Kernel-based Dependency Measures and Hypothesis Testing

Zoltán Szabó – CMAP, École Polytechnique

Carnegie Mellon University

November 27, 2017

Outline

- Motivation.
- Diverse set of domains: kernel.
- Dependency measures:
  - Kernel canonical correlation analysis.
  - Maximum mean discrepancy.
  - Hilbert-Schmidt independence criterion.
- Hypothesis testing.

Dependency Measures as Objective Functions

Outlier-robust image registration [Kybic, 2004, Neemuchwala et al., 2007]

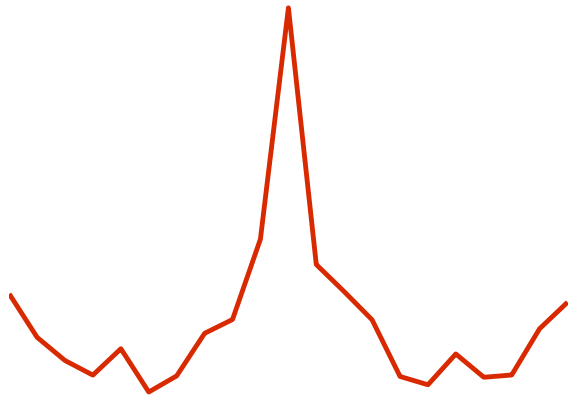
Given two images:


Goal: find the transformation which takes the right one to the left.

Outlier-robust image registration [Kybic, 2004, Neemuchwala et al., 2007]

Given two images:





Goal: find the transformation which takes the right one to the left.

Outlier-robust image registration: equations

- Reference image: $\mathbf{y}_{\text{ref}}$,
- test image: $\mathbf{y}_{\text{test}}$,
- possible transformations: $\boldsymbol{\Theta}$.

Objective:

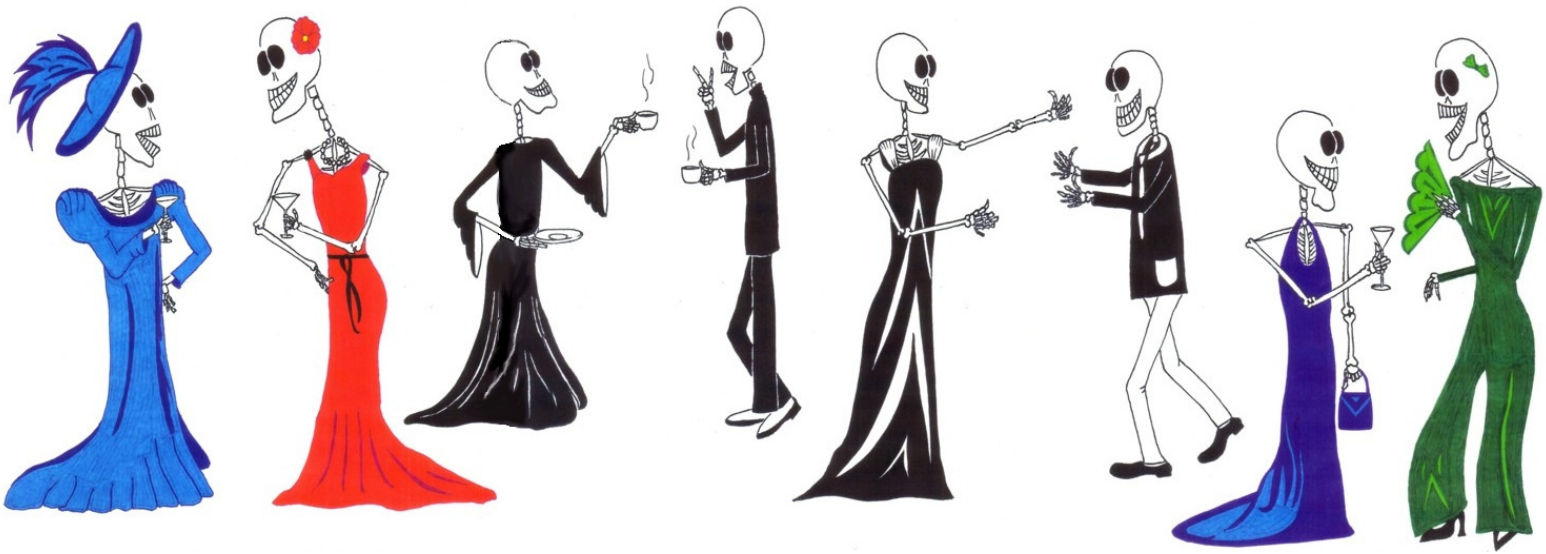
$$J(\boldsymbol{\theta}) = \underbrace{I(\mathbf{y}_{\text{ref}}, \mathbf{y}_{\text{test}}(\boldsymbol{\theta}))}_{\text{similarity}} \to \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}.$$

In the example: I=KCCA.

Independent subspace analysis [Cardoso, 1998] $\xleftarrow{\text{ext.}}$ ICA

Cocktail party problem:

- independent groups of people / music bands,
- observation = mixed sources.



ISA equations

Observation:

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right].$$

Goal: $\hat{\mathbf{s}}$ from $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. Assumptions:

- independent groups: $I\left(\mathbf{s}^1, \ldots, \mathbf{s}^M\right) = 0$,
- $\mathbf{s}^m$-s: non-Gaussian,
- $\mathbf{A}$: invertible.

ISA solution

Find $\mathbf{W}$ which makes the estimated components independent:

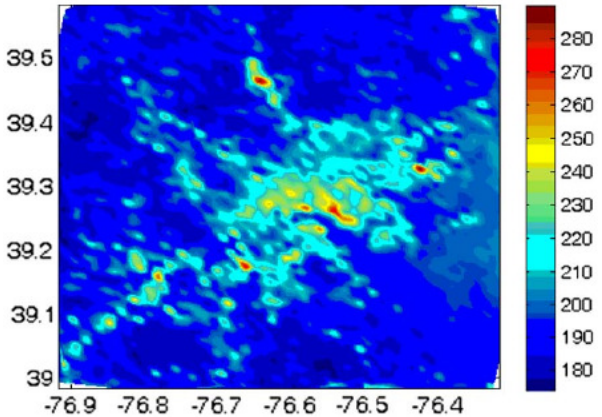
$$\mathbf{y} = \mathbf{W}\mathbf{x} = \left[\mathbf{y}^1; \ldots; \mathbf{y}^M\right],$$

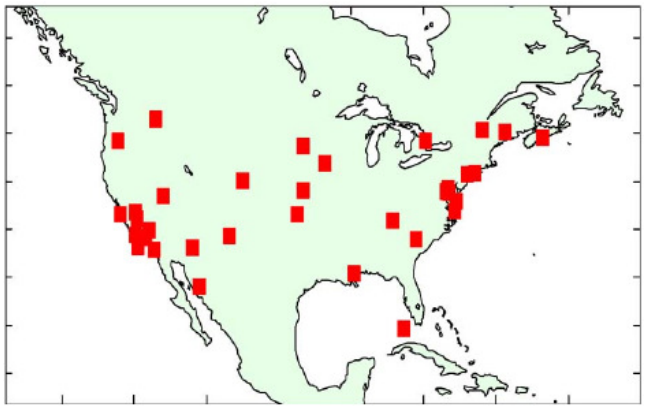
$$J(\mathbf{W}) = I\left(\mathbf{y}^1, \ldots, \mathbf{y}^M\right) \to \min_{\mathbf{W}}.$$

Distribution regression [Póczos et al., 2013, Szabó et al., 2016]. Sustainability

- **Goal**: aerosol prediction = air pollution $\rightarrow$ climate.

- Prediction using labelled bags:
  - bag := multi-spectral satellite measurements over an area,
  - label := local aerosol value.





Objects in the bags

- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, ...

Objects in the bags

- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, ...
- Wider context (statistics): point estimation tasks.

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\underbrace{\left(\mu_{\hat{\mathbb{P}}_i} \right)}_{\text{feature of } \hat{\mathbb{P}}_i} - y_i \right]^2 + \lambda \left\| f \right\|_{\mathcal{H}}^2 .$$

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\mu_{\hat{\mathbb{P}}_i}\right) - y_i \right]^2 + \lambda \left\| f \right\|_{\mathcal{H}_K}^2 .$$

- Prediction:

$$\hat{y}\left(\hat{\mathbb{P}}\right) = \mathbf{g}^T (\mathbf{G} + \ell\lambda \mathbf{I})^{-1} \mathbf{y},$$
$$\mathbf{g} = \left[K\left(\mu_{\hat{\mathbb{P}}}, \mu_{\hat{\mathbb{P}}_i}\right)\right], \mathbf{G} = \left[K\left(\mu_{\hat{\mathbb{P}}_i}, \mu_{\hat{\mathbb{P}}_j}\right)\right], \mathbf{y} = [y_i].$$

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{\mathbb{P}}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{\mathbb{P}}_i$: bag from $\mathbb{P}_i$, $N := |\hat{\mathbb{P}}_i|$.
  - test bag: $\hat{\mathbb{P}}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\mu_{\hat{\mathbb{P}}_i}\right) - y_i \right]^2 + \lambda \left\| f \right\|_{\mathcal{H}_K}^2 .$$

- Prediction:

$$\hat{y}\left(\hat{\mathbb{P}}\right) = \mathbf{g}^T (\mathbf{G} + \ell\lambda \mathbf{I})^{-1} \mathbf{y},$$
$$\mathbf{g} = \left[K\left(\mu_{\hat{\mathbb{P}}}, \mu_{\hat{\mathbb{P}}_i}\right)\right], \mathbf{G} = \left[K\left(\mu_{\hat{\mathbb{P}}_i}, \mu_{\hat{\mathbb{P}}_j}\right)\right], \mathbf{y} = [y_i].$$

Inner product of distributions

$K\left(\mu_{\hat{\mathbb{P}}_i}, \mu_{\hat{\mathbb{P}}_j}\right) = ?$

Feature selection

- **Goal**: find
  - the feature subset (# of rooms, criminal rate, local taxes)
  - most relevant for house price prediction (y).

Feature selection: equations

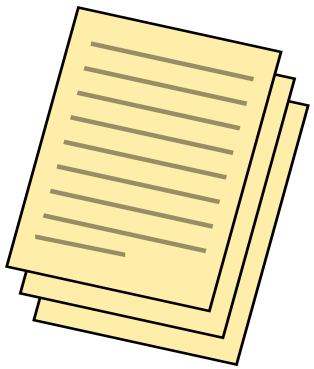
- Features: $x^1, \ldots, x^F$. Subset: $S \subseteq \{1, \ldots, F\}$.
- MaxRelevance - MinRedundancy principle [Peng et al., 2005]:

$$J(S) = \frac{1}{|S|} \sum_{i \in S} I\left(x^i, y\right) - \frac{1}{|S|^2} \sum_{i,j \in S} I\left(x^i, x^j\right) \to \max_{S \subseteq \{1,\ldots,F\}}.$$

Hypothesis Testing

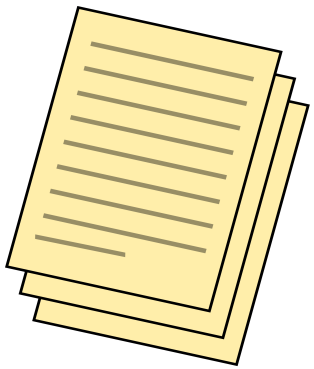
Example-1 (2-sample testing): NLP

- Given: 2 categories of documents (Bayesian inference, neuroscience).
- Task:
  - test their distinguishability,
  - most discriminative words $\longrightarrow$ interpretability.



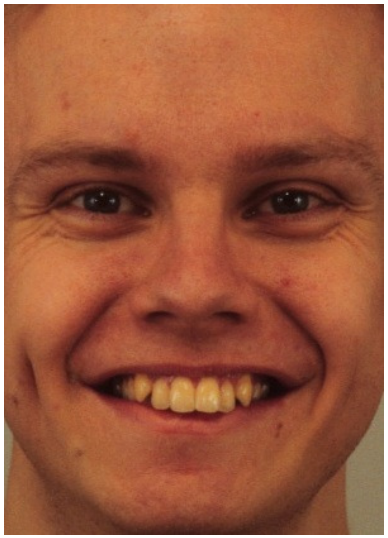
Example-1 (2-sample testing): NLP

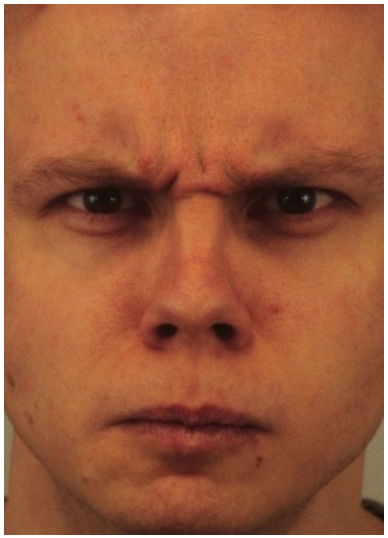
- Given: 2 categories of documents (Bayesian inference, neuroscience).
- Task:
  - test their distinguishability,
  - most discriminative words $\longrightarrow$ interpretability.



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

Example-2 (2-sample testing): computer vision

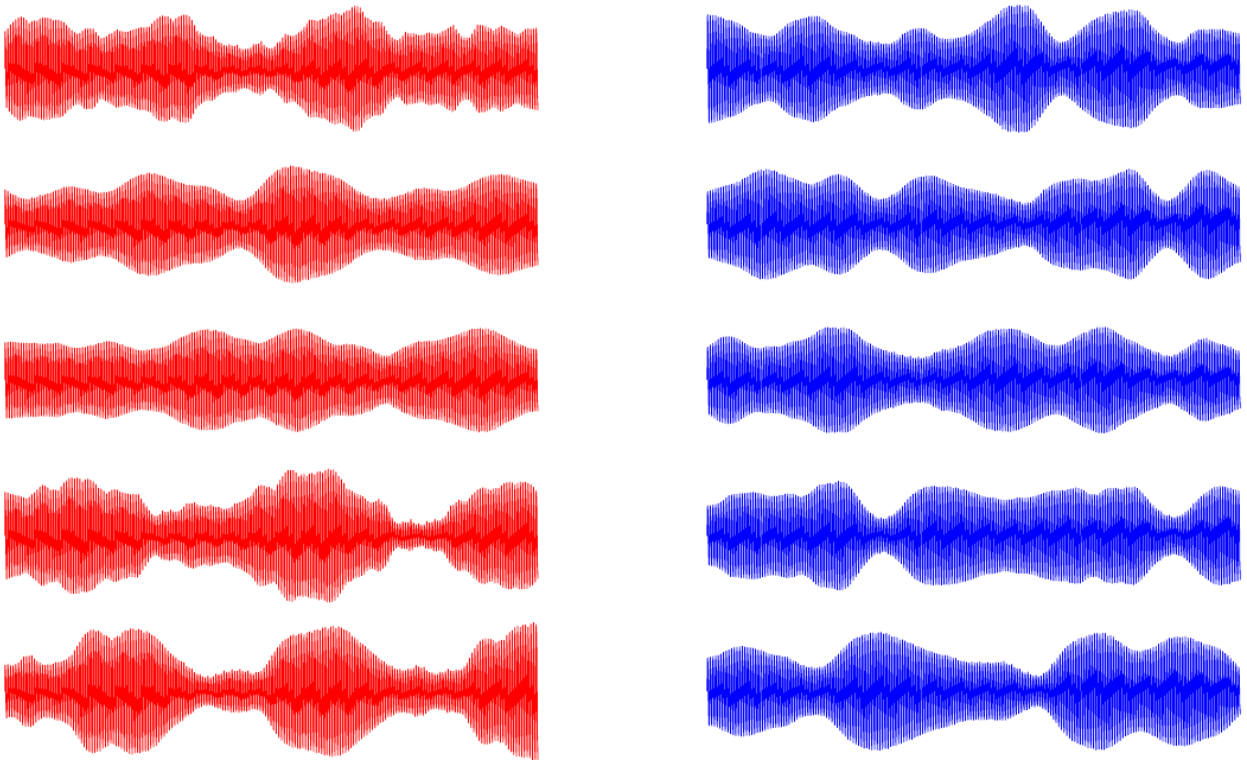




- Given: two sets of faces (happy, angry).
- Task:
  - check if they are different,
  - determine the most discriminative features/regions.

