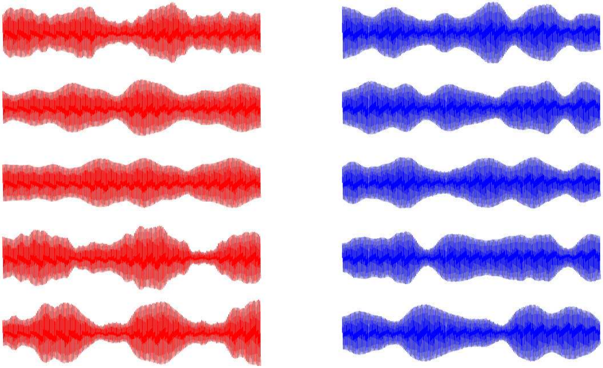
Hypothesis Testing with Kernels

Zoltán Szabó (Gatsby Unit, UCL)

PRNI, Trento
June 22, 2016

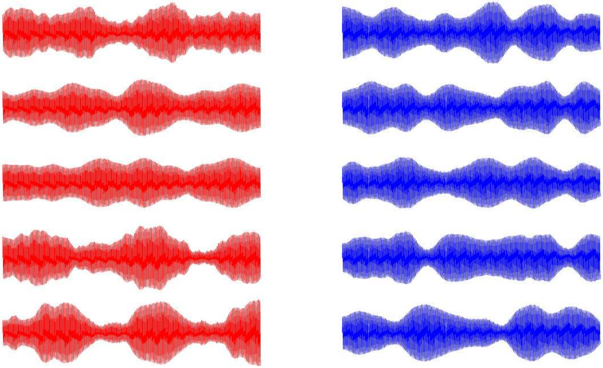
Motivation: detecting differences in AM signals

- Amplitude modulation:
  - simple technique to transmit voice over radio.
  - in the example: 2 songs.
- Fragments from $\text{song}_1 \sim \mathbb{P}_x$, $\text{song}_2 \sim \mathbb{P}_y$.



Motivation: detecting differences in AM signals

- Amplitude modulation:
  - simple technique to transmit voice over radio.
  - in the example: 2 songs.
- Fragments from song$_1 \sim \mathbb{P}_x$, song$_2 \sim \mathbb{P}_y$.



Question: $\mathbb{P}_x = \mathbb{P}_y$?

Motivation: discrete domain - 2-sample testing

- How do we compare distributions?
- Given: 2 sets of text fragments (fisheries, agriculture).

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, . . .

. . .

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.

. . .

Motivation: discrete domain - 2-sample testing

- How do we compare distributions?
- Given: 2 sets of text fragments (fisheries, agriculture).

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, . . .

. . .

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.

. . .

Do $\{x_i\}$ and $\{y_j\}$ come from the same distribution, i.e. $\mathbb{P}_x = \mathbb{P}_y$?

Motivation: discrete domain - independence testing

- How do we detect dependency? (paired samples)

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.

...

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.

...

Motivation: discrete domain - independence testing

- How do we detect dependency? (paired samples)

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.

...

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 paragraphs translations of the English ones, or have nothing to do with it, i.e. $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$?

Outline

1. RKHS based metric on probability distributions.
2. 2-sample testing:
    - Nonparametric.
    - Distance between distribution representations.

Outline

1. RKHS based metric on probability distributions.
2. 2-sample testing:
   - Nonparametric.
   - Distance between distribution representations.
3. Independence testing:
   - Dependency detection.
   - Distance between joint ($\mathbb{P}_{XY}$) and product of marginals ($\mathbb{P}_X \mathbb{P}_Y$).

Kernels

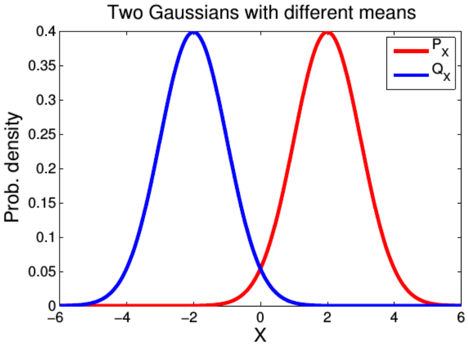
Kernels exist on essentially any data type:

- images, texts, graphs, time series, dynamical systems, ...

$\Rightarrow$ distribution representation, hypothesis testing: on all these domains.

