A linear-time adaptive nonparametric two-sample test

Zoltán Szabó (CMAP, École Polytechnique)

Wittawat Jitkrittum

Kacper Chwialkowski

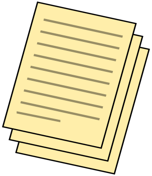
Arthur Gretton

Signal Processing and Machine Learning Seminar
Marseilles
March 24, 2017

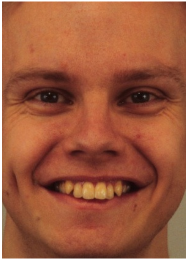
Motivating examples

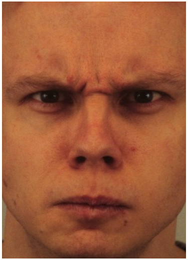
Motivating example-1: NLP

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




Motivating example-2: computer vision





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

One-page summary

- We propose a nonparametric t-test.
- It gives a reason why H_0 is rejected.
- It is
  - adaptive → high test power.
  - fast (linear time).

One-page summary

- We propose a nonparametric t-test.
- It gives a reason why H_0 is rejected.
- It is
  - adaptive → high test power.
  - fast (linear time).

Paper, code:

- NIPS [Jitkrittum et al., 2016].
- `https://github.com/wittawatj/interpretable-test`.

Two-sample test, distribution features

What is a two-sample test?

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

What is a two-sample test?

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

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

What is a two-sample test?

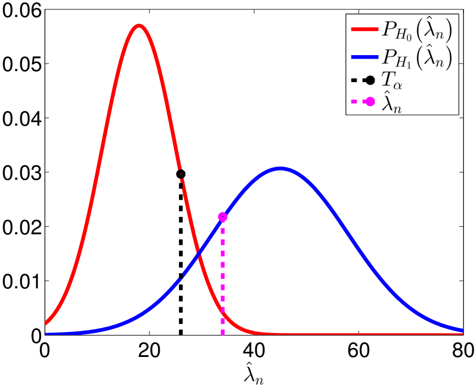
- Given:
  - $X = \{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$, $Y = \{\mathbf{y}_j\}_{j=1}^n \overset{i.i.d.}{\sim} \mathbb{Q}$.
  - Example: $\mathbf{x}_i = i^{th}$ happy face, $\mathbf{y}_j = j^{th}$ sad face.
- Problem: using X, Y test

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

- Assume $X, Y \subset \mathbb{R}^d$.

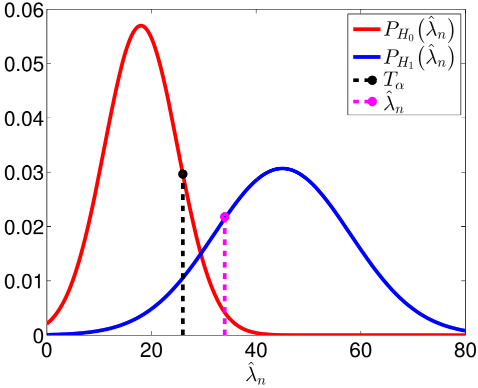
Ingredients of two-sample test

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



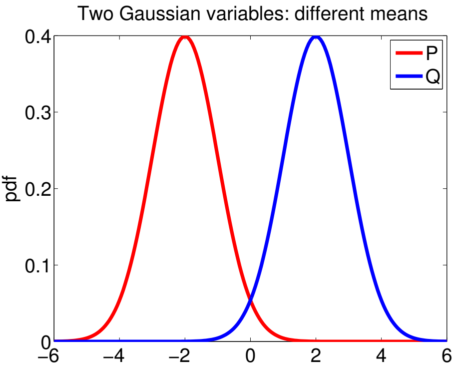
Ingredients of two-sample test

- 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: $P_{H_1}(T_\alpha < \hat{\lambda}_n) = P(\text{correctly rejecting } H_0) =: \text{power}$.