Example-3 (2-sample testing): audio

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



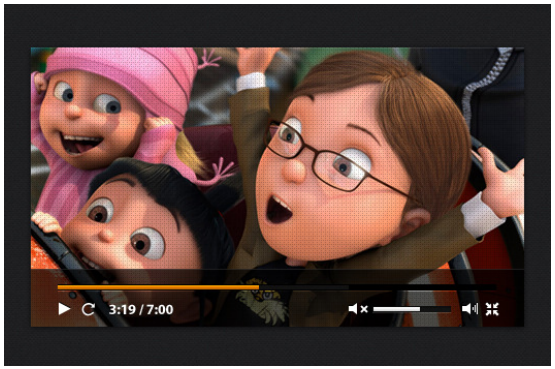
Example: independence testing-1

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

Example: independence testing-1

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

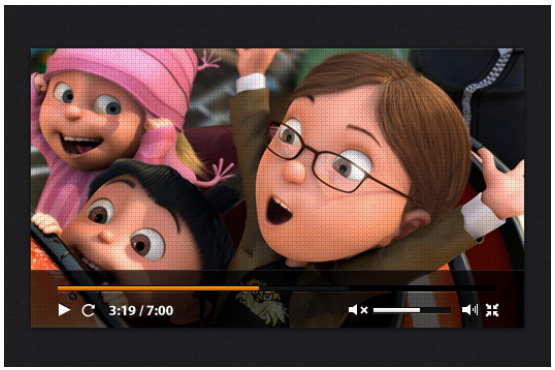
  - (video, caption) pairs



Example: independence testing-1

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

  - (video, caption) pairs



- $\{(x_i, y_i)\}_{i=1}^n \stackrel{?}{\longrightarrow} \mathbb{P}_{xy} = \mathbb{P}_x \mathbb{P}_y$.

Example: independence testing-2

- 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.

...

Example: independence testing-2

- 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. $\mathbb{P}_{xy} = \mathbb{P}_x \mathbb{P}_y$?

Diverse Set of Domains

'Classical' information theory

- Kullback-Leibler divergence:

$$KL\left(\mathbb{P},\mathbb{Q}\right)=\int_{\mathbb{R}^d}p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x.$$

'Classical' information theory

- Kullback-Leibler divergence:

$$KL\left(\mathbb{P},\mathbb{Q}\right)=\int_{\mathbb{R}^d}p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x.$$

- Mutual information:

$$I\left(\mathbb{P}\right)=KL\left(\mathbb{P},\otimes_{m=1}^{M}\mathbb{P}_m\right).$$

'Classical' information theory

- Kullback-Leibler divergence:

$$KL\left(\mathbb{P},\mathbb{Q}\right)=\int_{\mathbb{R}^d}p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x.$$

- Mutual information:

$$I\left(\mathbb{P}\right)=KL\left(\mathbb{P},\otimes_{m=1}^{M}\mathbb{P}_m\right).$$

Properties:

$I(\mathbb{P})\geqslant 0$.

'Classical' information theory

- Kullback-Leibler divergence:

$$KL\left(\mathbb{P},\mathbb{Q}\right)=\int_{\mathbb{R}^d}p(x)\log\left[\frac{p(x)}{q(x)}\right]\mathrm{d}x.$$

- Mutual information:

$$I\left(\mathbb{P}\right)=KL\left(\mathbb{P},\otimes_{m=1}^{M}\mathbb{P}_m\right).$$

Properties:

$I\left(\mathbb{P}\right)\geqslant 0$.
$I\left(\mathbb{P}\right)=0\Leftrightarrow\mathbb{P}=\otimes_{m=1}^{M}\mathbb{P}_m$.

'Classical' information theory

- Kullback-Leibler divergence:

$$KL(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} p(x) \log \left[\frac{p(x)}{q(x)} \right] \mathrm{d}x.$$

- Mutual information:

$$I(\mathbb{P}) = KL\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right).$$

Properties:

$I(\mathbb{P}) \geqslant 0$.
$I(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^{M} \mathbb{P}_m$.

Alternatives: Rényi, Tsallis, L^2 divergence... Typically: $\mathcal{X} = \mathbb{R}^d$.

Euclidean space $\longrightarrow$ inner product $\longrightarrow$ kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ leads to kernels.

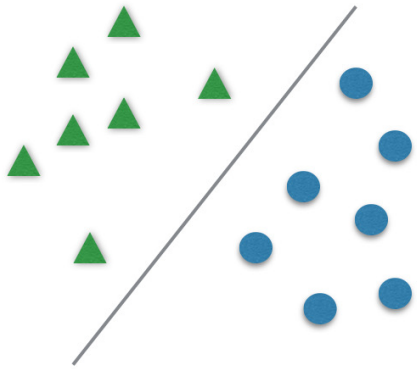
Euclidean space $\longrightarrow$ inner product $\longrightarrow$ kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ leads to kernels. Why?

Euclidean space $\longrightarrow$ inner product $\longrightarrow$ kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ leads to kernels. Why?

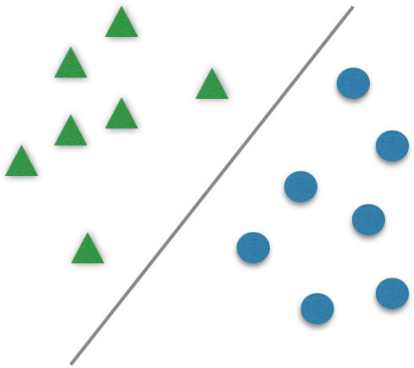
- Classification (SVM):

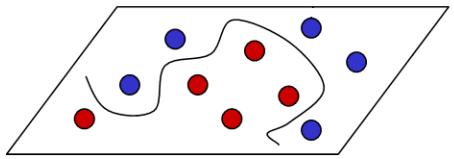


Euclidean space $\longrightarrow$ inner product $\longrightarrow$ kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$ leads to kernels. Why?

- Classification (SVM):



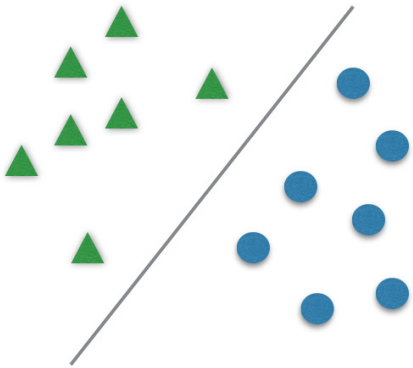


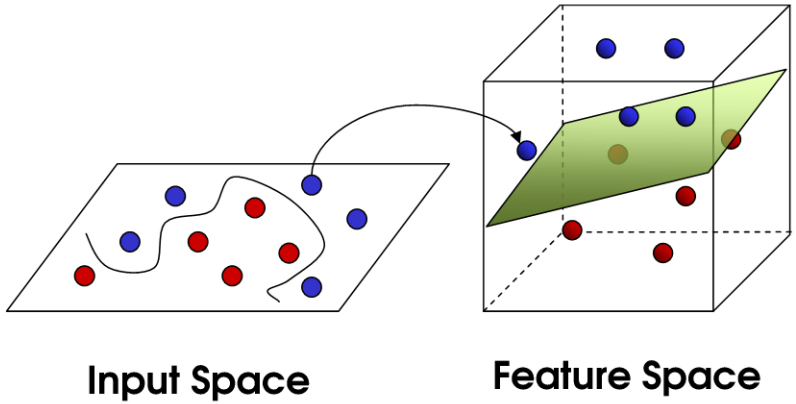
Input Space

Euclidean space $\longrightarrow$ inner product $\longrightarrow$ kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$ leads to kernels. Why?

- Classification (SVM):

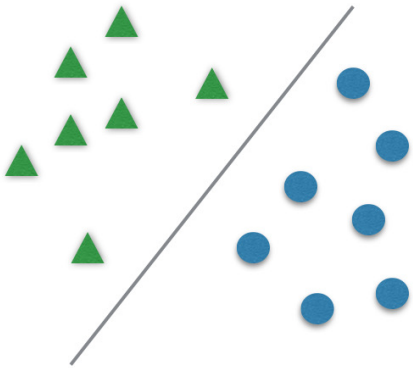


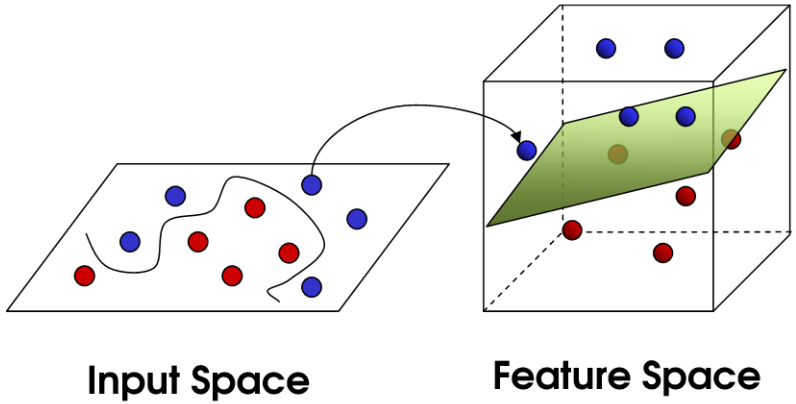


Euclidean space $\rightarrow$ inner product $\rightarrow$ kernel

Extension of $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$ leads to kernels. Why?

- Classification (SVM):





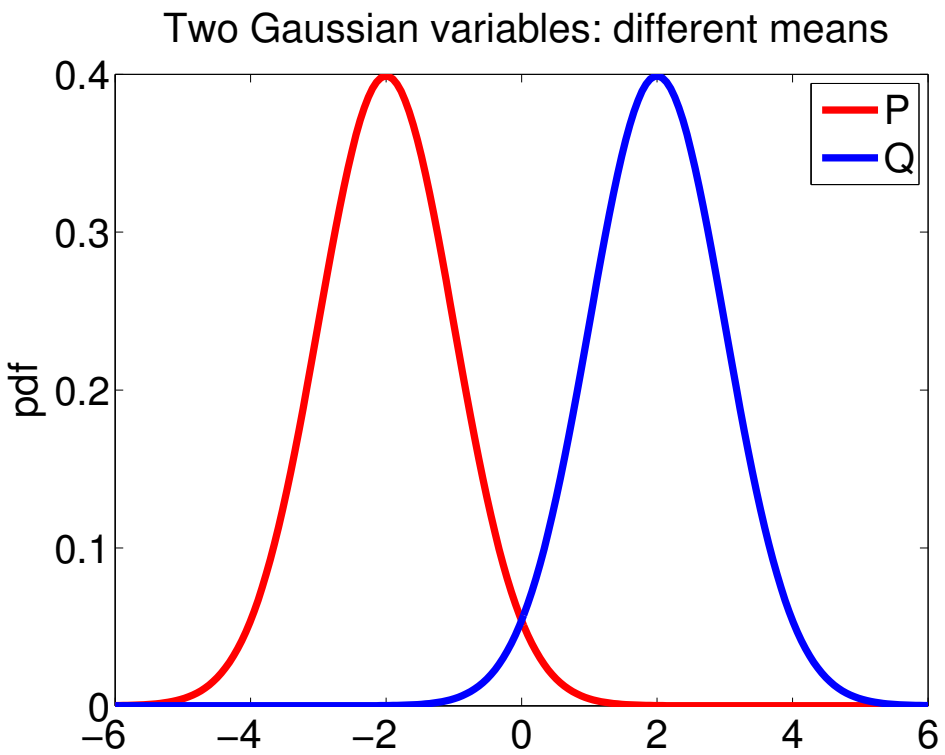
- Representation of distributions:

$$\mathbb{P} \mapsto \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\varphi(\mathbf{x}).$$

Example: $\varphi(\mathbf{x}) = \mathbf{x}$: mean.

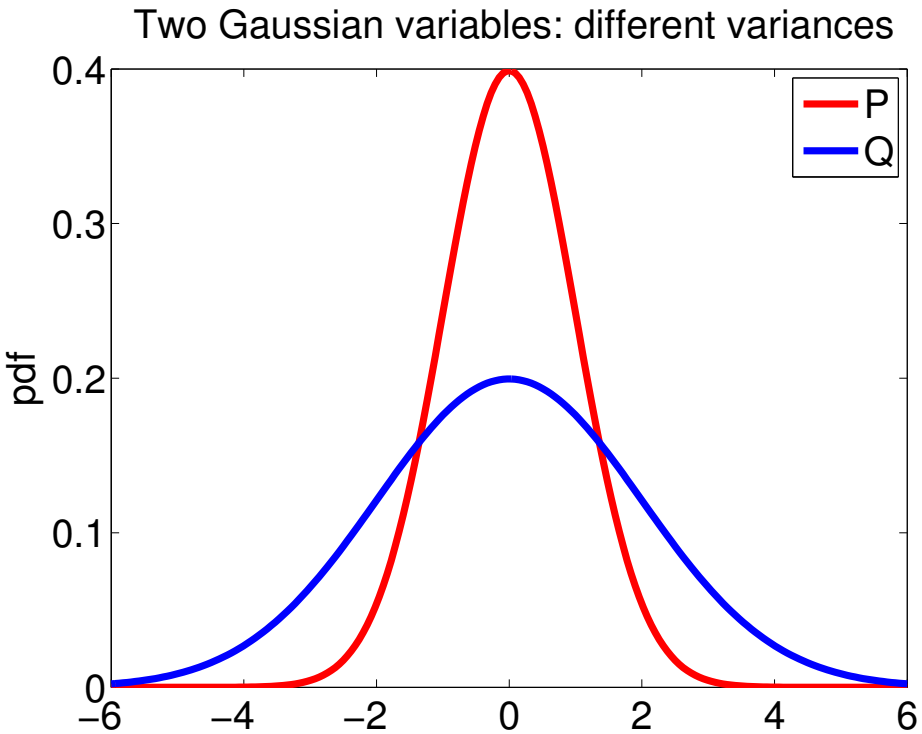
Representations of distributions: $\mathbb{E}X$

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



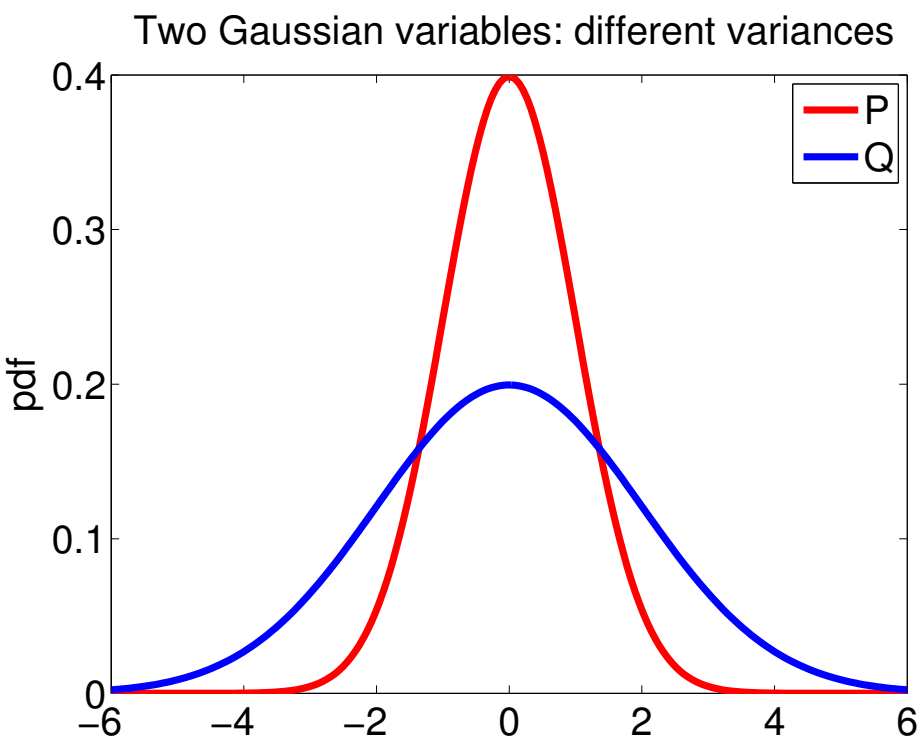
Representations of distributions: $\mathbb{E}X^2$

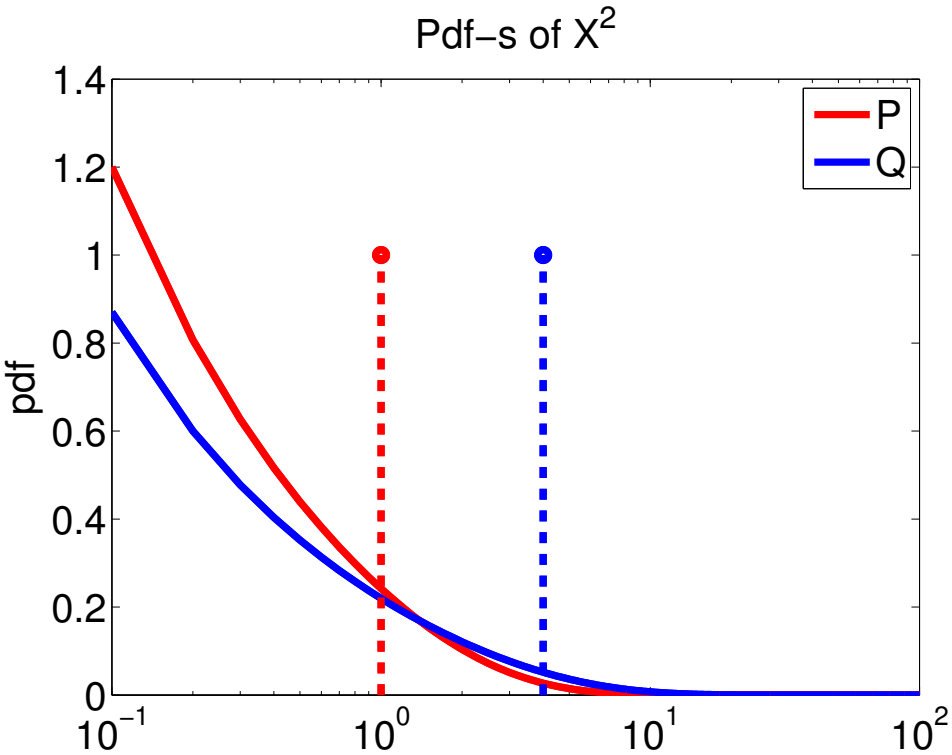
- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.



