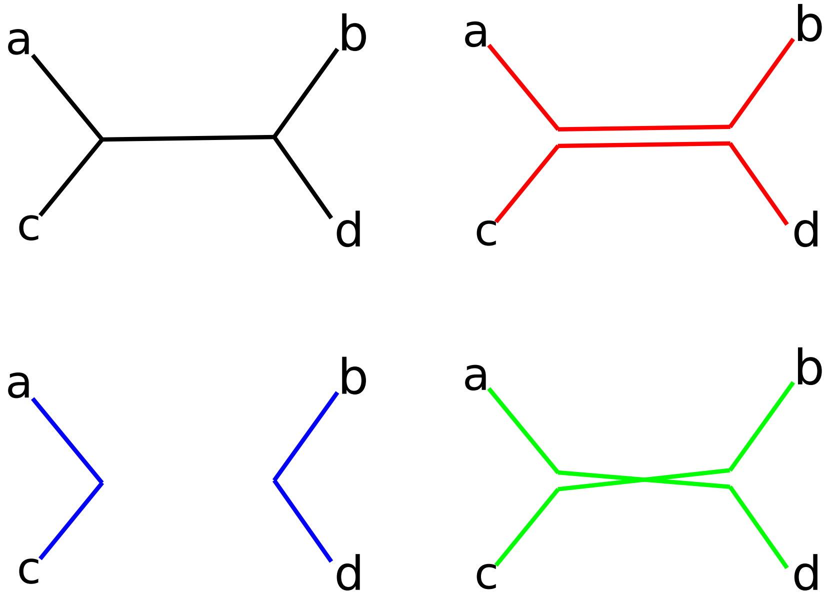




Y has no prior structure

- Add constraints to Y
  - Change Y^* $\rightarrow$ interpretability
  - Change Π^* $\rightarrow$ improved clustering

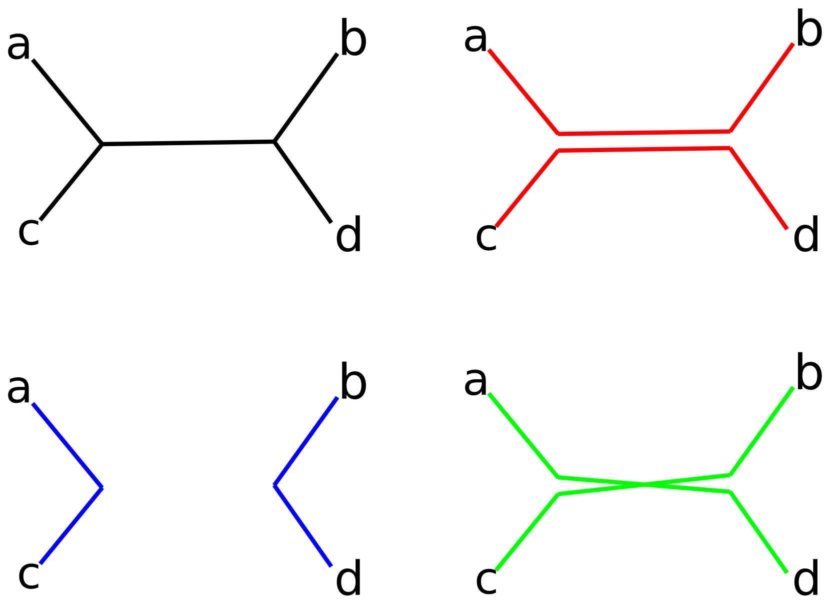
Numerical Taxonomy



- compute distance matrix, D
- $D_{ij} = \sqrt{Y_{ii} + Y_{jj} - 2Y_{ij}}$

- Four point condition:
- $D_{ab} + D_{cd} \leq \max\left(D_{ac} + D_{bd}, D_{ad} + D_{bc}\right) \quad \forall a, b, c, d$

Numerical Taxonomy



- compute distance matrix, D
- $D_{ij} = \sqrt{Y_{ii} + Y_{jj} - 2Y_{ij}}$

- Four point condition:
- $D_{ab} + D_{cd} \leq \max\left(D_{ac} + D_{bd}, D_{ad} + D_{bc}\right) \quad \forall a, b, c, d$
- Numerical taxonomy objective: $\min_{D_T} \|D - D_T\|^2$ where D_T is subject to the four point condition (NP hard, so approximation only) [Harb et al., 2005]
- From D_T to tree [Waterman et al., 1977]

Numerical Taxonomy Clustering

Require: $M \succeq \mathbf{0}$

Ensure: $(\Pi, Y) \approx (\Pi^*, Y^*)$ that max dependence s.t. 4-point condition

Initialize $Y = I$

Initialize Π using spectral clustering

while Convergence has not been reached **do**

Solve for Y given Π using closed form solution

Construct D such that $D_{ij} = \sqrt{Y_{ii} + Y_{jj} - 2Y_{ij}}$

Solve for $\min_{D_T} \|D - D_T\|^2$

Assign $Y = -\frac{1}{2} H(D_T \odot D_T) H$ (Hadamard product, next slide)

Update Π by changing labels to increase score [ICML07b]

end while

Numerical Taxonomy Clustering

Given a matrix of pairwise distances, D_T, we recover a centred kernel matrix,

$$HKH = H\left(D_T \circ D_T\right)H,$$

where $D_T \circ D_T$ denotes the Hadamard (entrywise) product.

Proof:

$$\begin{aligned} d^2(x_i, x_j) &= \left\|\phi(x_i) - \phi(x_j)\right\|^2 \\ &= k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j). \end{aligned}$$

Thus

$$k(x_i, x_j) = \frac{1}{2}\left(k(x_i, x_i) + k(x_j, x_j) - d_T^2(x_i, x_j)\right).$$

Numerical Taxonomy Clustering

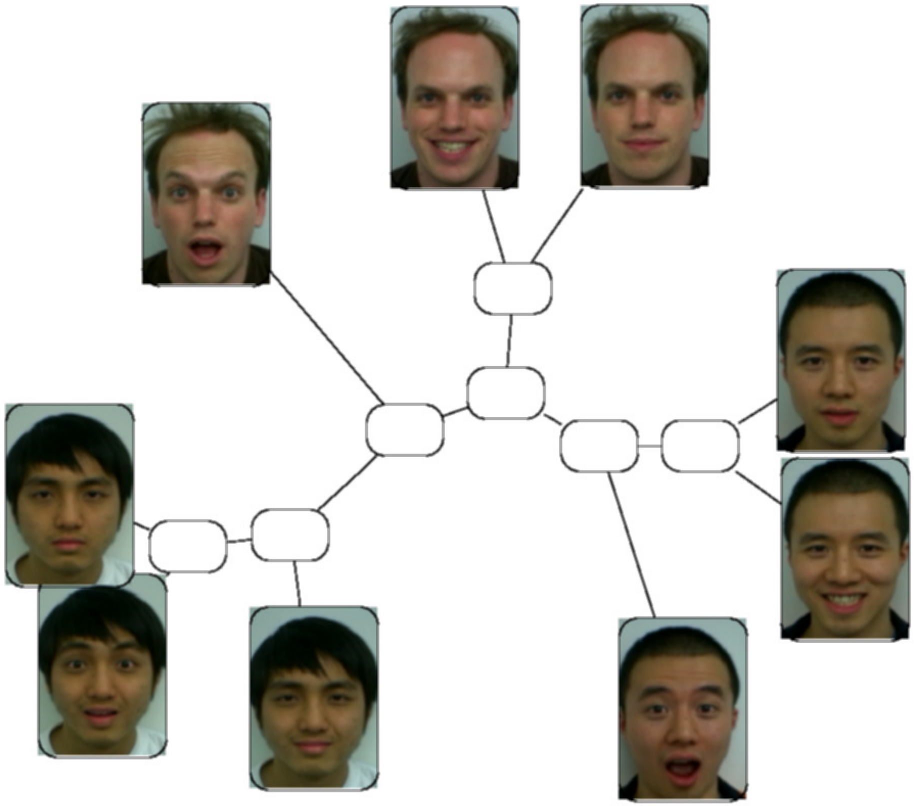
Writing this in matrix form,

$$K = \frac{1}{2}\left(\begin{bmatrix} \dots & k(x_1, x_1) & \dots \\ & \vdots & \\ \dots & k(x_m, x_m) & \dots \end{bmatrix} + \begin{bmatrix} \vdots & & \vdots \\ k(x_1, x_1) & \dots & k(x_m, x_m) \\ \vdots & & \vdots \end{bmatrix} - D_T \circ D_T \right).$$

Next, we use

$$H \begin{bmatrix} \dots & k(x_1, x_1) & \dots \\ & \vdots & \\ \dots & k(x_m, x_m) & \dots \end{bmatrix} = 0, \qquad \begin{bmatrix} \vdots & & \vdots \\ k(x_1, x_1) & \dots & k(x_m, x_m) \\ \vdots & & \vdots \end{bmatrix} H = 0,$$

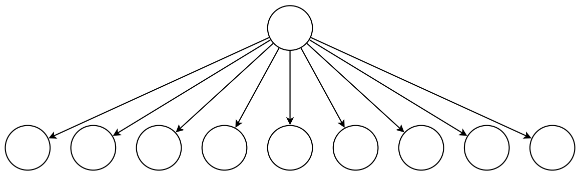
Attractive Scientist Dataset (1)



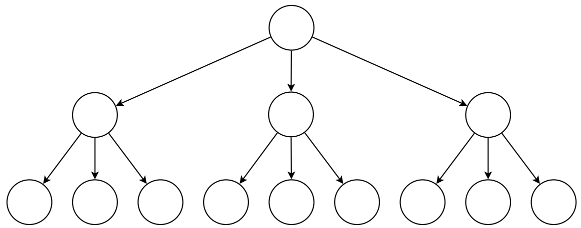
Face dataset and taxonomy discovered by the algorithm

Attractive Scientist Dataset (2)

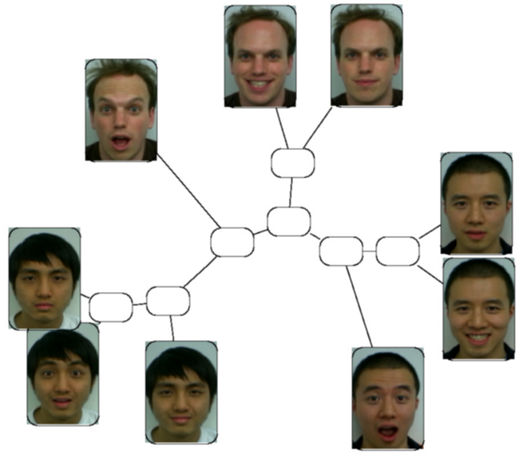
Conditional entropy scores for clusterings using [ICML07b]



flat (0.5180)

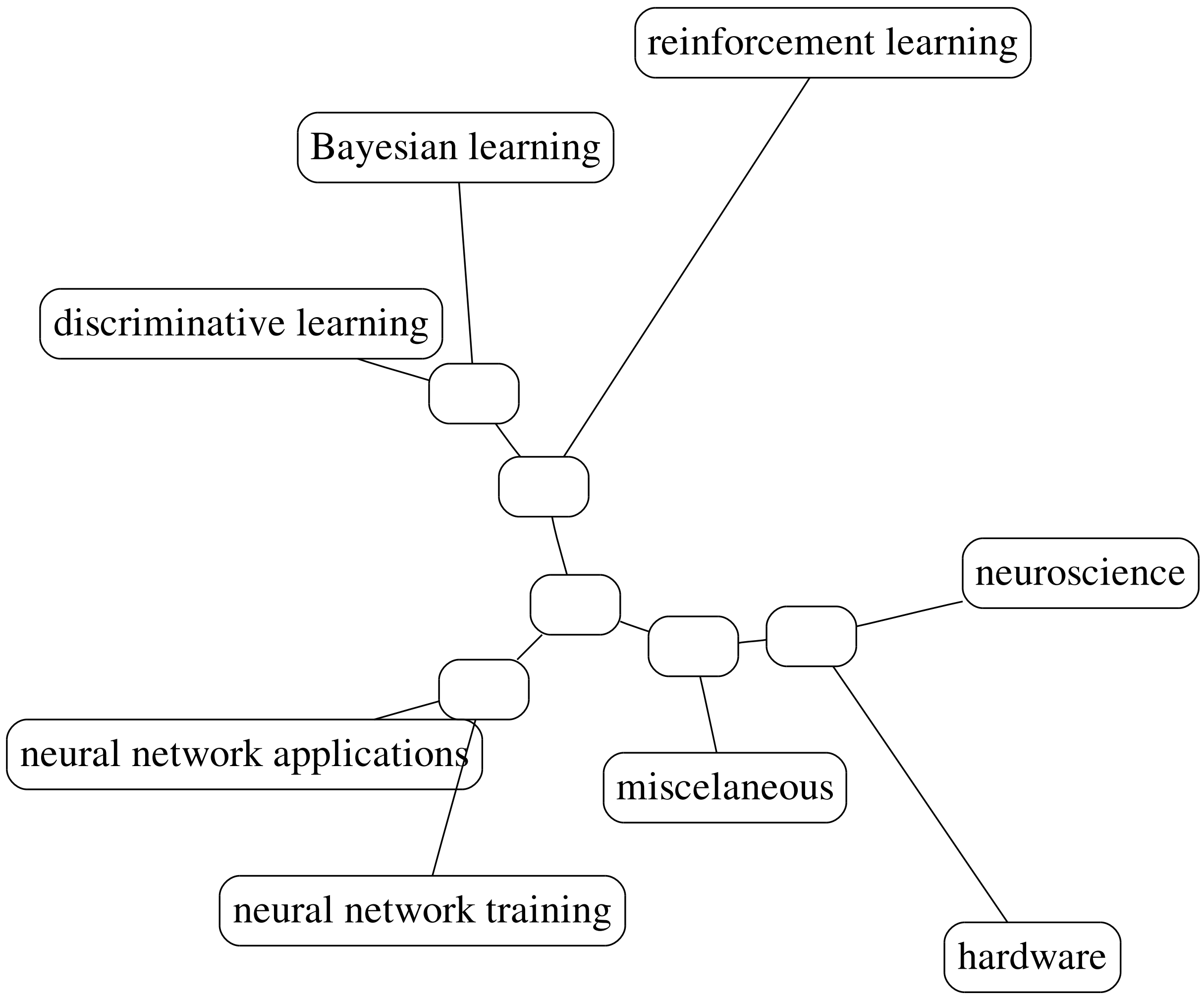


hierarchy (0.4970)



taxonomy (**0.2807**)

NIPS Articles



The taxonomy discovered for the NIPS dataset.

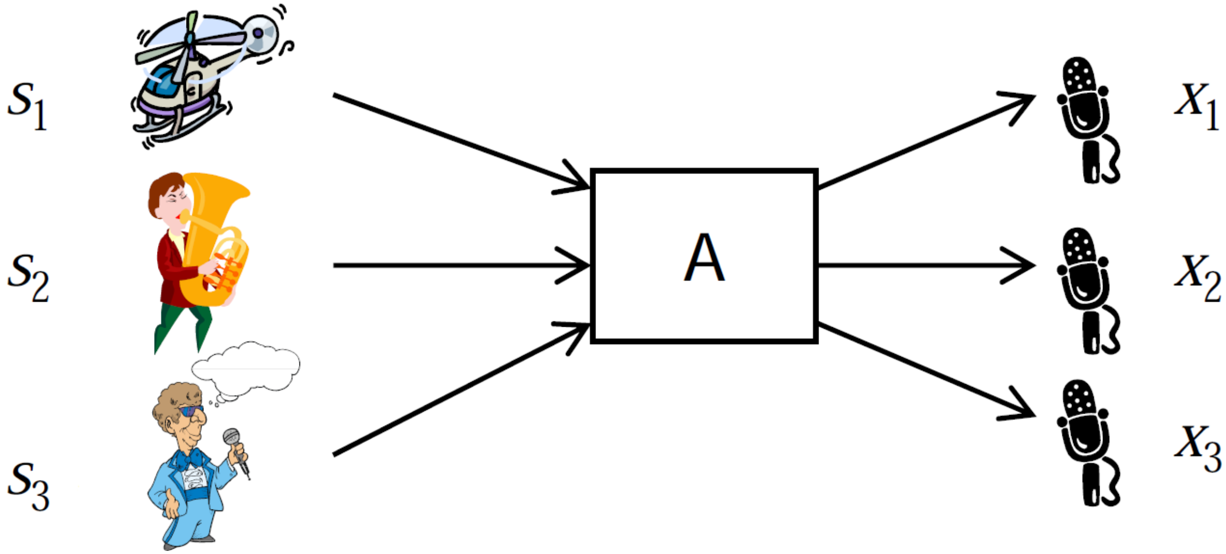
NIPS Articles: Categories

neurosci.	hardware	misc.	train-neural	app.-neural	reinforcement	discriminative	Bayesian
neurons	chip	memory	network	training	state	function	data
cells	circuit	dynamics	units	recognition	learning	error	model
model	analog	image	learning	network	policy	algorithm	models
cell	voltage	neural	hidden	speech	action	functions	distribution
visual	current	hopfield	networks	set	reinforcement	learning	gaussian
neuron	figure	control	input	word	optimal	theorem	likelihood
activity	vlsi	system	training	performance	control	class	parameters
synaptic	neuron	inverse	output	neural	function	linear	algorithm
response	output	energy	unit	networks	time	examples	mixture
firing	circuits	capacity	weights	trained	states	case	em
cortex	synapse	object	error	classification	actions	training	bayesian
stimulus	motion	field	weight	layer	agent	vector	posterior
spike	pulse	motor	neural	input	algorithm	bound	probability
cortical	neural	computational	layer	system	reward	generalization	density
frequency	input	network	recurrent	features	sutton	set	variables
orientation	digital	images	net	test	goal	approximation	prior
motion	gate	subjects	time	classifier	dynamic	bounds	log
direction	cmos	model	back	classifiers	step	loss	approach
spatial	silicon	associative	propagation	feature	programming	algorithms	matrix
excitatory	implementation	attractor	number	image	rl	dimension	estimation

Application 3: ICA

ICA: setting

Independent component analysis:

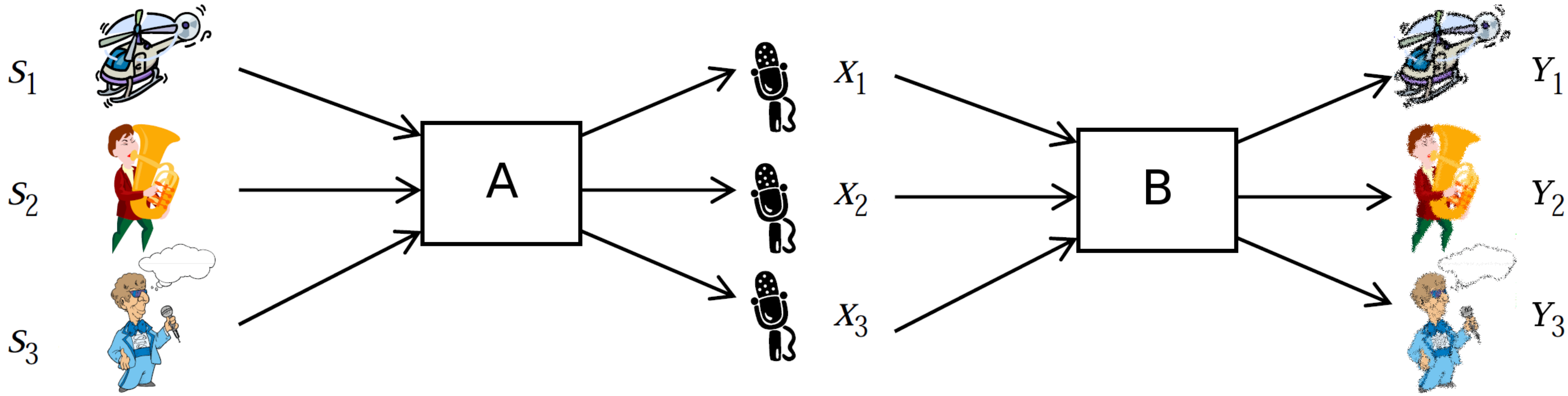


$$S \quad \times A \quad = X$$

- $\mathbf{s}$ a vector of l unknown, independent sources: $\mathbf{P_s} = \prod_{i=1}^{l} \mathbf{P}_{\mathbf{s}_i}$
- $\mathbf{x}$ vector of mixtures
- $\mathbf{A}$ is $l \times l$ mixing matrix (full rank)

ICA: setting

Independent component analysis:

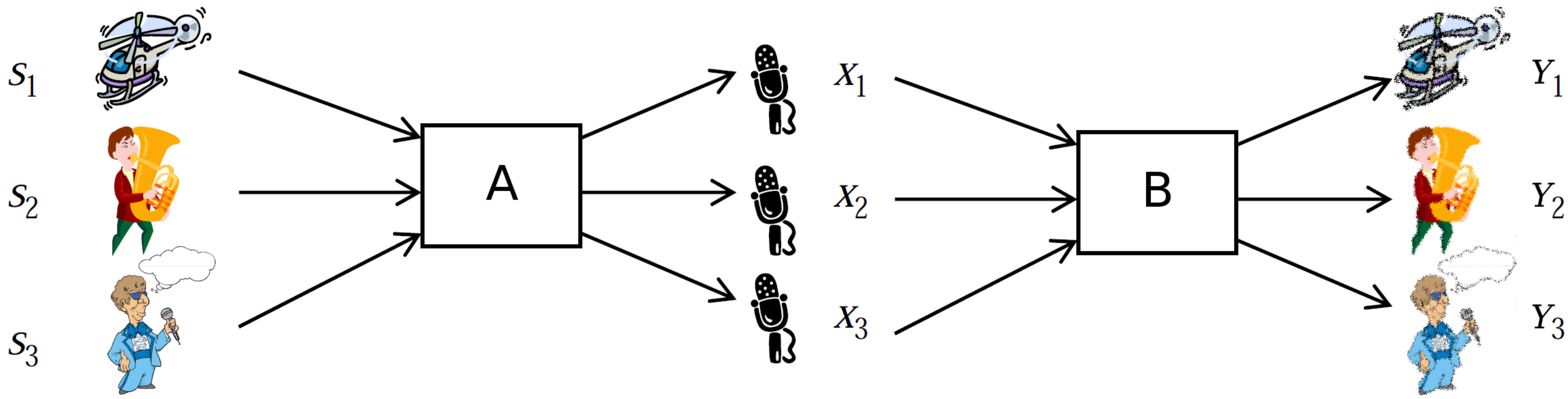


$$X \qquad \times B \qquad = Y$$

- $\mathbf{B}$ is estimated $\mathbf{A}^{-1}$, we solve for this
- $\mathbf{y}$ vector of estimated sources

ICA: setting

Independent component analysis:



$$X \qquad \times B \qquad = Y$$

- $\mathbf{B}$ is estimated $\mathbf{A}^{-1}$, we solve for this
- $\mathbf{y}$ vector of estimated sources

Neglect time dependence: m i.i.d. mixture observations

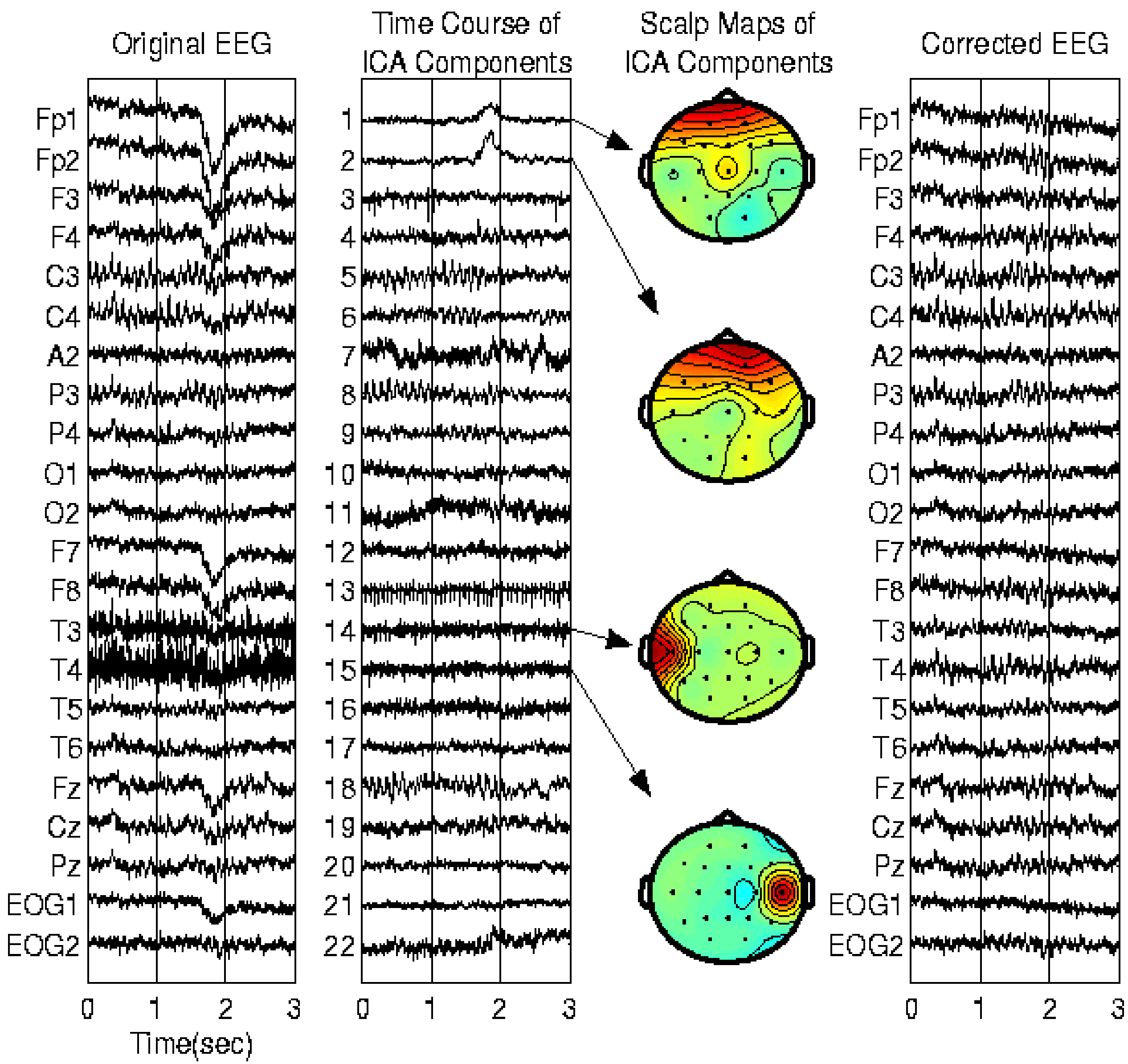
ICA: another example

- Mixtures X are original EEG

  [Jung et al., 2000]