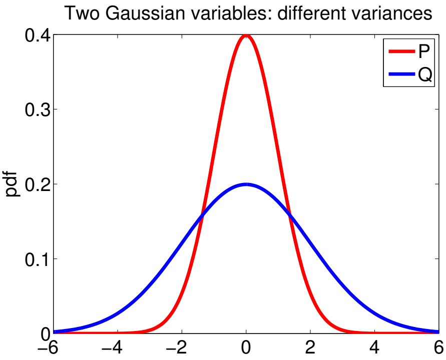
Towards representations of distributions: $\mathbb{E}X$

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



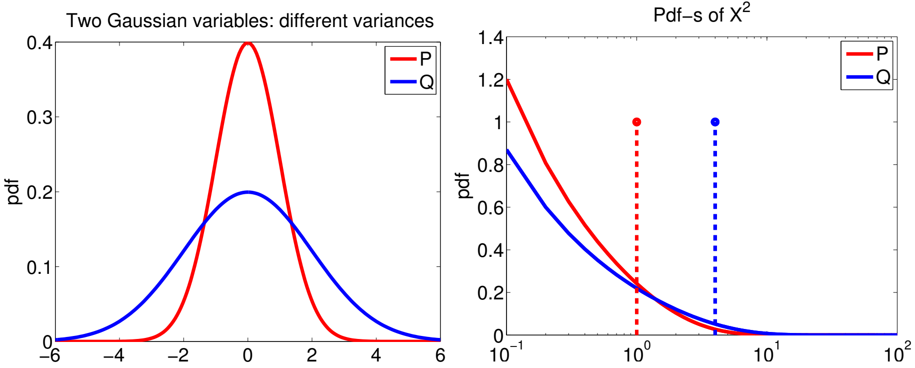
Towards representations of distributions: $\mathbb{E}X^2$

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



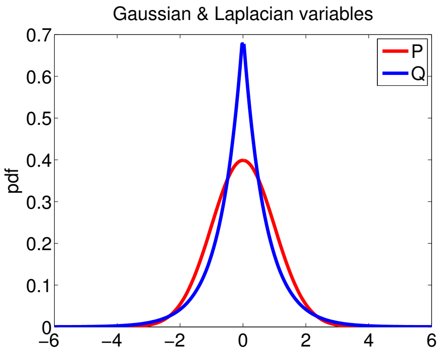
Towards representations of distributions: $\mathbb{E}X^2$

- 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: $\mathbf{x}$ and $\mathbf{x}'$ objects (images or texts).

Kernel: similarity between features

- Given: $\mathbf{x}$ and $\mathbf{x}'$ objects (images or texts).
- Question: how similar they are?

Kernel: similarity between features

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

$$\varphi_{\mathbf{x}} : \text{features of } \mathbf{x},$$
$$\varphi_{\mathbf{x}'} : \text{features of } \mathbf{x}'.$$

- Kernel: inner product of these features

$$k(\mathbf{x}, \mathbf{x}') := \langle \varphi_{\mathbf{x}}, \varphi_{\mathbf{x}'} \rangle .$$

Kernel examples on $\mathbb{R}^d$ ($\gamma > 0, p \in \mathbb{Z}^+$)

- Polynomial kernel:

$$k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p.$$

- Gaussian kernel:

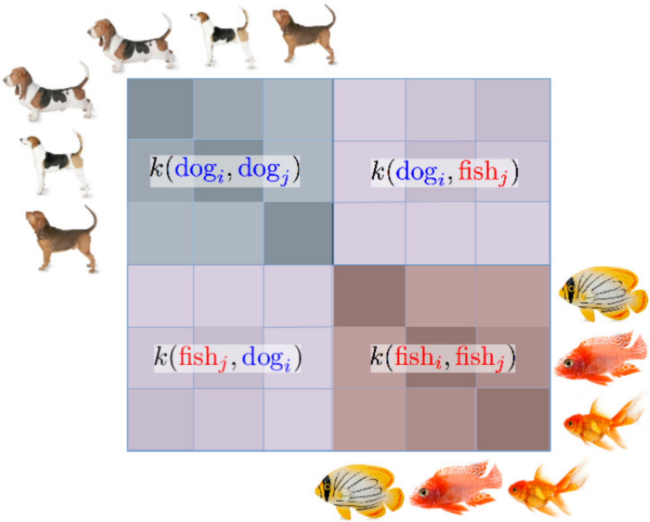
$$k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}.$$