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

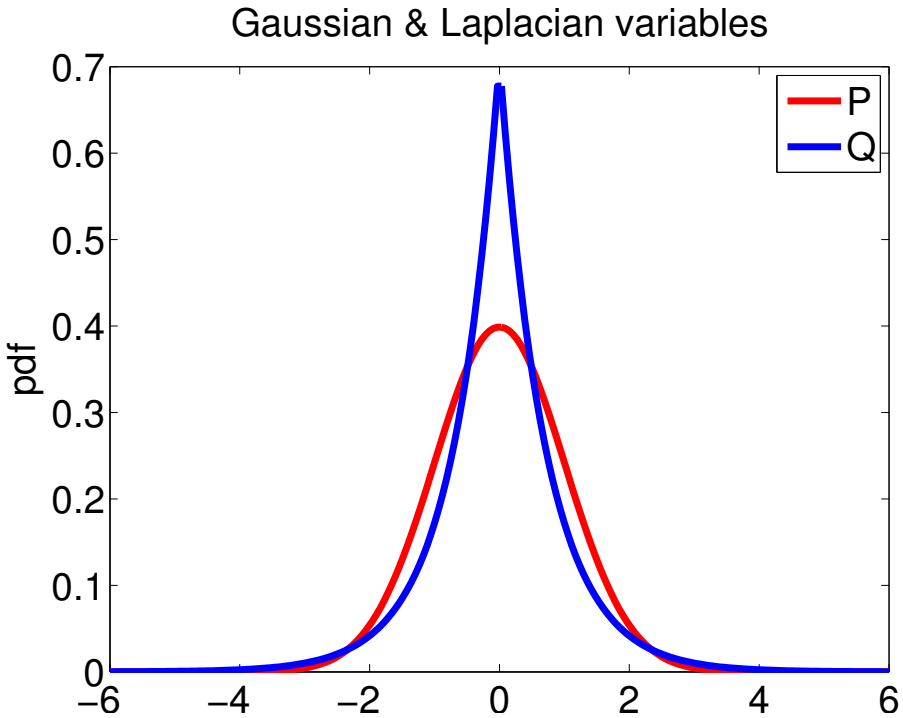
- Setup: 2 Gaussians; same means, different variances.
- Idea: look at the 2nd-order features of RVs.
- $\varphi_x = x^2 \Rightarrow$ difference in $\mathbb{E}X^2$.





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.



$\varphi(\mathbf{x}) = e^{i\langle \cdot, \mathbf{x} \rangle}$: characteristic function, $\mathcal{X} = \mathbb{R}^d$.

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

$$C_{xy} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right]$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

$$C_{xy} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

$$S = \|C_{xy}\|_F$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

$$C_{xy} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

$$C_{xy} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0 \leftrightarrow \text{ linear dependence.}$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

$$C_{xy} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0 \leftrightarrow \text{ linear dependence.}$$

- Covariance operator: take features of x and y

$$C_{xy} = \mathbb{E}_{xy}\big[\underbrace{(\varphi(x) - \mathbb{E}_x\varphi(x))}_{\text{centering in feature space}} \otimes (\psi(y) - \mathbb{E}_y\psi(y))\big]$$

Kernels: why? – continued

Idea: $\mathbb{P}_{xy} \mapsto C_{xy}$.

- Covariance matrix

$$C_{xy} = \mathbb{E}_{xy}\left[(x - \mathbb{E}x)(y - \mathbb{E}y)^T\right],$$

$$S = \|C_{xy}\|_F \stackrel{?}{=} 0 \leftrightarrow \text{ linear dependence.}$$

- Covariance operator: take features of x and y

$$C_{xy} = \mathbb{E}_{xy}\big[\underbrace{(\varphi(x) - \mathbb{E}_x\varphi(x))}_{\text{centering in feature space}} \otimes (\psi(y) - \mathbb{E}_y\psi(y))\big],$$

$$S = \|C_{xy}\|_{HS} =: \text{HSIC}(\mathbb{P}_{xy}).$$

We capture non-linear dependencies via φ, ψ!

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, ...).
- 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:

$$\varphi(x) : \text{features of } x,$$
$$\varphi(x') : \text{features of } x'.$$

- Kernel: inner product of these features

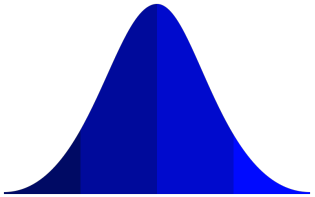
$$k(x, x') := \left\langle \varphi(x), \varphi(x') \right\rangle_{\mathcal{H}}.$$

RKHS: intuition

k defines an RKHS:

- $\mathcal{H}_k \subset \{f : \mathcal{X} \to \mathbb{R} \text{ functions}\}$. Elements:

$$\underbrace{k(\cdot, x)}_{} \in \mathcal{H}_k.$$

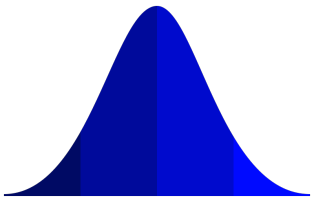


RKHS: intuition

k defines an RKHS:

- $\mathcal{H}_k \subset \{f : \mathcal{X} \to \mathbb{R} \text{ functions}\}$. Elements:

$$\underbrace{k(\cdot, x)}_{} \in \mathcal{H}_k.$$



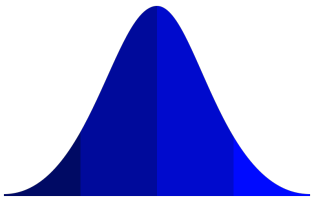
- $f = \sum_{i=1}^{n} c_i k(\cdot, x_i)$

RKHS: intuition

k defines an RKHS:

- $\mathcal{H}_k \subset \{f : \mathcal{X} \to \mathbb{R} \text{ functions}\}$. Elements:

$$\underbrace{k(\cdot, x)}_{} \in \mathcal{H}_k.$$



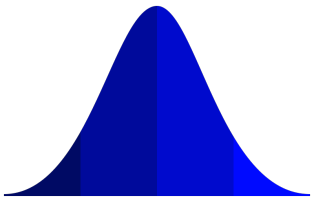
- $f = \sum_{i=1}^{n} c_i k(\cdot, x_i) \leftrightarrow \mathbf{c} \in \mathbb{R}^n$ [SVM dual].

RKHS: intuition

k defines an RKHS:

- $\mathcal{H}_k \subset \{f : \mathcal{X} \to \mathbb{R} \text{ functions}\}$. Elements:

$$\underbrace{k(\cdot, x)}_{} \in \mathcal{H}_k.$$



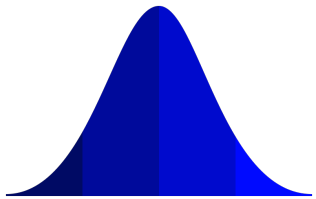
- $f = \sum_{i=1}^{n} c_i k(\cdot, x_i) \leftrightarrow \mathbf{c} \in \mathbb{R}^n$ [SVM dual].
- Inner product: $\Big\langle \underbrace{k(\cdot, x)}_{\varphi(x)}, \underbrace{k(\cdot, x')}_{\varphi(x')} \Big\rangle_{\mathcal{H}_k} = k(x, x')$ [evaluation].

RKHS: intuition

k defines an RKHS:

- $\mathcal{H}_k \subset \{f : \mathcal{X} \to \mathbb{R} \text{ functions}\}$. Elements:

$$\underbrace{k(\cdot, x)}_{} \in \mathcal{H}_k.$$

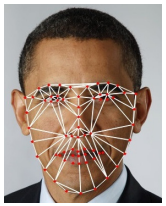


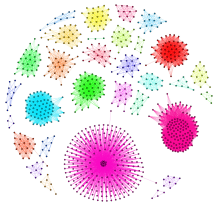
- $f = \sum_{i=1}^{n} c_i k(\cdot, x_i) \leftrightarrow \mathbf{c} \in \mathbb{R}^n$ [SVM dual].
- Inner product: $\left\langle \underbrace{k(\cdot, x)}_{\varphi(x)}, \underbrace{k(\cdot, x')}_{\varphi(x')} \right\rangle_{\mathcal{H}_k} = k(x, x')$ [evaluation].

Practically

$\mathcal{H}_k \ni f = f(\mathbf{c})$. Enough: $k(x, x')$. $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \succcurlyeq 0$.

Diverse set of domains, kernel examples



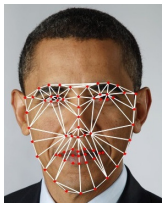


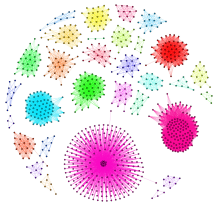
- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

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

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

Diverse set of domains, kernel examples





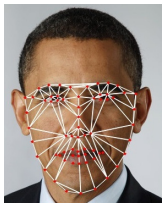
- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

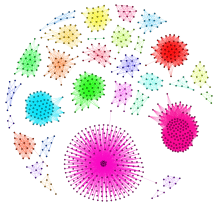
$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

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

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

Diverse set of domains, kernel examples





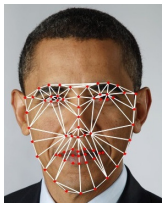
- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

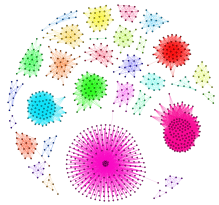
$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

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

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

Diverse set of domains, kernel examples





- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

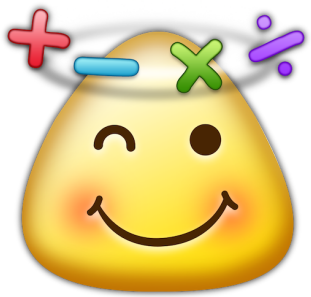
$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

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

- $\mathcal{X}$ = strings, texts:
  - r-spectrum kernel: # of common $\leqslant r$-substrings.
- $\mathcal{X}$ = time-series: dynamic time-warping.
- $\mathcal{X}$ = trees, graphs, dynamical systems, sets, permutations, ...

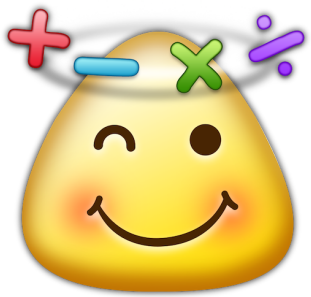
Independence measures

- Given: random variable $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $(x, y) \sim \mathbb{P}_{xy}$.
- Goal: measure the dependence of x and y.



Independence measures

- Given: random variable $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $(x, y) \sim \mathbb{P}_{xy}$.
- Goal: measure the dependence of x and y.
- Desiderata for a $Q(\mathbb{P}_{xy})$ independence measure [Rényi, 1959]:
  1. $Q(\mathbb{P}_{xy})$ is well-defined,
  2. $Q(\mathbb{P}_{xy}) \in [0, 1]$,
  3. $Q(\mathbb{P}_{xy}) = 0$ iff. $x \perp y$.
  4. $Q(\mathbb{P}_{xy}) = 1$ iff. $y = f(x)$ or $x = g(y)$.



Independence measures

- He showed:

$$Q(\mathbb{P}_{xy}) = \sup_{f,g} \operatorname{corr}(f(x), g(y)),$$

  satisfies 1-4.
- Too ambitious:
  - computationally intractable.
  - many functions.

Independence measures: restriction to continuous functions

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R}$, bounded continuous$\}$ would also work.
- Still too large!

Independence measures: restriction to continuous functions

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R}$, bounded continuous$\}$ would also work.
- Still too large!
- Idea:
  - certain $\mathcal{H}_k$ function classes are dense in $C_b(\mathcal{X})$.
  - computionally tractable.

Wanted

- Independence measure,
- distance,
- inner product

measures/estimates on probability distributions

without density estimation!

Kernel Canonical Correlation Analysis (KCCA)

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated RKHS-s: $\mathcal{H}_k$, $\mathcal{H}_\ell$.

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated RKHS-s: $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- KCCA measure of $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\rho_{\mathrm{KCCA}}(x, y) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \mathrm{corr}(f(x), g(y)),$$

$$\mathrm{corr}(f(x), g(y)) = \frac{\mathrm{cov}_{xy}(f(x), g(y))}{\sqrt{\mathrm{var}_x f(x)\, \mathrm{var}_y g(y)}}.$$

KCCA: estimation

Empirical estimate of KCCA from $\{(x_i, y_i)\}_{i=1}^n$:

$$\widehat{\rho_{\mathrm{KCCA}}}(x, y) = \sup_{\mathbf{c}\in\mathbb{R}^n, \mathbf{d}\in\mathbb{R}^n} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_x)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_y)^2 \mathbf{d}}}.$$

KCCA: estimation

Empirical estimate of KCCA from $\{(x_i, y_i)\}_{i=1}^n$:

$$\widehat{\rho_{\mathrm{KCCA}}}(x, y) = \sup_{\mathbf{c}\in\mathbb{R}^n, \mathbf{d}\in\mathbb{R}^n} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_x)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_y)^2 \mathbf{d}}}.$$

- Centered Gram matrix:

$$\mathbf{G}_x = [k(x_i, x_j)]_{i,j=1}^n,$$

$$\tilde{\mathbf{G}}_x = \underbrace{\mathbf{H}_n \mathbf{G}_x \mathbf{H}_n}_{\text{centering in feature space}}, \qquad \mathbf{H}_n = \mathbf{I}_n - \frac{\mathbf{1}_{n\times n}}{n}.$$

KCCA: estimation

Empirical estimate of KCCA from $\{(x_i, y_i)\}_{i=1}^n$:

$$\widehat{\rho_{\mathrm{KCCA}}}(x, y) = \sup_{\mathbf{c}\in\mathbb{R}^n, \mathbf{d}\in\mathbb{R}^n} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_x)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_y)^2 \mathbf{d}}}.$$

- Centered Gram matrix:

$$\mathbf{G}_x = [k(x_i, x_j)]_{i,j=1}^n,$$

$$\tilde{\mathbf{G}}_x = \underbrace{\mathbf{H}_n \mathbf{G}_x \mathbf{H}_n}_{\text{centering in feature space}}, \qquad \mathbf{H}_n = \mathbf{I}_n - \frac{\mathbf{1}_{n\times n}}{n}.$$

- $\mathbf{c}, \mathbf{d}$: come from a generalized eigenvalue problem ($\mathbf{A}\mathbf{v} = \lambda \mathbf{B}\mathbf{v}$).

KCCA: estimation

Empirical estimate of KCCA from $\{(x_i, y_i)\}_{i=1}^n$:

$$\widehat{\rho_{\mathrm{KCCA}}}(x, y) = \sup_{\mathbf{c}\in\mathbb{R}^n, \mathbf{d}\in\mathbb{R}^n} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_x \tilde{\mathbf{G}}_y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_x)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_y)^2 \mathbf{d}}}.$$

- Centered Gram matrix:

$$\mathbf{G}_x = [k(x_i, x_j)]_{i,j=1}^n,$$

$$\tilde{\mathbf{G}}_x = \underbrace{\mathbf{H}_n \mathbf{G}_x \mathbf{H}_n}_{\text{centering in feature space}}, \qquad \mathbf{H}_n = \mathbf{I}_n - \frac{\mathbf{1}_{n\times n}}{n}.$$

- $\mathbf{c}, \mathbf{d}$: come from a generalized eigenvalue problem ($\mathbf{A}\mathbf{v} = \lambda \mathbf{B}\mathbf{v}$).
- A bit of regularization: $\tilde{\mathbf{G}}_x \to \tilde{\mathbf{G}}_x + \kappa \mathbf{I}_n$, $\tilde{\mathbf{G}}_y \to \ldots$

KCCA as an independence measure

If $x \perp y$, then $\rho_{\mathrm{KCCA}}(x, y; \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$ [Bach and Jordan, 2002, Gretton et al., 2005].

