Data-Efficient Independence Testing with Analytic Kernel Embeddings

Zoltán Szabó – CMAP, École Polytechnique

Wittawat Jitkrittum

Arthur Gretton

PASADENA Seminar
Télécom ParisTech
May 17, 2017

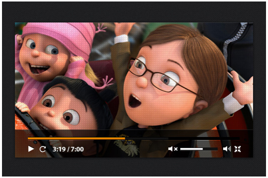
Motivation

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

Motivation

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

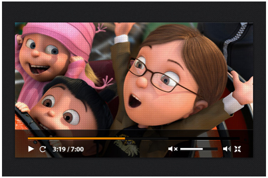
  - (video, caption) pairs



Motivation

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

  - (video, caption) pairs



- $\{(x_i, y_i)\}_{i=1}^n \xrightarrow{?} H_0 : \mathbb{P}_{xy} = \mathbb{P}_x \mathbb{P}_y, \; H_1 : \mathbb{P}_{xy} \neq \mathbb{P}_x \mathbb{P}_y.$

Covariance

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

- Covariance matrix

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

Covariance

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

- Covariance matrix

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

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

Covariance

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

- Covariance matrix

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

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

Covariance

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

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

$$C_{xy} = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right]$$

Covariance

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

$$C_{xy} = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right],$$

$$S = \|C_{xy}\|_{HS}.$$

Wishlist

Target:

$$C_{xy} = \mathbb{E}_{xy}\left[(\varphi(x) - \mathbb{E}_x\varphi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))\right], \quad S = \left\| C_{xy} \right\|_{HS}.$$

We need

- φ, $\mathbb{E}_x\varphi(x)$, $\otimes$, $\|\cdot\|_{HS}$.
- $S = 0 \overset{?}{\Leftrightarrow} x \perp y$.
- Estimator, fast?

One-page summary

- $\|C_{xy}\|_{HS}$:
  - Characterizes independence.
  - Estimation: slow $= \mathcal{O}\left(n^2\right)$.

One-page summary

- $\|C_{xy}\|_{HS}$:
  - Characterizes independence.
  - Estimation: slow $= \mathcal{O}\left(n^2\right)$.
- 'Sampled' $\|C_{xy}\|$:
  - Independence: ✓ in $\mathcal{O}(n)$-time.

One-page summary

- $\|C_{xy}\|_{HS}$:
  - Characterizes independence.
  - Estimation: slow $= \mathcal{O}\left(n^2\right)$.
- 'Sampled' $\|C_{xy}\|$:
  - Independence: ✓ in $\mathcal{O}(n)$-time.
  - Features: optimized for power.

One-page summary

- $\|C_{xy}\|_{HS}$:
  - Characterizes independence.
  - Estimation: slow $= \mathcal{O}\left(n^2\right)$.
- 'Sampled' $\|C_{xy}\|$:
  - Independence: ✓ in $\mathcal{O}(n)$-time.
  - Features: optimized for power.
  - ICML-2017: accepted [Jitkrittum et al., 2017].

Features, distribution representation: φ, $\mathbb{E}_x\varphi(x)$

Distribution $\mapsto$ function: examples

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z)$$

Distribution $\mapsto$ function: examples

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}} I_{(-\infty,z)}(x).$$

Distribution $\mapsto$ function: examples

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}} I_{(-\infty,z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(z) = \int e^{i\langle z,x\rangle} \mathrm{d}\mathbb{P}(x).$$

Distribution $\mapsto$ function: examples

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x\sim\mathbb{P}} I_{(-\infty,z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(z) = \int e^{i\langle z,x\rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(z) = \int e^{\langle z,x\rangle} \mathrm{d}\mathbb{P}\left(x\right).$$

Distribution $\mapsto$ function: examples

- Cumulative density function:

$$\mathbb{P} \mapsto F(z) = \mathbb{P}(x < z) = \mathbb{E}_{x \sim \mathbb{P}} I_{(-\infty, z)}(x).$$

- Characteristic function:

$$\mathbb{P} \mapsto c_{\mathbb{P}}(z) = \int e^{i\langle z, x\rangle} \mathrm{d}\mathbb{P}(x).$$

- Moment generating function:

$$\mathbb{P} \mapsto M_{\mathbb{P}}(z) = \int e^{\langle z, x\rangle} \mathrm{d}\mathbb{P}\left(x\right).$$

Pattern

$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$

Distribution $\mapsto$ function: examples

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

Distribution $\mapsto$ function: examples

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

2 questions

- How to choose φ?