Towards distribution features

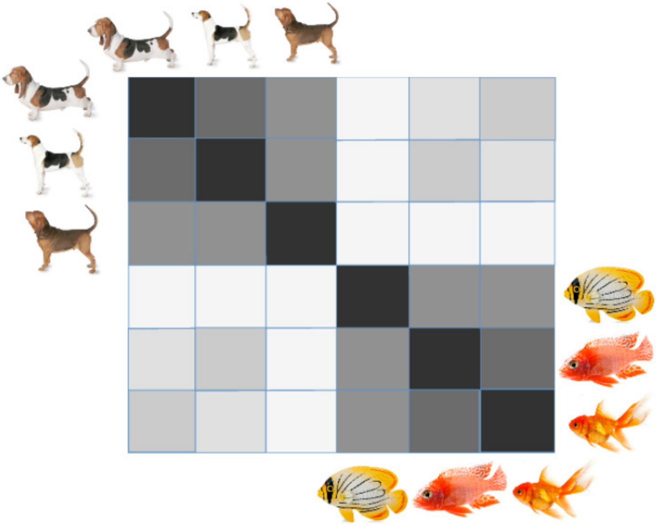
$\sim P$

$\sim Q$

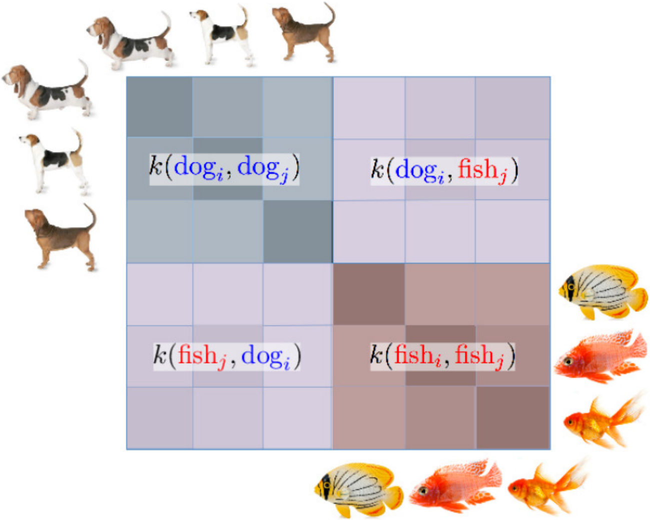
Towards distribution features



Towards distribution features



Towards distribution features



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

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

Kernel $\rightarrow$ distribution feature

- Kernel recall: $k(\mathbf{x}, \mathbf{x}') = \langle \varphi_{\mathbf{x}}, \varphi_{\mathbf{x}'} \rangle$.

Kernel $\to$ distribution feature

- Kernel recall: $k(\mathbf{x}, \mathbf{x}') = \langle \varphi_{\mathbf{x}}, \varphi_{\mathbf{x}'} \rangle$.
- Feature of $\mathbb{P}$ (mean embedding):

$$\mu_{\mathbb{P}} := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\varphi_{\mathbf{x}}].$$

Kernel $\rightarrow$ distribution feature

- Kernel recall: $k(\mathbf{x}, \mathbf{x}') = \langle \varphi_{\mathbf{x}}, \varphi_{\mathbf{x}'} \rangle$.
- Feature of $\mathbb{P}$ (mean embedding):

$$\mu_{\mathbb{P}} := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\varphi_{\mathbf{x}}].$$

- Previous quantity: unbiased estimate of

$$MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2 .$$

Kernel → distribution feature

- Kernel recall: $k(\mathbf{x}, \mathbf{x}') = \langle \varphi_{\mathbf{x}}, \varphi_{\mathbf{x}'} \rangle$.
- Feature of $\mathbb{P}$ (mean embedding):

$$\mu_{\mathbb{P}} := \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\varphi_{\mathbf{x}}].$$

- Previous quantity: unbiased estimate of

$$MMD^2(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2 .$$

- Valid test [Gretton et al., 2012]. Challenges:
  1. Threshold choice: 'ugly' asymptotics of $n\widehat{MMD}^2(\mathbb{P}, \mathbb{P})$.
  2. Test statistic: quadratic time complexity.
  3. Witness $\in \mathcal{H}(k)$: can be hard to interpret.

Linear-time tests

Linear-time 2-sample test

- Recall:

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

- Changing [Chwialkowski et al., 2015] 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 (a.s.). How do we estimate it? Distribution under H_0?

What is a random metric?

In short

It is a metric almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) = \rho(\mathbb{Q},\mathbb{P})$ almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) = \rho(\mathbb{Q},\mathbb{P})$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) \leqslant \rho(\mathbb{P},\mathbb{D}) + \rho(\mathbb{D},\mathbb{Q})$ almost surely.