KCCA as an independence measure

If $x \perp y$, then $\rho_{\mathrm{KCCA}}(x, y; \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$ [Bach and Jordan, 2002, Gretton et al., 2005].
- Enough: universal kernel = '$\mathcal{H}_k$ dense in $C_b(\mathcal{X})$'.

KCCA as an independence measure

If $x \perp y$, then $\rho_{\mathrm{KCCA}}(x, y; \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$ [Bach and Jordan, 2002, Gretton et al., 2005].
- Enough: universal kernel = '$\mathcal{H}_k$ dense in $C_b(\mathcal{X})$'.
- Example: Gaussian, Laplacian kernel.

ITE toolbox

- Estimators for
  - dependency measures ($\ni$ KCCA),
  - distances on distributions ($\ni$ MMD),
  - independence of random variables ($\ni$ HSIC),
  - and many more...
- Link:

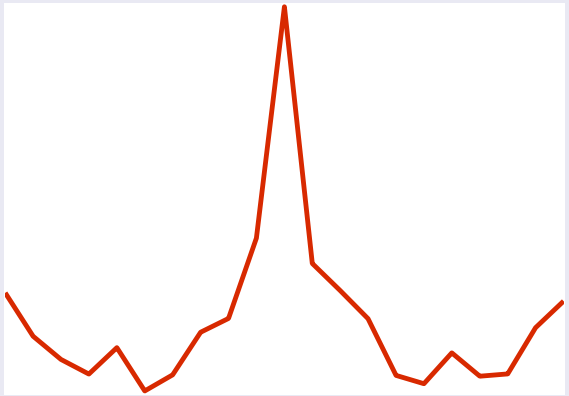
`https://bitbucket.org/szzoli/ite/`

ITE toolbox

- Estimators for
  - dependency measures (∋ KCCA),
  - distances on distributions (∋ MMD),
  - independence of random variables (∋ HSIC),
  - and many more...
- Link:

`https://bitbucket.org/szzoli/ite/`





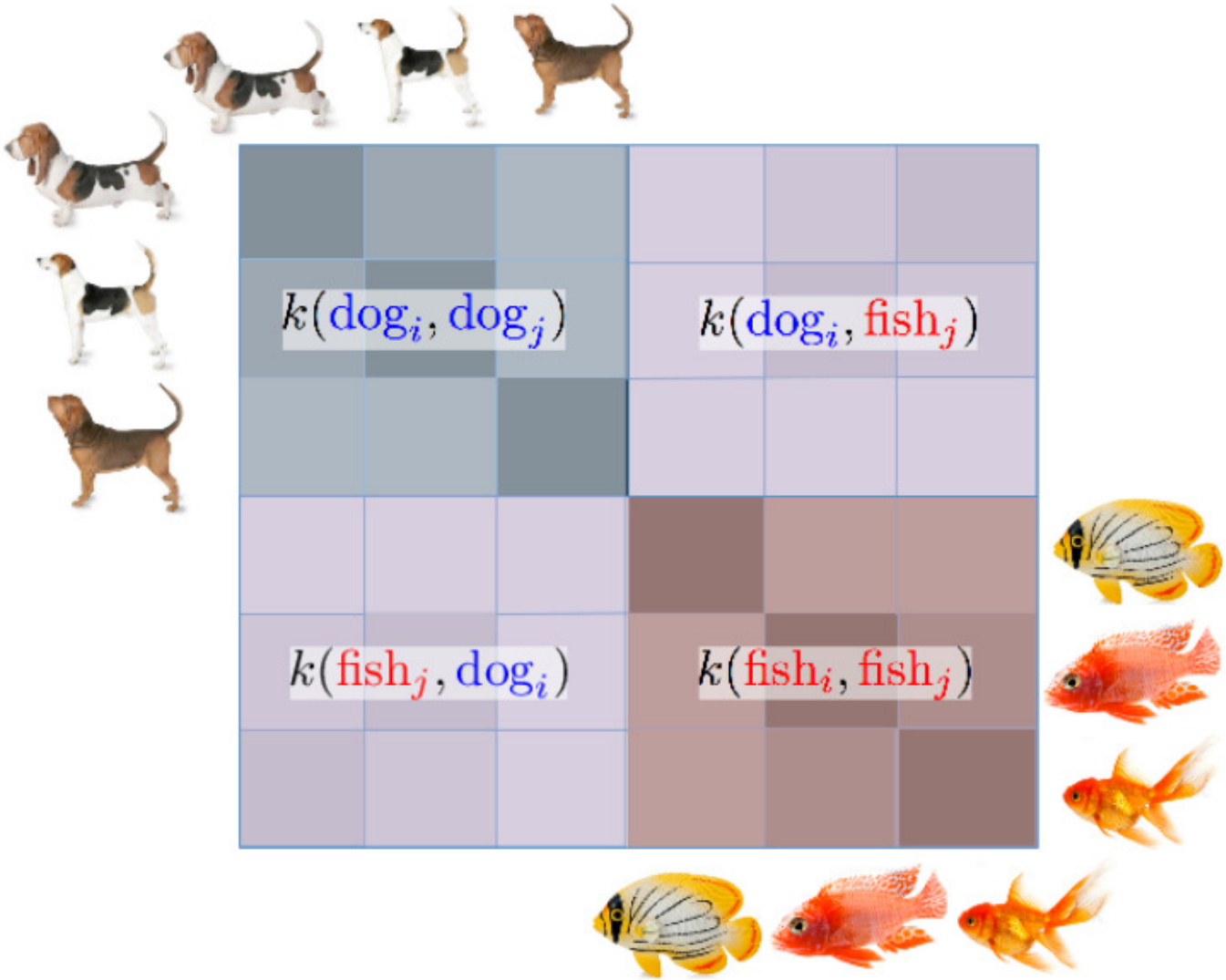
Maximum Mean Discrepancy (MMD)

MMD estimator: intuition

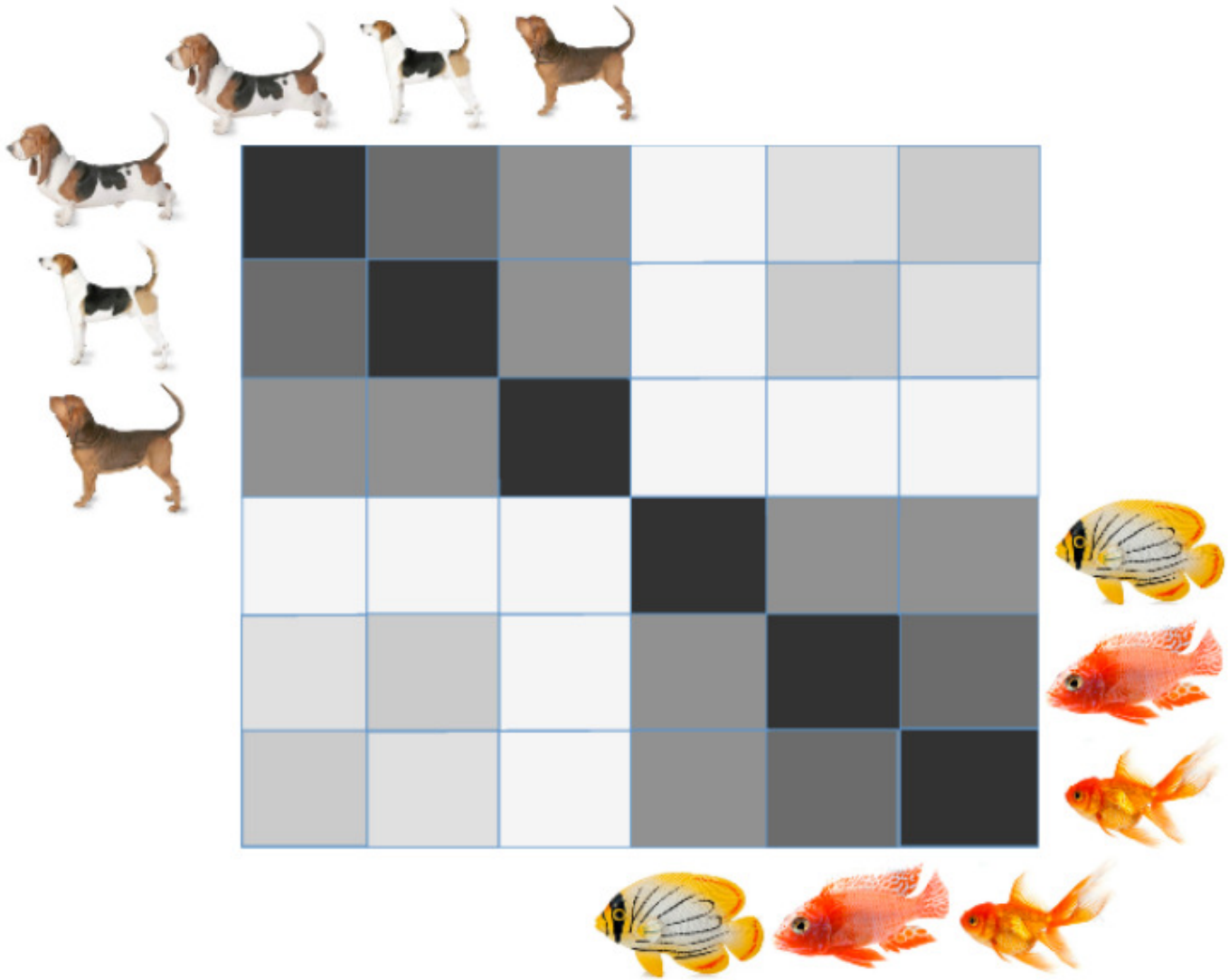
$\sim P$

$\sim Q$

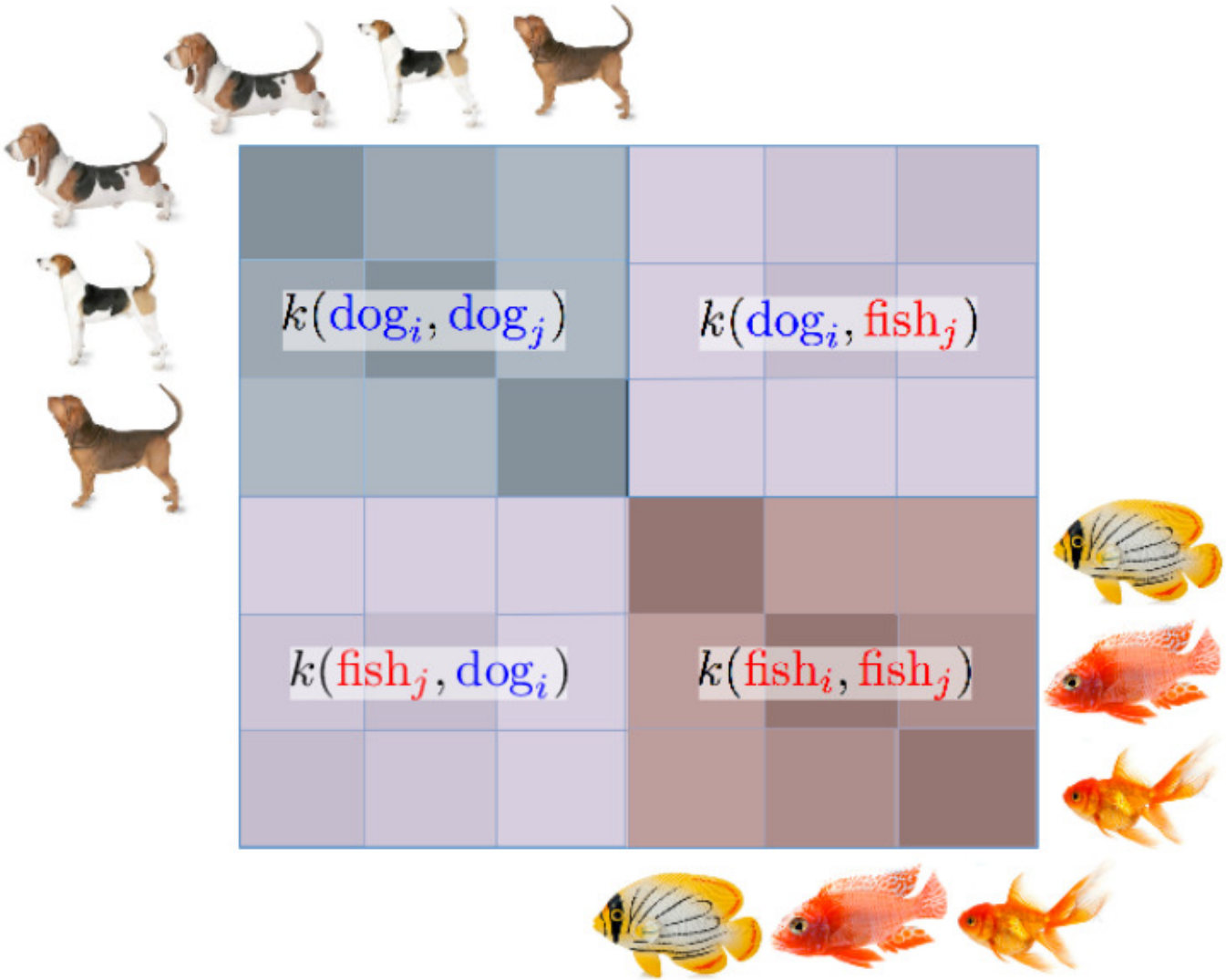
MMD estimator: intuition



MMD estimator: intuition



MMD estimator: intuition



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

† $\widehat{MMD}$ & $\widehat{HSIC}$ illustration credit: Arthur Gretton

MMD estimator

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\varphi(x)$.

MMD estimator

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\varphi(x)$.
- MMD = difference between feature means:

$$MMD^2(\mathbb{P},\mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k},$$

MMD estimator

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\varphi(x)$.
- MMD = difference between feature means:

$$MMD^2(\mathbb{P},\mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k},$$
$$\widehat{MMD^2_u}(\mathbb{P},\mathbb{Q}) = \overline{G_{\mathbb{P},\mathbb{P}}} + \overline{G_{\mathbb{Q},\mathbb{Q}}} - 2\overline{G_{\mathbb{P},\mathbb{Q}}}$$

using $\{x_i\}_{i=1}^m \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$ samples.

MMD estimator

- Feature of a distribution: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\varphi(x)$.
- MMD = difference between feature means:

$$MMD^2(\mathbb{P},\mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}_k},$$
$$\widehat{MMD^2_u}(\mathbb{P},\mathbb{Q}) = \overline{G_{\mathbb{P},\mathbb{P}}} + \overline{G_{\mathbb{Q},\mathbb{Q}}} - 2\overline{G_{\mathbb{P},\mathbb{Q}}}$$

  using $\{x_i\}_{i=1}^m \sim \mathbb{P}$, $\{y_j\}_{j=1}^n \sim \mathbb{Q}$ samples.
- Computational complexity: $\mathcal{O}\left((m+n)^2\right)$, quadratic.