- Estimated sources Y are ICA components

- Scalp map from B



ICA examples

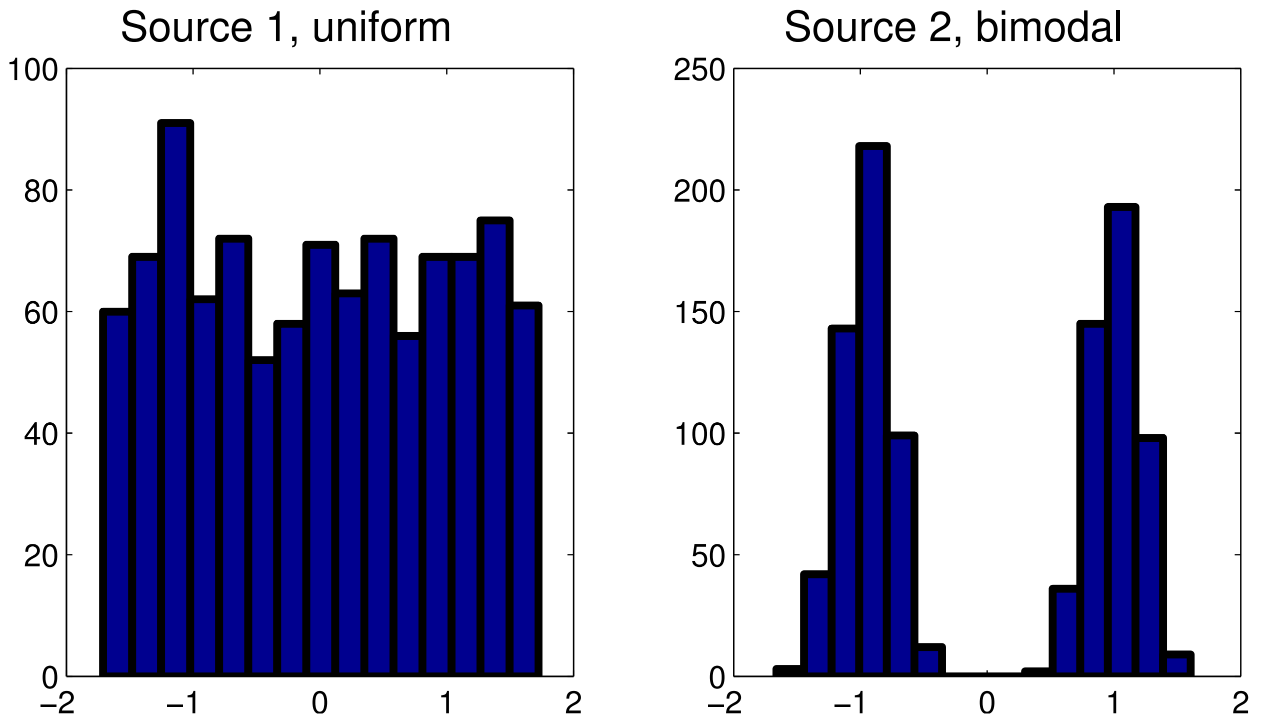
- We've seen:
  - Sounds mixed together ("cocktail party" problem) [Hyvärinen et al., 2001]
  - EEG recordings (brain, fetal heartbeat) [Jung et al., 2000, Stögbauer et al., 2004]

Warning: both the above examples violate the assumptions made in ICA (that the observations at each time are independent and identically distributed).

- Some further examples:
  - Extracting independent activity from fMRI [Calhoun et al., 2003]
  - Financial data [Kiviluoto and Oja, 1998]
  - Linear edge filters for image patch coding? (Possibly not: [Bethge, 2006])

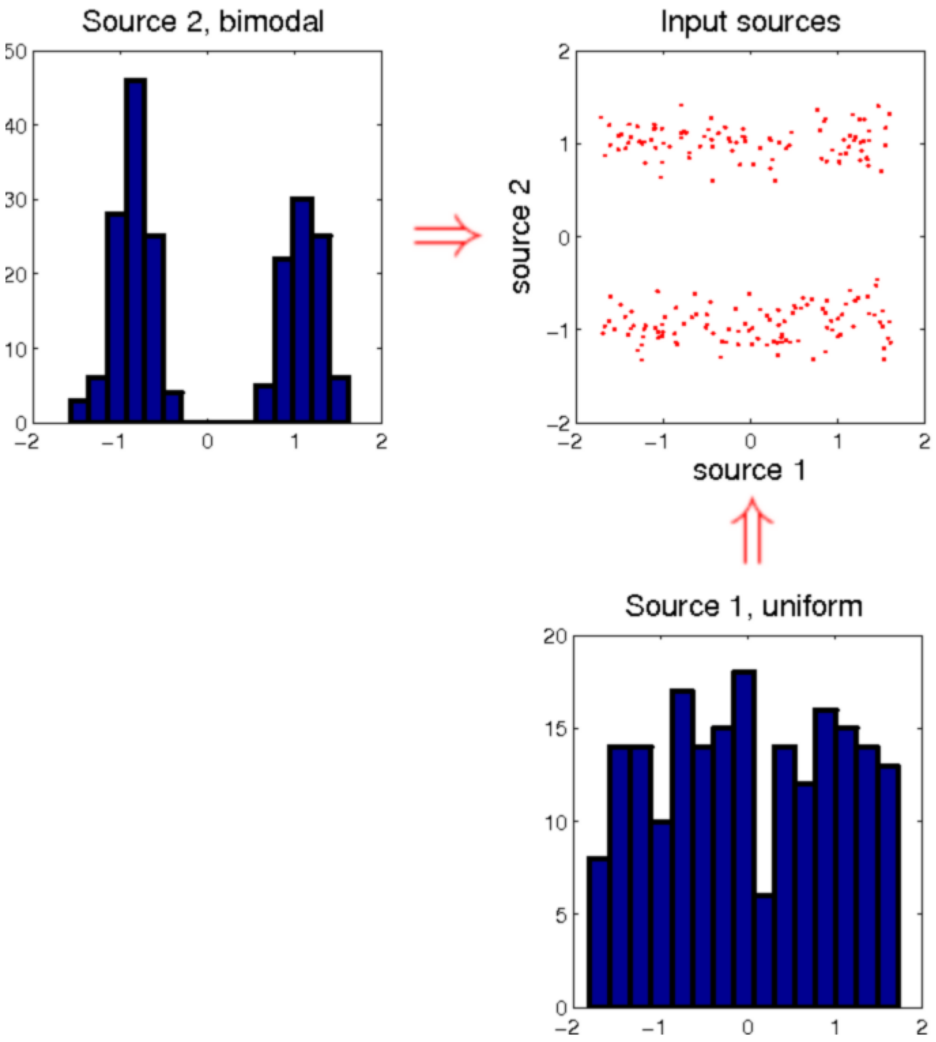
A toy example

- Two distributions: $\mathbf{P}_{\mathbf{s}_1}$ is uniform, $\mathbf{P}_{\mathbf{s}_2}$ is bimodal



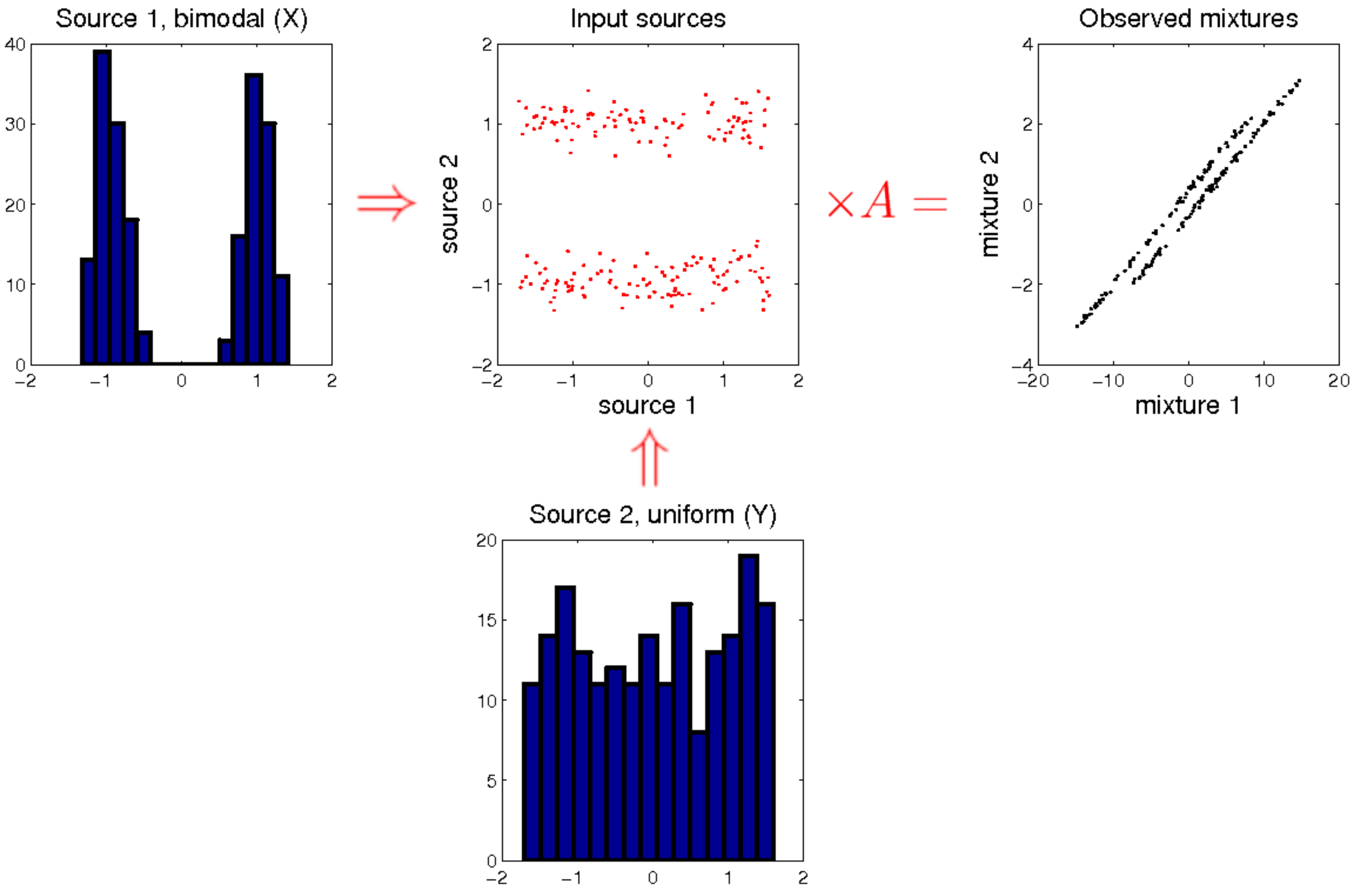
A toy example

- Two distributions: $\mathbf{P}_{\mathbf{s}_1}$ is uniform, $\mathbf{P}_{\mathbf{s}_2}$ is bimodal



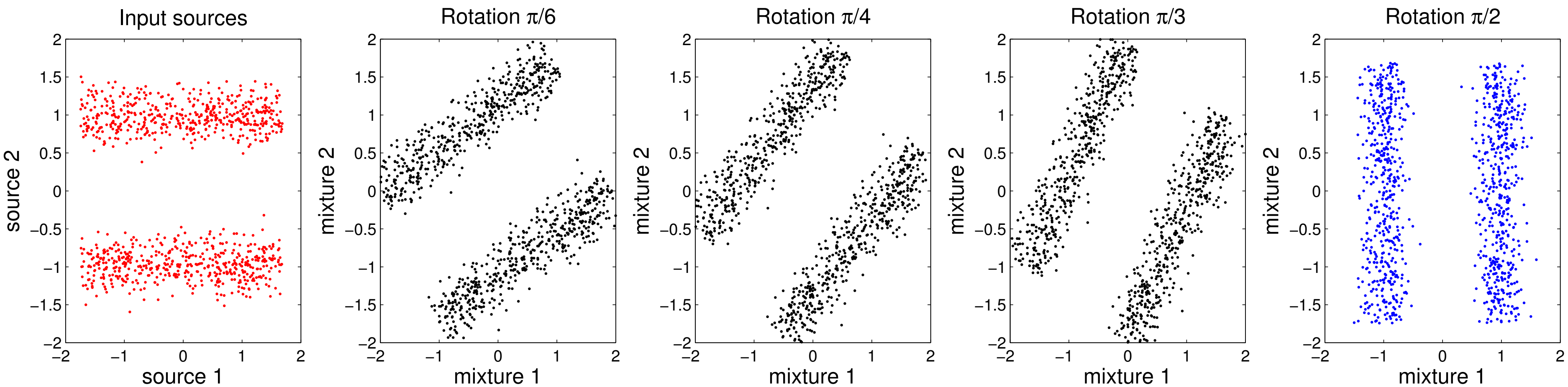
A toy example

- Two distributions: $\mathbf{P}_{\mathbf{s}_1}$ is uniform, $\mathbf{P}_{\mathbf{s}_2}$ is bimodal



First indeterminacy: ordering

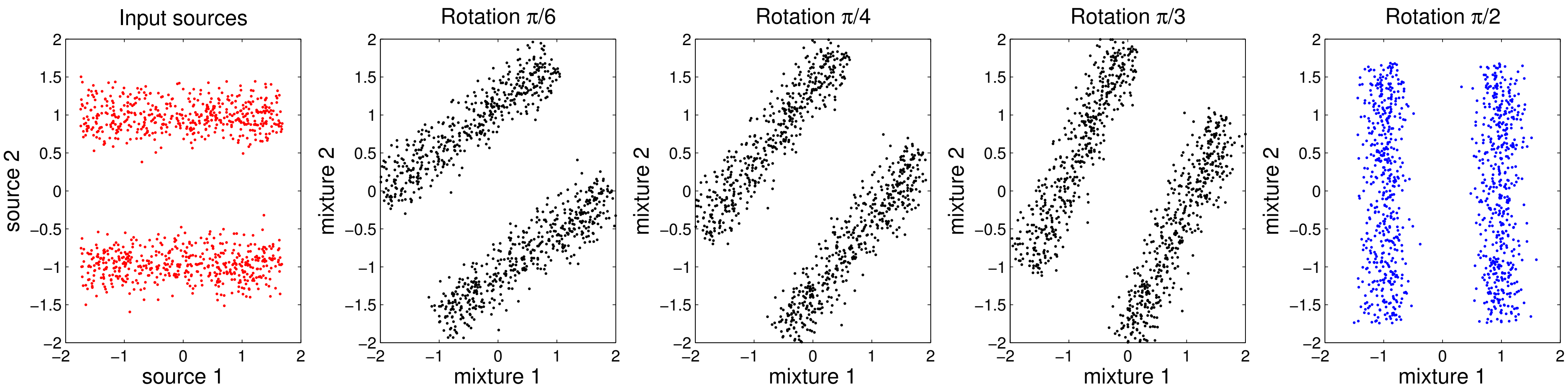
- Initial unmixed RVs in red



- Independent at rotation $\pi/2$

First indeterminacy: ordering

- Initial unmixed RVs in red

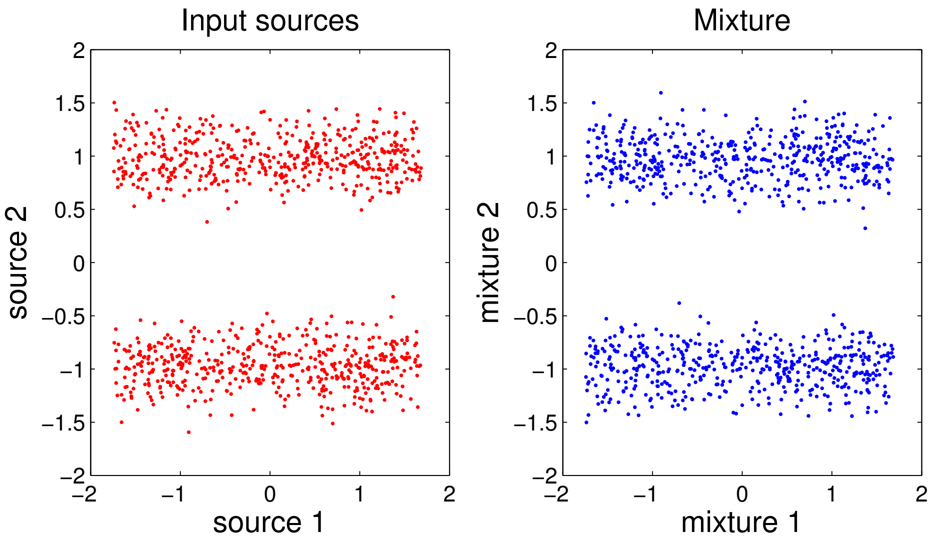


- Independent at rotation $\pi/2$

$\boxed{\text{Ignore source order}}$

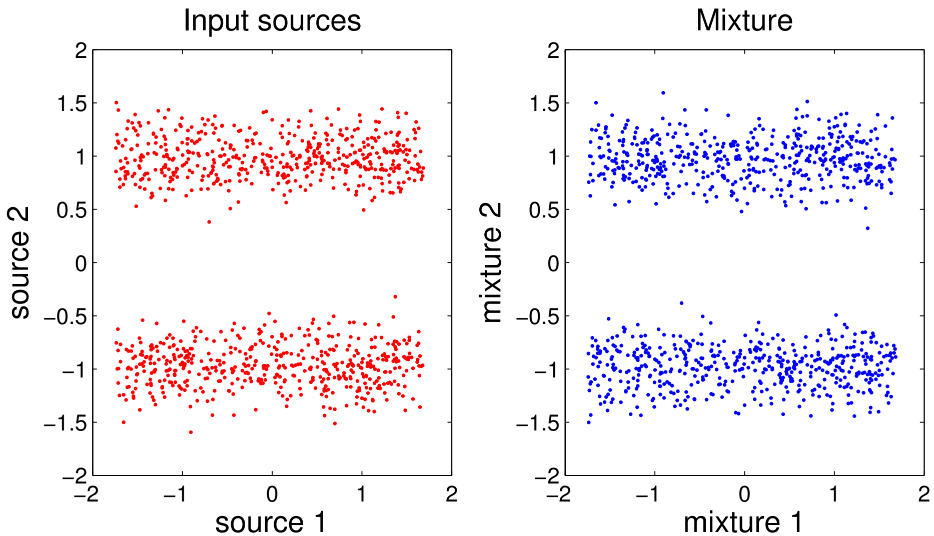
Second indeterminacy: sign

- Initial unmixed RVs in red
- Source 2 sign reversed in blue



Second indeterminacy: sign

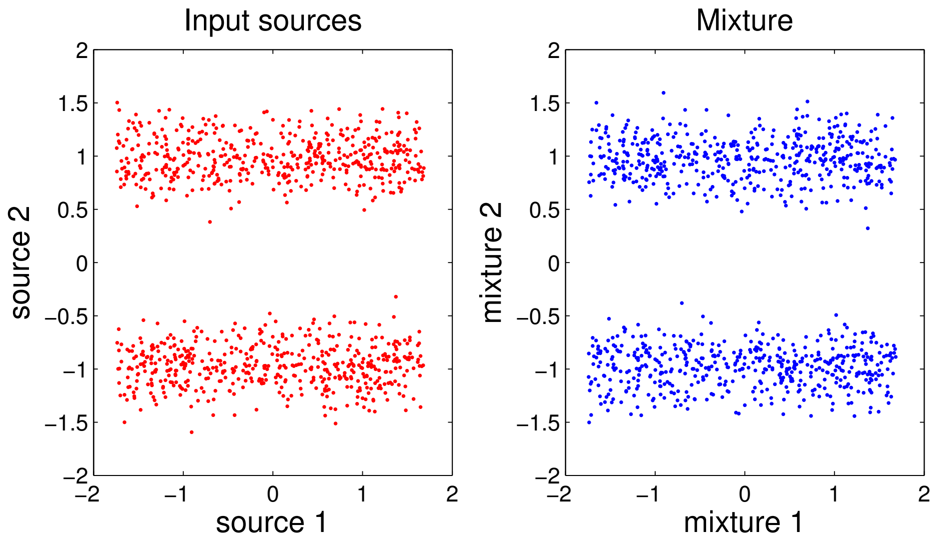
- Initial unmixed RVs in red
- Source 2 sign reversed in blue



Ignore source sign

Second indeterminacy: sign

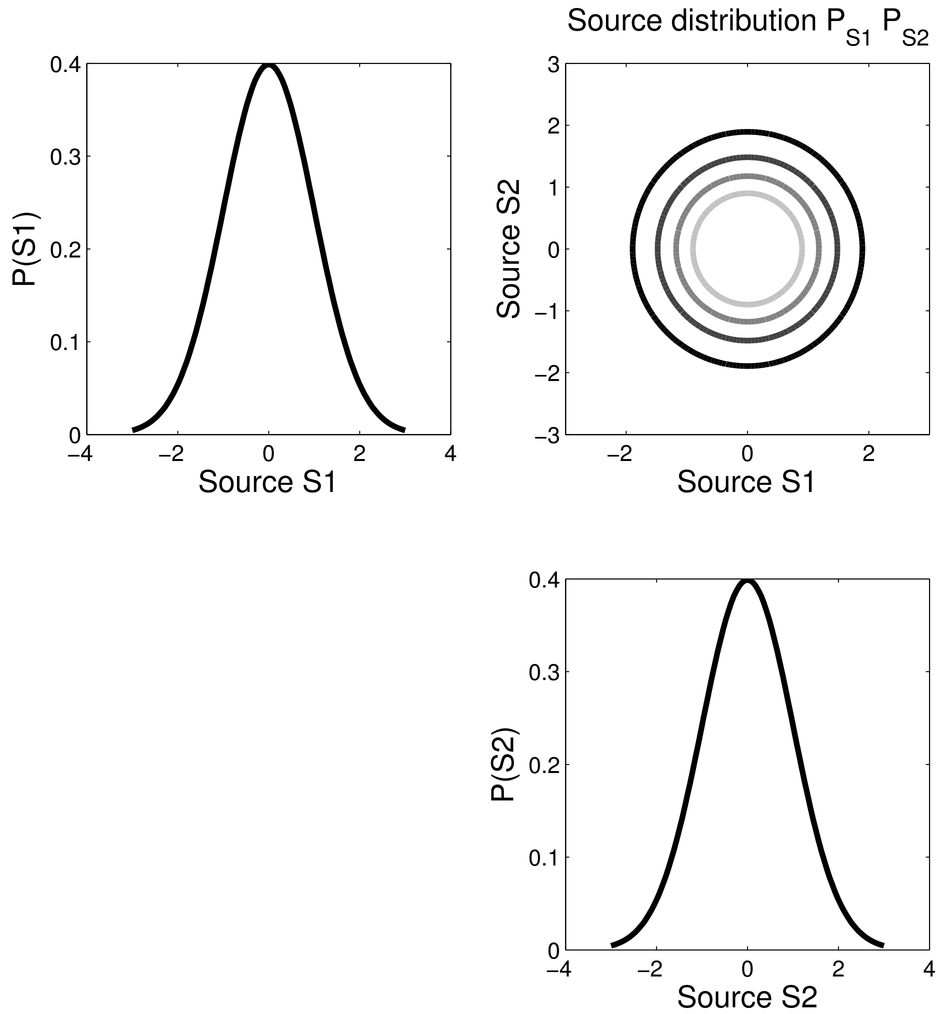
- Initial unmixed RVs in red
- Source 2 sign reversed in blue



- More generally: S_1 and S_2 independent iff aS_1 and S_2 independent for $a \neq 0$
  - Assume sources have unit variance

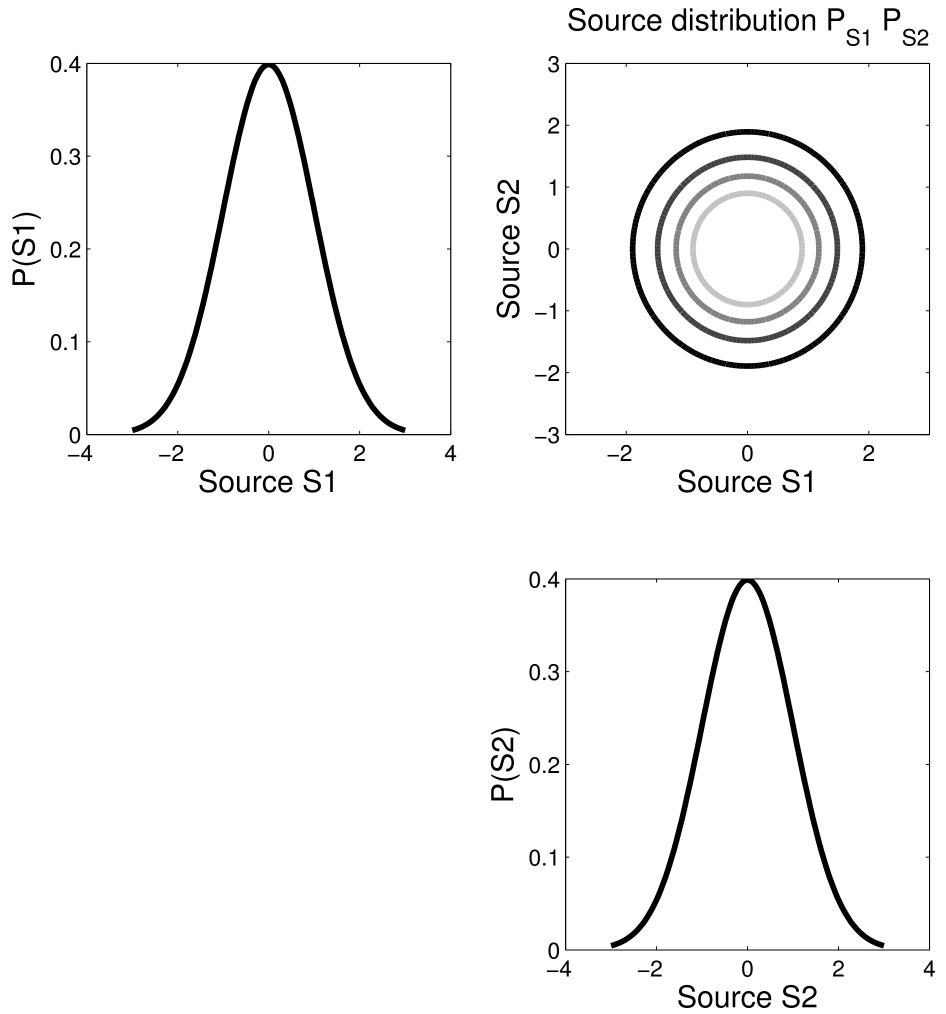
Third indeterminacy: Gaussians

Both sources Gaussian



Third indeterminacy: Gaussians

Both sources Gaussian



Meaningless to “unmix” Gaussians

Things that are impossible for ICA

Using independence alone, we cannot . . .