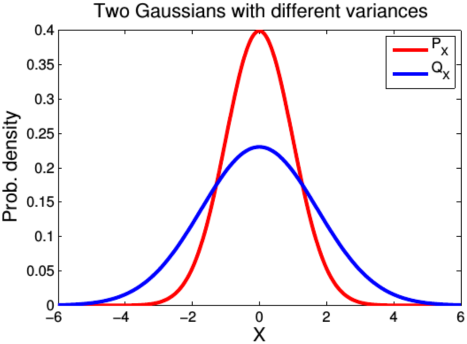
Towards representations of distributions: $\mathbb{E}X$

- Given: 2 Gaussians with different means.
- Solution: *t*-test.



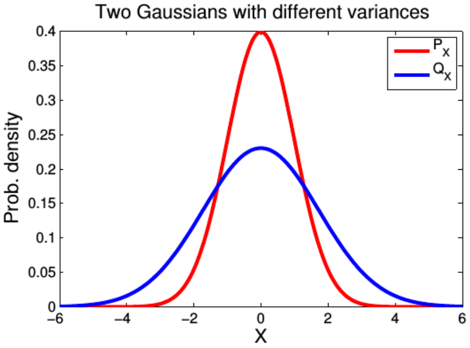
Towards representations of distributions: $\mathbb{E}X^2$

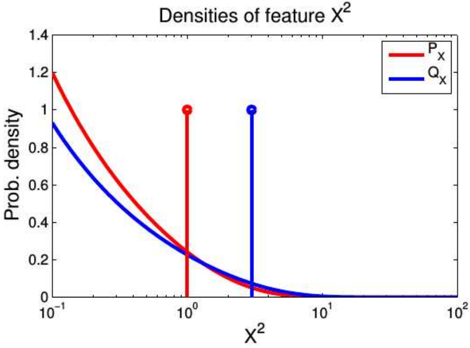
- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.



Towards representations of distributions: $\mathbb{E}X^2$

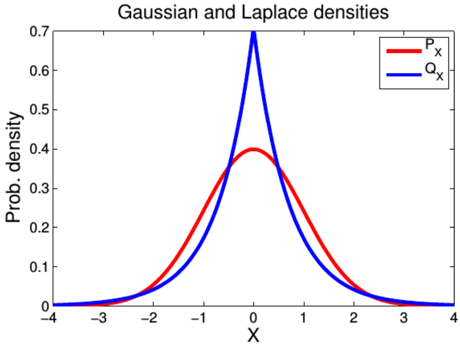
- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.
- $\varphi_x = x^2 \Rightarrow$ difference in $\mathbb{E}X^2$.





Towards representations of distributions: further moments

- Setup: a Gaussian and a Laplacian distribution.
- Challenge: their means *and* variances are the same.
- Idea: look at higher-order features.



Let us consider feature representations!

Kernel: similarity between features

- Given: x and $x' \in \mathcal{X}$ objects (images, texts, ...).

Kernel: similarity between features

- Given: x and $x' \in \mathcal{X}$ objects (images, texts, $\ldots$).
- Question: how similar they are?

Kernel: similarity between features

- Given: x and $x' \in \mathcal{X}$ objects (images, texts, ...).
- Question: how similar they are?
- Define features of the objects:

$$\begin{aligned} \varphi_x &: \text{features of } x, \\ \varphi_{x'} &: \text{features of } x'. \end{aligned}$$

- Kernel: inner product of these features

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

Kernel examples

- $\mathcal{X} = \mathbb{R}^d$:

$$k_p(x,y) = (\langle x,y\rangle + \gamma)^p, \qquad k_G(x,y) = e^{-\gamma\|x-y\|_2^2},$$

$$k_e(x,y) = e^{-\gamma\|x-y\|_2}, \qquad k_C(x,y) = 1 + \frac{1}{\gamma\|x-y\|_2^2}.$$

Kernel examples

- $\mathcal{X} = \mathbb{R}^d$:

$$k_p(x,y) = (\langle x,y\rangle + \gamma)^p, \qquad k_G(x,y) = e^{-\gamma\|x-y\|_2^2},$$
$$k_e(x,y) = e^{-\gamma\|x-y\|_2}, \qquad k_C(x,y) = 1 + \frac{1}{\gamma\left\|x-y\right\|_2^2}.$$

- $\mathcal{X}$ = texts, strings:
  - bag-of-word kernel,
  - r-spectrum kernel: # of common $\leqslant r$-substrings.