Hilbert-Schmidt Independence Criterion (HSIC)

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

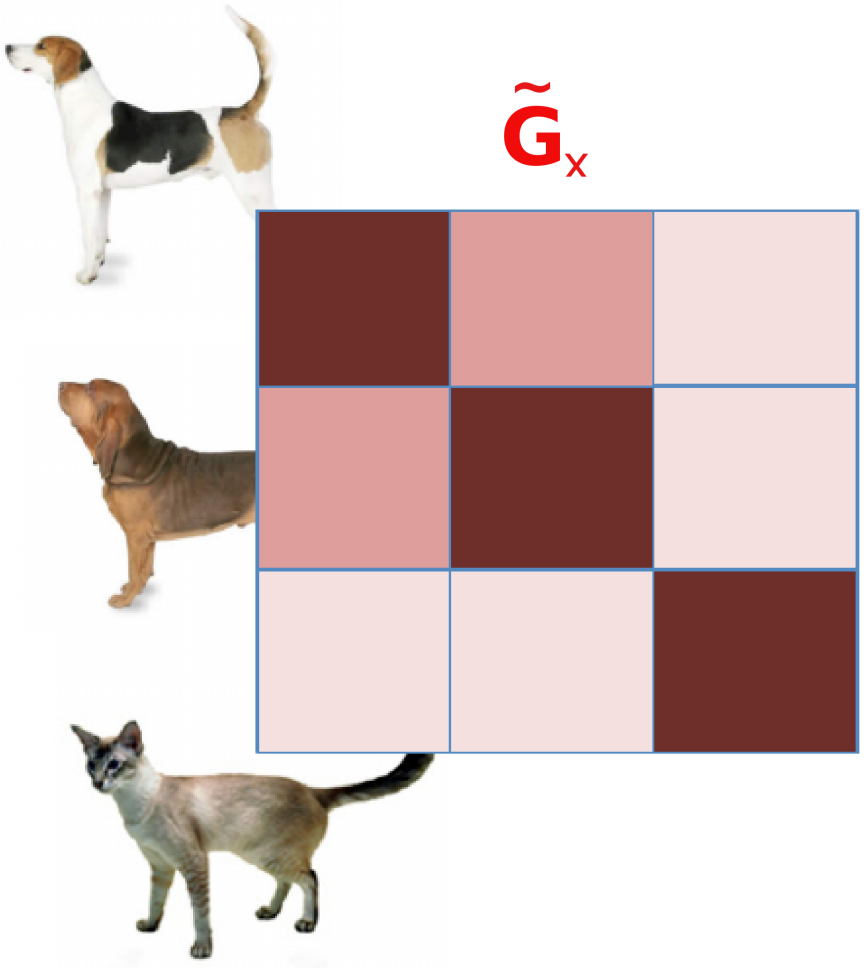
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

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

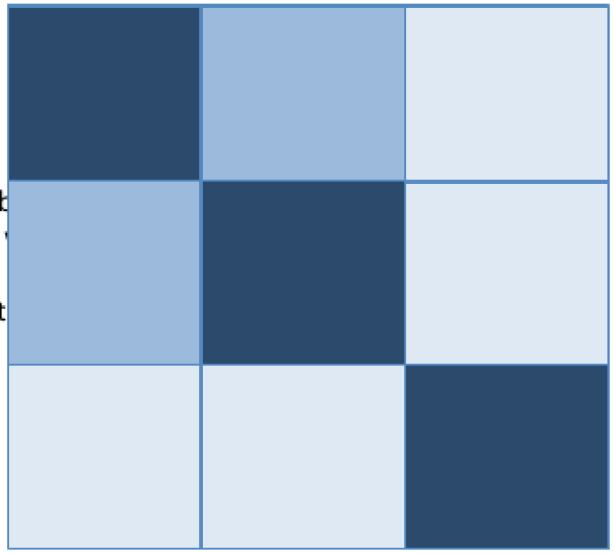


$\tilde{G}_x$

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.

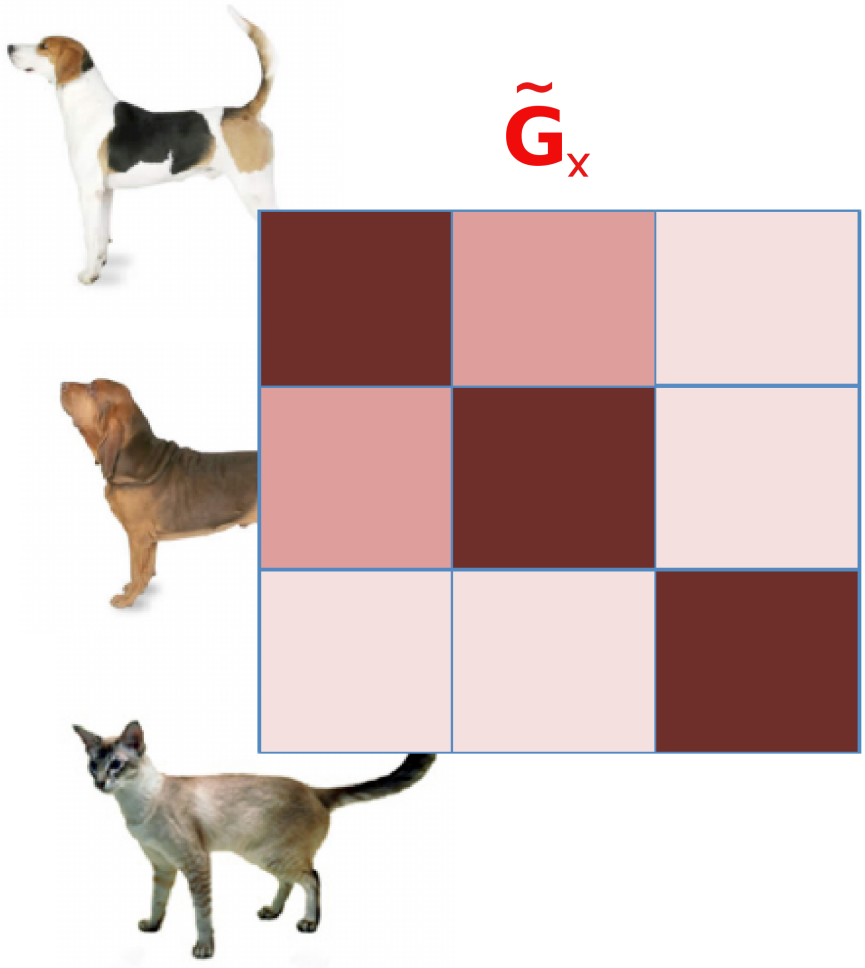
A large animal who slings slob
distinctive houndy odor, and
than to follow his nose. They
amount of exercise and ment

$\tilde{G}_y$



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.

HSIC intuition: Gram matrices



$\tilde{\mathbf{G}}_x$

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

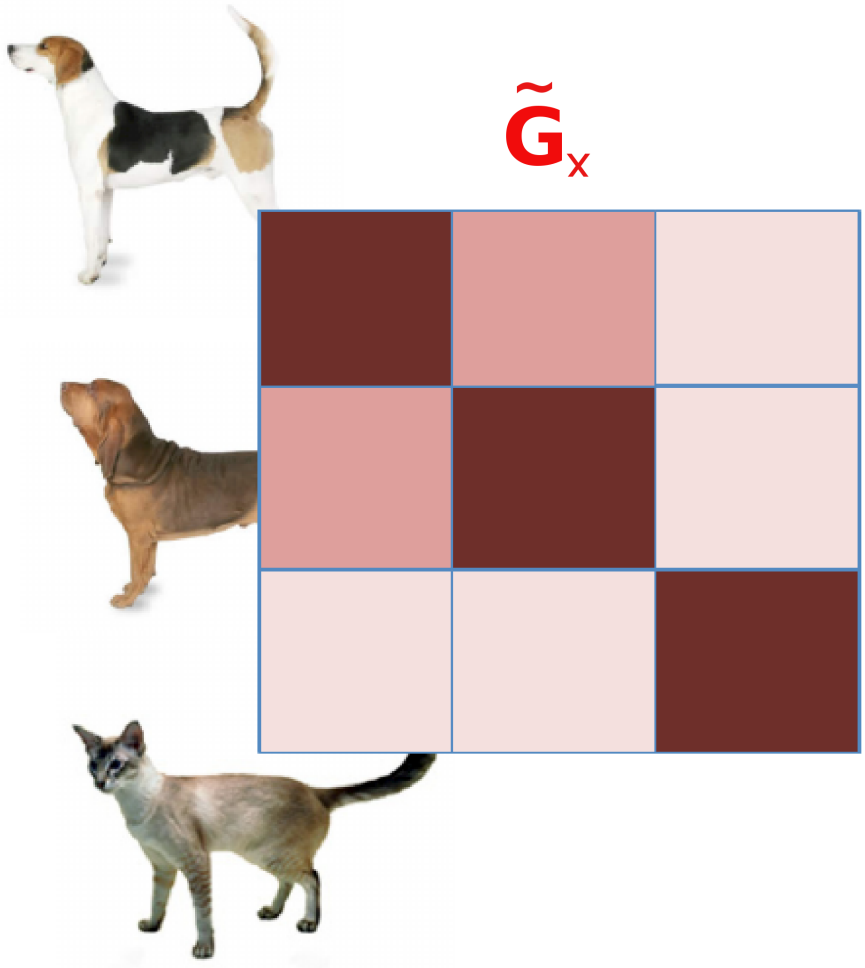
$\tilde{\mathbf{G}}_y$

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:

$$\widehat{HSIC^2} = \frac{1}{n^2} \left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F .$$

HSIC intuition: Gram matrices



$\tilde{\mathbf{G}}_x$

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

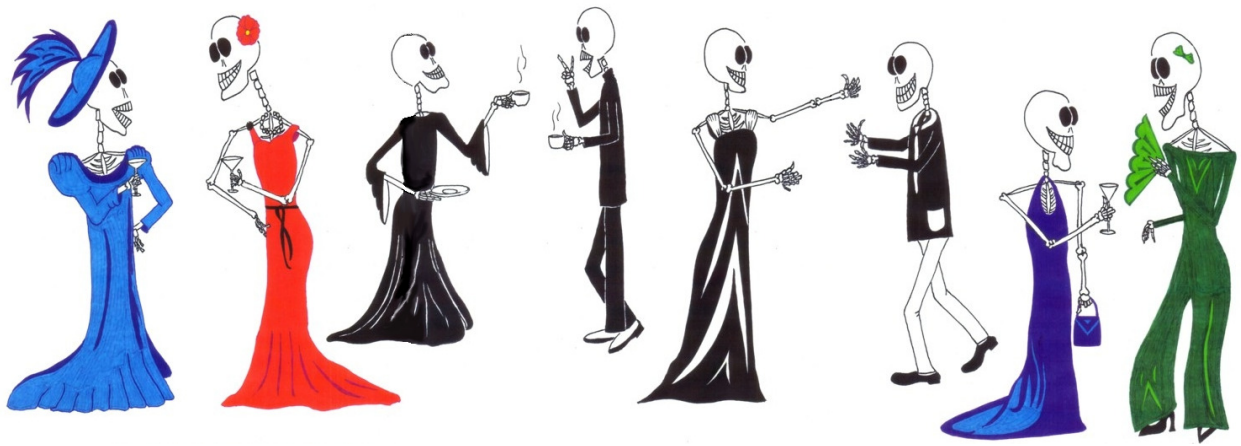
$\tilde{\mathbf{G}}_y$

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:

$$\widehat{HSIC^2} = \frac{1}{n^2}\left\langle \tilde{\mathbf{G}}_x, \tilde{\mathbf{G}}_y \right\rangle_F. \qquad HSIC(\mathbb{P}_{xy}) = MMD\left(\mathbb{P}_{xy}, \mathbb{P}_x \otimes \mathbb{P}_y\right).$$

Cocktail party: HSIC demo



ISA reminder

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right],$$

where $\mathbf{s}^m$-s are non-Gaussian & independent.

- Goal: $\{\mathbf{x}_t\}_{t=1}^T \to \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T$,

ISA reminder

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right],$$

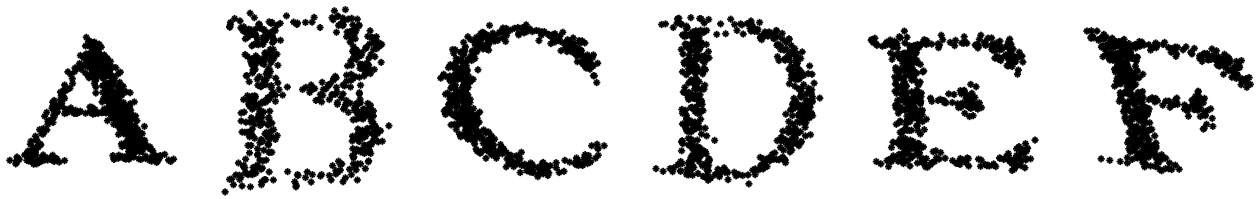
where $\mathbf{s}^m$-s are non-Gaussian & independent.

- Goal: $\{\mathbf{x}_t\}_{t=1}^T \to \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T$,
- Objective function:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x},$$
$$J(\mathbf{W}) = I\left(\hat{\mathbf{s}}^1, \ldots, \hat{\mathbf{s}}^M\right) \to \min_{\mathbf{W}}.$$

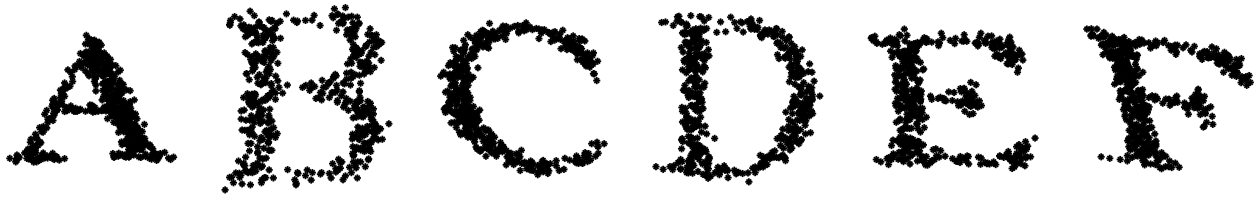
ISA: source, observation

- Hidden sources ($\mathbf{s}$):

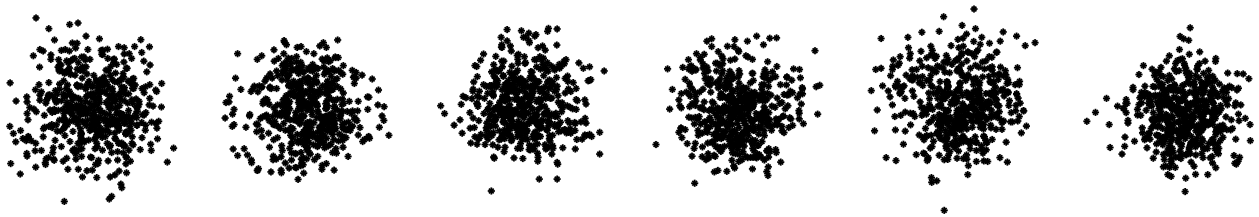


ISA: source, observation

- Hidden sources ($\mathbf{s}$):



- Observation ($\mathbf{x}$):



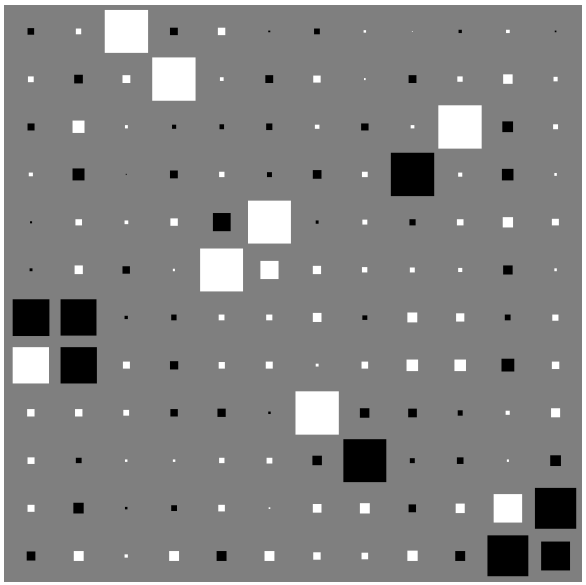
ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

- Performance ($\hat{\mathbf{W}}\mathbf{A}$), ambiguity:

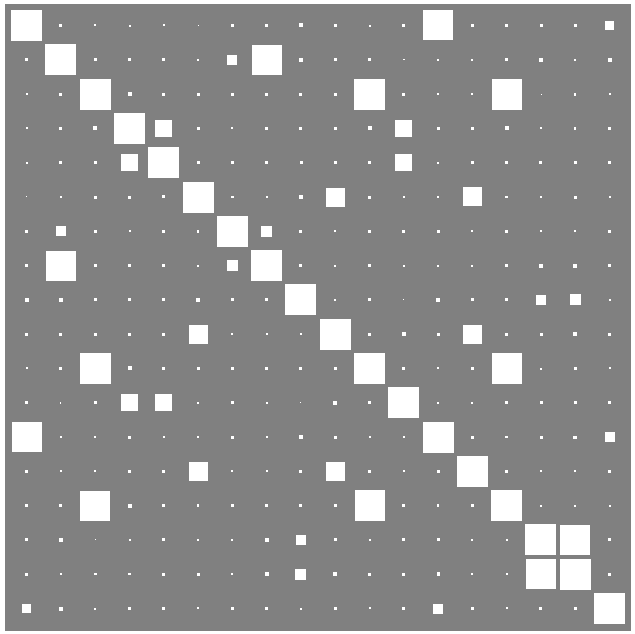


Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation.