Distribution $\mapsto$ function: examples

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

2 questions

- How to choose φ?
- How to interpret the integral?

Distribution $\mapsto$ function: examples

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

2 questions

- How to choose φ?
- How to interpret the integral?

Answers:

- We use kernels. $\rightarrow$ Computational tractability: ✓

Distribution $\mapsto$ function: examples

Wanted:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \varphi(x) \mathrm{d}\mathbb{P}(x).$$

2 questions

- How to choose φ?
- How to interpret the integral?

Answers:

- We use kernels. $\rightarrow$ Computational tractability: ✓
- Expectation: Bochner integral.

Kernel, RKHS definition(s) $\rightarrow$ feature: φ

Given: $\mathcal{X}$ set.

Definition

Kernel: $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.

Kernel, RKHS definition(s) $\rightarrow$ feature: φ

Given: $\mathcal{X}$ set.

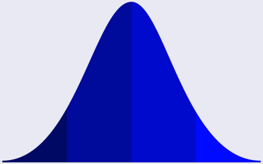
Definition

Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.

Definition

Reproducing kernel of an $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ Hilbert space,

- $k(\cdot, b) \in \mathcal{H}$,



Kernel, RKHS definition(s) $\rightarrow$ feature: φ

Given: $\mathcal{X}$ set.

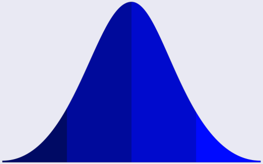
Definition

Kernel: $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.

Definition

Reproducing kernel of an $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ Hilbert space,

- $k(\cdot, b) \in \mathcal{H}$,



- $\langle f, k(\cdot, b) \rangle_{\mathcal{H}} = f(b)$.

Kernel, RKHS definition(s) $\rightarrow$ feature: φ

Given: $\mathcal{X}$ set.

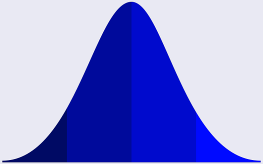
Definition

Kernel: $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.

Definition

Reproducing kernel of an $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ Hilbert space,

- $k(\cdot, b) \in \mathcal{H}$,



- $\langle f, k(\cdot, b) \rangle_{\mathcal{H}} = f(b)$. Note: $k(a,b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}$.

Kernel examples: $\mathcal{X} = \mathbb{R}^d$, $\theta > 0$

- Gaussian kernel:

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

Kernel examples: $\mathcal{X} = \mathbb{R}^d$, $\theta > 0$

- Gaussian kernel:

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

- Polynomial kernel:

$$k_p(a, b) = (\langle a, b \rangle + \theta)^p.$$

Kernel examples: $\mathcal{X} = \mathbb{R}^d$, $\theta > 0$

- Gaussian kernel:

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

- Polynomial kernel:

$$k_p(a,b) = \left(\langle a,b\rangle + \theta\right)^p.$$

- Matérn kernel:

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

Distribution representation, Bochner integral

- Kernel/mean embedding:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \underbrace{k(\cdot, x)}_{=\varphi(x) \in \mathcal{H}_k} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

Distribution representation, Bochner integral

- Kernel/mean embedding:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \underbrace{k(\cdot, x)}_{=\varphi(x) \in \mathcal{H}_k} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Existence: $\exists \mu_{\mathbb{P}} \Leftrightarrow \int \|k(\cdot, x)\|_{\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) < \infty$.

Distribution representation, Bochner integral

- Kernel/mean embedding:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \underbrace{k(\cdot, x)}_{=\varphi(x)\in\mathcal{H}_k} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Existence: $\exists \mu_{\mathbb{P}} \Leftrightarrow \int \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{=\sqrt{k(x,x)}} \mathrm{d}\mathbb{P}(x) < \infty.$

Distribution representation, Bochner integral

- Kernel/mean embedding:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int \underbrace{k(\cdot, x)}_{=\varphi(x)\in\mathcal{H}_k} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Existence: $\exists \mu_{\mathbb{P}} \Leftrightarrow \int \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{=\sqrt{k(x,x)}} \mathrm{d}\mathbb{P}(x) < \infty.$
  - Example: bounded k, e.g. Gaussian kernel.