Kernel examples

- $\mathcal{X} = \mathbb{R}^d$:

$$k_p(x, y) = (\langle x, y \rangle + \gamma)^p, \qquad k_G(x, y) = e^{-\gamma \|x-y\|_2^2},$$

$$k_e(x, y) = e^{-\gamma \|x-y\|_2}, \qquad k_C(x, y) = 1 + \frac{1}{\gamma \|x-y\|_2^2}.$$

- $\mathcal{X}$ = texts, strings:
  - bag-of-word kernel,
  - r-spectrum kernel: # of common $\leqslant r$-substrings.
- $\mathcal{X}$ = time-series: dynamic time-warping.

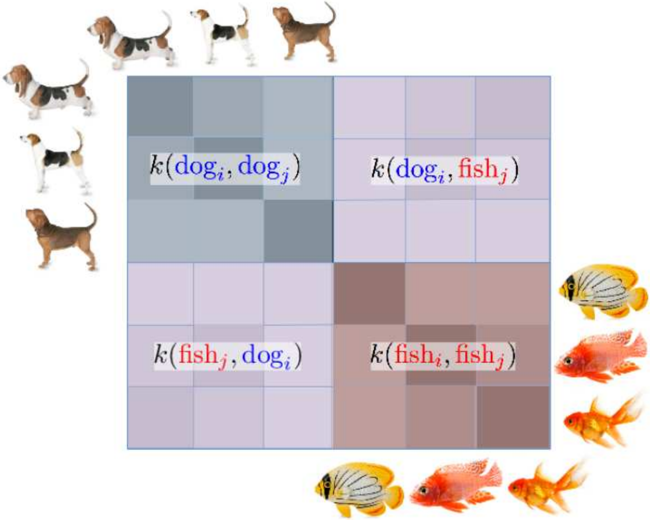
Two-sample testing

Ingredient: maximum mean discrepancy

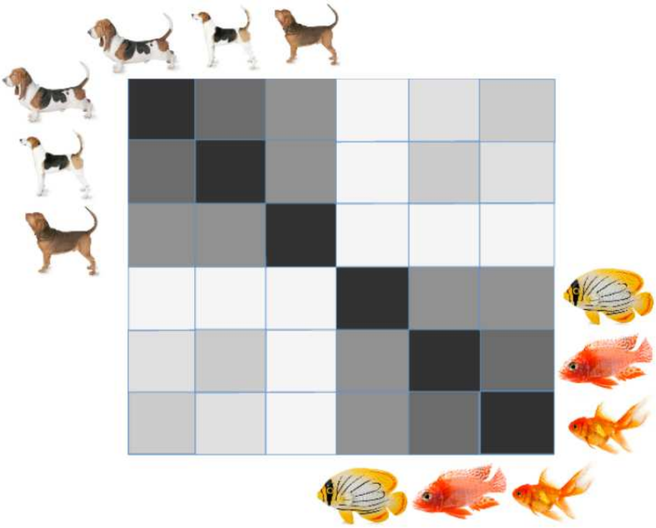
$\sim P$

$\sim Q$

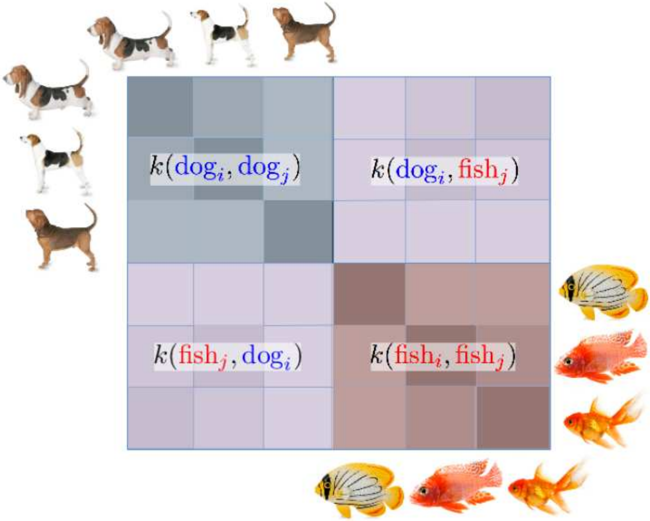
Ingredient: maximum mean discrepancy



Ingredient: maximum mean discrepancy



Ingredient: maximum mean discrepancy



$$\widehat{MMD}^2(\mathbb{P}, \mathbb{Q}) = \overline{K_{\mathbb{P},\mathbb{P}}} + \overline{K_{\mathbb{Q},\mathbb{Q}}} - 2\overline{K_{\mathbb{P},\mathbb{Q}}} \quad \text{(without diagonals in } \overline{K_{\mathbb{P},\mathbb{P}}}, \overline{K_{\mathbb{Q},\mathbb{Q}}}\text{)}$$

From kernel trick to mean trick

- Recall:
  - $\varphi_x \in \mathcal{H}$: feature of $x \in \mathcal{X}$.
  - Kernel: $k(x, x') = \langle \varphi_x, \varphi_{x'} \rangle_{\mathcal{H}}$.