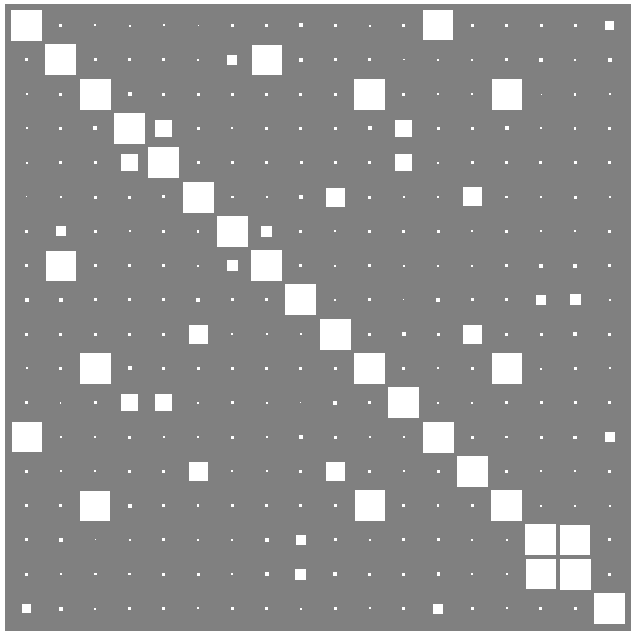
Conjecture: ISA separation theorem [Cardoso, 1998]

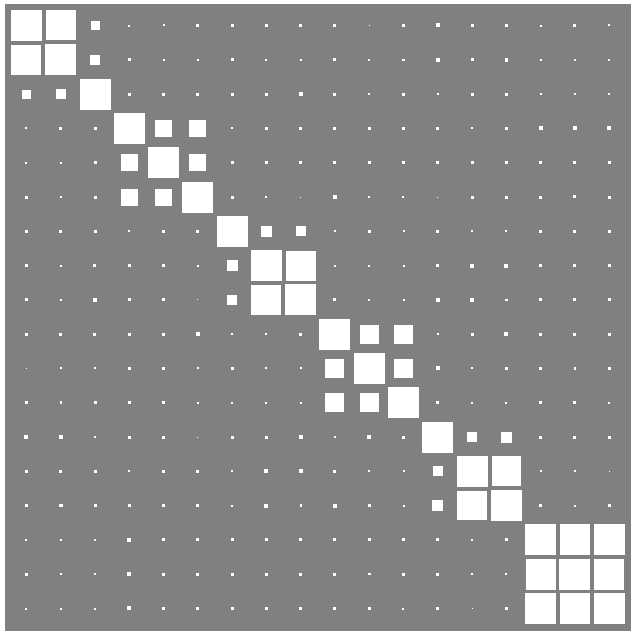
- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.



Conjecture: ISA separation theorem [Cardoso, 1998]

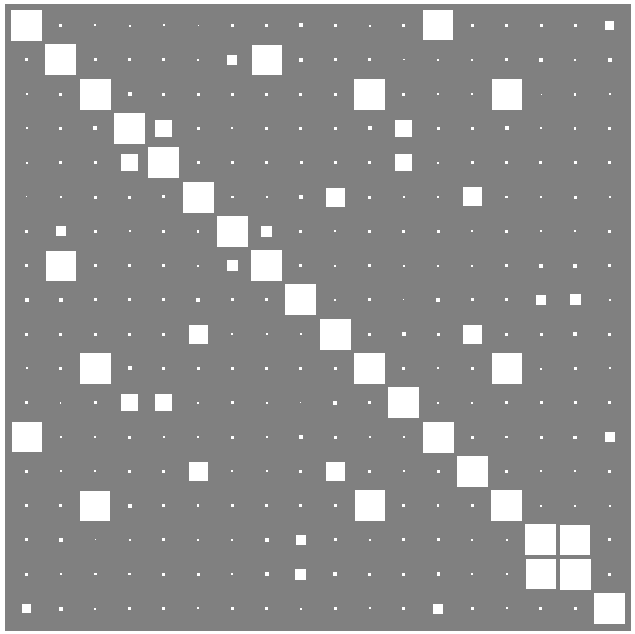
- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.

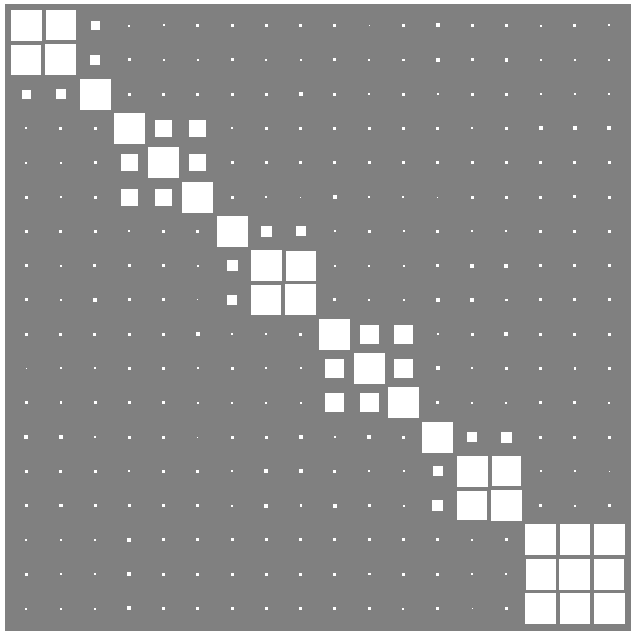




Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.

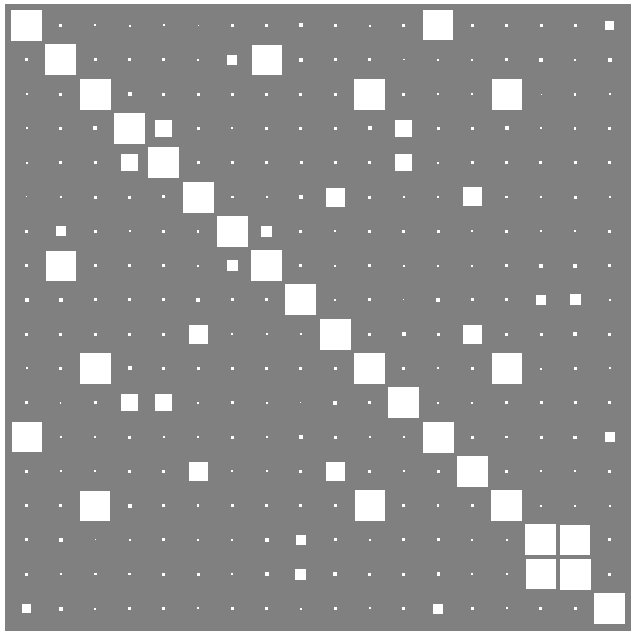


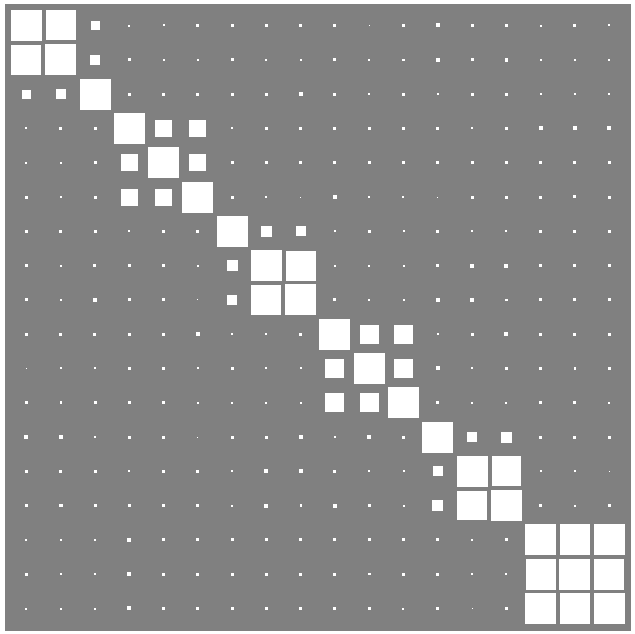


- Basis of the state-of-the-art ISA solvers.

Conjecture: ISA separation theorem [Cardoso, 1998]

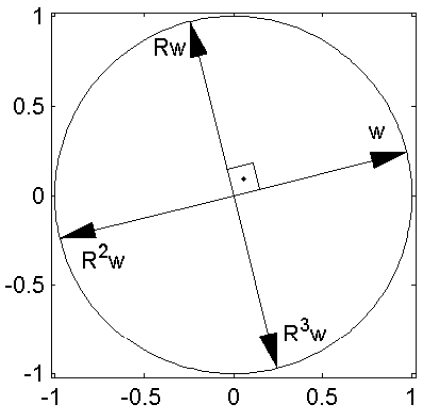
- ISA = ICA + permutation. $\widehat{HSIC}(\hat{s}_i, \hat{s}_j)$. Here: $dim(\mathbf{s}^m) = 3$.

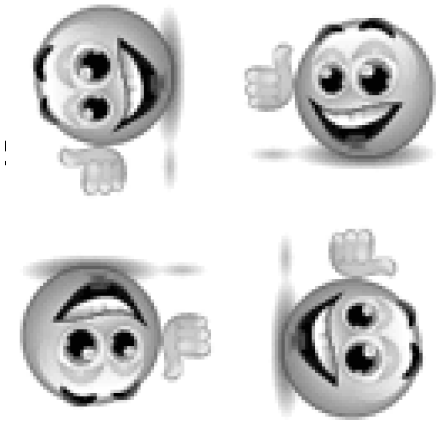




- Basis of the state-of-the-art ISA solvers.
- Sufficient conditions [Szabó et al., 2012]:
  - $\mathbf{s}^m$: spherical.

ISA separation theorem

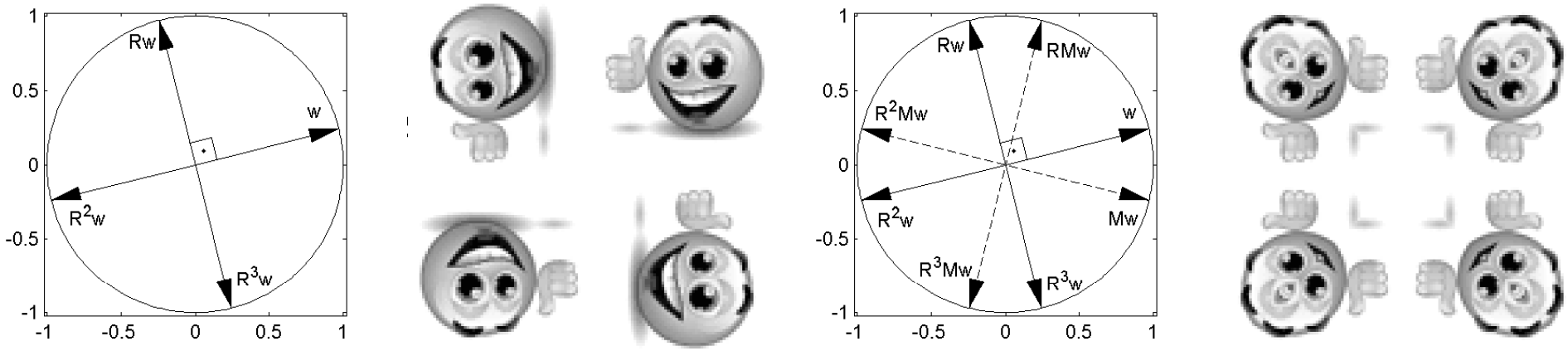




Invariance to

- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.

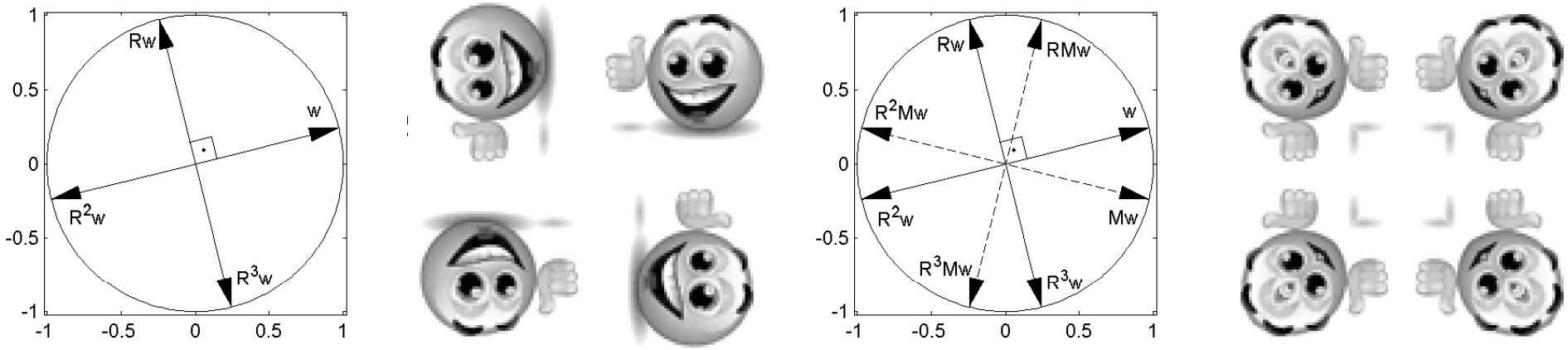
ISA separation theorem



Invariance to

- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.
- permutation and sign: $f(\pm u_1, \pm u_2) = f(\pm u_2, \pm u_1)$.

ISA separation theorem



Invariance to

- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.
- permutation and sign: $f(\pm u_1, \pm u_2) = f(\pm u_2, \pm u_1)$.
- L^p-spherical: $f(u_1, u_2) = h\left(\sum_i |u_i|^p\right) \quad (p > 0)$.

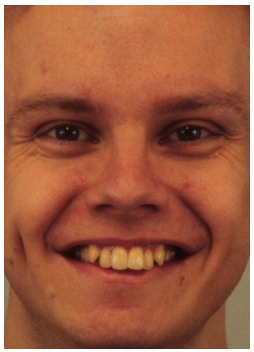
Mean embedding, MMD, HSIC: +applications & review

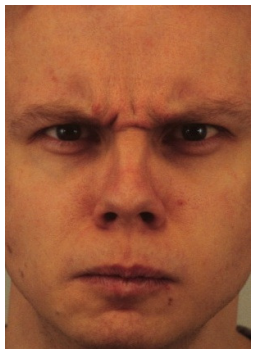
- Applications:
  - domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017],
  - interpretable machine learning [Kim et al., 2016],
  - kernel belief propagation [Song et al., 2011], kernel Bayes' rule [Fukumizu et al., 2013], model criticism [Lloyd et al., 2014],
  - approximate Bayesian computation [Park et al., 2016], probabilistic programming [Schölkopf et al., 2015],
  - distribution classification[Muandet et al., 2011], -regression [Szabó et al., 2016],
  - topological data analysis [Kusano et al., 2016],
  - post selection inference [Yamada et al., 2016],
  - causal inference [Mooij et al., 2016, Pfister et al., 2017, Strobl et al., 2017].
- Mean embedding review [Muandet et al., 2017].

Hypothesis Testing

Two-sample testing: recall

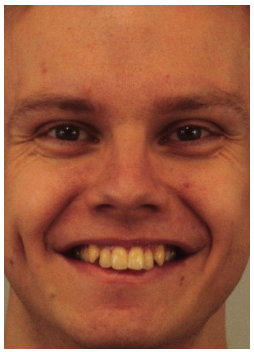
- Given:
  - $X = \{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \sim \mathbb{Q}$.
  - Example: $\mathbf{x}_i = i^{th}$ happy face, $\mathbf{y}_j = j^{th}$ sad face.

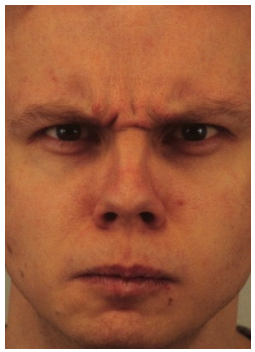




Two-sample testing: recall

- Given:
  - $X = \{\mathbf{x}_i\}_{i=1}^n \sim \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \sim \mathbb{Q}$.
  - Example: $\mathbf{x}_i = i^{th}$ happy face, $\mathbf{y}_j = j^{th}$ sad face.