What is a random metric?

In short

It is a metric almost surely.

In other words,

- $\rho(\mathbb{P},\mathbb{Q}) \geqslant 0$, $\rho(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) = \rho(\mathbb{Q},\mathbb{P})$ almost surely.
- $\rho(\mathbb{P},\mathbb{Q}) \leqslant \rho(\mathbb{P},\mathbb{D}) + \rho(\mathbb{D},\mathbb{Q})$ almost surely.

$\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J \subset \mathbb{R}^d$: reason of randomness.

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,
- *analytic:* $\mathbf{x} \mapsto k(\mathbf{x},\mathbf{y})$ *is analytic for any* $\mathbf{y} \in \mathbb{R}^d$.

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,
- *analytic:* $\mathbf{x} \mapsto k(\mathbf{x},\mathbf{y})$ *is analytic for any* $\mathbf{y} \in \mathbb{R}^d$.
- *characteristic:* μ *is injective,*

Result

Theorem

If k is

- *bounded:* $\sup_{\mathbf{x},\mathbf{x}'} k(\mathbf{x},\mathbf{x}') \leqslant B_k < \infty$,
- *analytic:* $\mathbf{x} \mapsto k(\mathbf{x},\mathbf{y})$ *is analytic for any* $\mathbf{y} \in \mathbb{R}^d$.
- *characteristic:* μ *is injective,*

then

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

is a metric a.s. w.r.t. $\{\mathbf{v}_j\}_{j=1}^{J}$.

Why do analytic features work? – proof idea

- μ is injective to analytic functions:
  - k: bounded, analytic $\Rightarrow$ elements of $\mathcal{H}_k$: analytic.
  - k: characteristic, bounded $\Rightarrow$ $\mu = \mu_k$: well-defined, injective.

Why do analytic features work? – proof idea

- μ is injective to analytic functions:
  - k: bounded, analytic $\Rightarrow$ elements of $\mathcal{H}_k$: analytic.
  - k: characteristic, bounded $\Rightarrow$ $\mu = \mu_k$: well-defined, injective.
- μ: characteristic $\Rightarrow$ for $\mathbb{P} \neq \mathbb{Q}$, $f := \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \neq 0$.

Why do analytic features work? – proof idea

- μ is injective to analytic functions:
  - k: bounded, analytic $\Rightarrow$ elements of $\mathcal{H}_k$: analytic.
  - k: characteristic, bounded $\Rightarrow$ $\mu = \mu_k$: well-defined, injective.
- μ: characteristic $\Rightarrow$ for $\mathbb{P} \neq \mathbb{Q}$, $f := \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \neq 0$.
- f: analytic, thus

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

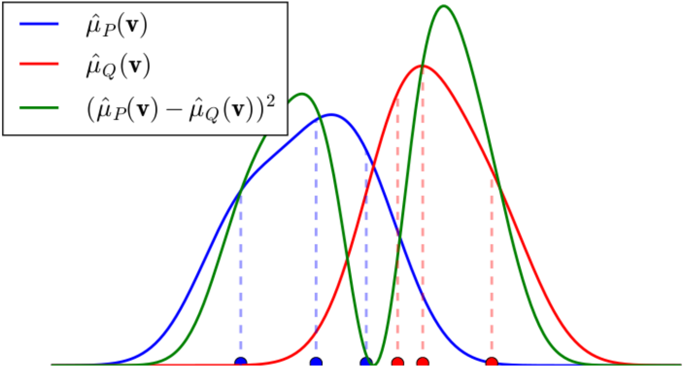
is a metric, a.s. w.r.t. $(\mathbf{v}_j \overset{i.i.d.}{\sim})$ $m \ll \lambda$. Reason: for an analytic $f \neq 0$, $m\{\mathbf{v} : f(\mathbf{v}) = 0\} = 0$.