Until now:

$$\mathbb{E}_{xy}\left[\left(\underbrace{\varphi(x)}_{\checkmark} - \underbrace{\mathbb{E}_x\varphi(x)}_{\checkmark}\right) \underbrace{\otimes}_{???} \left(\underbrace{\psi(y)}_{\checkmark} - \underbrace{\mathbb{E}_y\psi(y)}_{\checkmark}\right)\right]$$

Until now:

$$\mathbb{E}_{xy}\left[\left(\underbrace{\varphi(x)}_{\checkmark}-\underbrace{\mathbb{E}_x\varphi(x)}_{\checkmark}\right)\underbrace{\otimes}_{???}\left(\underbrace{\psi(y)}_{\checkmark}-\underbrace{\mathbb{E}_y\psi(y)}_{\checkmark}\right)\right]$$

Question: $a \otimes b, \mathcal{H}_1 \otimes \mathcal{H}_2$.

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a \left(b^T g\right)$$

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a\left(b^T g\right) = a \langle b, g \rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a\left(b^T g\right) = a \langle b, g \rangle \in \mathbb{R}^{d_1},$$

  $ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.
- Alternatively

$$\mathbb{R} \ni f^T \left(ab^T\right) g$$

  $ab^T : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ bilinear form.

Intuition of $a \otimes b$, $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a \left(b^T g\right) = a \langle b, g \rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

- Alternatively

$$\mathbb{R} \ni f^T \left(ab^T\right) g = \langle f, a \rangle \langle g, b \rangle$$

$ab^T : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ bilinear form.

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$: Hilbert spaces.
- $a \otimes b$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}, \quad (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2.$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$: Hilbert spaces.
- $a \otimes b$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}, \quad (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2.$$

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2 \right\}.$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$: Hilbert spaces.
- $a \otimes b$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}, \quad (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2.$$

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2 \right\}.$$

- Define inner product on $\mathcal{L}$, extended by linearity

$$\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle := \langle a_1, a_2 \rangle_{\mathcal{H}_1} \langle b_1, b_2 \rangle_{\mathcal{H}_2}.$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$: Hilbert spaces.
- $a \otimes b$ is the bilinear form:

$$(a \otimes b)(f, g) := \langle f, a \rangle_{\mathcal{H}_1} \langle g, b \rangle_{\mathcal{H}_2}, \quad (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2.$$

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2 \right\}.$$

- Define inner product on $\mathcal{L}$, extended by linearity

$$\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle := \langle a_1, a_2 \rangle_{\mathcal{H}_1} \langle b_1, b_2 \rangle_{\mathcal{H}_2}.$$

- $\mathcal{H}_1 \otimes \mathcal{H}_2$: completion of $\mathcal{L}$.

Tensor product of RKHSs

Theorem ([Berlinet and Thomas-Agnan, 2004])

- *Let* $\mathcal{H}_1 := \mathcal{H}_k$, $\mathcal{H}_2 := \mathcal{H}_\ell$ *RKHSs with kernel* k *and* ℓ.
- *Then* $\mathcal{H}_k \otimes \mathcal{H}_\ell$ *is RKHS with kernel*

$$k \otimes \ell : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R},$$
$$(k \otimes \ell)\left((x_1, y_1), (x_2, y_2)\right) := k(x_1, x_2)\ell(y_1, y_2).$$

Tensor product of RKHSs

Theorem ([Berlinet and Thomas-Agnan, 2004])

- *Let* $\mathcal{H}_1 := \mathcal{H}_k$, $\mathcal{H}_2 := \mathcal{H}_\ell$ *RKHSs with kernel* k *and* ℓ*.*
- *Then* $\mathcal{H}_k \otimes \mathcal{H}_\ell$ *is RKHS with kernel*

$$k \otimes \ell : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R},$$
$$(k \otimes \ell)\left((x_1, y_1), (x_2, y_2)\right) := k(x_1, x_2)\ell(y_1, y_2).$$

Intuition:

- inner product on $\mathcal{X}$ and $\mathcal{Y}$ $\to$ inner product on $\mathcal{X} \times \mathcal{Y}$.
- $\mathcal{X}$ = video, $\mathcal{Y}$ = caption.