- recover signal order,
- recover signal sign (or amplitude) ,
- separate multiple Gaussians.

Things that are impossible for ICA

Using independence alone, we cannot ...

- recover signal order,
- recover signal sign (or amplitude) ,
- separate multiple Gaussians.

We can recover

$$B^* = PDA^{-1}$$

- P is a permutation matrix
- D diagonal, $d_{ii} \in \{-1, 1\}$

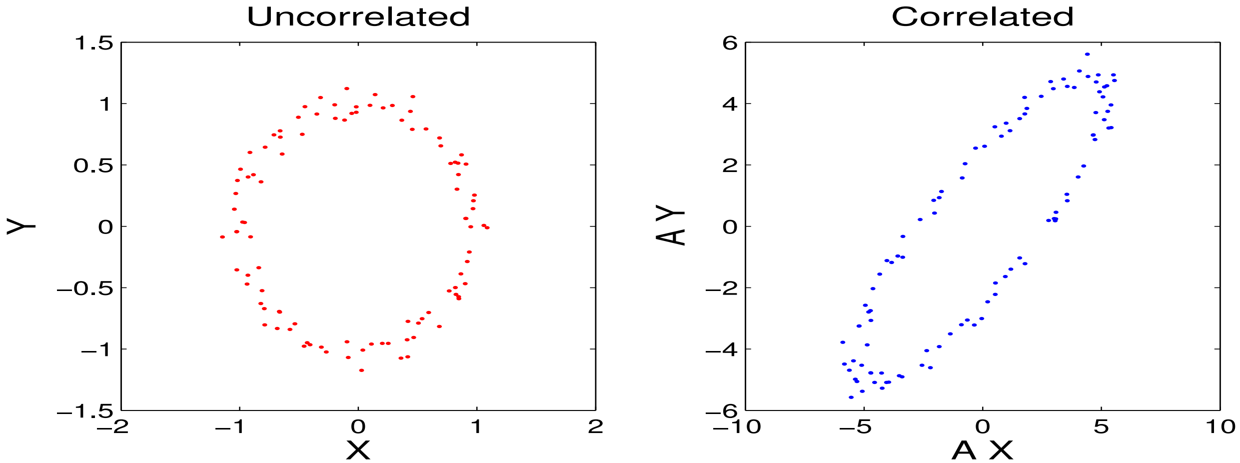
(as long as no more than one Gaussian source)

First step in ICA: decorrelate

- Idea: remove all dependencies of order 2 between mixtures $\mathbf{x}$

First step in ICA: decorrelate

- Idea: remove all dependencies of order 2 between mixtures $\mathbf{x}$



First step in ICA: decorrelate

- Idea: remove all dependencies of order 2 between mixtures $\mathbf{x}$
- New signals have unit covariance:

$$\mathbf{t} = \mathbf{B}_w \mathbf{x} \qquad\qquad \mathbf{C}_t = \mathbf{I}$$

- We thus break up $\mathbf{B}$ as follows:

$$\mathbf{B} = \mathbf{B}_r \mathbf{B}_w$$

  - $\mathbf{B}_w$ is a whitening matrix
  - $\mathbf{B}_r$ is remaining demixing operation
- Use the SVD of mixture covariance $\mathbf{C}_x = \mathbf{U}\Lambda\mathbf{U}^\top$:

$$\mathbf{B}_w = \Lambda^{-1/2}\mathbf{U}^\top$$

First step in ICA: decorrelate

Write C_y (size $l \times l$) as the covariance of $\mathbf{t}$.

$$C_t = m^{-1} T T^\top \qquad \text{where} \qquad T = \mathbf{B}_w X$$

We want to ensure

$$\begin{aligned} I &= C_t \\ &= m^{-1} \mathbf{B}_w X X^\top \mathbf{B}_w{}^\top \\ &= \mathbf{B}_w C_x \mathbf{B}_w{}^\top \end{aligned}$$

First step in ICA: decorrelate

Write C_y (size $l \times l$) as the covariance of **t**.

$$C_t = m^{-1}TT^\top \qquad \text{where} \qquad T = \mathbf{B}_w X$$

We want to ensure

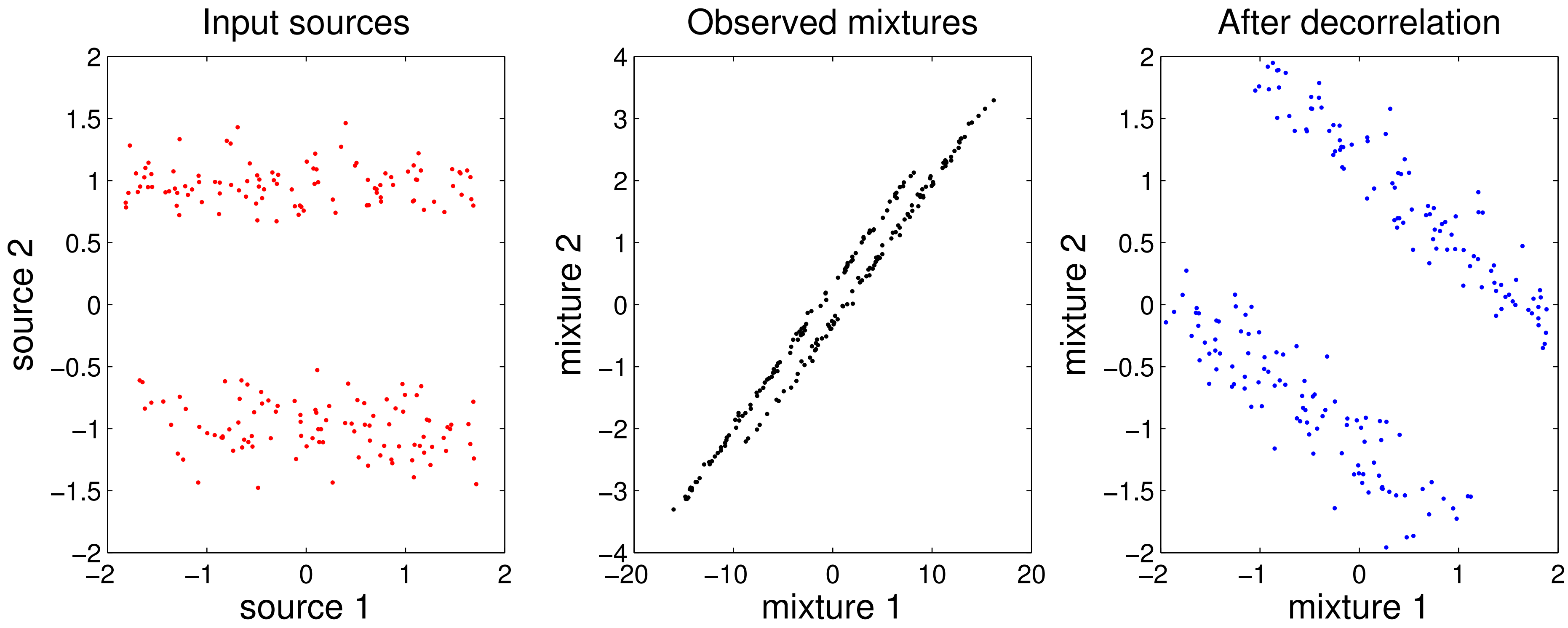
$$\begin{aligned} I &= C_t \\ &= m^{-1}\mathbf{B}_w X X^\top \mathbf{B}_w{}^\top \\ &= \mathbf{B}_w C_x \mathbf{B}_w{}^\top \end{aligned}$$

Write the SVD of $C_x = U\Lambda U^\top$. Write $\mathbf{B}_w = \Lambda^{-1/2}U^\top$. Then

$$\begin{aligned} C_t &= \Lambda^{-1/2}U^\top C_x U \Lambda^{-1/2} \\ &= \Lambda^{-1/2}U^\top U\Lambda U^\top U \Lambda^{-1/2} \\ &= I \end{aligned}$$

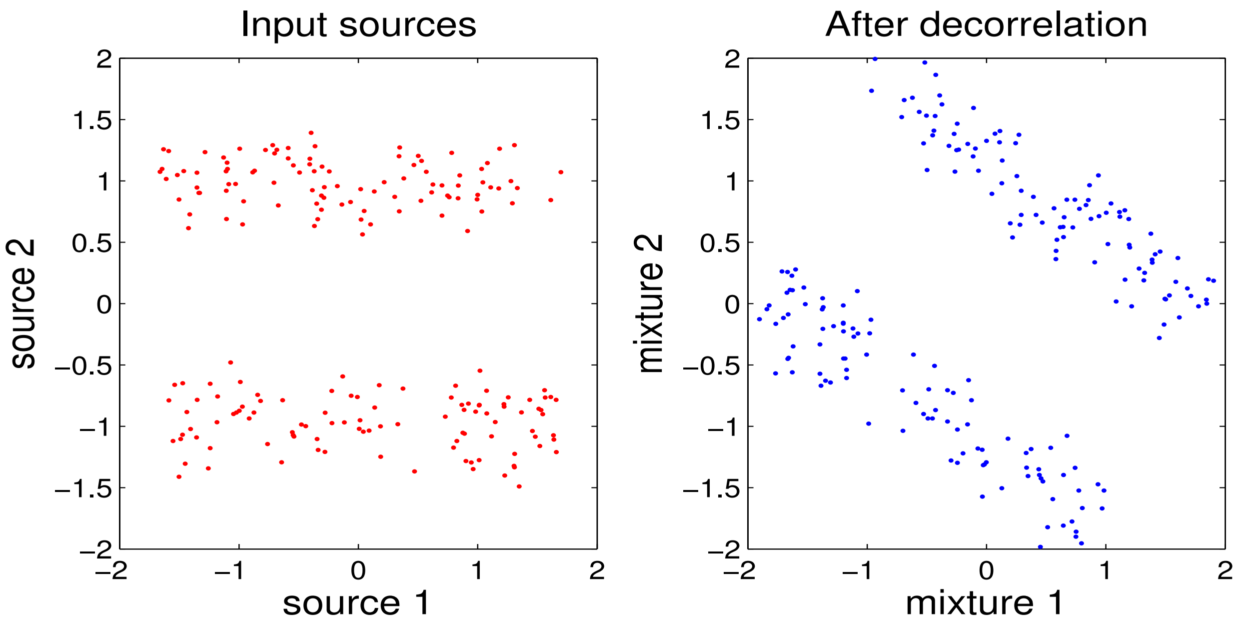
What does decorrelation achieve?

- Two distributions: $\mathbf{P}_{\mathbf{s}_1}$ is uniform, $\mathbf{P}_{\mathbf{s}_2}$ is bimodal



Problem remaining: *rotation*

- Assume correlation has already been removed
- To recover original signal, need to rotate



- In remainder: unmixing matrix $\mathbf{B}$ is rotation,

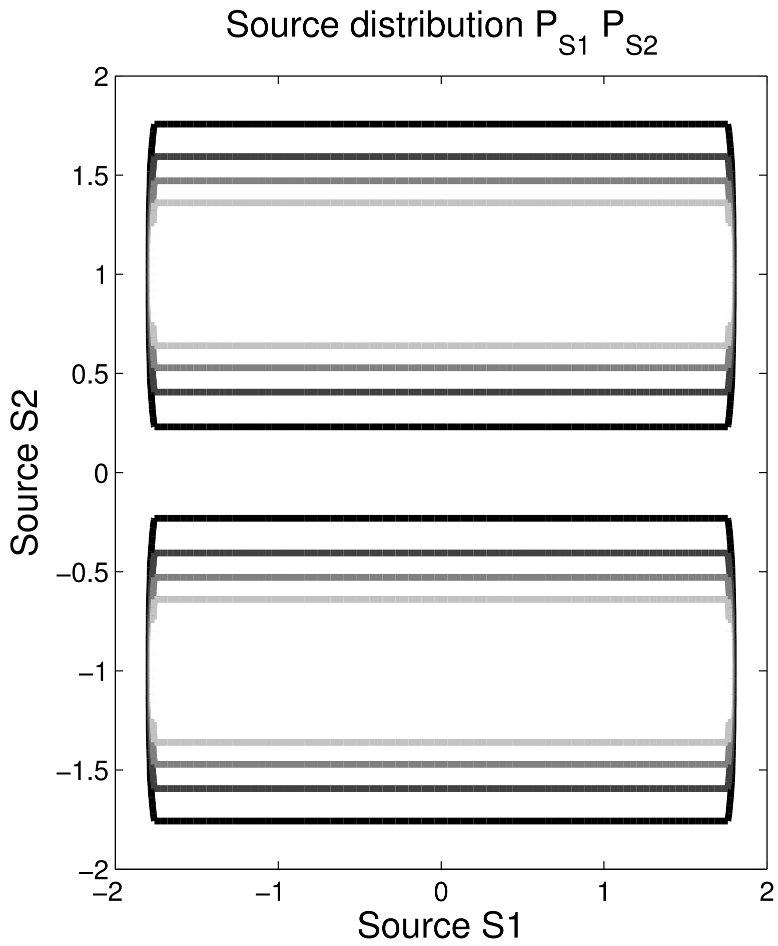
$$\mathbf{B}^{\top}\mathbf{B} = \mathbf{I}$$

ICA: maximum likelihood

- "ICA" using model parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\mathsf{s}})$
- Interpretation: assume we are given the source densities $\hat{\mathsf{P}}_{\mathsf{s}}$, so we only need to find $\mathbf{B}$.

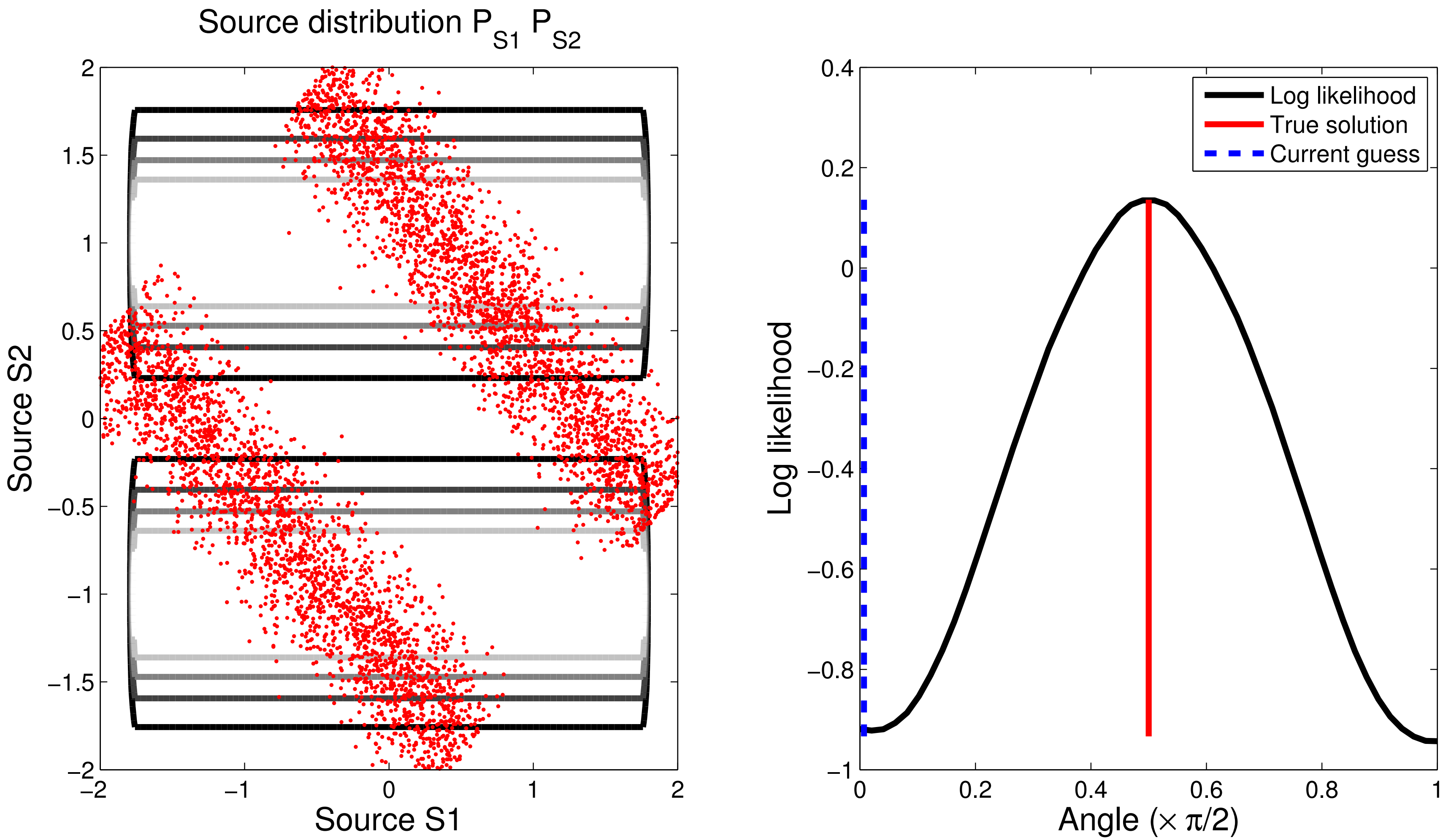
ICA: maximum likelihood

- "ICA" using model parametrised by $(\mathbf{B}, \hat{\mathbf{P}}_{\mathbf{s}})$



ICA: maximum likelihood

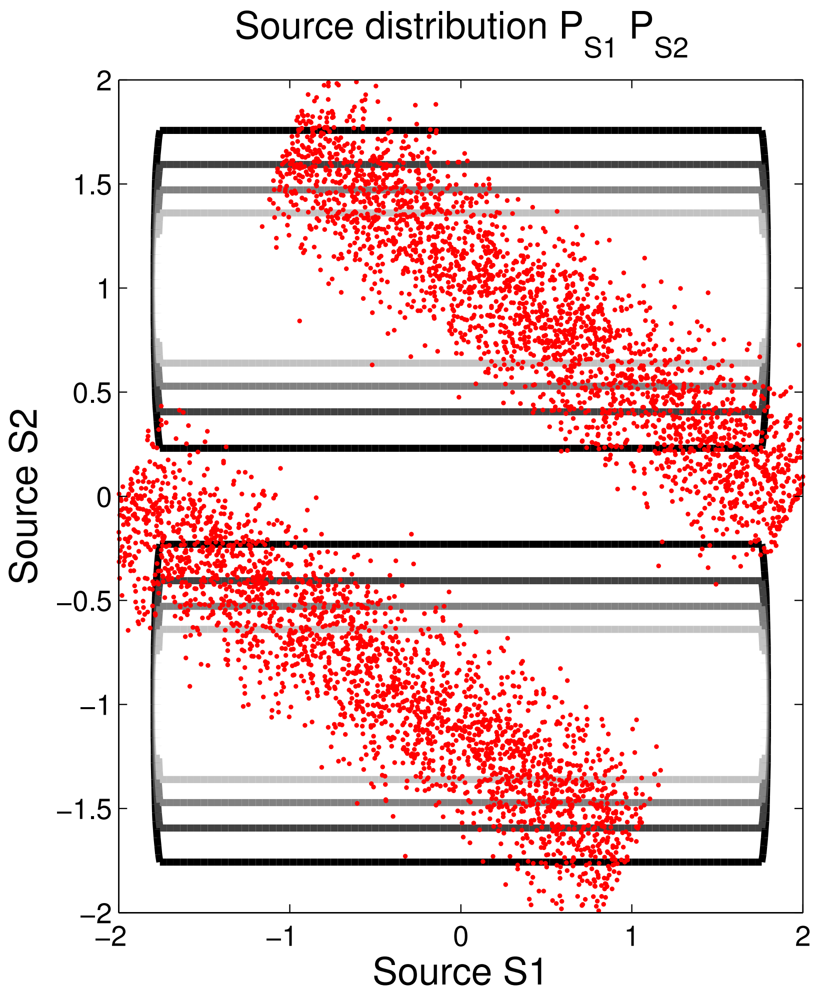
- “ICA” using model parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_\mathsf{s})$

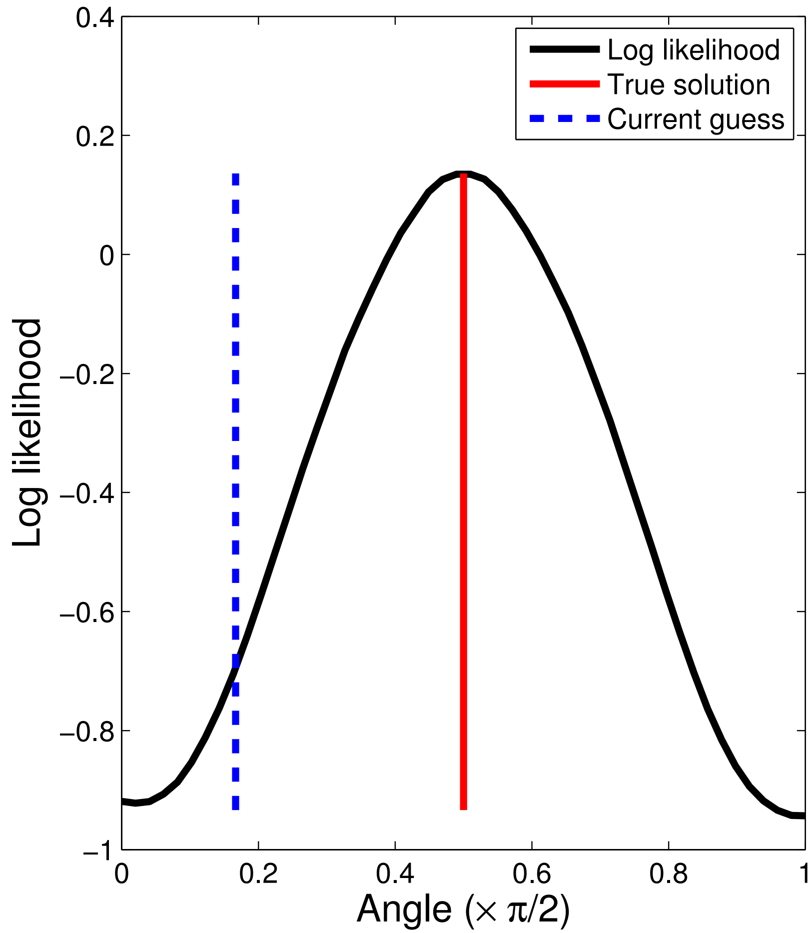


Unmixing angle for B: 0

ICA: maximum likelihood

- "ICA" using model parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\mathsf{s}})$