Estimation

Compute

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

where $\hat{\mu}_{\mathbb{P}}(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i,\mathbf{v})$. Example using $k(\mathbf{x},\mathbf{v}) = e^{-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}}$:



Estimation – continued

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

Estimation – continued

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

$$= \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i,\mathbf{v}_j) - \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{y}_i,\mathbf{v}_j)\right]^2$$

Estimation – continued

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

$$= \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}k(\mathbf{x}_i,\mathbf{v}_j) - \frac{1}{n}\sum_{i=1}^{n}k(\mathbf{y}_i,\mathbf{v}_j)\right]^2 = \frac{1}{J}\sum_{j=1}^{J}(\bar{\mathbf{z}}_n)_j^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

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

Estimation – continued

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

$$= \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}k(\mathbf{x}_i,\mathbf{v}_j) - \frac{1}{n}\sum_{i=1}^{n}k(\mathbf{y}_i,\mathbf{v}_j)\right]^2 = \frac{1}{J}\sum_{j=1}^{J}(\bar{\mathbf{z}}_n)_j^2 = \frac{1}{J}\bar{\mathbf{z}}_n^T\bar{\mathbf{z}}_n,$$

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

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

Normalized version gives tractable null

- Modified test statistic:

$$\hat{\lambda}_n = n\bar{\mathbf{z}}_n^T \boldsymbol{\Sigma}_n^{-1} \bar{\mathbf{z}}_n,$$

  where $\boldsymbol{\Sigma}_n = cov\left(\{\mathbf{z}_i\}_{i=1}^n\right)$.
- Under H_0:
  - $\hat{\lambda}_n \xrightarrow{w} \chi^2(J)$. $\Rightarrow$ Easy to get the $(1-\alpha)$-quantile!

Our idea

Idea

- Until this point: test locations ($\mathcal{V}$) are fixed.
- Instead: choose $\theta = \{\mathcal{V}, \sigma\}$ to

  maximize lower bound on the test power.

Idea

- Until this point: test locations ($\mathcal{V}$) are fixed.
- Instead: choose $\theta = \{\mathcal{V}, \sigma\}$ to

  maximize lower bound on the test power.

Theorem (Lower bound on power, for large n)

Test power $\geqslant L(\lambda_n)$; L: explicit function, increasing.

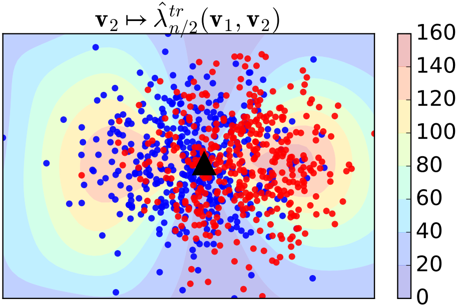
- Here,
  - $\lambda_n = n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$: population version of $\hat{\lambda}_n$.
  - $\boldsymbol{\mu} = \mathbb{E}_{\mathbf{xy}}[\mathbf{z}_1]$, $\boldsymbol{\Sigma} = \mathbb{E}_{\mathbf{xy}}\left[(\mathbf{z}_1 - \boldsymbol{\mu})(\mathbf{z}_1 - \boldsymbol{\mu})^T\right]$.

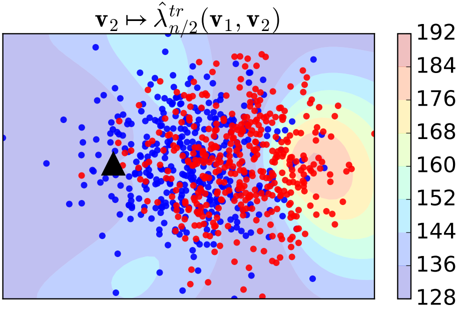
Non-convexity, informative features

- 2D problem:

$$\mathbb{P} := \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbb{Q} := \mathcal{N}(\mathbf{e}_1, \mathbf{I}).$$

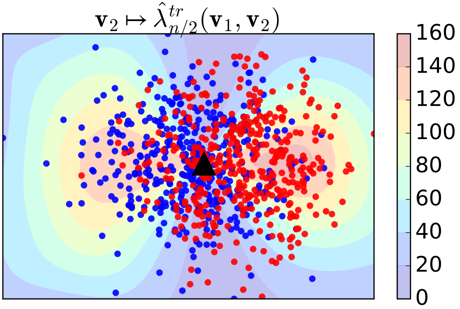
- $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2\}$. Fix $\mathbf{v}_1$ to ▲.
- $\mathbf{v}_2 \mapsto \hat{\lambda}_n(\{\mathbf{v}_1, \mathbf{v}_2\})$: contour plot.