From kernel trick to mean trick

- Recall:
  - $\varphi_x \in \mathcal{H}$: feature of $x \in \mathcal{X}$.
  - Kernel: $k(x, x') = \langle \varphi_x, \varphi_{x'} \rangle_{\mathcal{H}}$.
- Mean embedding:
  - Feature of $\mathbb{P}$: $\mu_{\mathbb{P}} := \mathbb{E}_{x\sim\mathbb{P}}[\varphi_x] \in \mathcal{H}(k)$.
  - Inner product: $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \mathbb{E}_{x\sim\mathbb{P}, x'\sim\mathbb{Q}} k(x, x')$.

From kernel trick to mean trick

- Recall:
  - $\varphi_x \in \mathcal{H}$: feature of $x \in \mathcal{X}$.
  - Kernel: $k(x, x') = \langle \varphi_x, \varphi_{x'} \rangle_{\mathcal{H}}$.
- Mean embedding:
  - Feature of $\mathbb{P}$: $\mu_{\mathbb{P}} := \mathbb{E}_{x \sim \mathbb{P}}[\varphi_x] \in \mathcal{H}(k)$.
  - Inner product: $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \mathbb{E}_{x \sim \mathbb{P}, x' \sim \mathbb{Q}} k(x, x')$.
- $\mu_{\mathbb{P}}$: well-defined for all distributions (bounded k).

Maximum mean discrepancy

Squared difference between feature means:

$$
\begin{aligned}
MMD^2(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} = \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2 \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \mathbb{E}_{\mathbb{P},\mathbb{P}} k(x, x') + \mathbb{E}_{\mathbb{Q},\mathbb{Q}} k(y, y') - 2\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(x, y).
\end{aligned}
$$

Maximum mean discrepancy

Squared difference between feature means:

$$\begin{aligned} MMD^2(\mathbb{P},\mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2 = \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2 \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \mathbb{E}_{\mathbb{P},\mathbb{P}} k(x,x') + \mathbb{E}_{\mathbb{Q},\mathbb{Q}} k(y,y') - 2\mathbb{E}_{\mathbb{P},\mathbb{Q}} k(x,y). \end{aligned}$$

Unbiased empirical estimate for $\{x_i\}_{i=1}^n \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$:

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

Two-sample test using MMD

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

Two-sample test using MMD

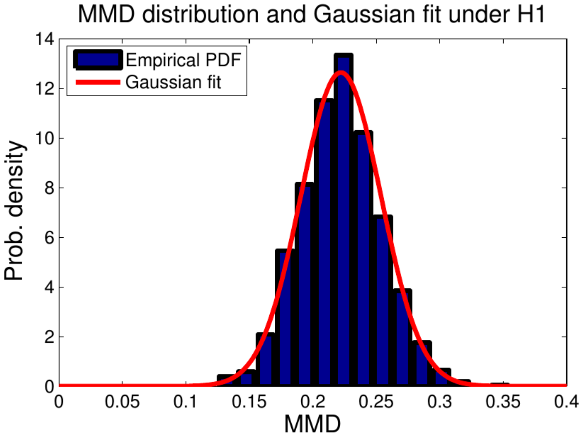
- Two hypotheses:
  - H_0 (null hypothesis): $\mathbb{P} = \mathbb{Q}$.
  - H_1 (alternative hypothesis): $\mathbb{P} \neq \mathbb{Q}$.
- Observation: $\{x_i\}_{i=1}^n \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$.
- Decision: if $\widehat{MMD}^2(\mathbb{P}, \mathbb{Q})$ is 'far from 0' $\Rightarrow$ reject H_0.