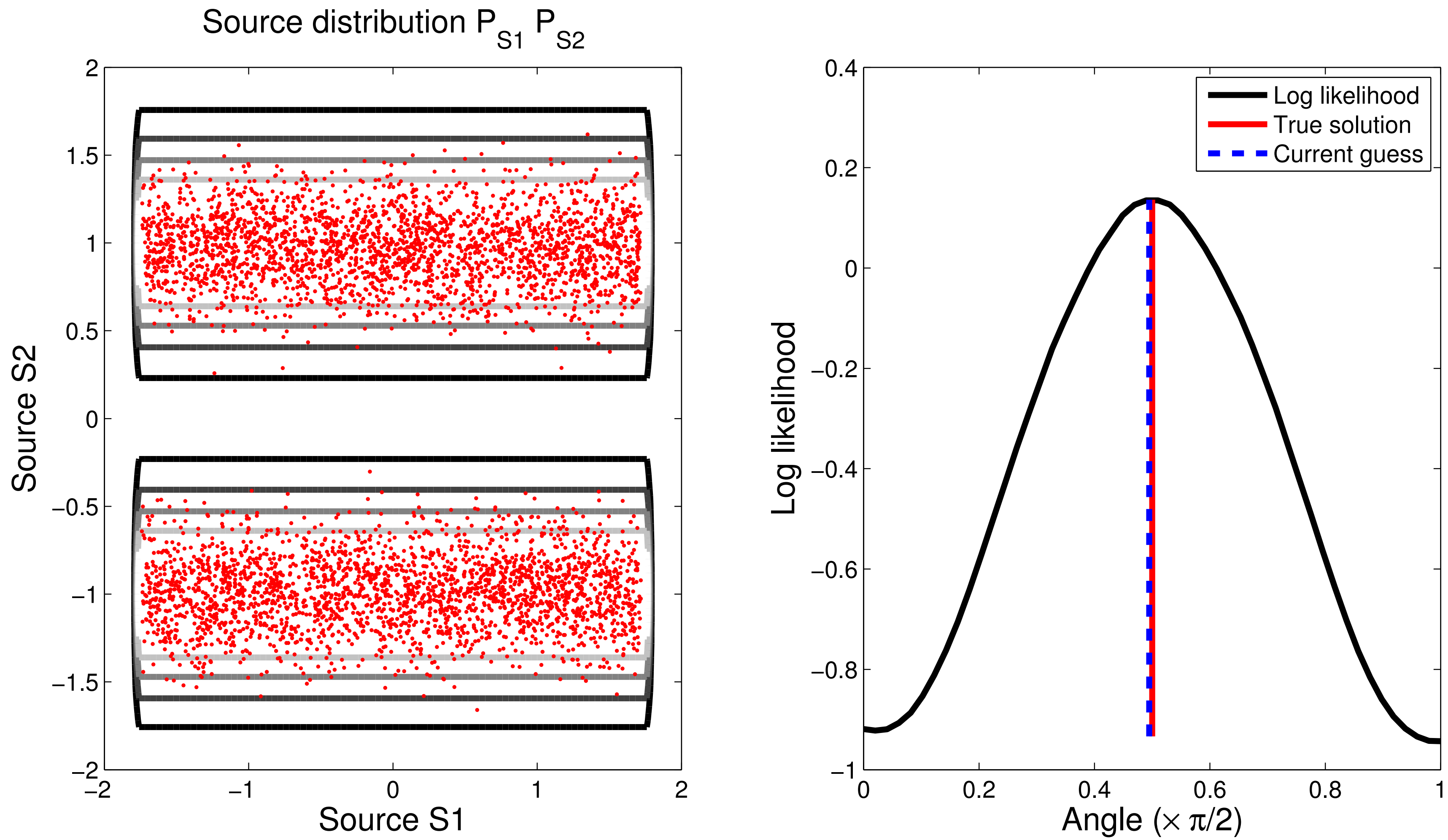




Unmixing angle for B: $\pi/12$

ICA: maximum likelihood

- "ICA" using model parametrised by $(\mathbf{B}, \hat{\mathbf{P}}_{\mathbf{s}})$



Unmixing angle for B: $\pi/4$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\mathsf{s}})$
  - Model must have $\hat{\mathsf{P}}_{\mathsf{s}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{\mathsf{s}_i}$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\mathsf{s}})$
  - Model must have $\hat{\mathsf{P}}_{\mathsf{s}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{\mathsf{s}_i}$
- We use the relation:

$$\begin{aligned} \mathbf{x} &= A\mathbf{s} \\ \mathsf{P}_{\mathsf{x}}(\mathbf{x}) &= \det(A^{-1})\mathsf{P}_{\mathsf{s}}(A^{-1}\mathbf{x}) \end{aligned} \tag{14}$$

- Thus our **estimated** density of observations is

$$\hat{\mathsf{P}}_{\mathsf{x}} = \det(\mathbf{B})\, \hat{\mathsf{P}}_{\mathsf{s}}(\mathbf{B}\mathsf{x})$$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\mathsf{s}})$
  - Model must have $\hat{\mathsf{P}}_{\mathsf{s}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{\mathsf{s}_i}$
- We use the relation:

$$
\begin{aligned}
\mathbf{x} &= A\mathbf{s} \\
\mathsf{P}_{\mathsf{x}}(\mathbf{x}) &= \det(A^{-1})\mathsf{P}_{\mathsf{s}}(A^{-1}\mathbf{x})
\end{aligned}
$$

- Thus, our **estimated** density of observations is

$$
\hat{\mathsf{P}}_{\mathsf{x}} = \cancel{\det(\mathbf{B})}\, \hat{\mathsf{P}}_{\mathsf{s}}(\mathbf{B}\mathsf{x})
$$

ICA: maximum likelihood

- We have a model for the observations, parametrised by $(\mathbf{B}, \hat{\mathsf{P}}_{\mathsf{s}})$
  - Model must have $\hat{\mathsf{P}}_{\mathsf{s}} = \prod_{i=1}^{l} \hat{\mathsf{P}}_{\mathsf{s}_i}$
- Our **estimated** density of observations is

$$\hat{\mathsf{P}}_{\mathsf{x}} = \hat{\mathsf{P}}_{\mathsf{s}}(\mathbf{B}\mathsf{x})$$

- Maximise the expected log likelihood, ($\mathbf{B}_{i,:}$ is ith row)

$$L := \mathbf{E}_{\mathsf{x}}\left[\log \hat{\mathsf{P}}_{\mathsf{x}}\right] = \sum_{i=1}^{l} \mathbf{E}_{\mathsf{x}} \log \hat{\mathsf{P}}_{\mathsf{s}_i}(\mathbf{B}_{i,:}\mathsf{x})$$

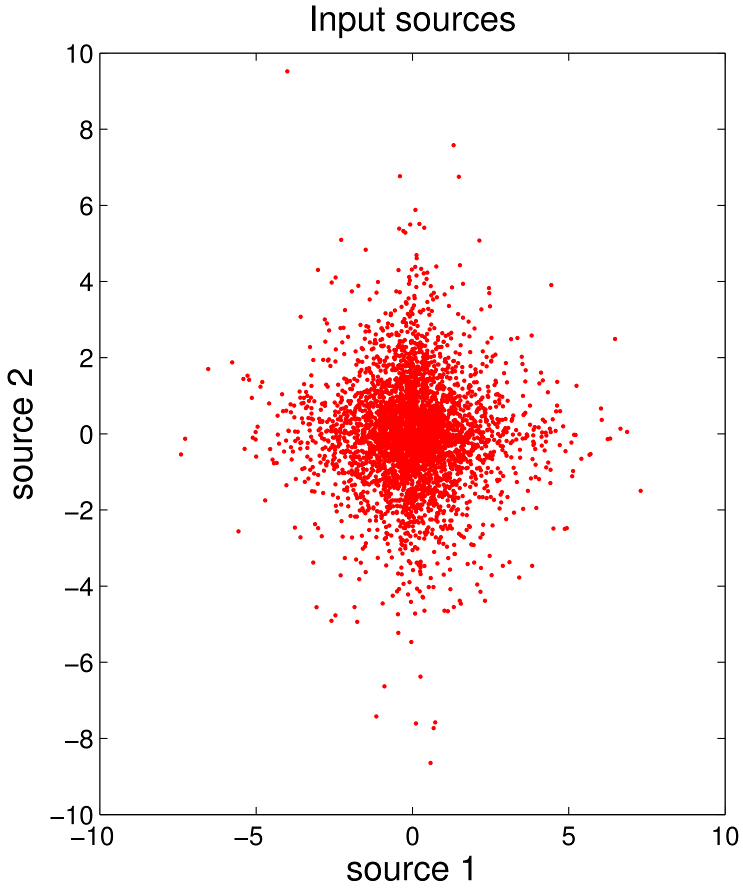
- Finite sample version:

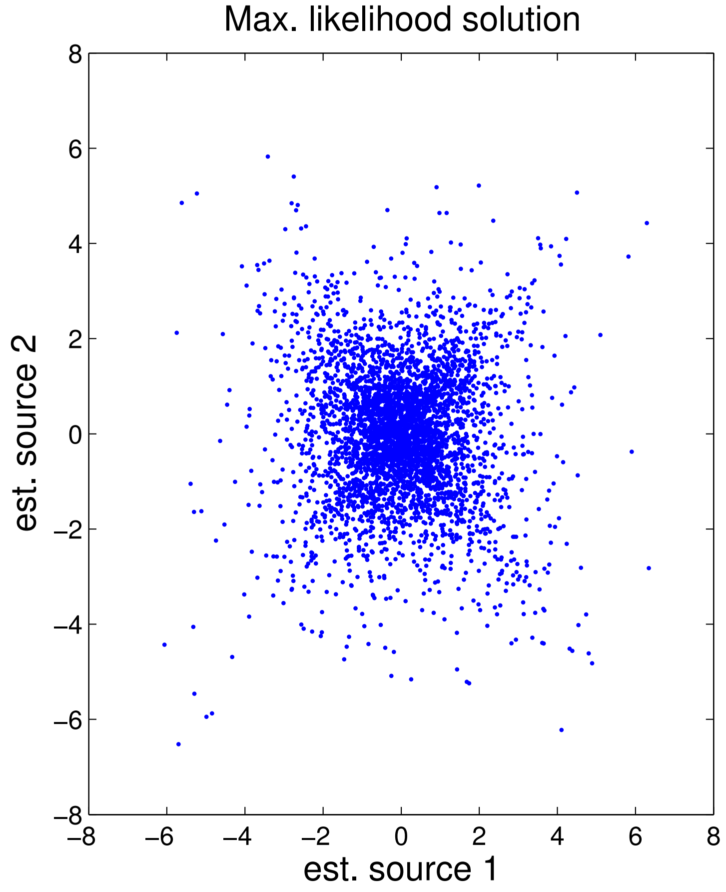
$$L_{\text{emp}} = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{l} \log \hat{\mathsf{P}}_{\mathsf{s}_i}(\mathbf{B}_{i,:} X_{:,j})$$

Notation: $X_{:,j}$ is jth column.

Maximum likelihood: where it fails

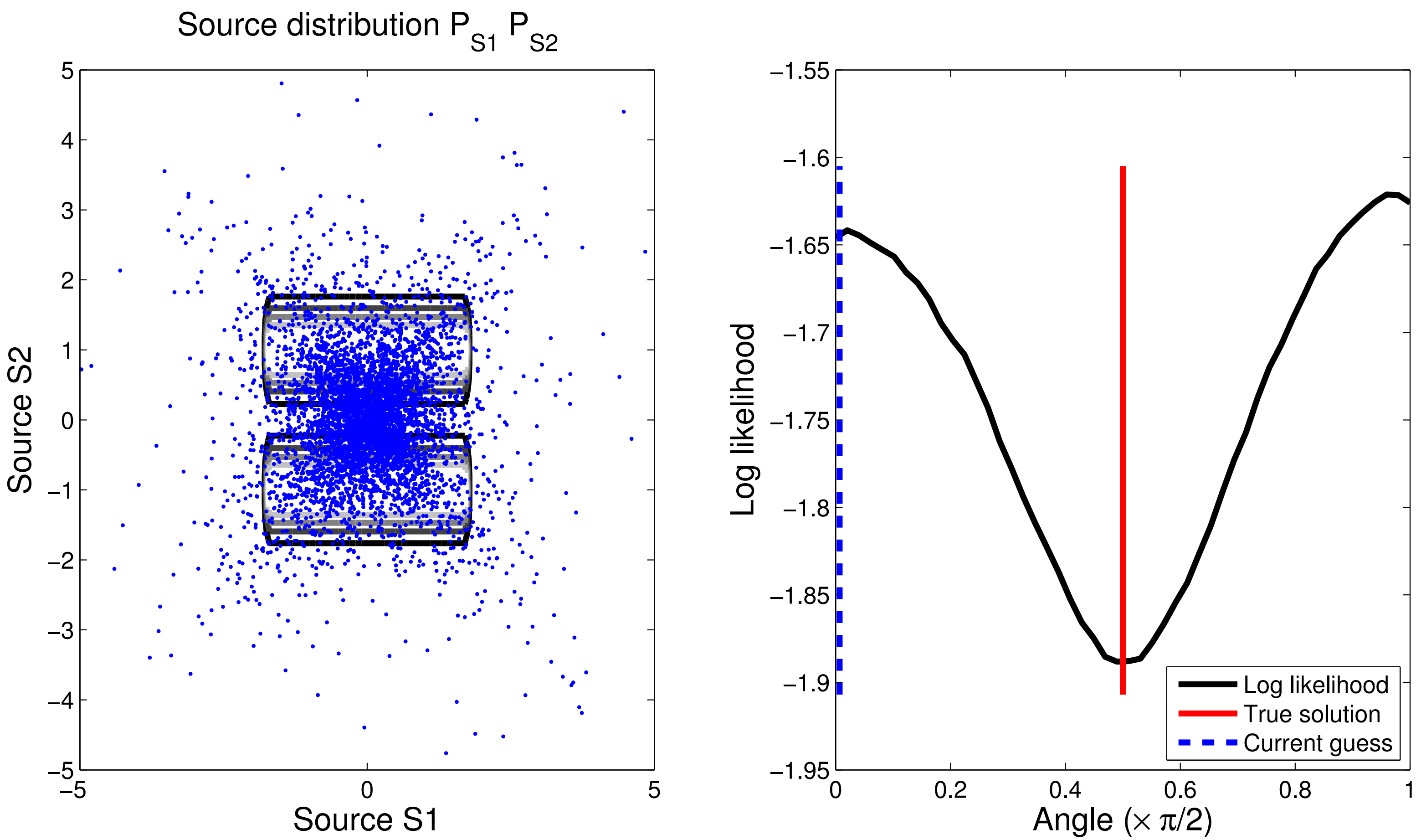
- Model as before, but true source densities are Laplace.
- Why is this wrong?





Maximum likelihood: where it fails

- Model as before, but true source densities are Laplace.
- Why is this wrong?



Another failure mode: Gaussians revisited

Setting:

- $\mathbf{s}$ are two independent, unit variance Gaussians.
- Unmixing matrix B is orthogonal

The density of the mixture $\mathbf{x}$ is proportional to

$$\hat{\mathbf{P}}_{\mathbf{x}} = \mathbf{P}_{\mathbf{s}}(B\mathbf{x}) \propto \exp\left(-\mathbf{x}^\top B^\top C_s^{-1} B\mathbf{x}\right).$$

- C_s is diagonal with equal entries, hence B commutes with C_s^{-1}.
- $B^\top B = I$
- Hence: $\hat{\mathbf{P}}_{\mathbf{x}}$ constant wrt B

We cannot recover independent Gaussians when they are mixed with a rotation matrix.

Back to original setting: independence

- A model-free approach to ICA: use an objective function (contrast function) $\phi(\mathbf{y})$ which measures "closeness to independence".

Back to original setting: independence

- A model-free approach to ICA: use an objective function (contrast function) $\phi(\mathbf{y})$ which measures "closeness to independence".
- Ideally: contrast $\phi(\mathbf{y}) = 0$ if and only if all components of $\mathbf{y}$ mutually independent:

$$\mathbf{P}_{\mathbf{y}} = \prod_{i=1}^{l} \mathbf{P}_{\mathrm{y}_i}.$$

  – Under our mixing assumptions: $\mathbf{y}$ are original sources $\mathbf{s}$ besides permutations, sign swaps

Back to original setting: independence

- A model-free approach to ICA: use an objective function (contrast function) $\phi(\mathbf{y})$ which measures "closeness to independence".
- Ideally: contrast $\phi(\mathbf{y}) = 0$ if and only if all components of $\mathbf{y}$ mutually independent:

$$\mathbf{P}_{\mathbf{y}} = \prod_{i=1}^{l} \mathbf{P}_{\mathrm{y}_i}.$$

  – Under our mixing assumptions: $\mathbf{y}$ are original sources $\mathbf{s}$ besides permutations, sign swaps

- How it's *really* used: contrast should be "smallest" when random variables are "most independent"

Mutual information

- A widely used contrast function: The mutual information,

$$I(\mathbf{y}) = D_{\mathrm{KL}}\left(\mathsf{P}_{\mathbf{y}} \middle\| \prod_{i=1}^{l} \mathsf{P}_{\mathrm{y}_i}\right) = \int \log\left(\frac{\mathsf{P}_{\mathbf{y}}}{\prod_{i=1}^{l} \mathsf{P}_{\mathrm{y}_i}}\right) d\mathsf{P}_{\mathbf{y}}$$

- $D_{\mathrm{KL}} \geq 0$ with equality iff $\mathsf{P}_{\mathbf{y}} = \prod_{i=1}^{l} \mathsf{P}_{\mathrm{y}_i}$

Mutual information

- A widely used contrast function: The mutual information,

$$I(\mathbf{y}) = D_{\mathrm{KL}}\left(\mathbf{P_y}\middle\|\prod_{i=1}^{l}\mathbf{P}_{\mathrm{y}_i}\right) = \int \log\left(\frac{\mathbf{P_y}}{\prod_{i=1}^{l}\mathbf{P}_{\mathrm{y}_i}}\right) d\mathbf{P_y}$$

- $D_{\mathrm{KL}} \geq 0$ with equality iff $\mathbf{P_y} = \prod_{i=1}^{l}\mathbf{P}_{\mathrm{y}_i}$
- Simplification: when $\mathbf{B}$ is a rotation,

$$D_{\mathrm{KL}}\left(\mathbf{P_y}\middle\|\prod_{i=1}^{l}\mathbf{P}_{\mathrm{y}_i}\right) = \sum_{i=1}^{l} h\left(\mathrm{y}_i\right) - h\left(\mathbf{x}\right) - \log\det\mathbf{B}.$$

where $h(\mathrm{y}) = -\mathbf{E}_{\mathrm{y}}\log(\mathbf{P}_{\mathrm{y}}(y))$

Proof: Given $\mathbf{y} = \mathbf{Bx}$

$$\mathbf{P_y}(\mathbf{y}) = \det(\mathbf{B}^{-1})\mathbf{P_x}(\mathbf{B}^{-1}\mathbf{y}) = \det(\mathbf{B}^{-1})\mathbf{P_x}(\mathbf{x})$$

and $\det(\mathbf{B}^{-1}) = (\det(\mathbf{B}))^{-1}$

Mutual information

- A widely used contrast function: The mutual information,

$$I(\mathbf{y}) = D_{\mathrm{KL}}\left(\mathbf{P}_{\mathbf{y}} \middle\| \prod_{i=1}^{l} \mathbf{P}_{\mathrm{y}_i}\right) = \int \log\left(\frac{\mathbf{P}_{\mathbf{y}}}{\prod_{i=1}^{l} \mathbf{P}_{\mathrm{y}_i}}\right) d\mathbf{P}_{\mathbf{y}}$$

- $D_{\mathrm{KL}} \geq 0$ with equality iff $\mathbf{P}_{\mathbf{y}} = \prod_{i=1}^{l} \mathbf{P}_{\mathrm{y}_i}$
- Simplification: when $\mathbf{B}$ is a rotation,

$$D_{\mathrm{KL}}\left(\mathbf{P}_{\mathbf{y}} \middle\| \prod_{i=1}^{l} \mathbf{P}_{\mathrm{y}_i}\right) = \sum_{i=1}^{l} h\left(\mathrm{y}_i\right) - \underbrace{h\left(\mathbf{x}\right) - \log\det\mathbf{B}.}_{\text{constant}}$$

where $h(\mathrm{y}) = -\mathbf{E}_{\mathrm{y}} \log(\mathbf{P}_{\mathrm{y}}(y))$

Mutual information

- A widely used contrast function: The mutual information,

$$I(\mathbf{y}) = D_{\mathrm{KL}}\left(\mathsf{P}_{\mathbf{y}} \middle\| \prod_{i=1}^{l} \mathsf{P}_{y_i}\right) = \int \log\left(\frac{\mathsf{P}_{\mathbf{y}}}{\prod_{i=1}^{l} \mathsf{P}_{y_i}}\right) d\mathsf{P}_{\mathbf{y}}$$

- $D_{\mathrm{KL}} \geq 0$ with equality iff $\mathsf{P}_{\mathbf{y}} = \prod_{i=1}^{l} \mathsf{P}_{y_i}$
- Simplification: when $\mathbf{B}$ is a rotation,

$$D_{\mathrm{KL}}\left(\mathsf{P}_{\mathbf{y}} \middle\| \prod_{i=1}^{l} \mathsf{P}_{y_i}\right) = \sum_{i=1}^{l} h(y_i) - \underbrace{h(\mathbf{x}) - \log\det\mathbf{B}}_{\text{constant}}.$$

where $h(y) = -\mathbf{E}_{y}\log(\mathsf{P}_{y}(y))$

$$\text{Contrast: } \phi_{KL}(\mathbf{y}) := \sum_{i=1}^{l} h(y_i)$$

Maximum likelihood revisited

- Mutual information contrast: minimize

$$\phi_{KL}(\mathbf{y}) := \sum_{i=1}^{l} -\mathbf{E}_{\mathsf{y}_i} \log(\mathbf{P}_{\mathsf{y}_i}(y_i))$$

- Maximum likelihood: maximize

$$\begin{aligned} L &:= \sum_{i=1}^{l} \mathbf{E}_{\mathbf{x}} \log \hat{\mathbf{P}}_{\mathsf{s}_i}(\mathbf{B}_{i,:}\mathbf{x}) \\ &= \sum_{i=1}^{l} \mathbf{E}_{\mathsf{y}_i} \log(\mathbf{P}_{\mathsf{y}_i}(y_i)) \end{aligned}$$

- Same thing! The difference is in approach:
  - For max. likelihood we assumed a model $\hat{\mathbf{P}}_{\mathbf{s}}$
  - Now we (ideally...) assume no model for $\mathbf{P}_{\mathbf{y}}$