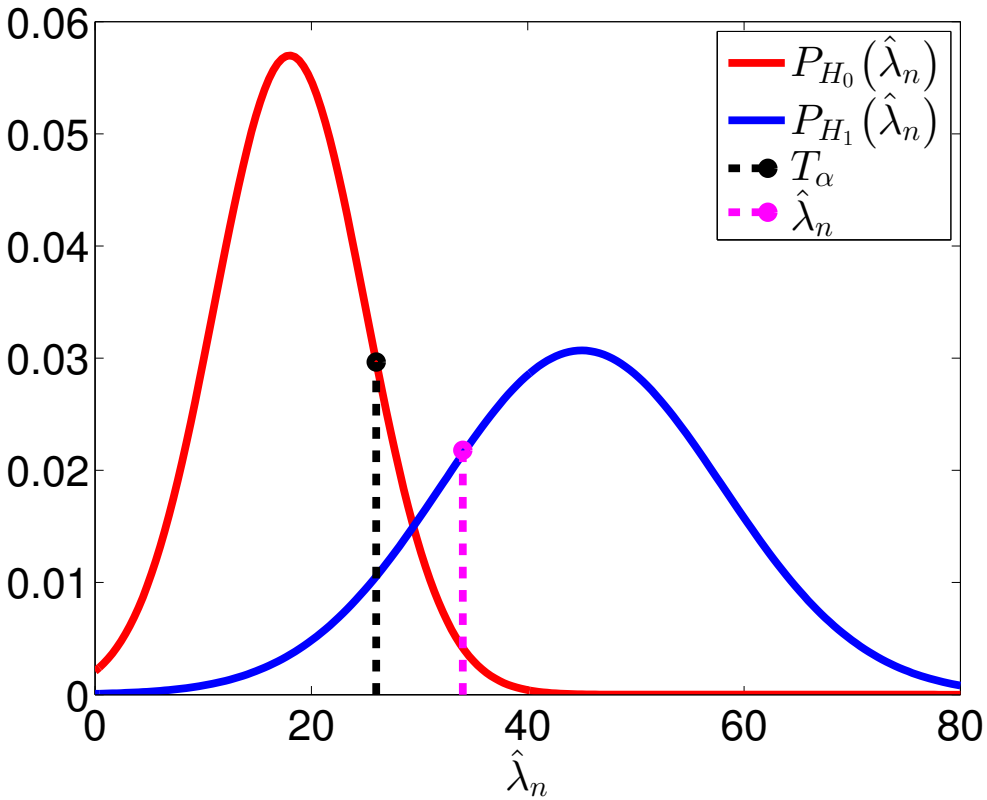


- Problem: using X, Y test

$$H_0 : \mathbb{P} = \mathbb{Q}, \text{ vs}$$
$$H_1 : \mathbb{P} \neq \mathbb{Q}.$$

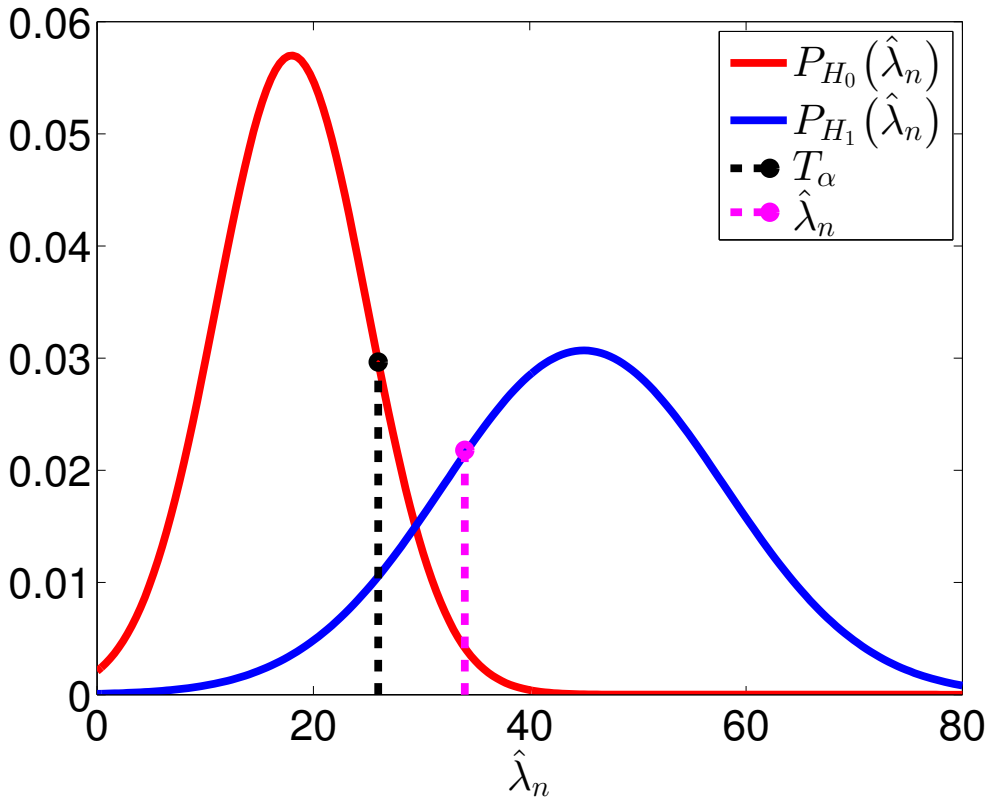
Ingredients of two-sample test

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Significance level: $\alpha = 0.01$.
- Under H_0: $P_{H_0}(\underbrace{\hat{\lambda}_n \leqslant T_\alpha}_{\text{correctly accepting } H_0}) = 1 - \alpha$.



Ingredients of two-sample test

- Test statistic: $\hat{\lambda}_n = \hat{\lambda}_n(X, Y)$, random.
- Significance level: $\alpha = 0.01$.
- Under H_0: $P_{H_0}(\underbrace{\hat{\lambda}_n \leqslant T_\alpha}_{\text{correctly accepting } H_0}) = 1 - \alpha$.
- Under H_1: $P_{H_1}(T_\alpha < \hat{\lambda}_n) = P(\text{correctly rejecting } H_0) =: \text{power}$.



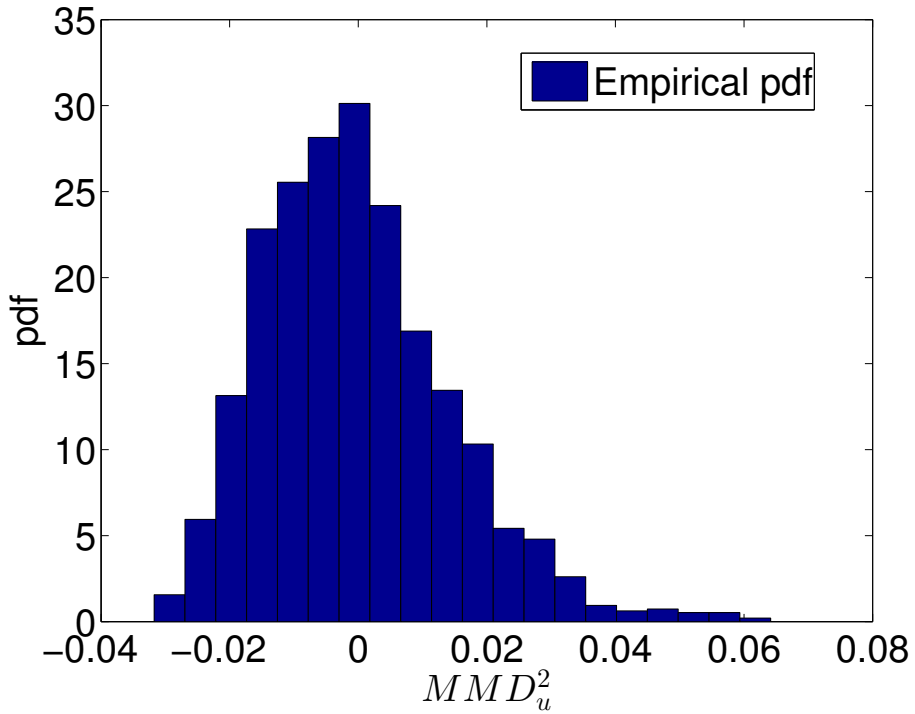
Two-sample test using MMD asymptotics: H_0

Under H_0 [Gretton et al., 2007, Gretton et al., 2012] $\xrightarrow{\text{U-statistics}}$

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

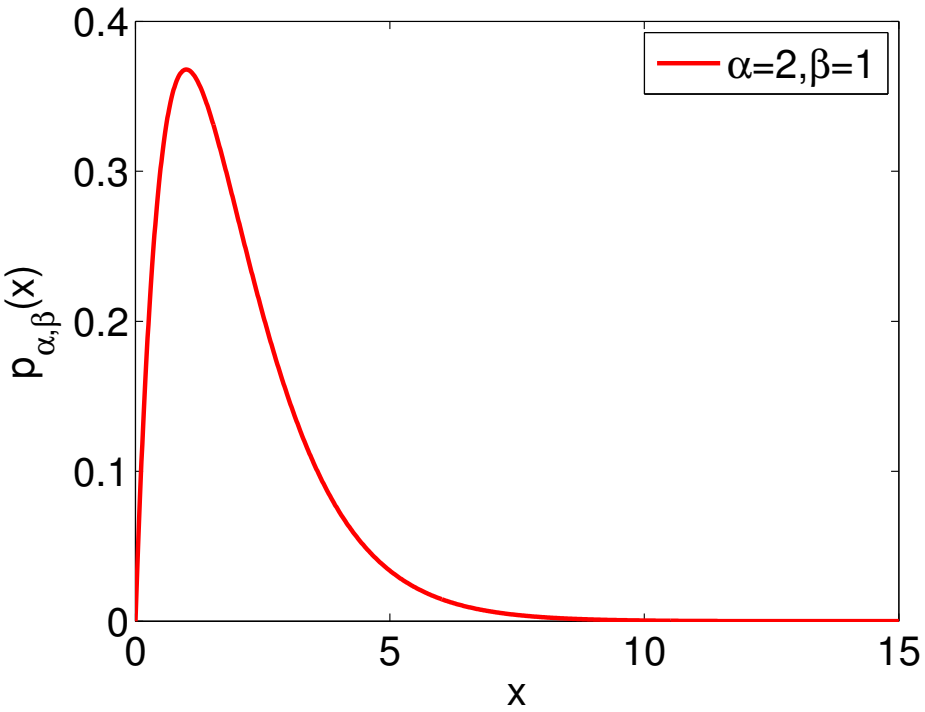
where $z_i \sim N(0,2)$ i.i.d.,

$$\int_{\mathcal{X}} \tilde{k}(x,x')v_i(x)\mathrm{d}\mathbb{P}(x) = \lambda_i v_i(x'),\ \tilde{k}(x,x') = \left\langle \varphi(x) - \mu_{\mathbb{P}}, \varphi(x') - \mu_{\mathbb{P}} \right\rangle_{\mathcal{H}_k}.$$



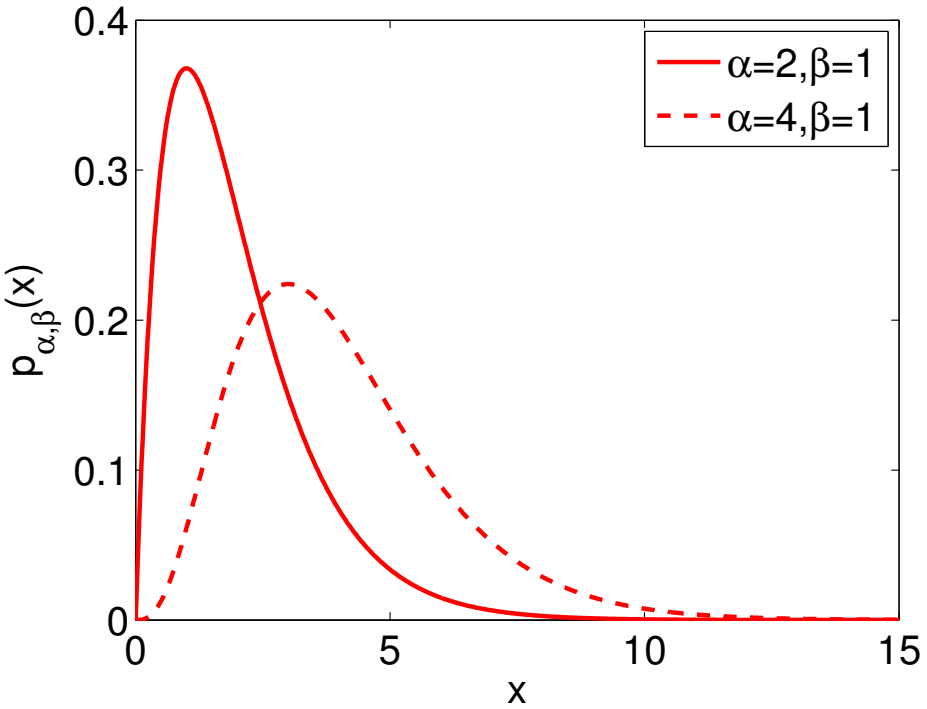
Null approximations; test statistics: quadratic in time

- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



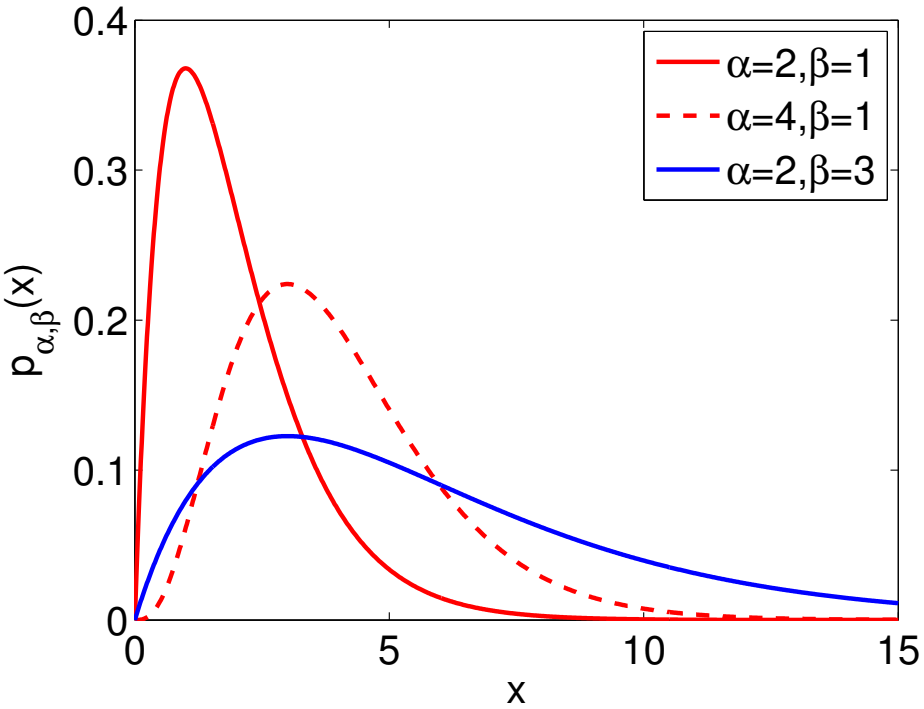
Null approximations; test statistics: quadratic in time

- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



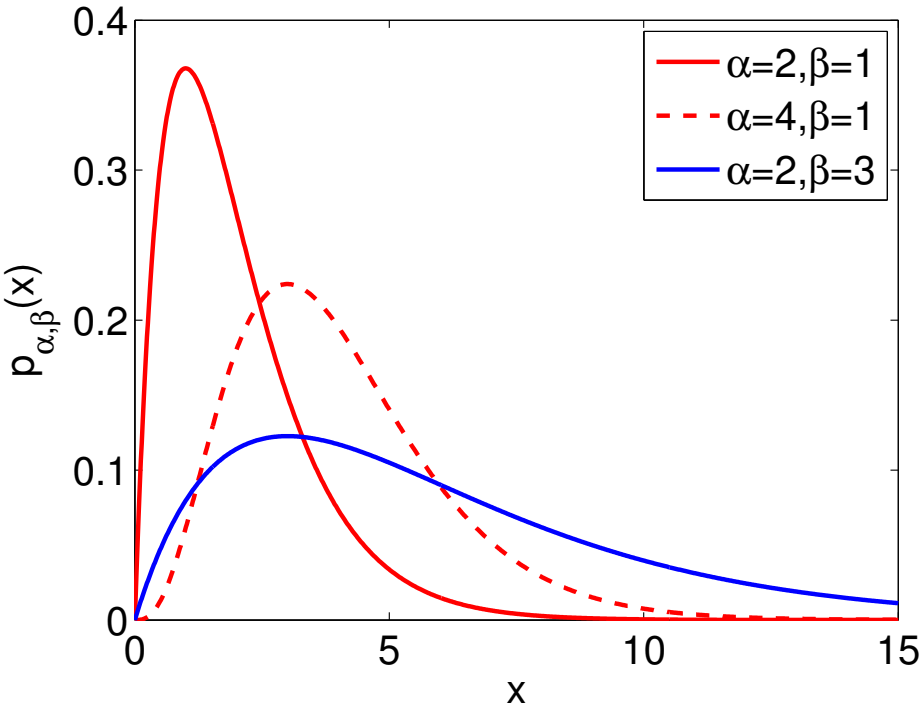
Null approximations; test statistics: quadratic in time

- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



Null approximations; test statistics: quadratic in time

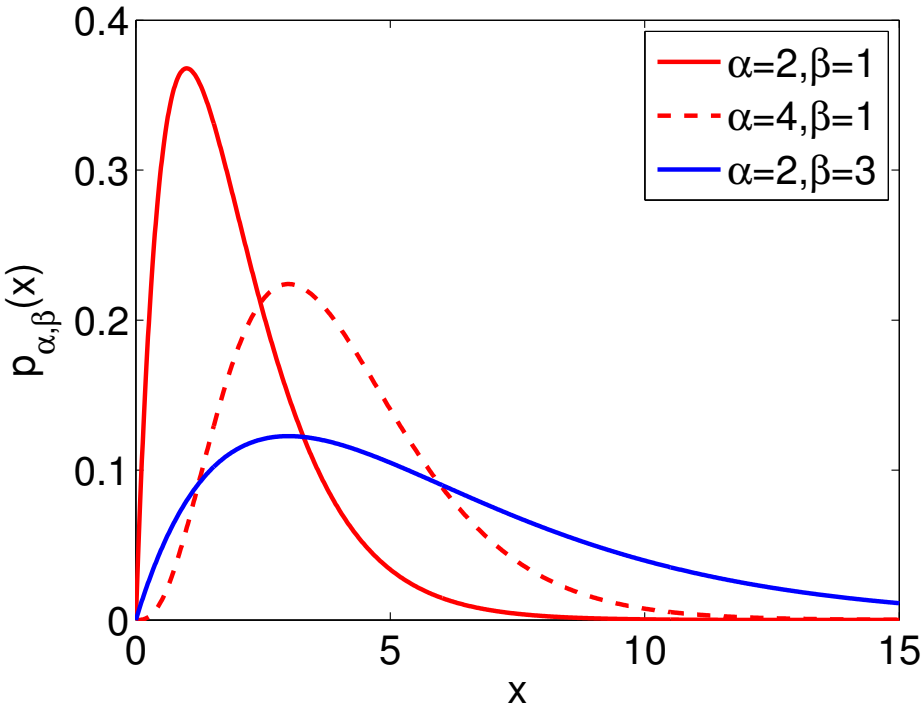
- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



  - truncated expansion [Gretton et al., 2009].