Hilbert-Schmidt independence criterion (HSIC)

HSIC:

$$\begin{aligned} HSIC(x, y) &:= \|C_{xy}\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell} \\ &= \left\|\mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x \mathbb{P}_y}\right\|_{\mathcal{H}_{k \otimes \ell}}. \end{aligned}$$

Hilbert-Schmidt independence criterion (HSIC)

HSIC:

$$\begin{aligned} HSIC(x,y) &:= \|C_{xy}\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell} \\ &= \left\|\mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x \mathbb{P}_y}\right\|_{\mathcal{H}_{k\otimes\ell}}. \end{aligned}$$

- Naming: $\mathcal{H}_k \otimes \mathcal{H}_\ell \simeq HS(\mathcal{H}_\ell, \mathcal{H}_k)$, 'Frobenius norm'.

Hilbert-Schmidt independence criterion (HSIC)

HSIC:

$$HSIC(x,y) := \|C_{xy}\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$$
$$= \left\|\mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x \mathbb{P}_y}\right\|_{\mathcal{H}_{k\otimes\ell}}.$$

- Naming: $\mathcal{H}_k \otimes \mathcal{H}_\ell \simeq HS(\mathcal{H}_\ell, \mathcal{H}_k)$, 'Frobenius norm'.
- Measure for $\mu_{\mathbb{P}_{xy}} = \mu_{\mathbb{P}_x \mathbb{P}_y} \Rightarrow k \otimes \ell$: 'characteristic' matters.

Hilbert-Schmidt independence criterion (HSIC)

HSIC:

$$\begin{aligned} HSIC(x, y) &:= \|C_{xy}\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell} \\ &= \left\| \mu_{\mathbb{P}_{xy}} - \mu_{\mathbb{P}_x \mathbb{P}_y} \right\|_{\mathcal{H}_{k \otimes \ell}}. \end{aligned}$$

- Naming: $\mathcal{H}_k \otimes \mathcal{H}_\ell \simeq HS(\mathcal{H}_\ell, \mathcal{H}_k)$, 'Frobenius norm'.
- Measure for $\mu_{\mathbb{P}_{xy}} = \mu_{\mathbb{P}_x \mathbb{P}_y} \Rightarrow k \otimes \ell$: 'characteristic' matters.
- [Gretton, 2015]: k, ℓ: characteristic, translation-invariant $\Rightarrow$ ✓

Local summary

- Mean embedding: distribution representation,

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x).$$

Characteristic means: $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective.

Local summary

- Mean embedding: distribution representation,

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x).$$

  Characteristic means: $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective.
- Cross-covariance operator:

$$C_{xy} = \mathbb{E}_{xy} \left[\varphi(x) \otimes \psi(y) \right] - \mu_x \otimes \mu_y.$$

Local summary

- Mean embedding: distribution representation,

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\mathbb{P}(x).$$

  Characteristic means: $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective.
- Cross-covariance operator:

$$C_{xy} = \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right] - \mu_x \otimes \mu_y.$$

- HSIC:

$$HSIC(x, y) = \|C_{xy}\|_{HS}.$$

  k, ℓ: characteristic $\Rightarrow$ independence measure.

Characteristic property

Well-understood for

- Continuous bounded translation-invariant kernels on $\mathbb{R}^d$:

$$k(x,y) = k_0(x-y), k_0 \in C_b(\mathbb{R}^d).$$

Characteristic property

Well-understood for

- Continuous bounded translation-invariant kernels on $\mathbb{R}^d$:

$$k(x, y) = k_0(x - y), k_0 \in C_b(\mathbb{R}^d).$$

- In this case (Bochner's theorem):

$$k_0(z) = \int_{\mathbb{R}^d} e^{-i\langle z,\omega\rangle} \mathrm{d}\Lambda(\omega),$$
$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} = \|c_{\mathbb{P}} - c_{\mathbb{Q}}\|_{L^2(\Lambda)}.$$

Characteristic property

Well-understood for

- Continuous bounded translation-invariant kernels on $\mathbb{R}^d$:

$$k(x,y) = k_0(x-y),\ k_0 \in C_b(\mathbb{R}^d).$$

- In this case (Bochner's theorem):

$$k_0(z) = \int_{\mathbb{R}^d} e^{-i\langle z,\omega\rangle} \mathrm{d}\Lambda(\omega),$$

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} = \|c_{\mathbb{P}} - c_{\mathbb{Q}}\|_{L^2(\Lambda)}.$$

Theorem ([Sriperumbudur et al., 2010])

k is characteristic iff. $supp(\Lambda) = \mathbb{R}^d$.

Translation-invariant kernels on $\mathbb{R}$

For Poisson kernel: $\sigma \in (0,1)$.