Contrast functions with fixed nonlinearities

- Entropies hard to compute/optimize: replace with

$$\phi_f(\mathbf{y}) = \sum_{i=1}^{l} \mathbf{E}_{\mathsf{y}_i}(f(y_i))$$

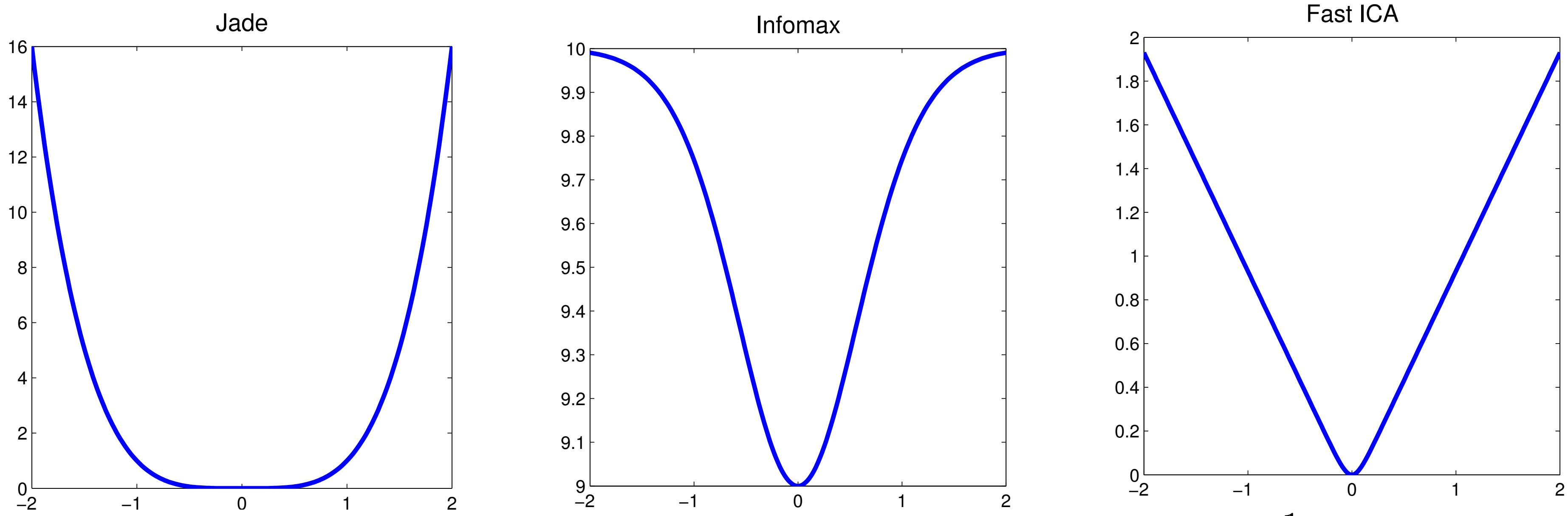
for some other nonlinear $f(y)$

Contrast functions with fixed nonlinearities

- Entropies hard to compute/optimize: replace with

$$\phi_f(\mathbf{y}) = \sum_{i=1}^{l} \mathbf{E}_{\mathrm{y}_i}(f(y_i))$$

for some other nonlinear $f(y)$



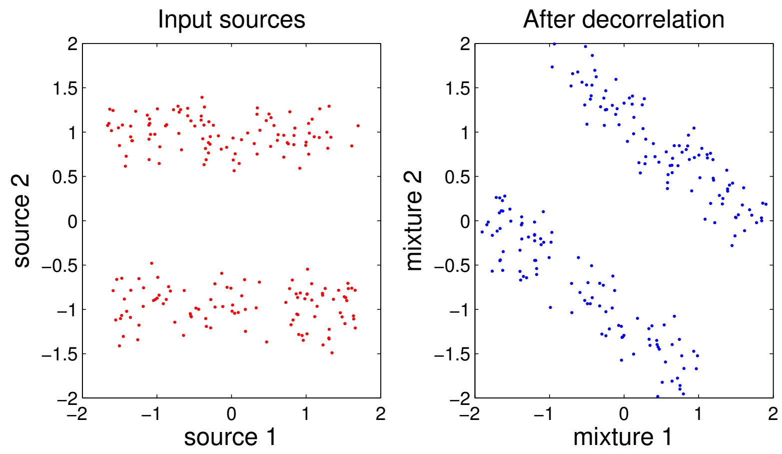
$f(y) = y^4$

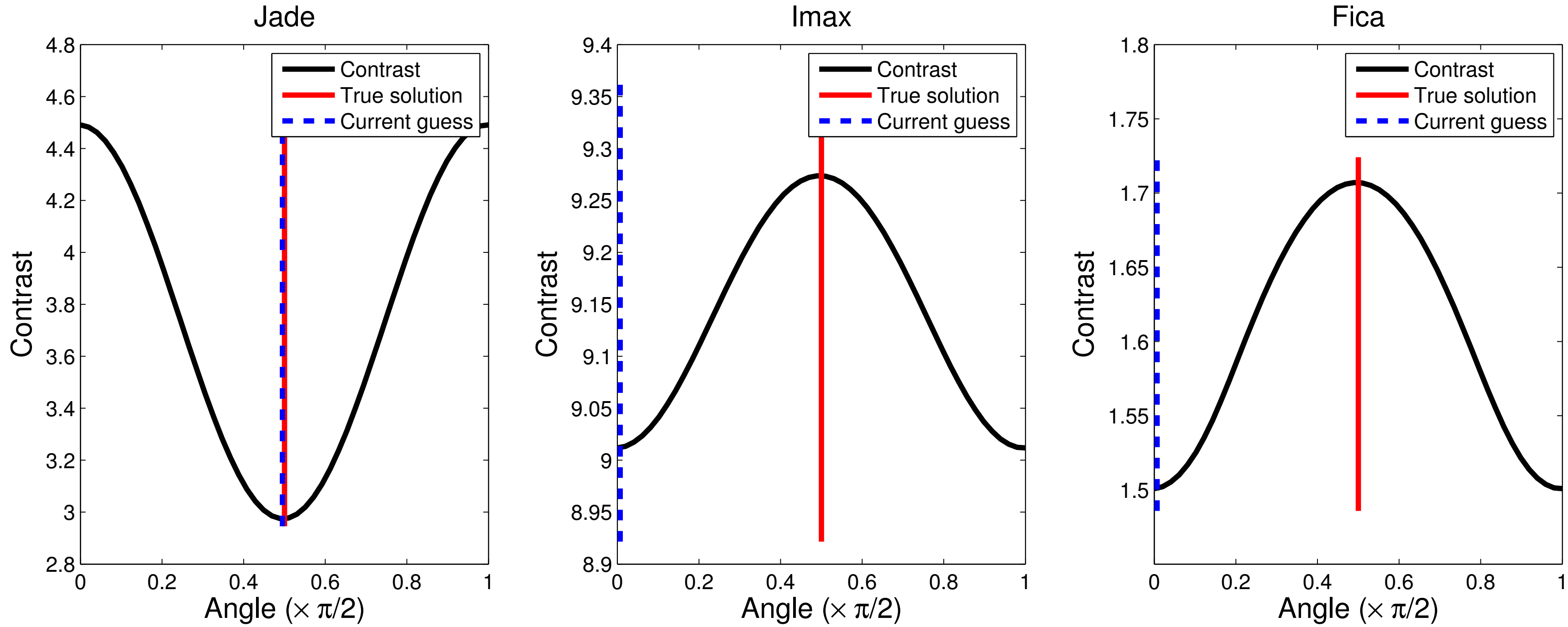
$f(y) = a - exp(-y^2/2)\mathrm{sech}^2(y)$

$f(y) = \dfrac{1}{a} \log \cosh(ay),$

Our example again

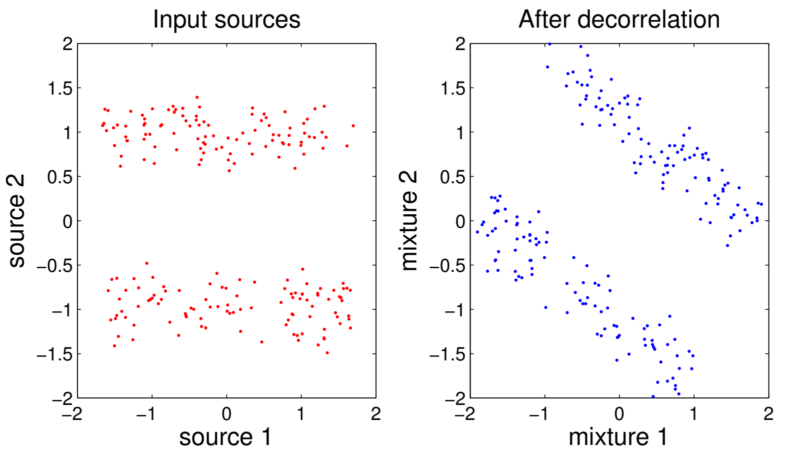
Recall: minimize contrast.





Our example again

Recall: minimize contrast.



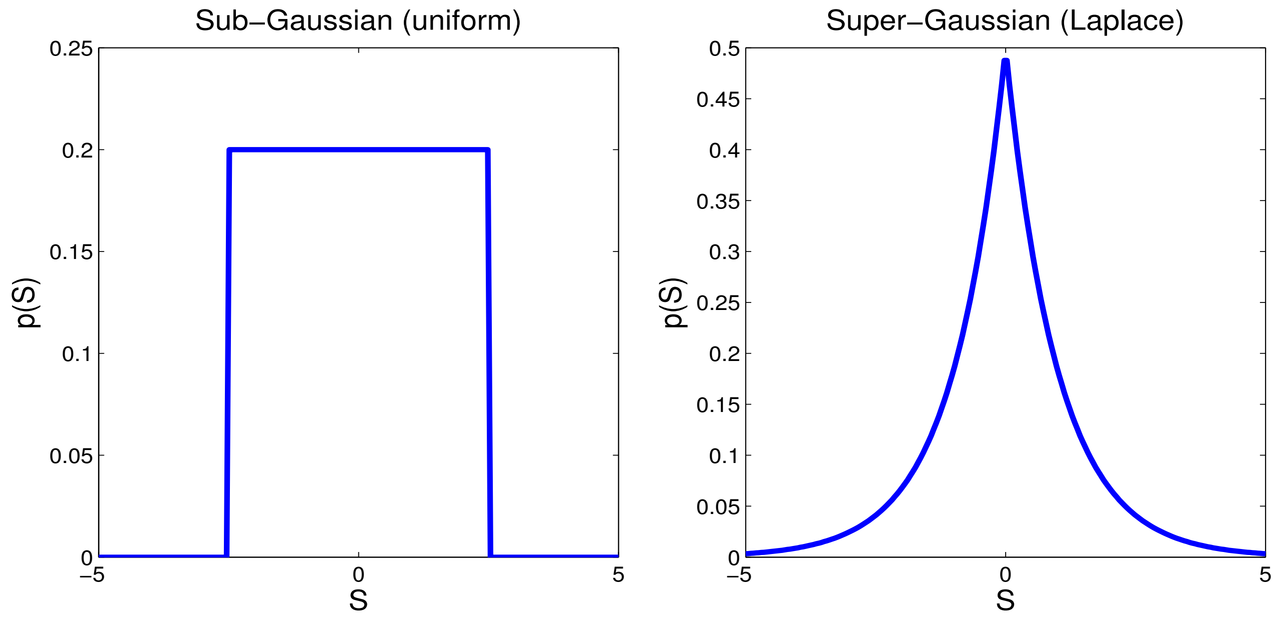
$\boxed{\text{What went wrong?}}$

Kurtosis: an important concept

- Kurtosis definition: when mean is zero,

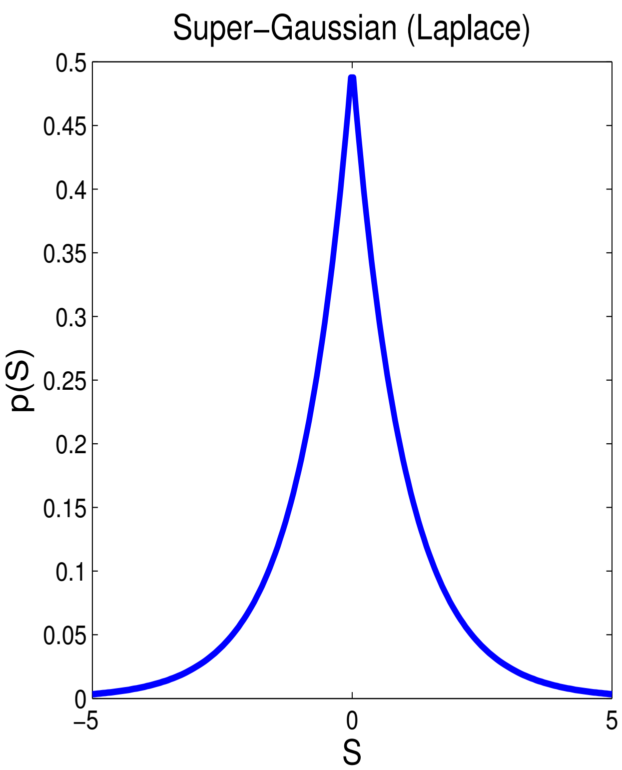
$$\kappa_4 = \mathbf{E}\left(x^4\right) - 3\left(\mathbf{E}\left(x^2\right)\right)^2.$$

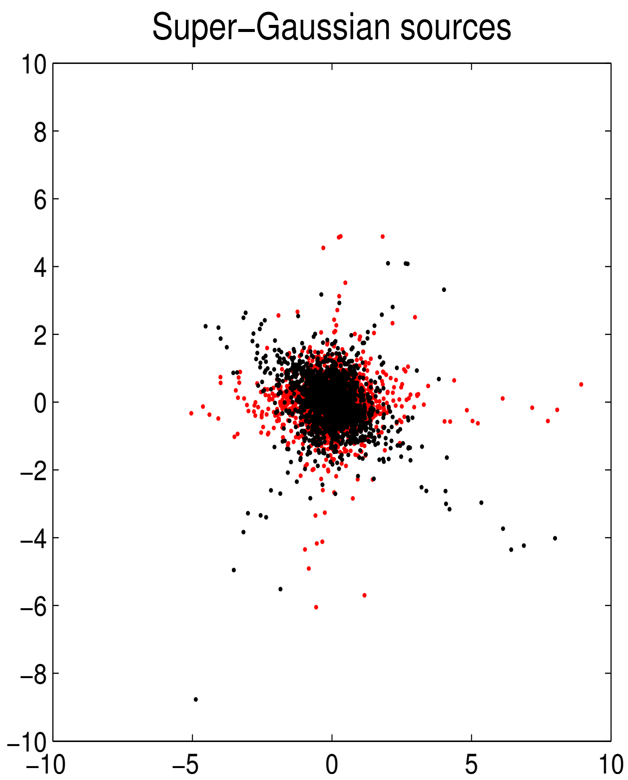
- Source densities can be super-Gaussian (positive kurtosis) or sub-Gaussian (negative kurtosis)
- Zero kurtosis does not mean Gaussian!



Demo: contrasts with fixed nonlinearities

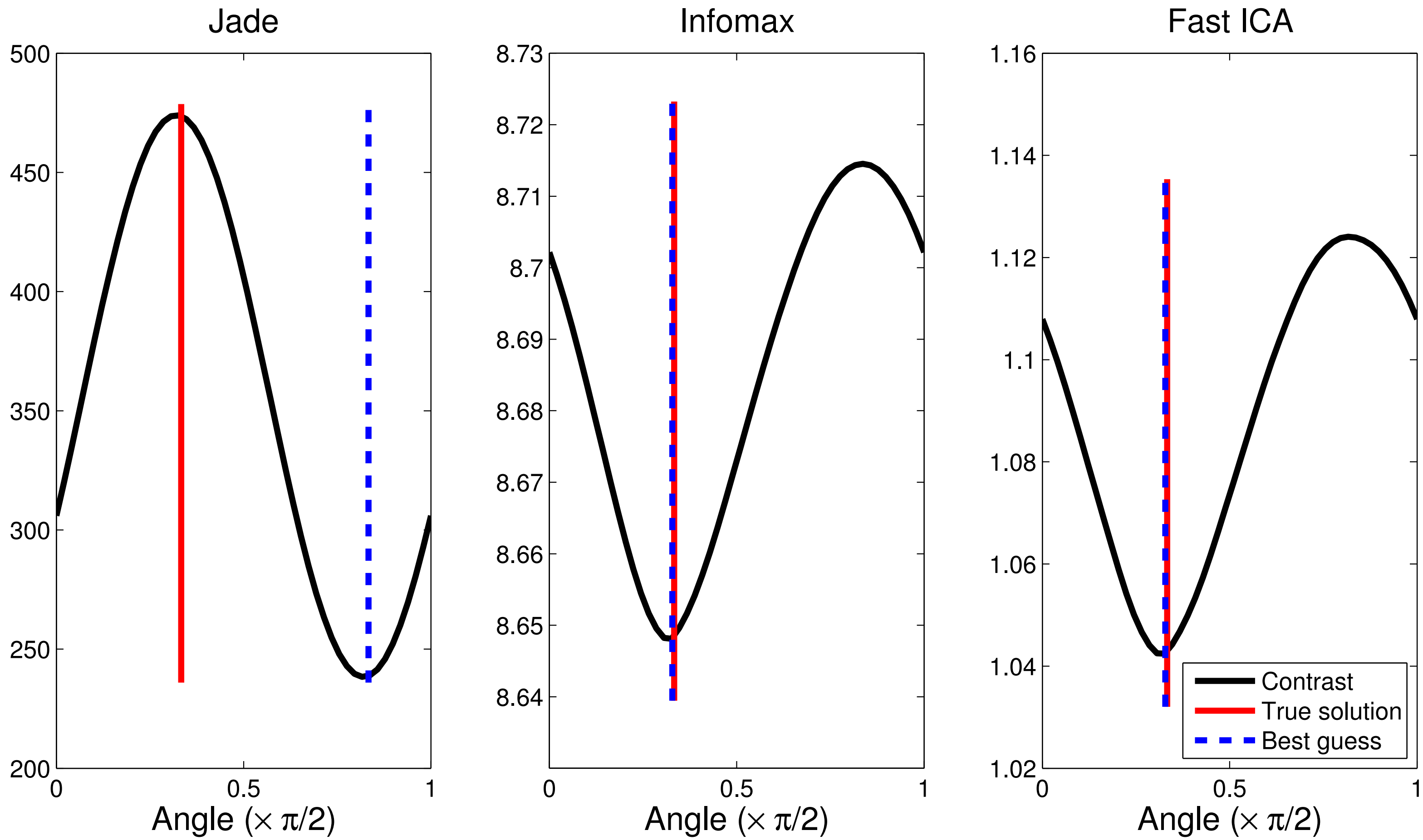
- Super-Gaussian (Laplace) sources
- Unmixed sources in red
- Mixture (angle $\pi/6$) in black





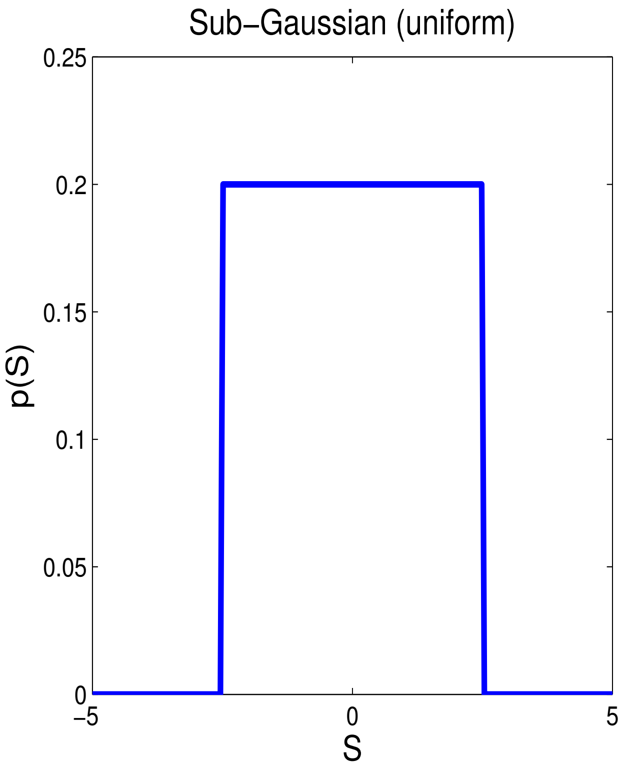
Demo: contrasts with fixed nonlinearities

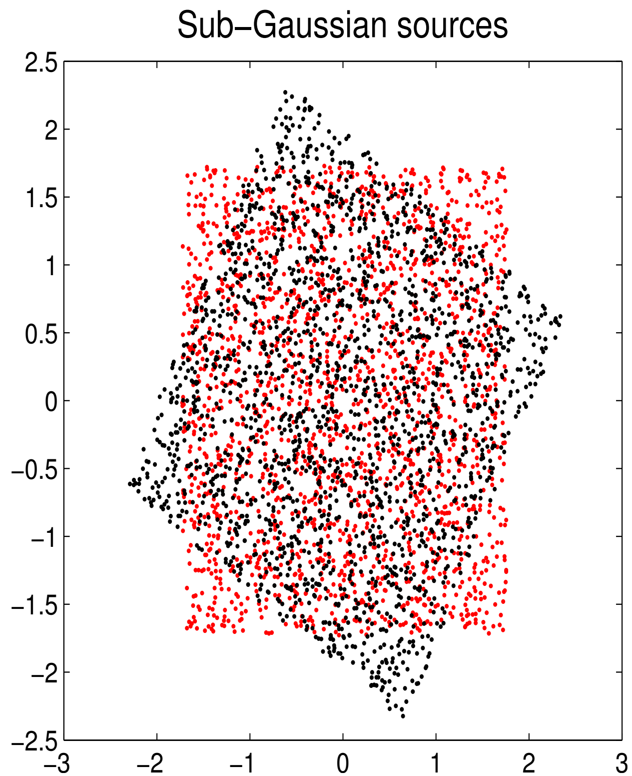
- Super-Gaussian results for Jade, Infomax, and Fast ICA



Demo: contrasts with fixed nonlinearities

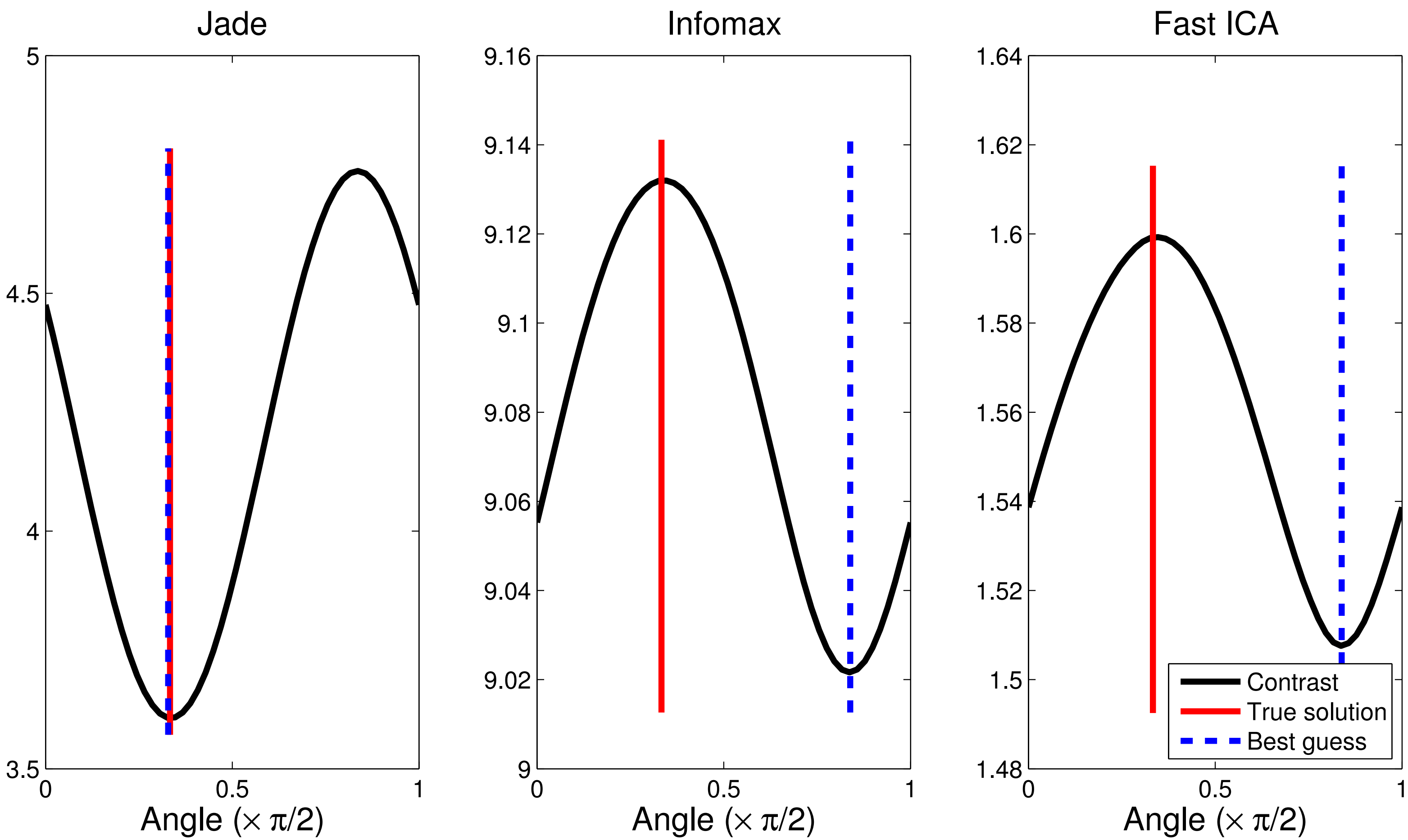
- Sub-Gaussian (Uniform) sources
- Unmixed sources in red
- Mixture (angle $\pi/6$) in black





Demo: contrasts with fixed nonlinearities

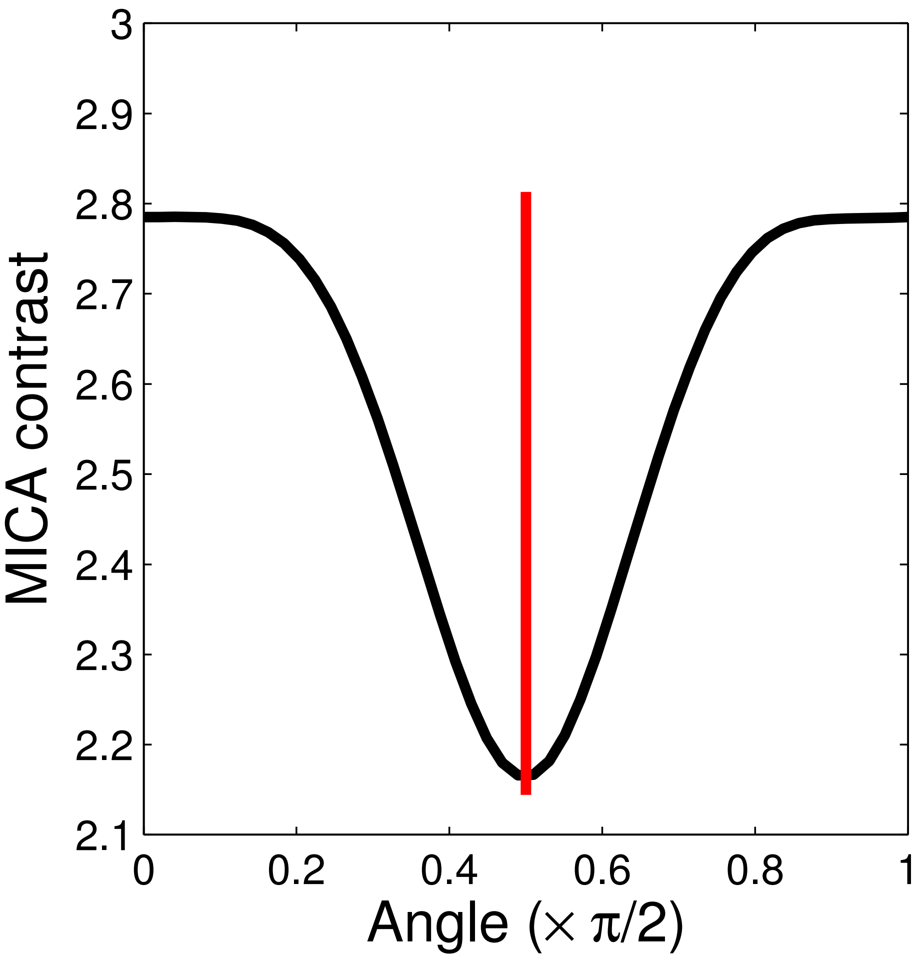
- Sub-Gaussian results for Jade, Infomax, and Fast ICA

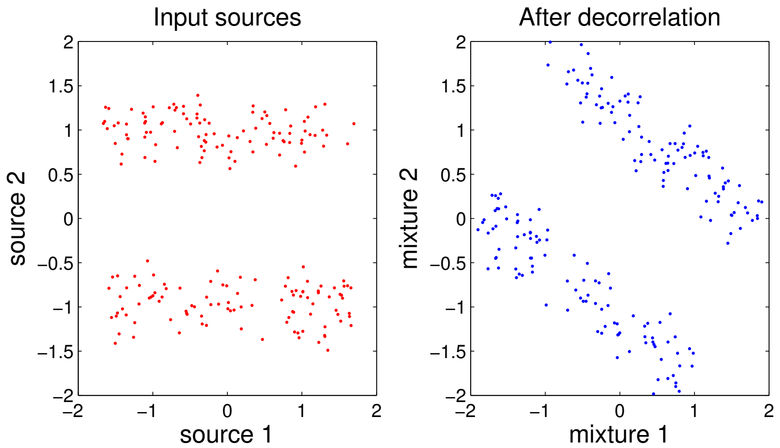


Care needed when using fixed contrasts!