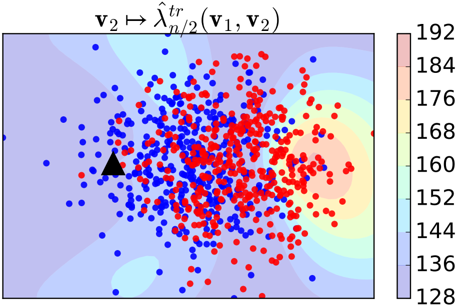




Non-convexity, informative features

- Nearby locations: do not increase discrimininability.
- Non-convexity: reveals multiple ways to capture the difference.





Convergence of the λ_n estimator

But λ_n is unknown. Split (X, Y) into (X_{tr}, Y_{tr}) and (X_{te}, Y_{te}).

1. Locations, kernel parameter: $\hat{\theta} = \arg\max_{\theta} \hat{\lambda}^{tr}_{\frac{n}{2}}(\theta)$.

Convergence of the λ_n estimator

But λ_n is unknown. Split (X, Y) into (X_{tr}, Y_{tr}) and (X_{te}, Y_{te}).

1. Locations, kernel parameter: $\hat{\theta} = \arg\max_{\theta} \hat{\lambda}^{tr}_{\frac{n}{2}}(\theta)$.
2. Test statistic: $\hat{\lambda}^{te}_{\frac{n}{2}}\left(\hat{\theta}\right)$.

Convergence of the λ_n estimator

Theorem (Guarantee on objective approximation, $\gamma_n \rightarrow 0$)

$$\sup_{\mathcal{V},\mathcal{K}} \left| \bar{\mathbf{z}}_n^T (\boldsymbol{\Sigma}_n + \gamma_n)^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| = \mathcal{O}\left(n^{-\frac{1}{4}}\right).$$

Convergence of the λ_n estimator

Theorem (Guarantee on objective approximation, $\gamma_n \to 0$)

$$\sup_{\mathcal{V},\mathcal{K}} \left| \bar{\mathbf{z}}_n^T (\boldsymbol{\Sigma}_n + \gamma_n)^{-1} \bar{\mathbf{z}}_n - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right| = \mathcal{O}\left(n^{-\frac{1}{4}}\right).$$

Examples:

$$\mathcal{K} = \left\{ k_\sigma(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}} : \sigma > 0 \right\},$$

$$\mathcal{K} = \left\{ k_{\mathbf{A}}(\mathbf{x},\mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^T \mathbf{A} (\mathbf{x}-\mathbf{y})} : \mathbf{A} \succ 0 \right\}.$$

Proof idea

- Lower bound on the test power:
  - $|\hat{\lambda}_n - \lambda_n| \lesssim \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}\|_F$.
  - Bound the r.h.s. by Hoeffding inequality $\Rightarrow P(|\hat{\lambda}_n - \lambda_n| \geqslant t)$.
  - By reparameterization: $P(\hat{\lambda}_n \geqslant T_\alpha)$ bound.

Proof idea

- Lower bound on the test power:
  - $|\hat{\lambda}_n - \lambda_n| \lesssim \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2 + \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}\|_F$.
  - Bound the r.h.s. by Hoeffding inequality $\Rightarrow P(|\hat{\lambda}_n - \lambda_n| \geqslant t)$.
  - By reparameterization: $P(\hat{\lambda}_n \geqslant T_\alpha)$ bound.
- Uniformly $\hat{\lambda}_n \approx \lambda_n$:
  - Reduction to bounding $\sup_{\mathcal{V},\mathcal{K}} \|\bar{\mathbf{z}}_n - \boldsymbol{\mu}\|_2$, $\sup_{\mathcal{V},\mathcal{K}} \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}\|_F$.
  - Empirical processes, Dudley entropy bound.

Numerical demos

Parameter settings

- Gaussian kernel (σ). $\alpha = 0.01$. $J = 1$. Repeat 500 trials.
- Report

$$P(\text{reject } H_0) \approx \frac{\#\text{times } \hat{\lambda}_n > T_\alpha \text{ holds}}{\#\text{trials}}.$$

- Compare 4 methods
  - **ME-full**: Optimize $\mathcal{V}$ and Gaussian bandwidth σ.
  - **ME-grid**: Optimize σ. Random $\mathcal{V}$ [Chwialkowski et al., 2015].
  - **MMD-quad**: Test with quadratic-time MMD [Gretton et al., 2012].
  - **MMD-lin**: Test with linear-time MMD [Gretton et al., 2012].
- Optimize kernels to power in MMD-lin, MMD-quad.