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\|x\|}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\|j\|}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\|j\|}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

Translation-invariant kernels on $\mathbb{R}$

For Poisson kernel: $\sigma \in (0,1)$.

kernel name	k_0	$\widehat{k_0}(\omega)$	$supp\left(\widehat{k_0}\right)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\|x\|}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\|j\|}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\|j\|}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

For $x \in \mathbb{R}^d$: $k_0(x) = \prod_{j=1}^d k_0(x_j)$, $\widehat{k_0}(\omega) = \prod_{j=1}^d \widehat{k_0}(\omega_j)$.

Estimation of HSIC

- Estimate:

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

Estimation of HSIC

- Estimate:

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

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

Estimation of HSIC

- Estimate:

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

$$\mathbf{G}_x = [k(x_i, x_j)]_{i,j=1}^n, \ \tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}, \ \mathbf{H} = \mathbf{I} - \frac{\mathbf{E}}{n}.$$

Estimation of HSIC

- Estimate:

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

$$\mathbf{G}_x = [k(x_i, x_j)]_{i,j=1}^n, \; \tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}, \; \mathbf{H} = \mathbf{I} - \frac{\mathbf{E}}{n}.$$

- It is a plug-in estimator: $\widehat{HSIC^2} = \left\| \hat{C}_{xy} \right\|_{HS}^2$.

Estimation of HSIC

- Estimate:

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

$$\mathbf{G}_x = [k(x_i, x_j)]_{i,j=1}^n, \ \tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}, \ \mathbf{H} = \mathbf{I} - \frac{\mathbf{E}}{n}.$$

- It is a plug-in estimator: $\widehat{HSIC^2} = \left\| \hat{C}_{xy} \right\|_{HS}^2$.
- Computational time: $\mathcal{O}\left(n^2\right)$.

Estimation of HSIC

- Estimate:

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

$$\mathbf{G}_x = [k(x_i, x_j)]_{i,j=1}^n, \; \tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}, \; \mathbf{H} = \mathbf{I} - \frac{\mathbf{E}}{n}.$$

- It is a plug-in estimator: $\widehat{HSIC}^2 = \left\| \hat{C}_{xy} \right\|_{HS}^2$.
- Computational time: $\mathcal{O}\left(n^2\right)$.

In short

HSIC: captures independence, slow to estimate.

'Sampled' HSIC

Idea [Jitkrittum et al., 2017]

Use different norm of the witness function (u):

$$HSIC(x,y) = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}, \quad u(v,w) = \mu_{xy}(v,w) - \mu_x(v)\mu_y(w),$$

Idea [Jitkrittum et al., 2017]

Use different norm of the witness function (u):

$$HSIC(x,y) = \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}, \quad u(v,w) = \mu_{xy}(v,w) - \mu_x(v)\mu_y(w),$$

$$FSIC(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(v_j,w_j)}, \qquad \mathcal{V} = \{(v_j,w_j)\}_{j=1}^{J},$$

Idea [Jitkrittum et al., 2017]

Use different norm of the witness function (u):

$$HSIC(x,y) = \left\|\mu_{xy} - \mu_x \otimes \mu_y\right\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}, \quad u(v,w) = \mu_{xy}(v,w) - \mu_x(v)\mu_y(w),$$

$$FSIC(x,y) = \sqrt{\frac{1}{J}\sum_{j=1}^{J} u^2(v_j,w_j)}, \qquad \mathcal{V} = \{(v_j,w_j)\}_{j=1}^{J},$$

$$= \left\|u\right\|_{L^2(\mathcal{V})}.$$

FSIC: covariance view

Recall

$HSIC = \|C_{xy}\|_{HS} = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{k\otimes\ell}.$

By rewriting

$$
\begin{aligned}
u(v,w) &= \mu_{xy}(v,w) - \mu_x(v)\mu_y(w) \\
&= \mathbb{E}_{xy}[k(x,v)\ell(y,w)] - \mathbb{E}_x[k(x,v)]\mathbb{E}_y[\ell(y,w)] \\
&= cov_{xy}\left(k(x,v), \ell(y,w)\right).
\end{aligned}
$$

FSIC: covariance view

Recall

$HSIC = \|C_{xy}\|_{HS} = \|\mu_{xy} - \mu_x \otimes \mu_y\|_{k\otimes\ell}.$

By rewriting

$$\begin{aligned} u(v,w) &= \mu_{xy}(v,w) - \mu_x(v)\mu_y(w) \\ &= \mathbb{E}_{xy}[k(x,v)\ell(y,w)] - \mathbb{E}_x[k(x,v)]\mathbb{E}_y[\ell(y,w)] \\ &= cov_{xy}\left(k(x,v), \ell(y,w)\right). \end{aligned}$$