Contrast functions using entropy estimates

- Simplest option: convolve with spline kernel, then compute discrete entropy via space partition [Pham, 2004]

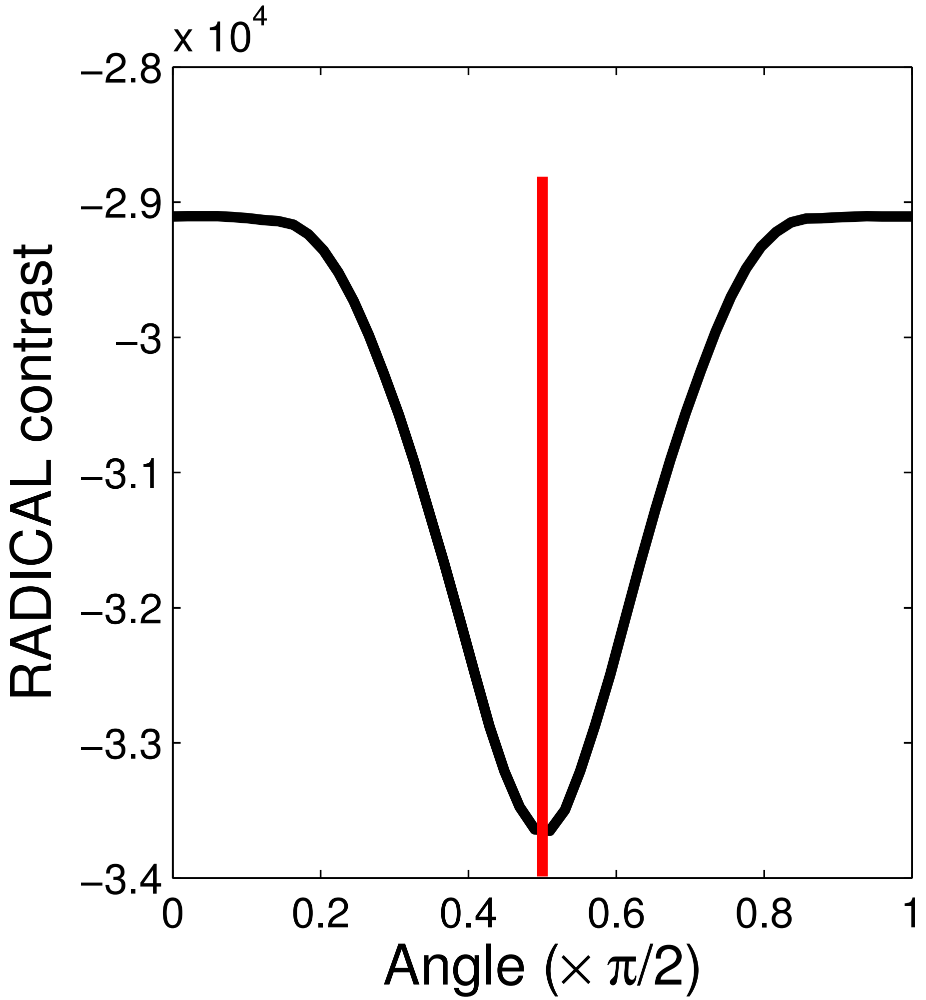


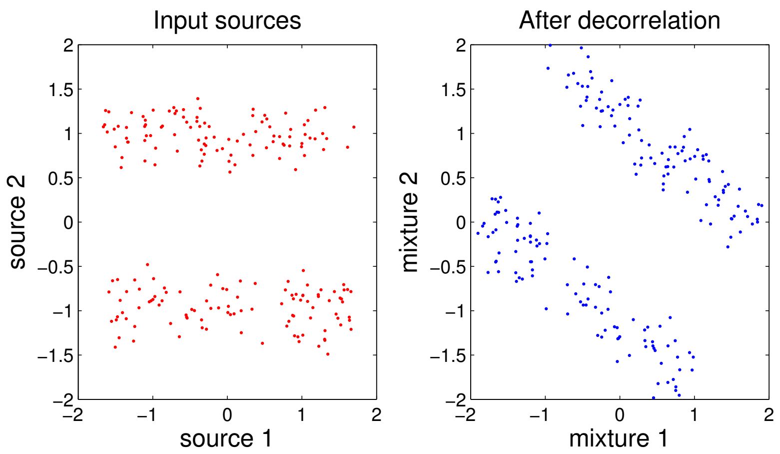


Contrast functions using spacings entropy estimate

- More sophisticated option: spacings estimate of entropy

[Learned-Miller and Fisher III, 2003]



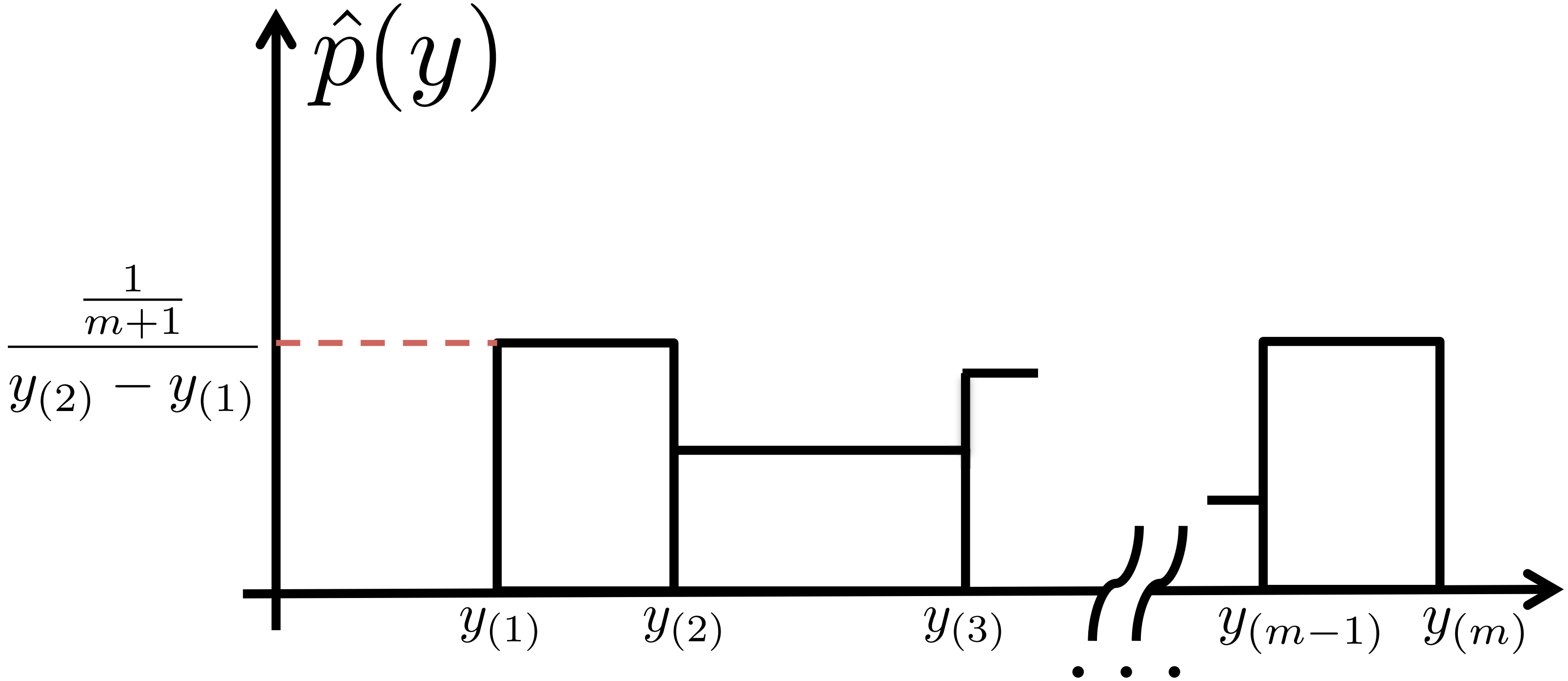


Contrast functions using spacings entropy estimate

- More sophisticated option: spacings estimate of entropy

  [Learned-Miller and Fisher III, 2003]

- Sort sample $Y_1, \ldots, Y_m$ in increasing order: $Y_{(i)} \leq Y_{(i+1)}$
- Prob. density estimate based on spacings
- Idea: prob. mass between adjacent samples $y_{(i)}, y_{(i+1)}$ is $\approx (m+1)^{-1}$



Contrast functions using spacings entropy estimate

- More sophisticated option: spacings estimate of entropy

  [Learned-Miller and Fisher III, 2003]

- Sort sample $Y_1, \ldots, Y_m$ in increasing order: $Y_{(i)} \leq Y_{(i+1)}$
- Prob. density estimate based on spacings

$$\hat{\mathbf{P}}(y; Y_1, \ldots, Y_m) = \frac{1}{(m+1)(Y_{(i+1)} - Y_{(i)})}, \qquad Y_{(i)} \leq y < Y_{(i+1)}$$

- Entropy estimate based on spacings

$$\hat{h}(Y) = \frac{1}{m-1} \sum_{i=1}^{m-1} \log(m+1)(Y_{(i+1)} - Y_{(i)})$$

Contrast functions using spacings entropy estimate

Proof:

$$
\begin{aligned}
H(Y) &= -\int_{-\infty}^{\infty} p(y)\log p(y)dy \\
&\approx -\sum_{i=0}^{m}\int_{y_{(i)}}^{y_{(i+1)}} \hat{p}(y)\log \hat{p}(y)dy \\
&= -\sum_{i=0}^{m}\int_{y_{(i)}}^{y_{(i+1)}} \frac{(m+1)^{-1}}{y_{(i+1)}-y_{(i)}}\log \frac{(m+1)^{-1}}{y_{(i+1)}-y_{(i)}}dy \\
&= -\sum_{i=0}^{m}(m+1)^{-1}\log \frac{(m+1)^{-1}}{y_{(i+1)}-y_{(i)}} \\
&\approx -\sum_{i=1}^{m-1}(m-1)^{-1}\log \frac{(m+1)^{-1}}{y_{(i+1)}-y_{(i)}} \\
&= \sum_{i=1}^{m-1}(m-1)^{-1}\log \left[(m+1)\left(y_{(i+1)}-y_{(i)}\right)\right]
\end{aligned}
$$

Contrast functions using spacings entropy estimate

- More sophisticated option: spacings estimate of entropy

  [Learned-Miller and Fisher III, 2003]

- Sort sample $Y_1, \ldots, Y_m$ in increasing order: $Y_{(i)} \leq Y_{(i+1)}$
- Prob. density estimate based on spacings

$$\hat{\mathsf{P}}(y; Y_1, \ldots, Y_m) = \frac{1}{(m+1)(Y_{(i+1)} - Y_{(i)})}, \qquad Y_{(i)} \leq y < Y_{(i+1)}$$

- Entropy estimate based on spacings

$$\hat{h}(Y) = \frac{1}{m-1} \sum_{i=1}^{m-1} \log(m+1)(Y_{(i+1)} - Y_{(i)})$$

- Smoothing: add "extra" mixture points (noisy copies of original mixtures)
- Hard to optimize

Other independence measures as contrasts

- Why mutual information?
  - Same as maximum likelihood (good if model is correct)
  - Contrast function is sum of entropies: fast
- Other independence measures?

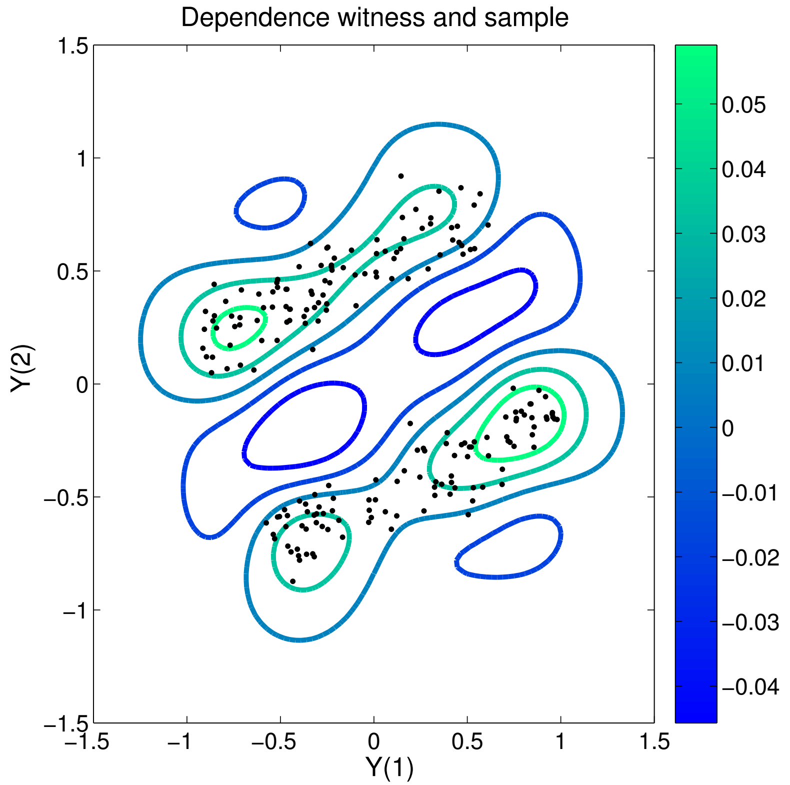
Other independence measures as contrasts

- Why mutual information?
  - Same as maximum likelihood (good if model is correct)
  - Contrast function is sum of entropies: fast
- Other independence measures?
- Most common: kernel/characteristic function-based
  - Characteristic function-based ICA [Eriksson and Koivunen, 2003, Chen and Bickel, 2005]
  - Kernel ICA (covariance): COCO, KMI, HSIC [Gretton et al., 2005, Shen et al., 2007, 2009]
  - Kernel ICA (correlation): KCCA, KGV [Bach and Jordan, 2002]
- HSIC same as characteristic function-based (for the purposes of ICA) [Shen et al., 2009]

Kernel contrast function: HSIC

- Dependence measure:

$$\mathrm{HSIC}(\mathbf{P}_{UV}, F) := \left(\sup_{f \in F} \left[\mathbf{E}_{UV} f - \mathbf{E}_U \mathbf{E}_V f \right] \right)^2$$



HSIC: empirical expression

- Empirical HSIC:

$$\mathrm{HSIC} := \frac{1}{m^2}\mathrm{tr}(KHLH)$$

  - K Gram matrix for $(u_1, \ldots, u_m)$
  - L Gram matrix for $(v_1, \ldots, v_m)$
  - Centering $H = I - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top$

Contrast functions: a small selection

Contrast function summary

- Sum of expectations of a fixed nonlinearity
  - Fast ICA, Infomax, Jade
- Sum of entropies/mutual information...
  - ... using fast, smoothed entropy estimates
  - ... using spacings/k-nn entropy estimates
- Kernel/characteristic function dependence measures

Contrast functions: a small selection

Contrast function summary

- Sum of expectations of a fixed nonlinearity
  - Fast ICA, Infomax, Jade
- Sum of entropies/mutual information…
  - … using fast, smoothed entropy estimates
  - … using spacings/k-nn entropy estimates
- Kernel/characteristic function dependence measures

How do we optimize?

Optimization (Jacobi)

- For two signals, the rotation is expressed

$$\mathbf{B} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

- Higher dimensions, eg for $l = 3$,

$$\mathbf{B} := \begin{bmatrix} \cos(\theta_z) & -\sin(\theta_z) & 0 \\ \sin(\theta_z) & \cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \cos(\theta_y) & 0 & -\sin(\theta_y) \\ 0 & 1 & 0 \\ \sin(\theta_y) & 0 & \cos(\theta_y) \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) \\ 0 & \sin(\theta_x) & \cos(\theta_x) \end{bmatrix}$$

- Coordinate descent, exhaustive search, etc...

Optimization (Newton)

- Unmixing matrix B satisfies $B^\top B = I$
- Local parameterisation Ω about B: at iteration k,

$$B_{k+1} = B_k \exp(\Omega) \qquad \Omega = -\Omega^\top$$

- How to choose direction and size of Ω?

Optimization (Newton)

- Unmixing matrix B satisfies $B^\top B = I$
- Local parameterisation Ω about B: at iteration k,

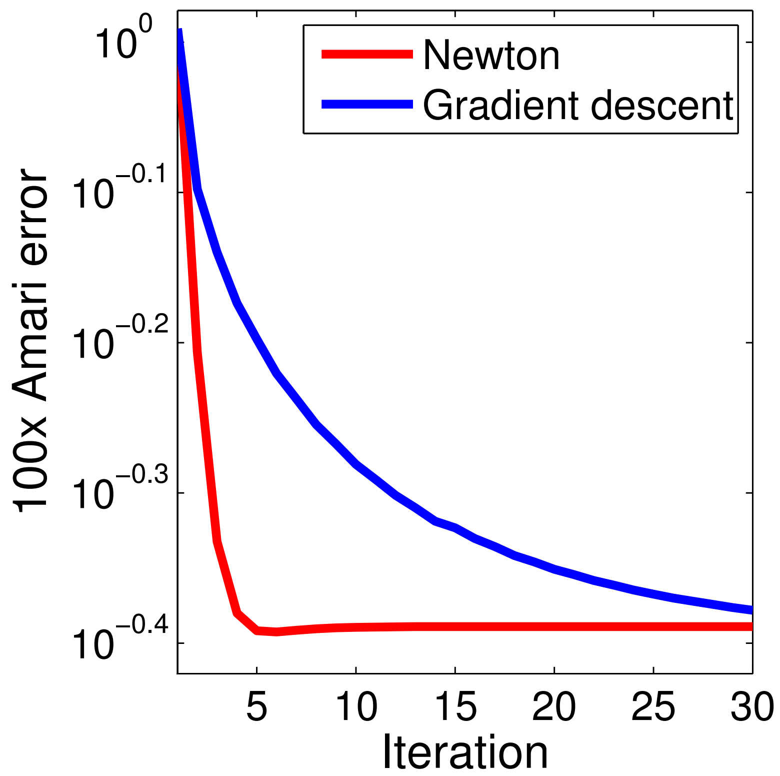
$$B_{k+1} = B_k \exp(\Omega) \qquad \Omega = -\Omega^\top$$

- How to choose direction and size of Ω?
- Write $\widetilde{\Omega} \in \mathbb{R}^{l(l-1)/2}$ the unique entries of Ω
- Newton-like method: solve the linear system for $\widetilde{\Omega} \in \mathbb{R}^{l(l-1)/2}$

$$\mathcal{H}_{B_k}(\phi)\widetilde{\Omega} = -\nabla_{B_k}(\phi)$$

  - $\nabla_{B_k}(\phi)$ is gradient of ϕ wrt $\widetilde{\Omega}$
  - $\mathcal{H}_{B_k}(\phi)$ is Hessian of ϕ wrt $\widetilde{\Omega}$
- Approximate Hessian as diagonal: FastICA [Shen and Hüper, 2006]

Gradient descent vs Newton



What if we have time dependence?

- We can get extra information from sources not being i.i.d.
- Mixture $\mathbf{x}(t)$ now stationary random process, depends on $\mathbf{x}(t-\tau)$
- Define mixture covariances

$$\mathbf{C}_0 = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t)), \qquad \mathbf{C}_\tau = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t-\tau)),$$

  - $\mathbf{C}_\tau$ independent of t (stationarity)

What if we have time dependence?

- We can get extra information from sources not being i.i.d.
- Mixture $\mathbf{x}(t)$ now stationary random process, depends on $\mathbf{x}(t-\tau)$
- Define mixture covariances

$$\mathbf{C}_0 = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t)), \qquad \mathbf{C}_\tau = \mathbf{E}(\mathbf{x}(t)\mathbf{x}(t-\tau)),$$

  – $\mathbf{C}_\tau$ independent of t (stationarity)

- Decorrelate:

$$\mathbf{B}\mathbf{C}_0\mathbf{B}^\top = \Lambda \qquad \mathbf{B}\mathbf{C}_\tau\mathbf{B}^\top = \widetilde{\Lambda}$$

  – Λ and $\widetilde{\Lambda}$ diagonal

- Combining both requirements:

$$\mathbf{B}\mathbf{C}_0\mathbf{C}_\tau^{-1} = \left(\Lambda\widetilde{\Lambda}^{-1}\right)\mathbf{B}$$

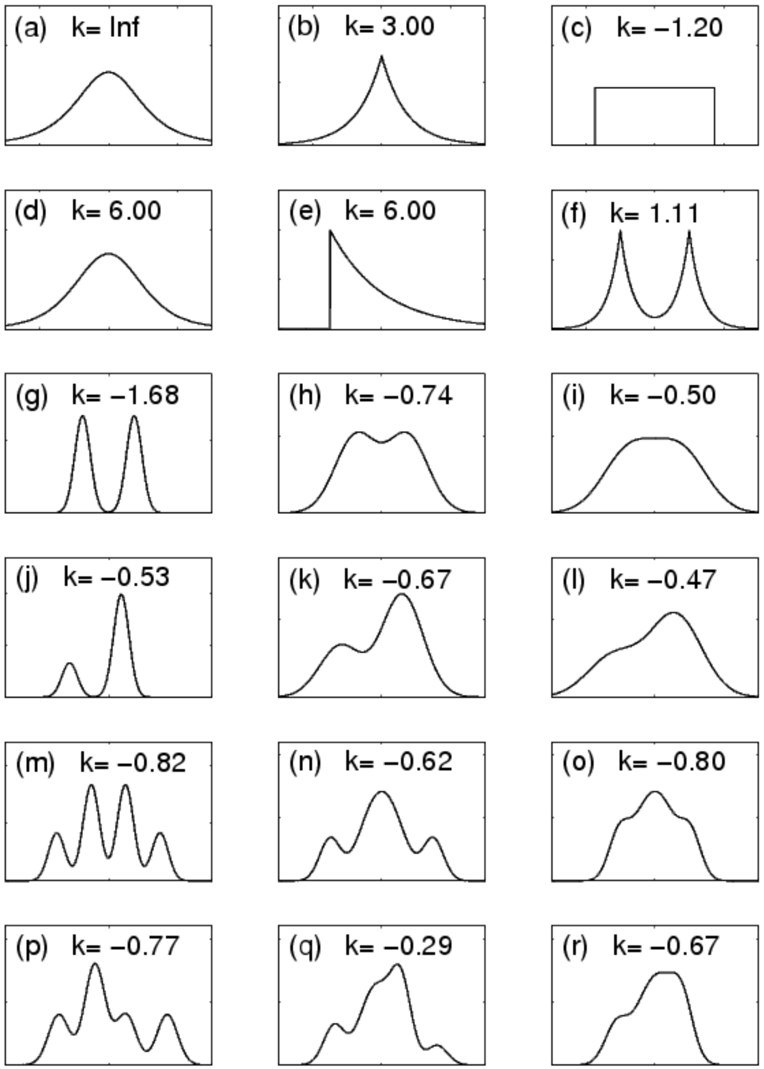
- Greater number of delays: joint diagonalisation

What’s the best method?