NLP: discrimination of document categories

- 5903 NIPS papers (1988-2015).
- Keyword-based category assignment into 4 groups:
  - Bayesian inference, Deep learning, Learning theory, Neuroscience
- $d = 2000$ nouns. TF-IDF representation.

Problem	n^{te}	**ME-full**	ME-grid	MMD-quad	MMD-lin
1. Bayes-Bayes	215	.012	.018	.022	.008
2. Bayes-Deep	216	.954	.034	.906	.262
3. Bayes-Learn	138	.990	.774	1.00	.238
4. Bayes-Neuro	394	1.00	.300	.952	.972
5. Learn-Deep	149	.956	.052	.876	.500
6. Learn-Neuro	146	.960	.572	1.00	.538

- Performance of ME-full $[\mathcal{O}(n)]$ is comparable to MMD-quad $[\mathcal{O}(n^2)]$.

NLP: most/least discriminative words

- Aggregating over trials; example: '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.

NLP: most/least discriminative words

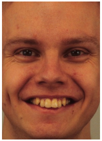
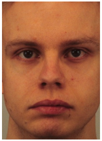
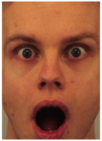
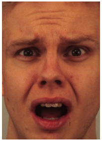
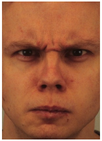
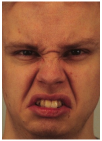
- Aggregating over trials; example: 'Bayes-Neuro'.

- Least dicriminative ones:

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

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

Problem	n^{te}	**ME-full**	ME-grid	MMD-quad	MMD-lin
$\pm$ vs. $\pm$	201	.010	.012	.018	.008
$+$ vs. $-$	201	.998	.656	1.00	.578

- Learned test location (averaged) = 

Summary

- We proposed a nonparametric t-test:
  - linear time,
  - adaptive → high-power ($\approx$ 'MMD-quad'),
- 2 demos: discriminating
  - documents of different categories,
  - positive/negative emotions.

Summary

- Extension (independence testing):

  `https://arxiv.org/abs/1610.04782`

  `https://github.com/wittawatj/fsic-test`

Thank you for the attention!

Acknowledgements: This work was supported by the Gatsby Charitable Foundation.

Contents

- Characteristic functions, infinite J.
- Number of locations (J).
- MMD: IPM representation.
- Estimation of MMD^2.
- Computational complexity: (J, n, d)-dependence.

Characteristic functions, infinite J

- Characteristic functions – poor choice:

$$\rho_2(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\phi_{\mathbb{P}}(\mathbf{v}_j) - \phi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

Characteristic functions, infinite J

- Characteristic functions – poor choice:

$$\rho_2(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\phi_{\mathbb{P}}(\mathbf{v}_j) - \phi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

- [Moulines et al., 2007]:

$$\rho_3(\mathbb{P},\mathbb{Q}) := \frac{n_x n_y}{n}\left\| C^{-\frac{1}{2}}(\mu_{\mathbb{Q}} - \mu_{\mathbb{P}})\right\|_{\mathcal{H}_k},$$

$$C = \frac{n_x}{n_x+n_y}C_{xx} + \frac{n_y}{n_x+n_y}C_{yy} : \text{ pooled covariance operator.}$$

Characteristic functions, infinite J

- Characteristic functions – poor choice:

$$\rho_2(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\phi_{\mathbb{P}}(\mathbf{v}_j) - \phi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

- [Moulines et al., 2007]:

$$\rho_3(\mathbb{P},\mathbb{Q}) := \frac{n_x n_y}{n}\left\| C^{-\frac{1}{2}}(\mu_{\mathbb{Q}} - \mu_{\mathbb{P}})\right\|_{\mathcal{H}_k},$$

$$C = \frac{n_x}{n_x+n_y}C_{xx} + \frac{n_y}{n_x+n_y}C_{yy}: \text{ pooled covariance operator.}$$

Computational cost: high (cubic).