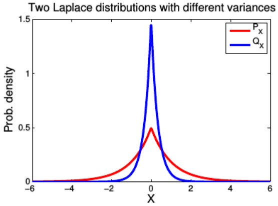
Two-sample test using MMD

- Two hypotheses:
  - H_0 (null hypothesis): $\mathbb{P} = \mathbb{Q}$.
  - H_1 (alternative hypothesis): $\mathbb{P} \neq \mathbb{Q}$.
- Observation: $\{x_i\}_{i=1}^n \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$.
- Decision: if $\widehat{MMD}^2(\mathbb{P}, \mathbb{Q})$ is 'far from 0' $\Rightarrow$ reject H_0.
- Threshold $=$? $\xrightarrow{\text{one answer}}$ asymptotic distribution of $\widehat{MMD}^2$.

Two-sample test using MMD: H_1

Under H_1 ($\mathbb{P} \neq \mathbb{Q}$): asymptotic distribution of $\widehat{MMD}^2$ is Gaussian.





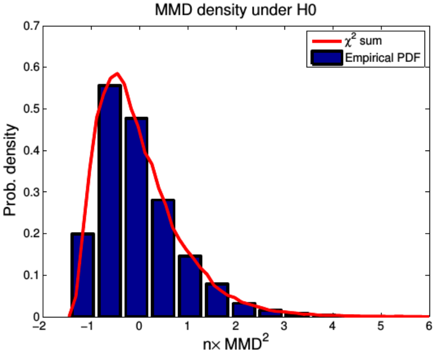
Two-sample test using MMD: H_0

Under H_0 ($\mathbb{P} = \mathbb{Q}$): asymptotic distribution is

$$n\widehat{MMD}^2(\mathbb{P}, \mathbb{P}) \sim \sum_{i=1}^{\infty} \lambda_i (z_i^2 - 2),$$

where $z_i \sim N(0, 2)$ i.i.d.,

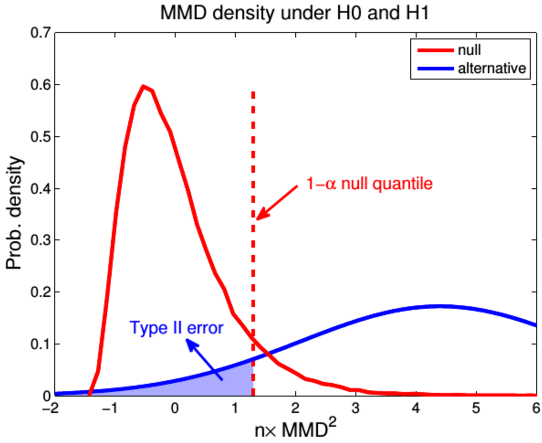
$$\int_{\mathcal{X}} \tilde{k}(x, x') v_i(x) \mathrm{d}\mathbb{P}(x) = \lambda_i v_i(x'), \; \tilde{k}(x, x') = \langle \varphi_x - \mu_{\mathbb{P}}, \varphi_{x'} - \mu_{\mathbb{P}} \rangle_{\mathcal{H}}.$$



Two-sample test using MMD: threshold

To the decision:

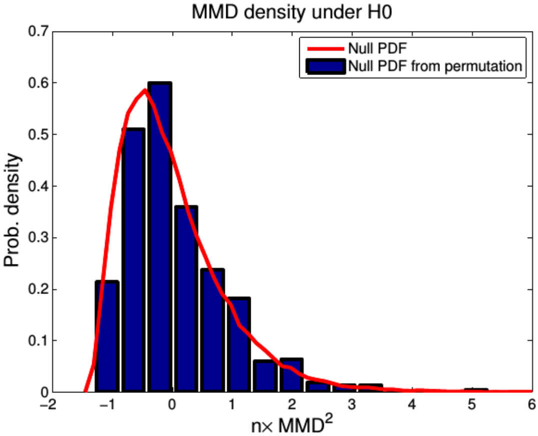
- given that $\mathbb{P} = \mathbb{Q}$,
- we want threshold T such that $\mathbb{P}(n\widehat{MMD}^2 > T) \leqslant 0.05 =: \alpha$.