A basic benchmark

- $l = 8$ sources
- $m = 40,000$ samples
- Benchmark data from

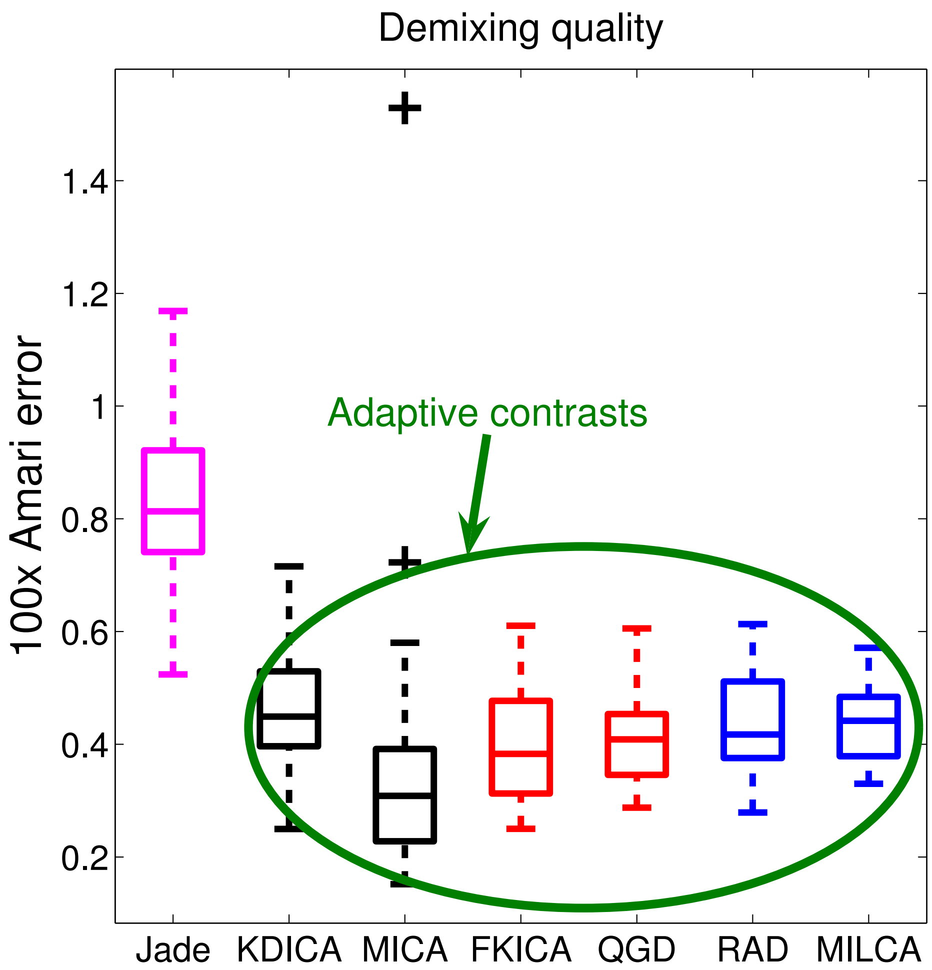
  [Bach and Jordan, 2002]
- Average over 24 repetitions



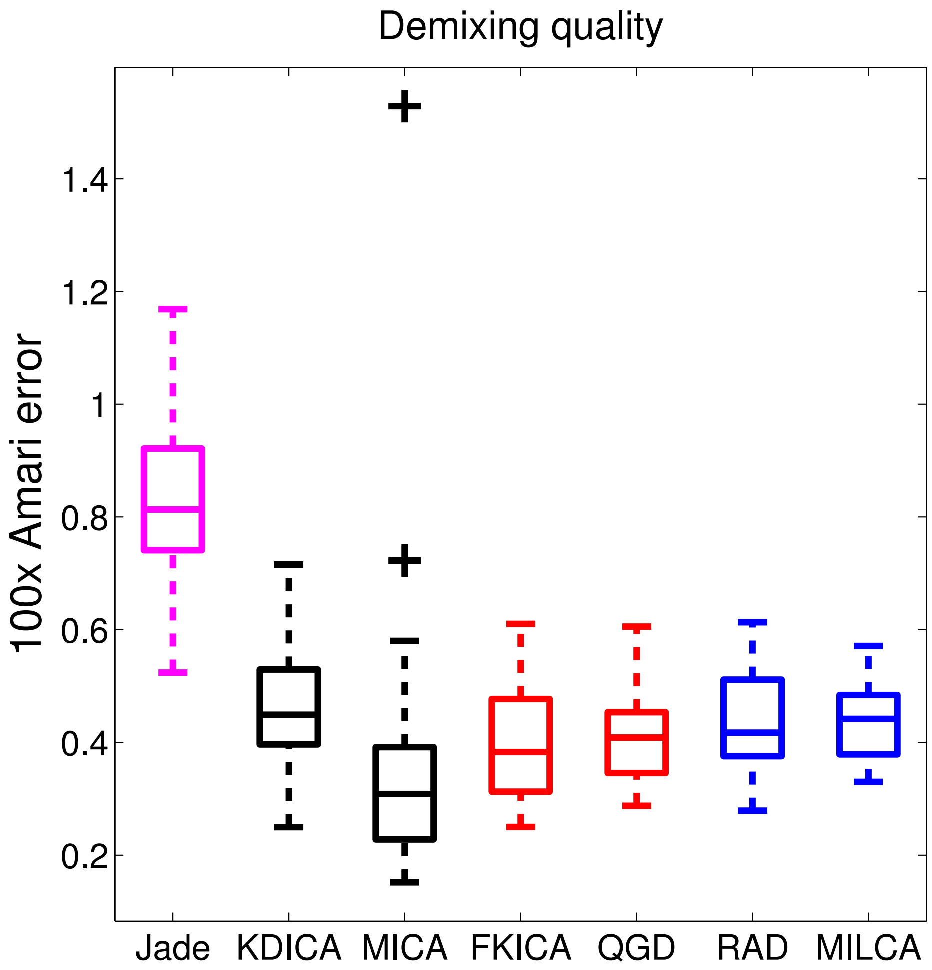
A basic benchmark: results

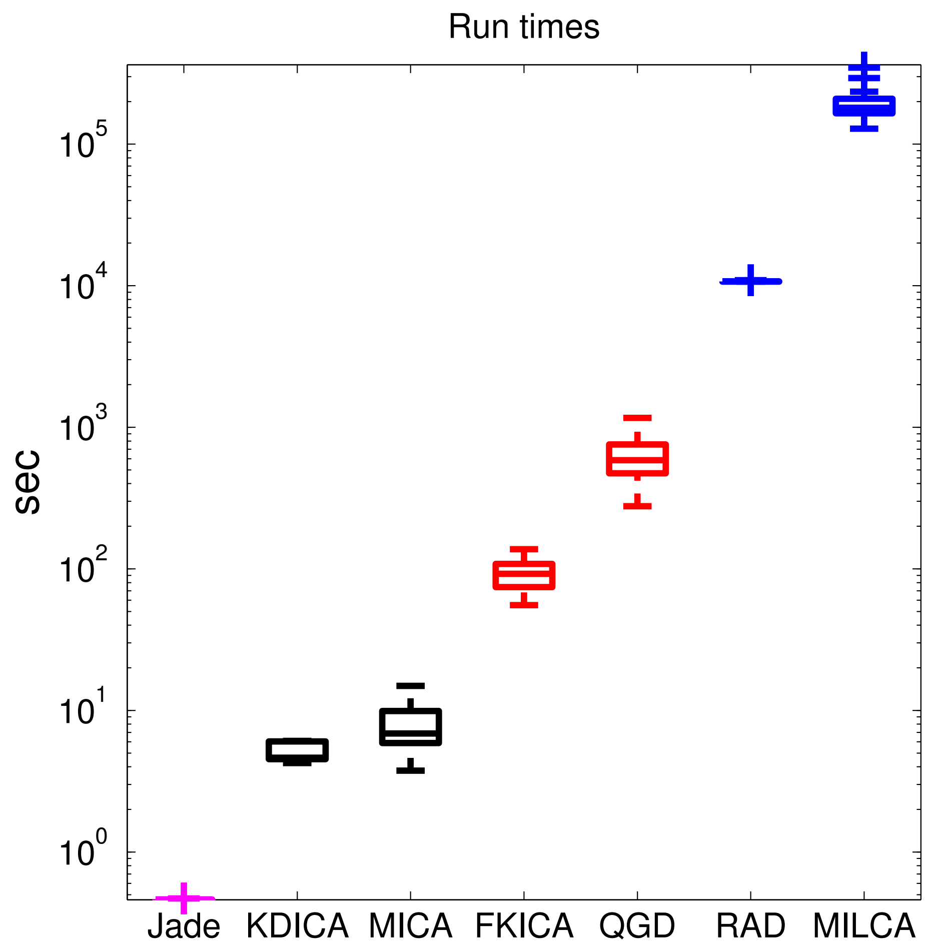
A basic benchmark: results

Adaptive contrasts outperform fixed nonlinearities



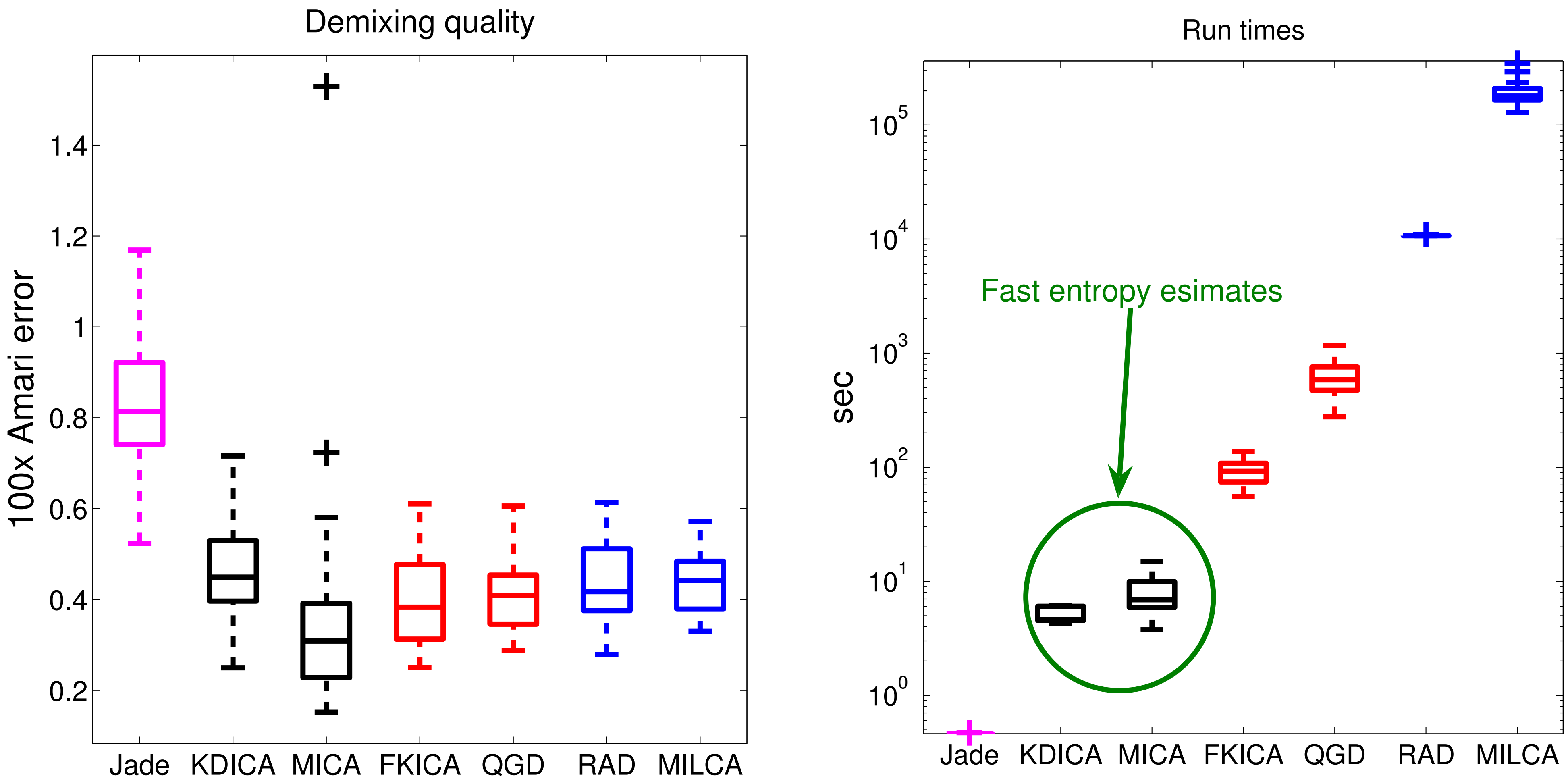
A basic benchmark: computational cost





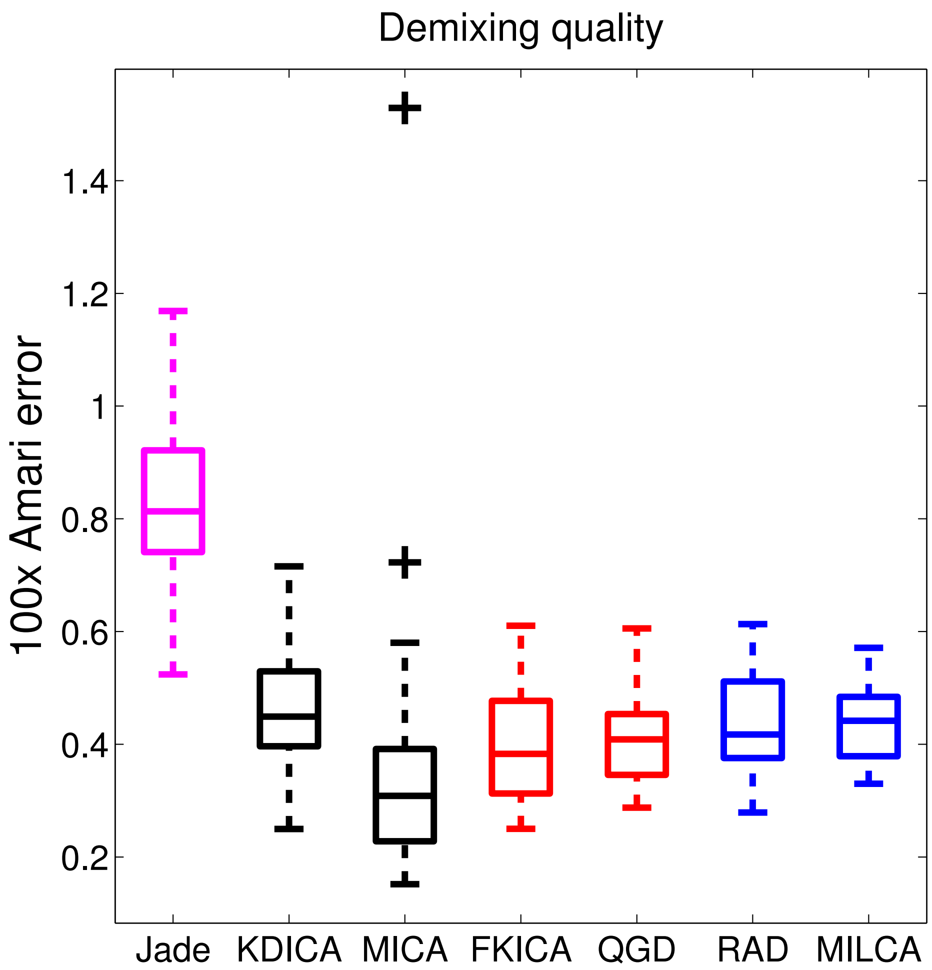
A basic benchmark: computational cost

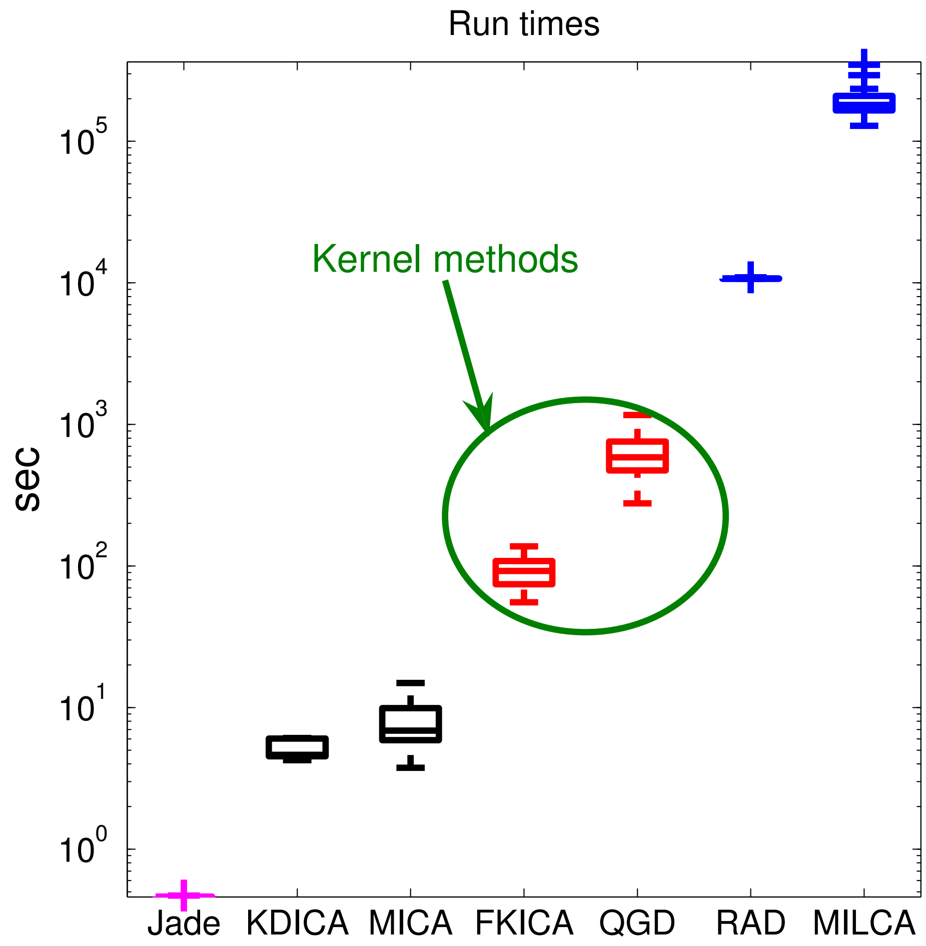
Best runtime (adaptive): fast entropy estimates



A basic benchmark: computational cost

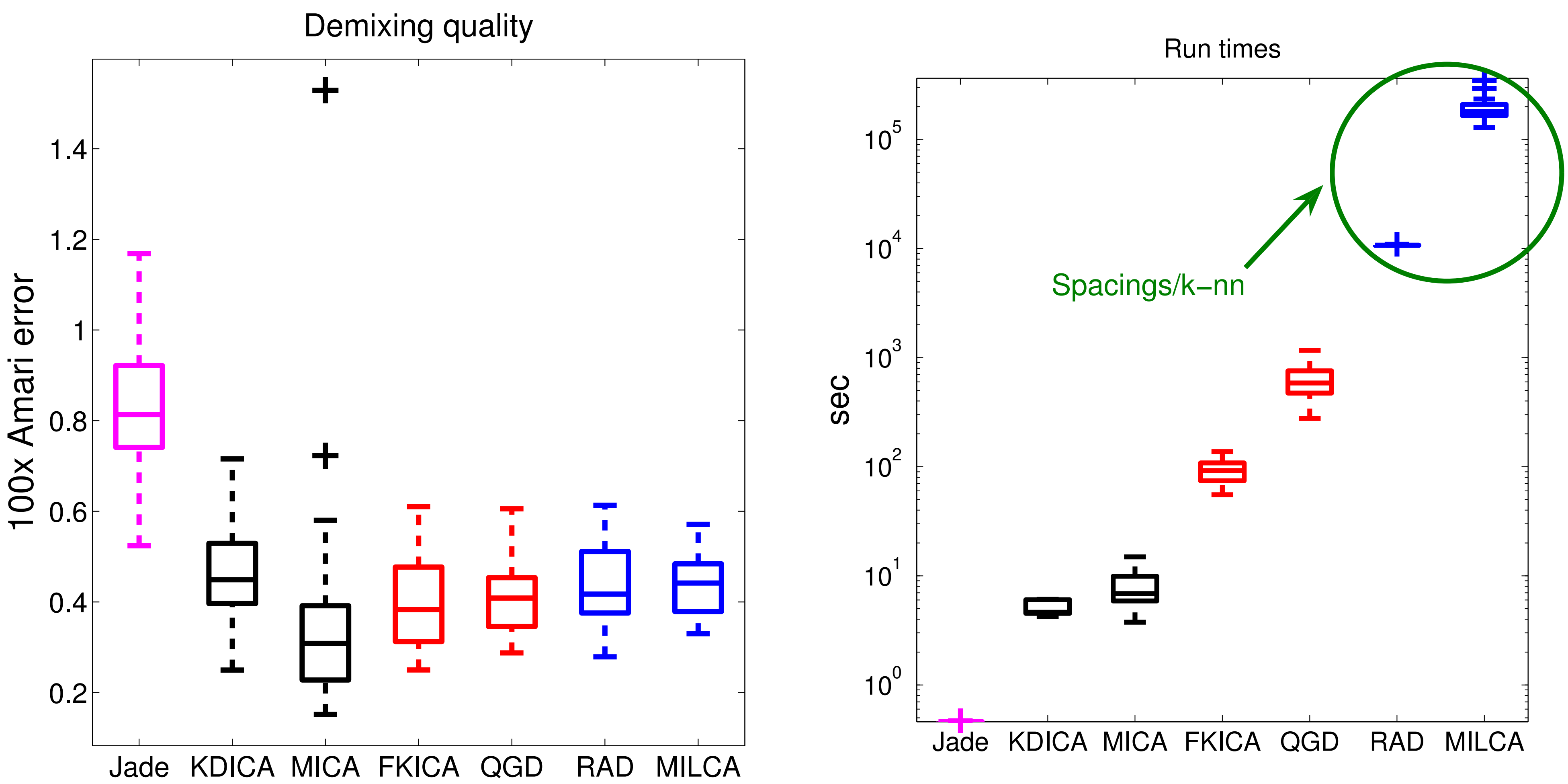
Kernel methods: Newton outperforms Gradient Descent





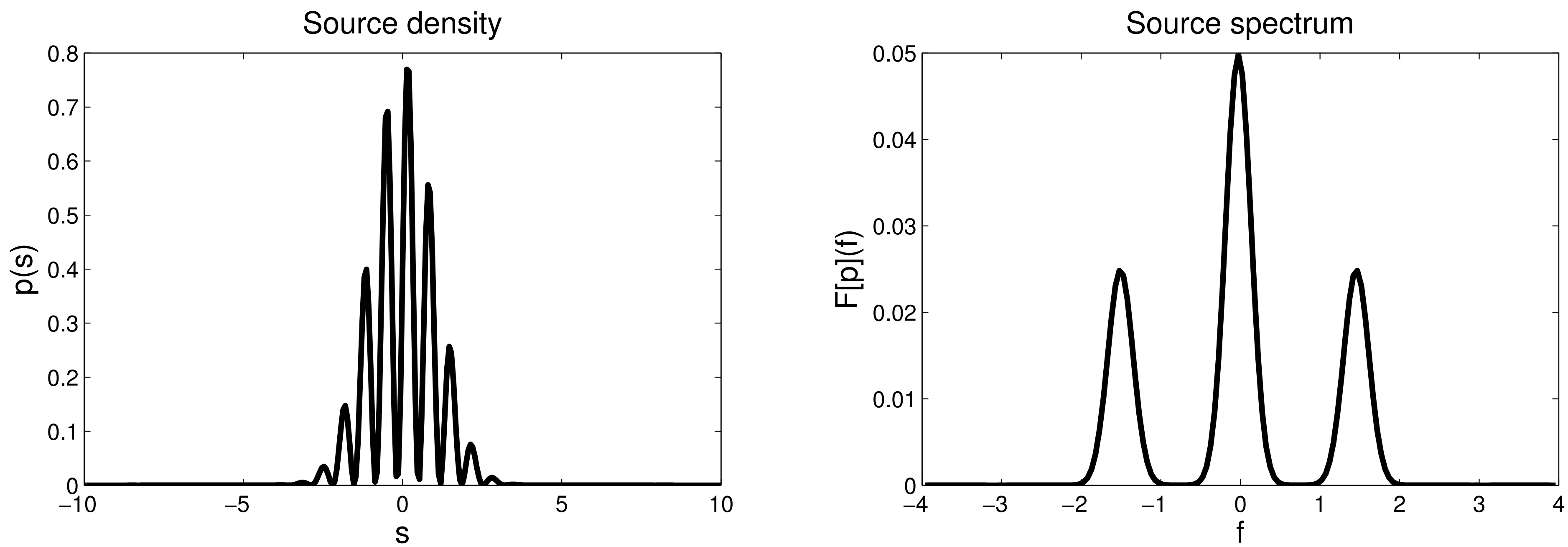
A basic benchmark: computational cost

Spacings/k-nn entropy contrasts slowest

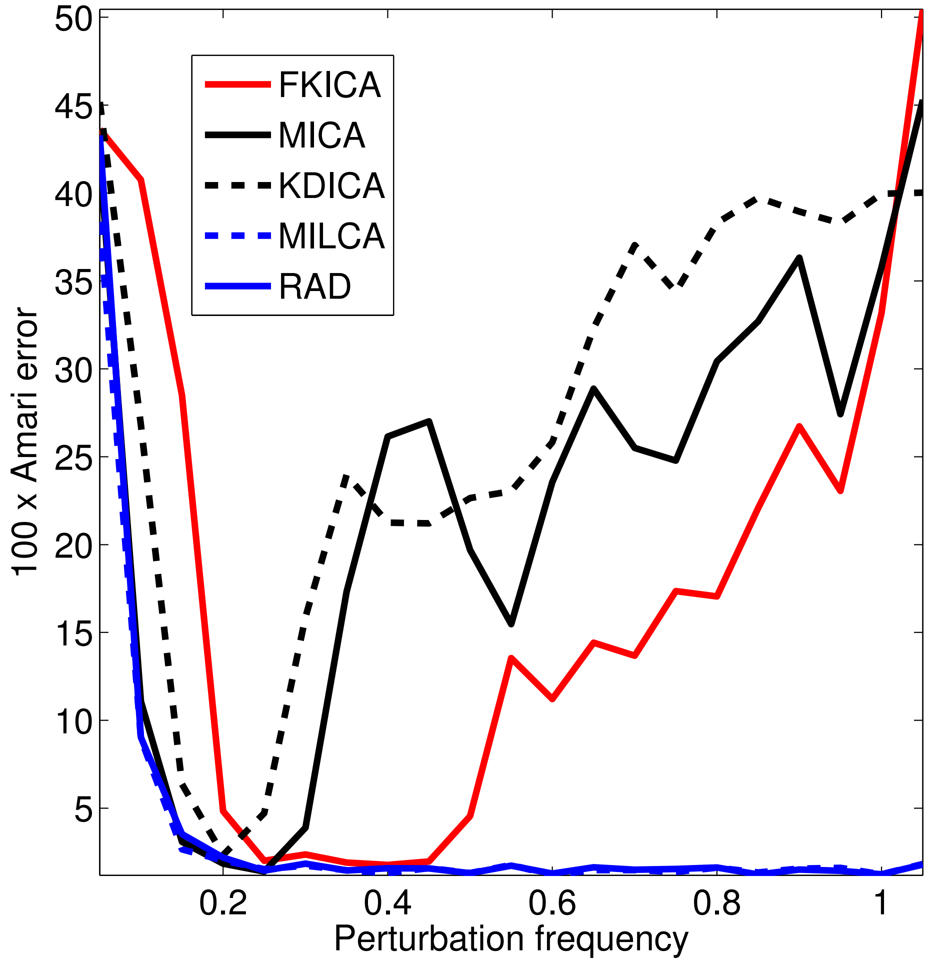


High frequency perturbations

- Two sources, sinusoidal perturbations to Gaussian
- Random mixing angle.
- Results averaged over 25 datasets, $m = 1000$