Null approximations; test statistics: quadratic in time

- Small sample size: permutation test.
- Medium sample size:
  - gamma approximation:



  - truncated expansion [Gretton et al., 2009].
- Large sample size:
  - online techniques [Gretton et al., 2012] (large *var*),
  - recent linear methods (soon).

Independence testing with HSIC

Similary story [Gretton et al., 2008, Pfister et al., 2016]:

- Null asymptotics:

$$\sum_{i=1}^{\infty} \lambda_i z_i^2, \quad z_i \sim N(0,1).$$

- In practice: permutation-test/gamma-approximation.

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - *var* $\searrow$

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - *var* $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013].

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - *var* $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013].
- Goodness-of-fit [Chwialkowski et al., 2016].

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - *var* $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013].
- Goodness-of-fit [Chwialkowski et al., 2016].
- Time-series:
  - independence (stationary $\rightarrow$ shift) [Chwialkowski and Gretton, 2014],
  - wild bootstrap: [Chwialkowski et al., 2014, Rubenstein et al., 2016].

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - *var* $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013].
- Goodness-of-fit [Chwialkowski et al., 2016].
- Time-series:
  - independence (stationary $\rightarrow$ shift) [Chwialkowski and Gretton, 2014],
  - wild bootstrap: [Chwialkowski et al., 2014, Rubenstein et al., 2016].
- block-HSIC [Zhang et al., 2017]:
  - RFF acceleration.

Related work

- 2-sample testing: block-MMD [Zaremba et al., 2013]
  - *var* $\searrow$
- 3-variable interaction [Sejdinovic et al., 2013].
- Goodness-of-fit [Chwialkowski et al., 2016].
- Time-series:
  - independence (stationary $\rightarrow$ shift) [Chwialkowski and Gretton, 2014],
  - wild bootstrap: [Chwialkowski et al., 2014, Rubenstein et al., 2016].
- block-HSIC [Zhang et al., 2017]:
  - RFF acceleration.
- Conditional independence & RFF [Strobl et al., 2017].

Linear-time Tests

Linear-time 2-sample test [Chwialkowski et al., 2015]

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

Linear-time 2-sample test [Chwialkowski et al., 2015]

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

- Idea: change this to

$$\rho(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}$$

with random $\{\mathbf{v}_j\}_{j=1}^{J}$ test locations.

Linear-time 2-sample test [Chwialkowski et al., 2015]

- Recall:

$$MMD(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}.$$

- Idea: change this to

$$\rho(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\mu_{\mathbb{P}}(\mathbf{v}_j) - \mu_{\mathbb{Q}}(\mathbf{v}_j)]^2}$$

with random $\{\mathbf{v}_j\}_{j=1}^{J}$ test locations.

ρ is a metric for Gaussian k (bounded, analytic, characteristic), a.s.

Estimation

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

$$\bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}[\underbrace{k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)}_{\mathbf{z}_i}]_{j=1}^{J} \in \mathbb{R}^J.$$

Estimation

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

$$\bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}\underbrace{[k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)]_{j=1}^{J}}_{\mathbf{z}_i} \in \mathbb{R}^J.$$

- Good news: estimation is linear in n!
- Bad news: intractable null distr. $= \sqrt{n}\widehat{\rho^2}(\mathbb{P},\mathbb{P}) \xrightarrow{d}$ sum of J correlated χ^2.

Estimation

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

$$\bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}\underbrace{[k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)]_{j=1}^{J}}_{\mathbf{z}_i} \in \mathbb{R}^J.$$

- Good news: estimation is linear in n!
- Bad news: intractable null distr. $= \sqrt{n}\widehat{\rho^2}(\mathbb{P},\mathbb{P}) \xrightarrow{d}$ sum of J correlated χ^2.

- Modified statistic (whitening):

$$\hat{\lambda}_n = n\bar{\mathbf{z}}_n^T\boldsymbol{\Sigma}_n^{-1}\bar{\mathbf{z}}_n, \qquad \boldsymbol{\Sigma}_n = cov\left(\{\mathbf{z}_i\}_{i=1}^n\right).$$

Estimation

$$\widehat{\rho^2}(\mathbb{P},\mathbb{Q}) = \frac{1}{J}\sum_{j=1}^{J}[\hat{\mu}_{\mathbb{P}}(\mathbf{v}_j) - \hat{\mu}_{\mathbb{Q}}(\mathbf{v}_j)]^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

$$\bar{\mathbf{z}}_n = \frac{1}{n}\sum_{i=1}^{n}[\underbrace{k(\mathbf{x}_i,\mathbf{v}_j) - k(\mathbf{y}_i,\mathbf{v}_j)}_{\mathbf{z}_i}]_{j=1}^{J} \in \mathbb{R}^J.$$

- Good news: estimation is linear in n!
- Bad news: intractable null distr. $= \sqrt{n}\widehat{\rho^2}(\mathbb{P},\mathbb{P}) \xrightarrow{d}$ sum of J correlated χ^2.

- Modified statistic (whitening):

$$\hat{\lambda}_n = n\bar{\mathbf{z}}_n^T\boldsymbol{\Sigma}_n^{-1}\bar{\mathbf{z}}_n, \qquad \boldsymbol{\Sigma}_n = cov\left(\{\mathbf{z}_i\}_{i=1}^n\right).$$

- Under H_0: $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$. $\Rightarrow$ Easy to get the $(1-\alpha)$-quantile!

Optimize locations and kernel

- Test locations ($\{\mathbf{v}_j\}_{j=1}^J$), kernel parameters: fixed.
- Idea [Jitkrittum et al., 2016]: optimize them for a power proxy

$$\lambda_n = n\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} \quad \text{(population } \hat{\lambda}_n\text{)}.$$

In practice

Power is similar to the $\mathcal{O}(n^2)$ MMD, but in $\mathcal{O}(n)$ time!

Optimize locations and kernel

- Test locations ($\{\mathbf{v}_j\}_{j=1}^J$), kernel parameters: fixed.
- Idea [Jitkrittum et al., 2016]: optimize them for a power proxy

$$\lambda_n = n\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} \quad \text{(population } \hat{\lambda}_n\text{)}.$$

In practice

Power is similar to the $\mathcal{O}(n^2)$ MMD, but in $\mathcal{O}(n)$ time!

- Independence testing [Jitkrittum et al., 2017a].

Optimize locations and kernel

- Test locations ($\{\mathbf{v}_j\}_{j=1}^J$), kernel parameters: fixed.
- Idea [Jitkrittum et al., 2016]: optimize them for a power proxy

$$\lambda_n = n\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z} \quad (\text{population } \hat{\lambda}_n).$$

In practice

Power is similar to the $\mathcal{O}(n^2)$ MMD, but in $\mathcal{O}(n)$ time!

- Independence testing [Jitkrittum et al., 2017a].
- Goodness-of-fit [Jitkrittum et al., 2017b]: Best Paper Award @ NIPS-2018!

Demo-1 = NLP: most/least discriminative words

- NIPS papers (1988-2015). TF-IDF repr. 'Bayes-Neuro'.
- Most discriminative words:

  spike, markov, cortex, dropout, recurr, iii, gibb.

  - learned test locations: highly interpretable,
  - 'markov', 'gibb' ($\Leftarrow$ Gibbs): Bayesian inference,
  - 'spike', 'cortex': key terms in neuroscience.

Demo-1 = NLP: most/least discriminative words

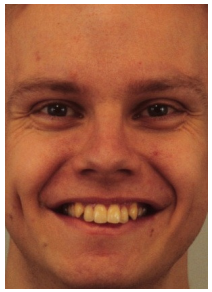
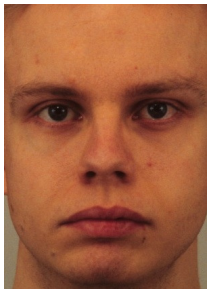
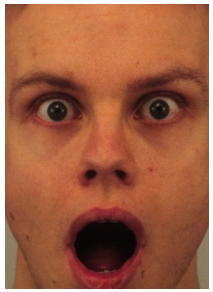
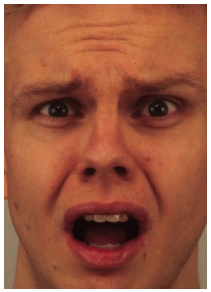
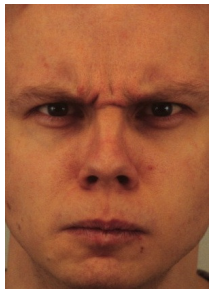
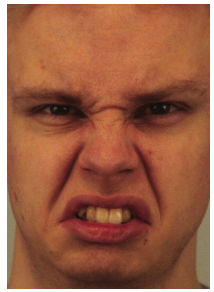
- NIPS papers (1988-2015). TF-IDF repr. 'Bayes-Neuro'.

- Least dicriminative ones:

  circumfer, bra, dominiqu, rhino, mitra, kid, impostor.

Demo-2: Distinguish positive/negative emotions

- Karolinska Directed Emotional Faces (KDEF) [Lundqvist et al., 1998].
- 70 actors = 35 females and 35 males.
- $d = 48 \times 34 = 1632$. Grayscale. Pixel features.

$+$:    $-$:   

happy neutral surprised afraid angry disgusted

- Learned test location (averaged) = 

Summary

- Dependency measures, distance: KCCA, HSIC, MMD.
- Tool: kernels.

Summary

- Dependency measures, distance: KCCA, HSIC, MMD.
- Tool: kernels.
- Applications:
  - image registration, ISA, distribution regression, feature selection.
  - hypothesis testing.

Summary

- Dependency measures, distance: KCCA, HSIC, MMD.
- Tool: kernels.
- Applications:
  - image registration, ISA, distribution regression, feature selection.
  - hypothesis testing.
- MMD, HSIC: $\mathcal{O}(n^2)$.
- Linear-time tests ($\mathbb{R}^d$).

Thank you for the attention!


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


Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2017).
Domain generalization by marginal transfer learning.
Technical report.
`https://arxiv.org/abs/1711.07910`.


Cardoso, J.-F. (1998).
Multidimensional independent component analysis.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1941–1944.


Chwialkowski, K. and Gretton, A. (2014).
A kernel independence test for random processes.
In *International Conference on Machine Learning (ICML; JMLR W&CP)*, volume 32, page 14221430.


Chwialkowski, K., Ramdas, A., Sejdinovic, D., and Gretton, A. (2015).
Fast two-sample testing with analytic representations of probability measures.
In *Neural Information Processing Systems (NIPS)*, pages 1972–1980.


Chwialkowski, K., Sejdinovic, D., and Gretton, A. (2014).
A wild bootstrap for degenerate kernel tests.
In *Neural Information Processing Systems (NIPS)*, pages 3608–3616.


Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *International Conference on Machine Learning (ICML)*, pages 2606–2615.


Fukumizu, K., Song, L., and Gretton, A. (2013).
Kernel Bayes' rule: Bayesian inference with positive definite kernels.

Journal of Machine Learning Research, 14:3753–3783.


Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research, 13:723–773.


Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009).
A fast, consistent kernel two-sample test.
In *Neural Information Processing Systems (NIPS)*, pages 673–681.


Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007).
A kernel statistical test of independence.
In *Neural Information Processing Systems (NIPS)*, pages 585–592.


Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008).

A kernel statistical test of independence.
In *Neural Information Processing Systems (NIPS)*, pages 585–592.


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


Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016).
Interpretable distribution features with maximum testing power.
In *Neural Information Processing Systems (NIPS)*, pages 181–189.


Jitkrittum, W., Szabó, Z., and Gretton, A. (2017a).
An adaptive test of independence with analytic kernel embeddings.
In *International Conference on Machine Learning (ICML)*, volume 70, pages 1742–1751. PMLR.


Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017b).
A linear-time kernel goodness-of-fit test.
In *Advances in Neural Information Processing Systems (NIPS)*.


Kim, B., Khanna, R., and Koyejo, O. O. (2016).
Examples are not enough, learn to criticize! criticism for interpretability.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2280–2288.


Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016).
Persistence weighted Gaussian kernel for topological data analysis.
In *International Conference on Machine Learning (ICML)*, pages 2004–2013.


Kybic, J. (2004).
High-dimensional mutual information estimation for image registration.

In *IEEE International Conference on Image Processing (ICIP)*, pages 1779–1782.


Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2014).
Automatic construction and natural-language description of nonparametric regression models.
In *AAAI Conference on Artificial Intelligence*, pages 1242–1250.


Lundqvist, D., Flykt, A., and Öhman, A. (1998).
The Karolinska directed emotional faces-KDEF.
Technical report, ISBN 91-630-7164-9.


Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016).
Distinguishing cause from effect using observational data: Methods and benchmarks.
Journal of Machine Learning Research, 17:1–102.


Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2011).
Learning from distributions via support measure machines.
In *Neural Information Processing Systems (NIPS)*, pages 10–18.


Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017).
Kernel mean embedding of distributions: A review and beyond.
Technical report.
`https://arxiv.org/abs/1605.09522`.


Neemuchwala, H., Hero, A., Zabuawala, S., and Carson, P. (2007).
Image registration methods in high dimensional space.
International Journal of Imaging Systems and Technology, 16:130–145.


Park, M., Jitkrittum, W., and Sejdinovic, D. (2016).

K2-ABC: Approximate Bayesian computation with kernel embeddings.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 51, pages 51:398–407.


Peng, H., Long, F., and Ding, C. (2005).
Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8):1226–1238.


Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2016).
Kernel-based tests for joint independence.
Technical report.
(`https://arxiv.org/abs/1603.00285`).


Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2017).
Kernel-based tests for joint independence.

Journal of the Royal Statistical Society: Series B (Statistical Methodology).


Póczos, B., Singh, A., Rinaldo, A., and Wasserman, L. (2013).
Distribution-free distribution regression.
In *International Conference on AI and Statistics (AISTATS; JMLR W&CP)*, volume 31, pages 507–515.


Rényi, A. (1959).
On measures of dependence.
Acta Mathematica Academiae Scientiarum Hungaricae, 10:441–451.


Rubenstein, P. K., Chwialkowski, K. P., and Gretton, A. (2016).
A kernel test for three-variable interactions with random processes.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 637–646.


Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S., and Peters, J. (2015).
Computing functions of random variables via reproducing kernel Hilbert space representations.
Statistics and Computing, 25(4):755–766.


Sejdinovic, D., Gretton, A., and Bergsma, W. (2013).
A kernel test for three-variable interactions.
In *Neural Information Processing Systems (NIPS)*, pages 1124–1132.


Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. (2011).
Kernel belief propagation.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715.


Strobl, E. V., Visweswaran, S., and Zhang, K. (2017).
Approximate kernel-based conditional independence tests for fast non-parametric causal discovery.

Technical report.
https://arxiv.org/abs/1702.03877.


Szabó, Z., Póczos, B., and Lörincz, A. (2012).
Separation theorem for independent subspace analysis and its consequences.
Pattern Recognition, 45(4):1782–1791.


Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.
Journal of Machine Learning Research, 17(152):1–40.


Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2016).
Post selection inference with kernels.
Technical report.
(https://arxiv.org/abs/1610.03725).


Zaremba, W., Gretton, A., and Blaschko, M. (2013).
B-tests: Low variance kernel two-sample tests.

In *NIPS*, pages 755–763.


Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013).
Domain adaptation under target and conditional shift.
Journal of Machine Learning Research, 28(3):819–827.


Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2017).
Large-scale kernel methods for independence testing.
Statistics and Computing, pages 1–18.