Two-sample test using MMD: threshold

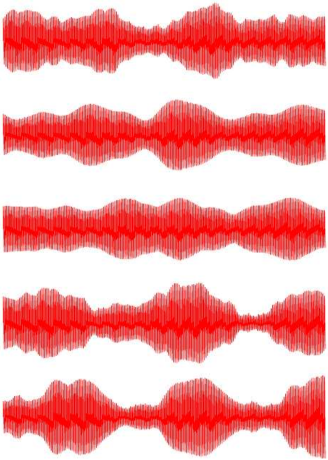
Task: $\mathbb{P}\left(n\widehat{MMD}^2 > T\right) \leqslant \alpha$. Solutions:

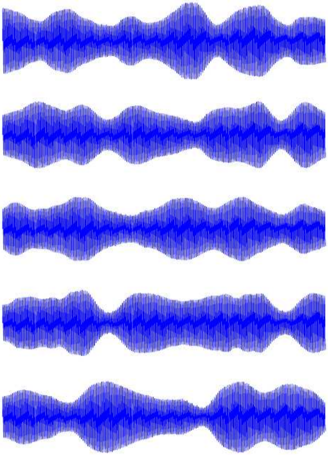
- permutation test: below,
- kernel eigenspectrum estimate: $\hat{\lambda}_i$.
- moment matching: Gamma approximation.



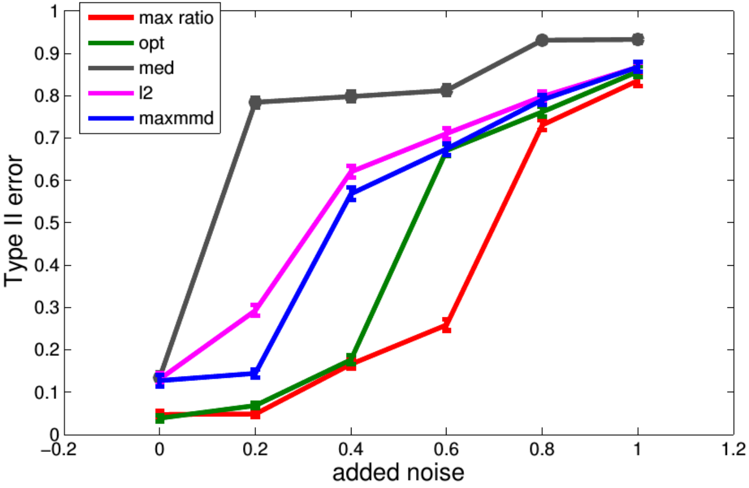
Demo: amplitude modulated signals

Question: $\mathbb{P}_x = \mathbb{P}_y$?





$n = 10,000$. Average over 4124 trials. Gaussian noise: added.



Independence testing

Independence testing

- Given:
  - 2 kernel-endowed domain: $(\mathcal{X}, k)$, $(\mathcal{Y}, \ell)$,
  - paired samples: $\{(x_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$.
- Hypotheses: $H_0 : \mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$, $H_1 : \mathbb{P}_{XY} \neq \mathbb{P}_X \mathbb{P}_Y$.

Independence testing

- Given:
  - 2 kernel-endowed domain: $(\mathcal{X}, k)$, $(\mathcal{Y}, \ell)$,
  - paired samples: $\{(x_i, y_i)\}_{i=1}^n \sim \mathbb{P}_{XY}$.
- Hypotheses: $H_0 : \mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$, $H_1 : \mathbb{P}_{XY} \neq \mathbb{P}_X\mathbb{P}_Y$.
- Statistics:

$$HSIC = MMD^2(\mathbb{P}_{XY}, \mathbb{P}_X\mathbb{P}_Y) = \left\|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}_X\mathbb{P}_Y}\right\|^2_{\mathcal{H}(\check{k})},$$

$$\check{k}((x,y),(x',y')) = k(x,x')\ell(y,y').$$

HSIC in terms of expectations

$$
\begin{aligned}
HSIC(\mathbb{P}_{XY}, \mathbb{P}_X \mathbb{P}_Y) &= \mathbb{E}_{xy}\mathbb{E}_{x'y'}[k(x,x')\ell(y,y')] \\
&+ \mathbb{E}_x\mathbb{E}_{x'}[k(x,x')]\mathbb{E}_y\mathbb{E}_{y'}[\ell(y,y')] \\
&- 2\mathbb{E}_{x'y'}\left[\mathbb{E}_x k(x,x')\mathbb{E}_y \ell(y,y')\right]
\end{aligned}
$$

Let us consider an example!

HSIC: intuition. $\mathcal{X}$: images, $\mathcal{Y}$: descriptions.