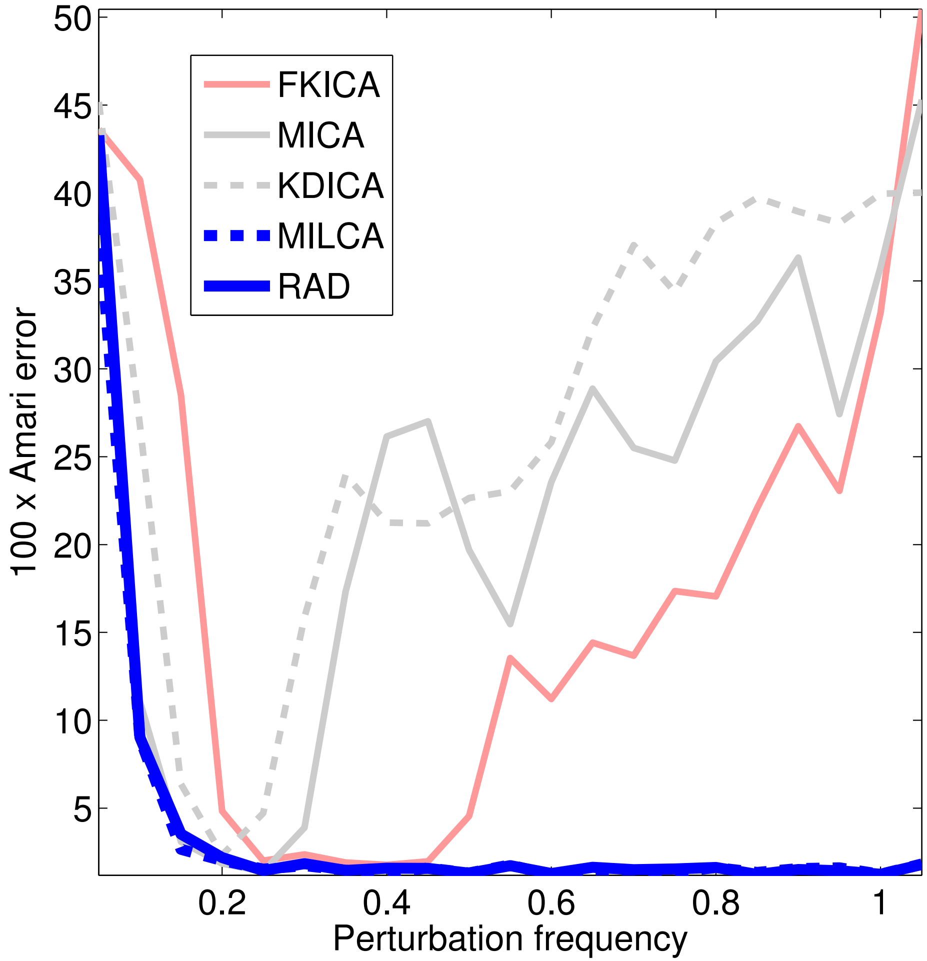
High frequency perturbations



High frequency perturbations

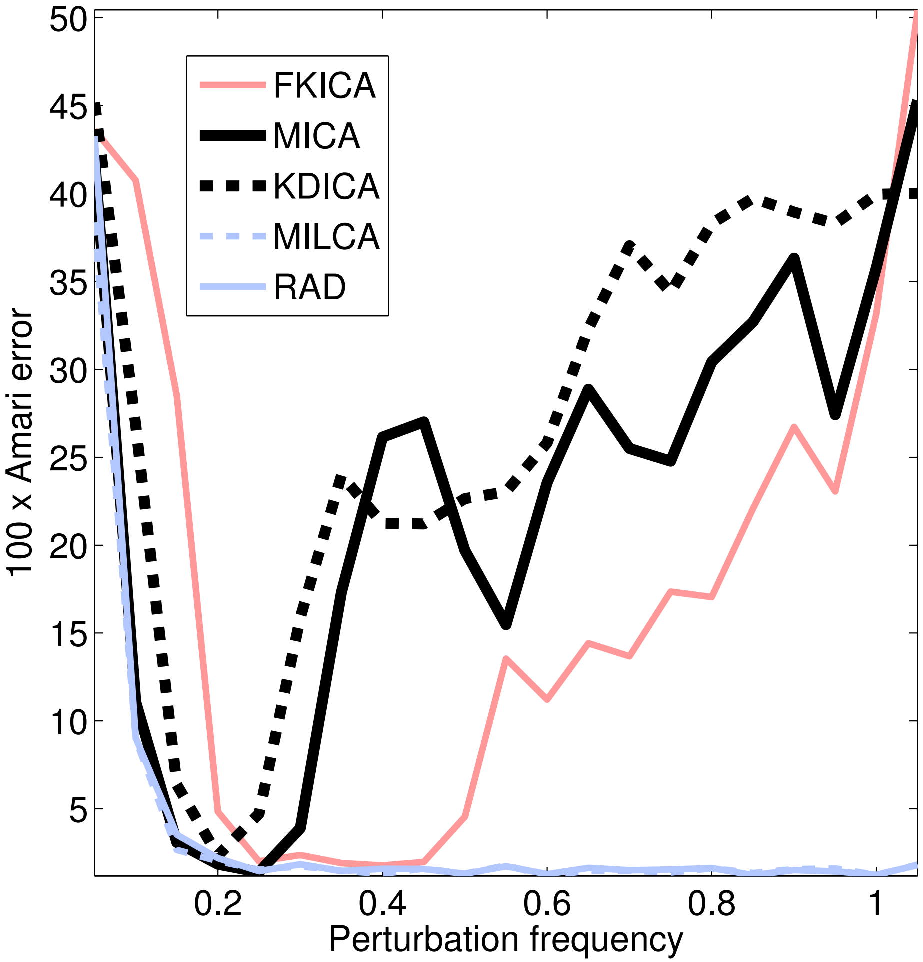
Spacings/k-nn methods perform best

(but slow)



High frequency perturbations

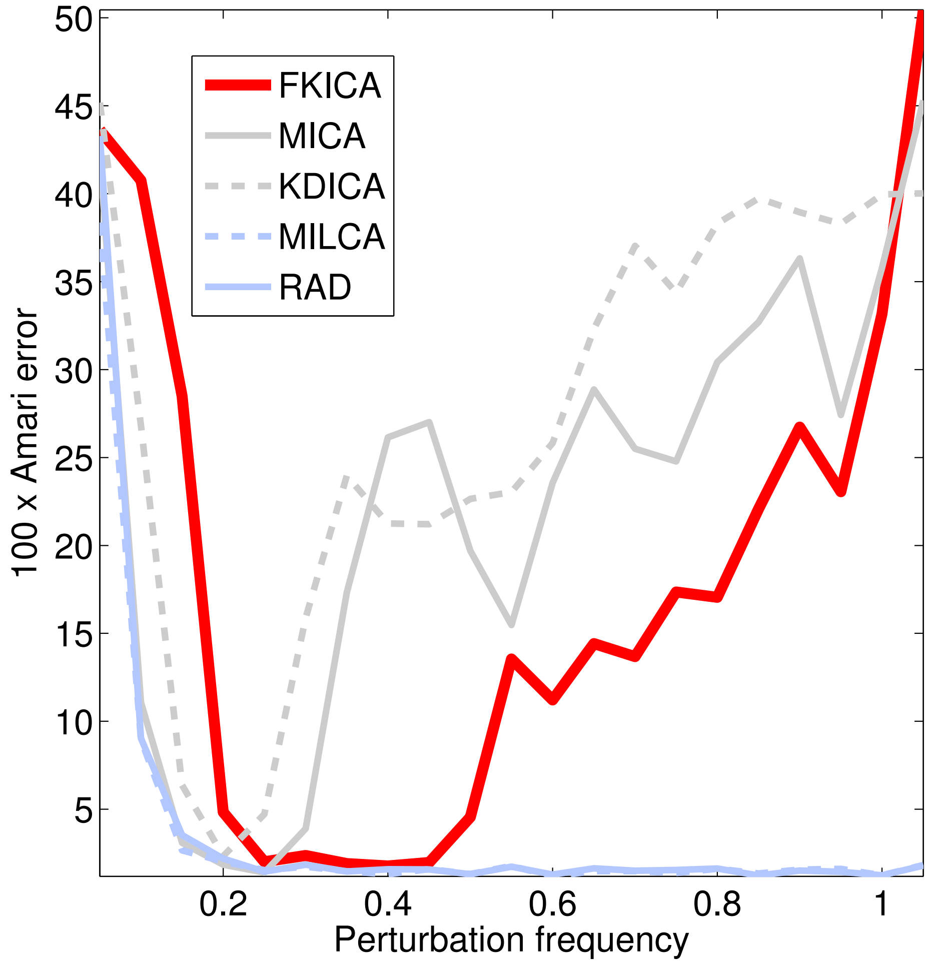
Fast entropy estimates: narrowest range



High frequency perturbations

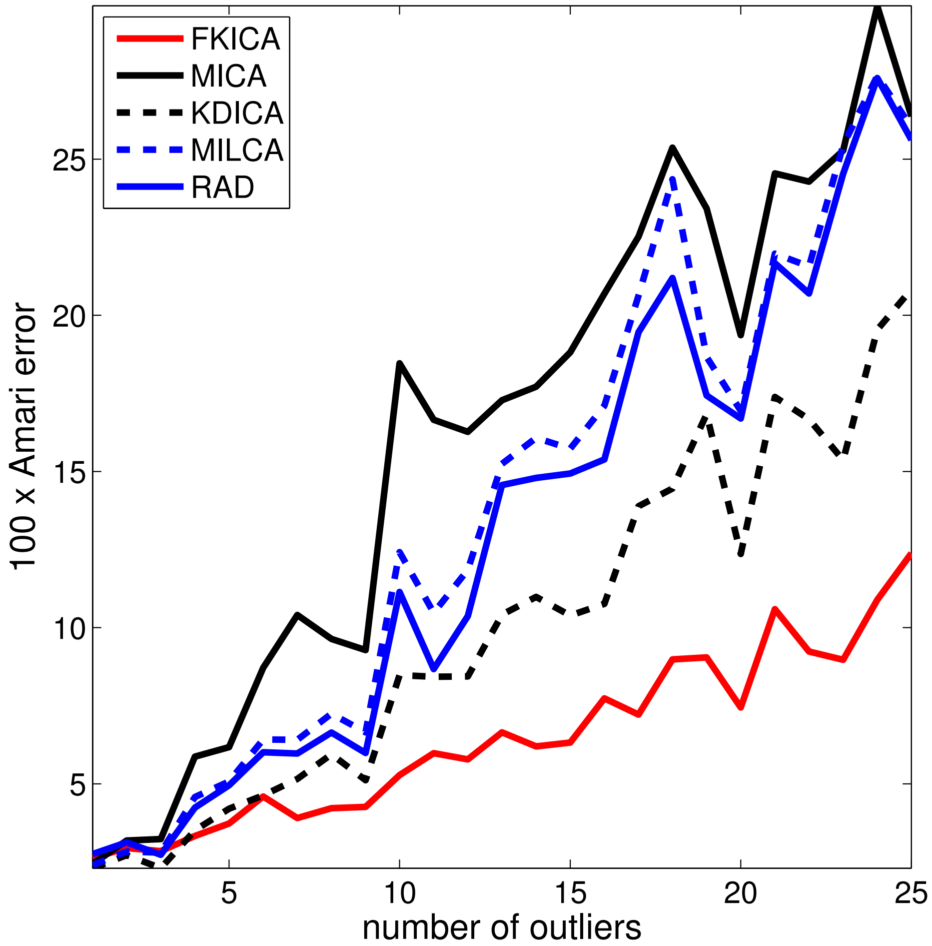
Fast Kernel ICA: peforms in between

(good performance/runtime tradeoff)



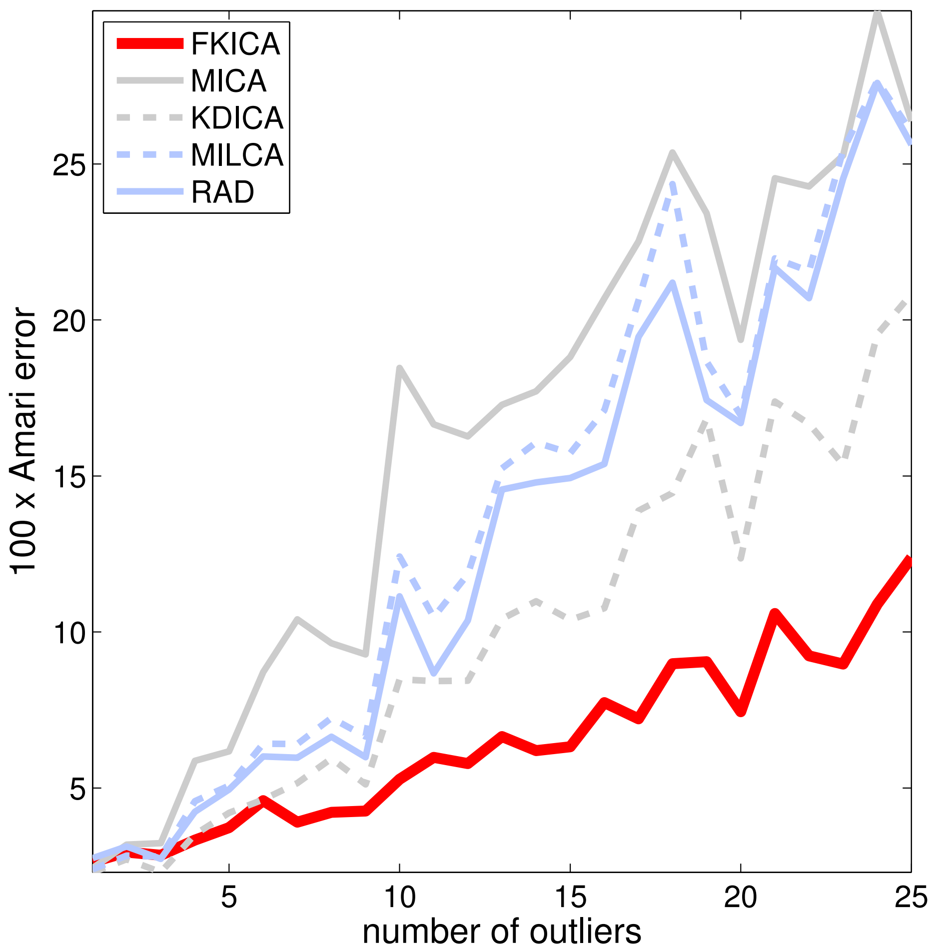
Outlier resistance

Two sources, outliers added to both *mixtures*



Outlier resistance

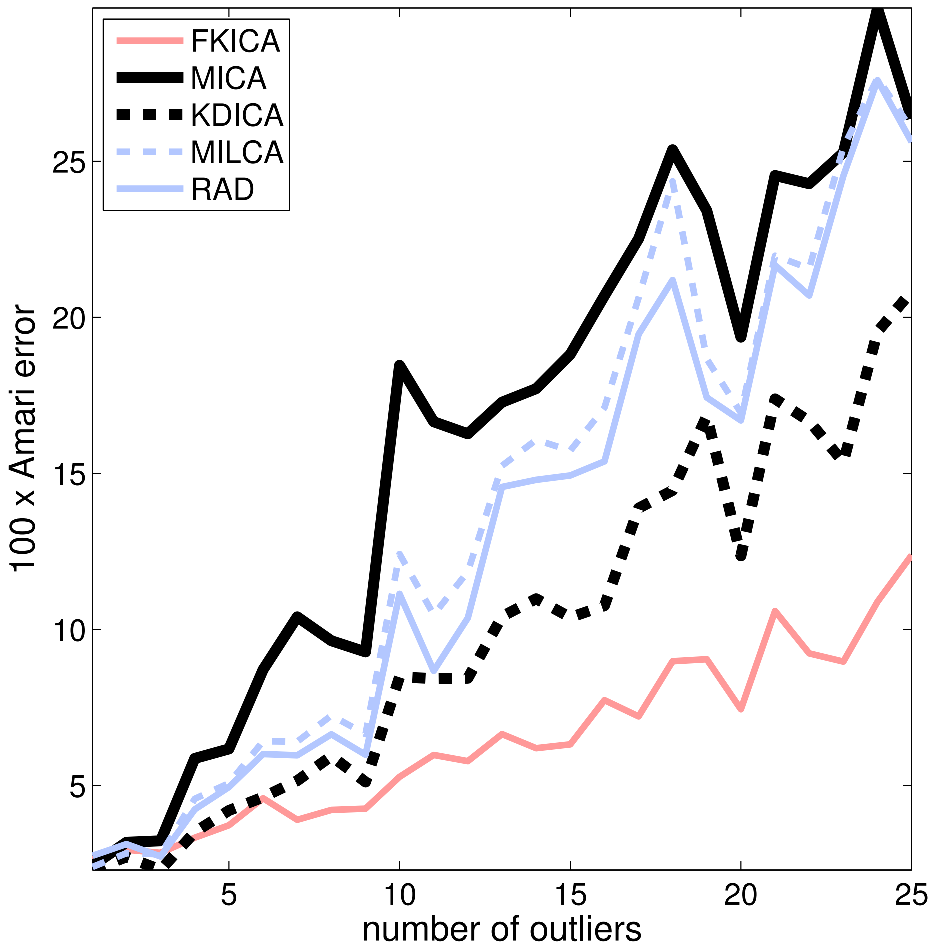
Kernel ICA performs best



Outlier resistance

Fast entropy estimates: less good

KDICA initialized with kernel ICA solution!



ICA algorithm choice

- Choosing kernel ICA approach
  - Fastest (by far): Fast ICA [Hyvärinen et al., 2001], Jade [Cardoso, 1998]
  - Good tradeoff between speed and performance: MICA [Pham, 2004]
  - Tricky cases (outliers, non-smooth sources): Fast KICA [Shen et al., 2007, 2009]
  - Small sample size: KGV very good [Bach and Jordan, 2002]

ICA algorithm choice

- Choosing kernel ICA approach
  - Fastest (by far): Fast ICA [Hyvärinen et al., 2001], Jade [Cardoso, 1998]
  - Good tradeoff between speed and performance: MICA [Pham, 2004]
  - Tricky cases (outliers, non-smooth sources): Fast KICA [Shen et al., 2007, 2009]
  - Small sample size: KGV very good [Bach and Jordan, 2002]
- Some further hints:
  - Use multiple restarts (non-convex)
  - Independence test to check answer

ICA algorithm choice

- Choosing kernel ICA approach
  - Fastest (by far): Fast ICA [Hyvärinen et al., 2001], Jade [Cardoso, 1998]
  - Good tradeoff between speed and performance: MICA [Pham, 2004]
  - Tricky cases (outliers, non-smooth sources): Fast KICA [Shen et al., 2007, 2009]
  - Small sample size: KGV very good [Bach and Jordan, 2002]
- Some further hints:
  - Use multiple restarts (non-convex)
  - Independence test to check answer
- Comparing (usually fixed contrast) algorithms:
  - One approach "better" than another?
  - Example: sources l very large, samples m small (wrt l), e.g. microarray data [Lee and Batzoglou, 2003]

Selected ICA references

- Start with Cardoso's excellent introduction [Cardoso, 1998], and the book by Hyvärninen *et al.* [Hyvärinen et al., 2001]
- Fast kernel ICA is described in [Shen et al., 2007, 2009]. Characteristic function-based ICA is described in [Eriksson and Koivunen, 2003, Chen and Bickel, 2005]. For earlier kernel ICA methods, see [Bach and Jordan, 2002, Gretton et al., 2005]
- Mutual information/entropy based: [Pham, 2004, Learned-Miller and Fisher III, 2003, Stögbauer et al., 2004, Chen, 2006]
- Classic algorithms for *time series* separation with second order methods (not covered much in this talk): [Molgedey and Schuster, 1994, Belouchrani et al., 1997]
- An important paper for optimising over orthogonal matrices: [Edelman et al., 1998]. The Newton-like method: [Hüper and Trumpf, 2004].

Conclusion

- With RKHS distribution embeddings, compare distributions in high dimensions and on structured objects
  - Easier than density estimation

Conclusion

- With RKHS distribution embeddings, compare distributions in high dimensions and on structured objects
  - Easier than density estimation
- It is easy to check whether distribution embeddings are unique
  - Characteristic kernel: check Fourier transform
  - Any difference in distributions detectable

Conclusion

- With RKHS distribution embeddings, compare distributions in high dimensions and on structured objects
  - Easier than density estimation
- It is easy to check whether distribution embeddings are unique
  - Characteristic kernel: check Fourier transform
  - Any difference in distributions detectable
- Can use HSIC dependence measure for feature relevance
  - Feature selection
  - Taxonomy fitting
- More: conditional dependence tests, independent component analysis, covariate shift correction,...

References from my publications

- MMD a distance between distributions [ISMB06, NIPS06a, JMLR10, JMLR12a]
  - high dimensionality
  - non-euclidean data (strings, graphs)
  - Nonparametric hypothesis tests
- Measure and test independence [ALT05, NIPS07a, NIPS07b, ALT08, JMLR10, JMLR12a]
- Characteristic RKHS: MMD a metric [NIPS07b, COLT08, NIPS08a]
  - Easy to check: does spectrum cover $\mathbb{R}^d$
- Applications:
  - Feature selection [ISMB07, ICML07a, JMLR12b]
  - Clustering and taxonomy discovery [ICML07b, NIPS08b]
  - Covariate shift correction [NIPS06b, Book Ch. 08] , testing conditional dependence [NIPS07b] , independent component analysis [JMLR05, Book Ch. 07, AISTATS07, IEEE TSP 09] , . . .

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.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

J. Bedo, C. Sanderson, and A. Kowalczyk. An efficient alternative to SVM based recursive feature elimination with applications in natural language processing and bioinformatics. In *Artificial Intelligence*, 2006.

A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

M. Bethge. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America A*, 23(6):1253–1268, 2006.

V. Calhoun, T. Adali, L. Hansen, J. Larsen, and J. Pekar. Ica of functional mri data: An overview. In *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 281–288, 2003.

J.-F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation*, 86(10):2009–2025, 1998.

A. Chen. Fast kernel density independent component analysis. In *Sixth International Conference on ICA and BSS*, volume 3889, pages 24–31, Berlin/Heidelberg, 2006. Springer-Verlag.

A. Chen and P. J. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

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

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA*, 103(15):5923–5928, Apr 2006.

J. Eriksson and K. Koivunen. Characteristic-function based independent component analysis. *Signal Processing*, 83(10):2195–2208, 2003.

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

A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

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

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

K. Hüper and J. Trumpf. Newton-like methods for numerical optimisation on manifolds. In *Proceedings of Thirty-eighth Asilomar Conference on Signals, Systems and Computers*, pages 136–139, 2004.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

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

T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. McKeown, V. Iragui, and T. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37:163–178, 2000.

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

K. Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In *In Proc. Int. Conf. on Neural Information Processing (ICONIP)*, volume 2, pages 895–898, 1998.

E. G. Learned-Miller and J. W. Fisher III. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

S.I. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biology*, 4(11):R76, 2003.

F. Li and Y. Yang. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 21(19):3741–3747, Oct 2005.

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

L. Molgedey and H. Schuster. Separation of a mixture of independent signals using time delayed correlation. *Physical Review Letters*, 72(23):3634–3637, 1994.

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

D.-T. Pham. Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 52(10):2690–2700, 2004.

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

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

H. Shen and K. Hüper. Newton-like methods for parallel independent component analysis. In *MLSP 16*, pages 283–288, Maynooth, Ireland, 2006.

H. Shen, S. Jegelka, and A. Gretton. Fast kernel ICA using an approximate Newton method. In *AISTATS 11*, pages 476–483. Microtome, 2007.

H. Shen, S. Jegelka, and A. Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57:3498 – 3511, 2009.

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

H. Stögbauer, A. Kraskov, S. Astakhov, and P. Grassberger. Least dependent component analysis based on mutual information. *Phys. Rev. E*, 70(6): 066123, 2004.

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

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

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. In *National Academy of Sciences*, volume 99, pages 6567–6572, 2002.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applicaitons to dna microarrays. *Stat Sci*, 18:104–117, 2003.

L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.