Smoothed characteristic functions

$$\color{red}{\psi_{\mathbb{P}}}(t) = \int_{\mathbb{R}^d} \phi_{\mathbb{P}}(\boldsymbol{\omega})\ell(t - \boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}, \quad t \in \mathbb{R}^d,$$

$$\rho_4(\mathbb{P}, \mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\psi_{\mathbb{P}}(\mathbf{v}_j) - \psi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

Smoothed characteristic functions

$$\psi_{\mathbb{P}}(\mathbf{t}) = \int_{\mathbb{R}^d} \phi_{\mathbb{P}}(\boldsymbol{\omega})\ell(\mathbf{t}-\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}, \quad \mathbf{t} \in \mathbb{R}^d,$$

$$\rho_4(\mathbb{P},\mathbb{Q}) := \sqrt{\frac{1}{J}\sum_{j=1}^{J}[\psi_{\mathbb{P}}(\mathbf{v}_j) - \psi_{\mathbb{Q}}(\mathbf{v}_j)]^2}.$$

It

- works,
- is more sensitive to differences in the frequency domain.

Number of locations (J)

- Small J:
  - often enough to detect the difference of $\mathbb{P}$ & $\mathbb{Q}$.
  - few distinguishing regions to reject H_0.
  - faster test.

Number of locations (J)

- Very large J:
  - test power need not increase monotonically in J (more locations $\Rightarrow$ statistic can gain in variance).
  - defeats the purpose of a linear-time test.

MMD: IPM representation

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

MMD: IPM representation

$$MMD^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}(k)} = \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f\rangle_{\mathcal{H}(k)}\right]^2$$

MMD: IPM representation

$$MMD^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}(k)} = \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f\rangle_{\mathcal{H}(k)}\right]^2$$

$$\overset{(*)}{=} \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{y\sim\mathbb{Q}} f(y)\right]^2.$$

MMD: IPM representation

$$MMD^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}(k)} = \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f\rangle_{\mathcal{H}(k)}\right]^2$$

$$\overset{(*)}{=} \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{y\sim\mathbb{Q}} f(y)\right]^2.$$

$(*)$ in details:

$$\langle \mu_{\mathbb{P}}, f\rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot,x)\mathrm{d}\mathbb{P}(x), f\right\rangle_{\mathcal{H}(k)}$$

MMD: IPM representation

$$MMD^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}(k)} = \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f\rangle_{\mathcal{H}(k)}\right]^2$$

$$\stackrel{(*)}{=} \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{y\sim\mathbb{Q}} f(y)\right]^2.$$

$(*)$ in details:

$$\langle \mu_{\mathbb{P}}, f\rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot,x)\mathrm{d}\mathbb{P}(x), f\right\rangle_{\mathcal{H}(k)} = \int \underbrace{\langle k(\cdot,x), f\rangle_{\mathcal{H}(k)}}_{=f(x)} \mathrm{d}\mathbb{P}(x)$$

MMD: IPM representation

$$MMD^2(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}(k)} = \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f\rangle_{\mathcal{H}(k)}\right]^2$$

$$\stackrel{(*)}{=} \left[\sup_{\|f\|_{\mathcal{H}(k)}\leqslant 1} \mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{y\sim\mathbb{Q}} f(y)\right]^2.$$

$(*)$ in details:

$$\langle \mu_{\mathbb{P}}, f\rangle_{\mathcal{H}(k)} = \left\langle \int k(\cdot,x)\mathrm{d}\mathbb{P}(x), f\right\rangle_{\mathcal{H}(k)} = \int \underbrace{\langle k(\cdot,x), f\rangle_{\mathcal{H}(k)}}_{=f(x)} \mathrm{d}\mathbb{P}(x)$$

$$= \mathbb{E}_{x\sim\mathbb{P}} f(x).$$

Estimation of MMD^2

Squared difference between feature means:

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

Estimation of MMD^2

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

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

Computational complexity

- Optimization & testing: linear in n.
- Testing: $\mathcal{O}\left(ndJ + nJ^2 + J^3\right)$.
- Optimization: $\mathcal{O}\left(ndJ^2 + J^3\right)$ per gradient ascent.

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 1981–1989.

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

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

Lundqvist, D., Flykt, A., and Öhman, A. (1998).
The Karolinska directed emotional faces-KDEF.

Technical report, ISBN 91-630-7164-9.

Moulines, É., Bach, F. R., and Harchaoui, Z. (2007).
Testing for homogenity with kernel Fisher discriminant analysis.
In *Neural Information Processing Systems (NIPS)*, pages 609–616.