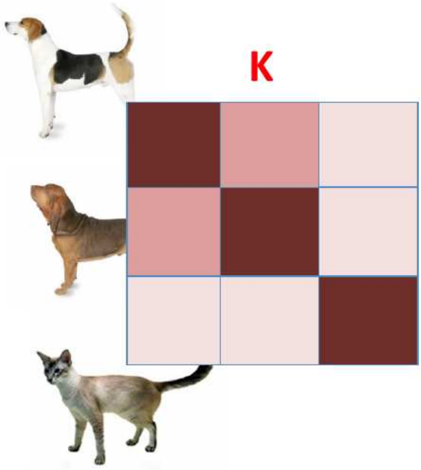
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

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.

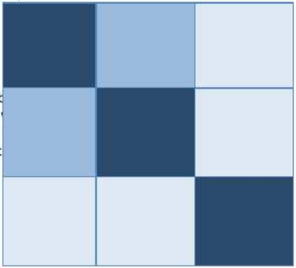
Text from dogtime.com and petfinder.com

HSIC intuition: Gram matrices



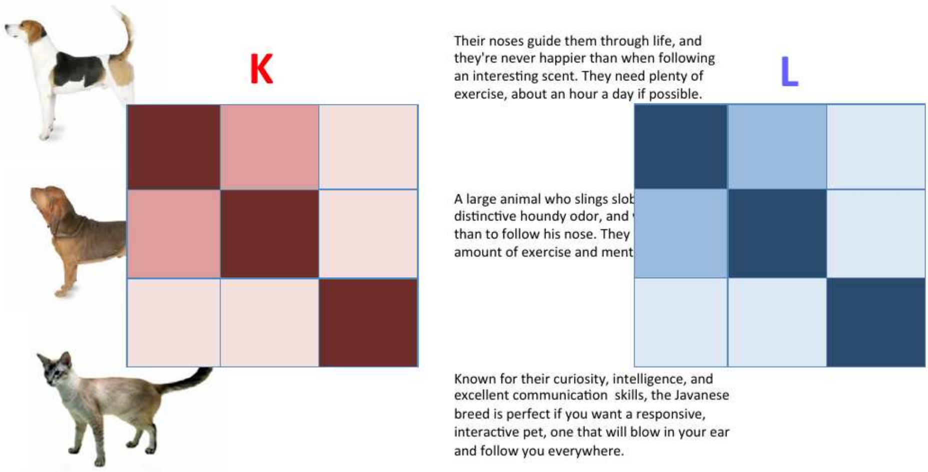
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

L



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.



Empirical estimate:

$$\widehat{HSIC}(\mathbb{P}_{XY}, \mathbb{P}_X\mathbb{P}_Y) = \frac{1}{n^2}(H{\color{red}K}H \circ H{\color{blue}L}H)_{++}, \quad H = I_n - n^{-1}11^T.$$

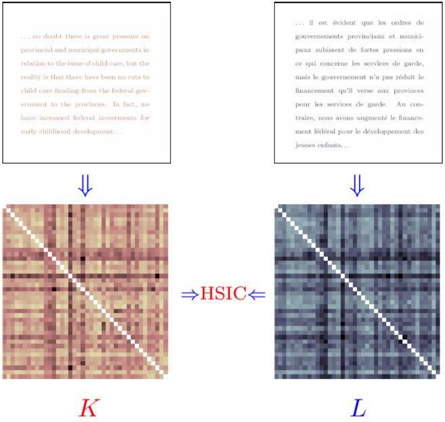
- Under H_0: $n\widehat{HSIC} \to \infty$-sum of weighted $\chi^2 \ldots$
- Permutation test:
  1. Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ with many π-s.
  2. Estimate the $(1-\alpha)$-quantile from the empirical CDF.

Demo: translation example

- 5-line extracts.
- kernel: bag-of-words, r-spectrum ($r = 5$)
- sample size: $n = 10$. repetitions: 300.

Results:

- r-spectrum: average Type-II error $= 0$ ($\alpha = 0.05$),
- bag-of-words: 0.18.



Summary

- Kernels on images, texts, graphs, time series, ...
- RKHS based metric on probability distributions.
- Applications:
  - 2-sample testing: MMD.
  - independence testing: HSIC.
- No density estimation.

Contents

- AM signals.
- Kernel examples.
- Universal kernel: definition, examples.
- MMD: IPM representation.
- HSIC: Where 'HS' is coming from?

AM signals

- s_i: i^{th} song.
- observation ($s \mapsto y$):

$$y(t) = \cos(\omega_c t)(As(t) + o_c) + n(t),$$

  where $n(t)$: Gaussian noise.
- The AM signals were sampled at 120kHz.