$\Rightarrow$ We picked the $(v,w)^{th}$ entry of

$$C_{xy} = \mathbb{E}_{xy}\left[\varphi(x) \otimes \psi(y)\right] - \mu_x \otimes \mu_y.$$

FSIC is an independence measure

Theorem

If k, ℓ are bounded, characteristic, analytic, then

$$FSIC(x, y) = 0 \Leftrightarrow x \perp y$$

almost surely w.r.t. $\mathcal{V}$.

FSIC is an independence measure

Theorem

If k, ℓ are bounded, characteristic, analytic, then

$$FSIC(x, y) = 0 \Leftrightarrow x \perp y$$

almost surely w.r.t. $\mathcal{V}$.

Examples:

- Gaussian: $k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma_1^2}}$, $\ell(y, y') = e^{-\frac{\|y-y'\|^2}{2\sigma_2^2}}$.

FSIC is an independence measure

Theorem

If k, ℓ are bounded, characteristic, analytic, then

$$FSIC(x,y) = 0 \Leftrightarrow x \perp y$$

almost surely w.r.t. $\mathcal{V}$.

Examples:

- Gaussian: $k(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma_1^2}}$, $\ell(y,y') = e^{-\frac{\|y-y'\|^2}{2\sigma_2^2}}$.
- Full Gaussian ($A \succ 0$, $B \succ 0$):

$$k\left(x,x'\right) = e^{-(x-x')^T A (x-x')}, \quad \ell\left(y,y'\right) = e^{-(y-y')^T B (y-y')}.$$

Empirical estimator for FSIC

$$FSIC^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} u^2(v_j,w_j), \quad u(v,w) = \mu_{xy}(v,w) - \mu_x(v)\mu_y(w),$$

$$\widehat{FSIC}^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} \hat{u}^2(v_j,w_j), \quad \hat{u}(v,w) = \widehat{\mu_{xy}}(v,w) - \underbrace{(\widehat{\mu_x\mu_y})(v,w)}_{:=\hat{\mu}_x(v)\hat{\mu}_y(w)},$$

$$= \frac{1}{J}\|\hat{\mathbf{u}}\|_2^2$$

Empirical estimator for FSIC

$$FSIC^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} u^2(v_j,w_j), \quad u(v,w) = \mu_{xy}(v,w) - \mu_x(v)\mu_y(w),$$

$$\widehat{FSIC}^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} \hat{u}^2(v_j,w_j), \quad \hat{u}(v,w) = \widehat{\mu_{xy}}(v,w) - \underbrace{(\widehat{\mu_x\mu_y})(v,w)}_{:=\hat{\mu}_x(v)\hat{\mu}_y(w)},$$

$$= \frac{1}{J}\left\|\hat{\mathbf{u}}\right\|_2^2, \qquad \hat{\mathbf{u}} = \left[\hat{u}(v_j,w_j)\right]_{j=1}^{J},$$

Empirical estimator for FSIC

$$FSIC^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} u^2(v_j,w_j), \quad u(v,w) = \mu_{xy}(v,w) - \mu_x(v)\mu_y(w),$$

$$\widehat{FSIC}^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} \hat{u}^2(v_j,w_j), \quad \hat{u}(v,w) = \widehat{\mu_{xy}}(v,w) - \underbrace{(\widehat{\mu_x\mu_y})(v,w)}_{:=\hat{\mu}_x(v)\hat{\mu}_y(w)},$$

$$= \frac{1}{J}\|\hat{\mathbf{u}}\|_2^2, \qquad \hat{\mathbf{u}} = [\hat{u}(v_j,w_j)]_{j=1}^{J},$$

where

$$\widehat{\mu_{xy}}(v,w) = \frac{1}{n}\sum_{i=1}^{n} k(x_i,v)\ell(y_i,w),$$

$$\widehat{\mu_x\mu_y}(v,w) = \frac{1}{n(n-1)}\sum_{i\neq j} k(x_i,v)\ell(y_j,w).$$