Kernel examples

$$k_G(a,b) = e^{-\frac{\|a-b\|_2^2}{2\theta^2}}, \qquad k_e(a,b) = e^{-\frac{\|a-b\|_2}{2\theta^2}},$$

$$k_C(a,b) = \frac{1}{1+\frac{\|a-b\|_2^2}{\theta^2}}, \quad k_t(a,b) = \frac{1}{1+\|a-b\|_2^\theta},$$

$$k_p(a,b) = (\langle a,b\rangle + \theta)^p, \; k_r(a,b) = 1 - \frac{\|a-b\|_2^2}{\|a-b\|_2^2 + \theta},$$

$$k_i(a,b) = \frac{1}{\sqrt{\|a-b\|_2^2 + \theta^2}},$$

$$k_{M,\frac{3}{2}}(a,b) = \left(1 + \frac{\sqrt{3}\,\|a-b\|_2}{\theta}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\theta}},$$

$$k_{M,\frac{5}{2}}(a,b) = \left(1 + \frac{\sqrt{5}\,\|a-b\|_2}{\theta} + \frac{5\,\|a-b\|_2^2}{3\theta^2}\right) e^{-\frac{\sqrt{5}\|a-b\|_2}{\theta}}.$$

Universal kernel: definition

Assume

- $\mathcal{X}$: compact, metric,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel is continuous.

Then

- Def-1: k is universal if $\mathcal{H}(k)$ is dense in $\left(C(\mathcal{X}), \|\cdot\|_\infty\right)$.

Universal kernel: definition

Assume

- $\mathcal{X}$: compact, metric,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel is continuous.

Then

- Def-1: k is universal if $\mathcal{H}(k)$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.
- Def-2: k is
  - characteristic, if $\mu : M_1^+(\mathcal{X}) \to \mathcal{H}(k)$ is injective.

Universal kernel: definition

Assume

- $\mathcal{X}$: compact, metric,
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel is continuous.

Then

- Def-1: k is universal if $\mathcal{H}(k)$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.
- Def-2: k is
  - characteristic, if $\mu : M_1^+(\mathcal{X}) \to \mathcal{H}(k)$ is injective.
  - universal, if μ is injective on the finite signed measures of $\mathcal{X}$.

Universal kernel: examples

On compact subsets of $\mathbb{R}^d$ ($\beta > 0$):

$$k(a,b) = e^{-\beta\|a-b\|_2^2},$$
$$k(a,b) = e^{-\beta\|a-b\|_1},$$
$$k(a,b) = e^{\beta\langle a,b\rangle}, (\beta > 0), \text{ or more generally}$$
$$k(a,b) = f(\langle a,b\rangle), \quad f(x) = \sum_{n=0}^{\infty} a_n x^n \quad (\forall a_n > 0).$$

Universal $\Rightarrow$ characteristic.

MMD: IPM represenation

Let $\mathcal{F} := \{f \in \mathcal{H}(k) : \|f\|_{\mathcal{H}} \leqslant 1\}$ be the unit ball in $\mathcal{H}$. Then

$$\begin{aligned}
MMD(\mathbb{P}, \mathbb{Q}; \mathcal{F}) &:= \sup_{f \in \mathcal{F}} [\mathbb{E}_{x \sim \mathbb{P}} f(x) - \mathbb{E}_{y \sim \mathbb{Q}} f(y)], \\
&= \sup_{f \in \mathcal{F}} [\langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} - \langle f, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \sup_{f \in \mathcal{F}} [\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.
\end{aligned}$$

HSIC: Where 'HS' is coming from?

Players: $(\mathcal{X},k)$, $(\mathcal{Y},\ell)$, $\mathbb{P}_{XY}$, $\mathbb{P}_X$, $\mathbb{P}_Y$; $C_{XY}:\mathcal{H}(\ell)\to\mathcal{H}(k)$.

$$\begin{aligned}
C_{XY} &= \mathbb{E}_{XY}[(\varphi_x-\mu_{\mathbb{P}_X})\otimes(\varphi_y-\mu_{\mathbb{P}_Y})],\\
\langle f, C_{XY}g\rangle_{\mathcal{H}(k)} &= \mathbb{E}_{XY}[f(x)-\mathbb{E}_X f(x)][g(y)-\mathbb{E}_Y g(y)], \forall f,g,\\
HSIC &= \|C_{XY}\|_{HS}^2.
\end{aligned}$$