Empirical estimator for FSIC

$$FSIC^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} u^2(v_j,w_j), \quad u(v,w) = \mu_{xy}(v,w) - \mu_x(v)\mu_y(w),$$

$$\widehat{FSIC}^2(x,y) = \frac{1}{J}\sum_{j=1}^{J} \hat{u}^2(v_j,w_j), \quad \hat{u}(v,w) = \widehat{\mu_{xy}}(v,w) - \underbrace{(\widehat{\mu_x\mu_y})(v,w)}_{:=\hat{\mu}_x(v)\hat{\mu}_y(w)},$$

$$= \frac{1}{J}\left\|\hat{\mathbf{u}}\right\|_2^2, \qquad \hat{\mathbf{u}} = \left[\hat{u}(v_j,w_j)\right]_{j=1}^{J},$$

where

$$\widehat{\mu_{xy}}(v,w) = \frac{1}{n}\sum_{i=1}^{n} k(x_i,v)\ell(y_i,w),$$

$$\widehat{\mu_x\mu_y}(v,w) = \frac{1}{n(n-1)}\sum_{i\neq j} k(x_i,v)\ell(y_j,w).$$

Computational complexity: $\mathcal{O}\left((d_x+d_y)Jn\right)$ = fast.

Empirical estimator for FSIC

For fixed (v, w) FSIC is a U-statistic:

$$\hat{u}(v, w) = \frac{2}{n(n-1)} \sum_{i<j} h_{v,w}\left((x_i, y_i), (x_j, y_j)\right),$$

$$h_{v,w}\left((x, y), (x', y')\right) = \frac{1}{2}\left[k(x, v) - k(x', v)\right]\left[\ell(y, w) - \ell(y', w)\right]$$

Empirical estimator for FSIC

For fixed (v,w) FSIC is a U-statistic:

$$\hat{u}(v,w) = \frac{2}{n(n-1)} \sum_{i<j} h_{v,w}\left((x_i,y_i),(x_j,y_j)\right),$$

$$h_{v,w}\left((x,y),(x',y')\right) = \frac{1}{2}\left[k(x,v)-k(x',v)\right]\left[\ell(y,w)-\ell(y',w)\right],$$

thus

Theorem (Asymptotic normality)

For any fixed locations $\mathcal{V} = \{(v_j,w_j)\}_{j=1}^{J}$, $\hat{\mathbf{u}} := \left[\hat{u}(v_j,w_j)\right]_{j=1}^{J}$

$$\sqrt{n}\left(\hat{\mathbf{u}} - \mathbf{u}\right) \xrightarrow{d} N(0,\Sigma),$$

$$\Sigma_{ij} = cov_{xy}\left(\hat{u}(v_i,w_i), \hat{u}(v_j,w_j)\right).$$

NFSIC = FSIC + whitening

- $n\widehat{FSIC}^2(x,y) = n\frac{\|\mathbf{u}\|_2^2}{J}$: asymptotically sum of correlated χ^2-s.

NFSIC = FSIC + whitening

- $n\widehat{FSIC}^2(x, y) = n\frac{\|\mathbf{u}\|_2^2}{J}$: asymptotically sum of correlated χ^2-s.
- Quantile: hard. $\Rightarrow$ With whitening trick:

Theorem

- *Under H_0: with $\gamma_n \to 0$*

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T \left(\hat{\Sigma}_n + \gamma_n I_J\right)^{-1} \hat{\mathbf{u}} \xrightarrow{d} \chi^2(J).$$

NFSIC = FSIC + whitening

- $n\widehat{FSIC}^2(x,y) = n\frac{\|\mathbf{u}\|_2^2}{J}$: asymptotically sum of correlated χ^2-s.
- Quantile: hard. $\Rightarrow$ With whitening trick:

Theorem

- *Under H_0: with $\gamma_n \to 0$*

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T \left(\hat{\Sigma}_n + \gamma_n I_J\right)^{-1} \hat{\mathbf{u}} \xrightarrow{d} \chi^2(J).$$

- *Under H_1: we get a consistent test (i.e., power $\to 1$).*

NFSIC can be estimated easily

Test statistic:

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T \left(\hat{\Sigma}_n + \gamma_n I_J\right)^{-1} \hat{\mathbf{u}}.$$

Estimator: no $n \times n$ Gram matrix

- $K := [k(v_i, x_j)] \in \mathbb{R}^{J \times n}$, $L := [\ell(w_i, y_j)] \in \mathbb{R}^{J \times n}$,
- $\hat{\Sigma}_n = \frac{\Gamma\Gamma^T}{n}$, $\Gamma = (KH_n) \circ (LH_n) - \hat{\mathbf{u}}1_n^T$, $\hat{\mathbf{u}} := \frac{(K \circ L)\mathbf{1}_n}{n-1} - \frac{(K1_n)\circ(L1_n)}{n(n-1)}$.

Computational time:

$$\mathcal{O}\left(J^3 + J^2 n + (d_x + d_y)Jn\right) = \text{fast}.$$

NFSIC can be estimated easily

Test statistic:

$$\hat{\lambda}_n = n\hat{\mathbf{u}}^T \left(\hat{\Sigma}_n + \gamma_n I_J\right)^{-1} \hat{\mathbf{u}}.$$

Estimator: no $n \times n$ Gram matrix

- $K := [k(v_i, x_j)] \in \mathbb{R}^{J \times n}$, $L := [\ell(w_i, y_j)] \in \mathbb{R}^{J \times n}$,
- $\hat{\Sigma}_n = \frac{\Gamma\Gamma^T}{n}$, $\Gamma = (KH_n) \circ (LH_n) - \hat{\mathbf{u}}1_n^T$, $\hat{\mathbf{u}} := \frac{(K \circ L)\mathbf{1}_n}{n-1} - \frac{(K1_n) \circ (L1_n)}{n(n-1)}$.

Computational time:

$$\mathcal{O}\left(J^3 + J^2 n + (d_x + d_y)Jn\right) = \text{fast}.$$

Code with demos:

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

Choosing the locations & kernel parameters

- Consistent test: for $\forall\ \mathcal{V} = \{(v_j, w_j\}_{j=1}^{J}$ and kernel parameters.

Choosing the locations & kernel parameters

- Consistent test: for $\forall\ \mathcal{V} = \{(v_j, w_j\}_{j=1}^{J}$ and kernel parameters.
- Choose the test-power proxy maximizers.

Theorem

Let $NFSIC^2(x, y) = \lambda_n = n\mathbf{u}^T\Sigma^{-1}\mathbf{u}$. For large n,

$$\textit{test power} \geqslant L(\lambda_n),$$

L: monotonically increasing.

Choosing the locations & kernel parameters

- Consistent test: for $\forall\ \mathcal{V} = \{(v_j, w_j)\}_{j=1}^{J}$ and kernel parameters.
- Choose the test-power proxy maximizers.

Theorem

Let $NFSIC^2(x, y) = \lambda_n = n\mathbf{u}^T\Sigma^{-1}\mathbf{u}$. For large n,

$$\textit{test power} \geqslant L(\lambda_n),$$

L: monotonically increasing.

- In practice: data-splitting.

Demo

Demo settings

- k, ℓ: Gaussian. $J = 10$.
- Report: rejection rate of H_0.
- Compare 6 methods:

Method	Description	Tuning	Test size	Complexity
NFSIC-opt	Studied	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heuristic	n	$\mathcal{O}(n^2)$
NyHSIC	Nyström + HSIC	Median heuristic	n	$\mathcal{O}(n)$
FHSIC	RFF + HSIC	Median heuristic	n	$\mathcal{O}(n)$
RDC	RFF + CCA	Median heuristic	n	$\mathcal{O}(n \log n)$

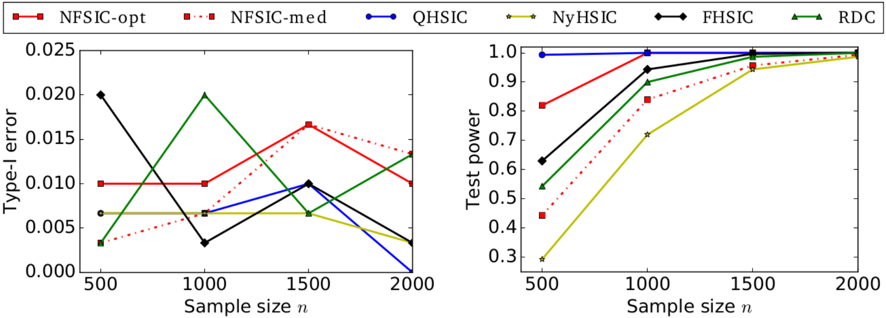
Demo-1: million song data

Song (x) vs. year of release (y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $x \in \mathbb{R}^{90=d_x}$: audio features.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.

Song (x) vs. year of release (y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $x \in \mathbb{R}^{90=d_x}$: audio features.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.



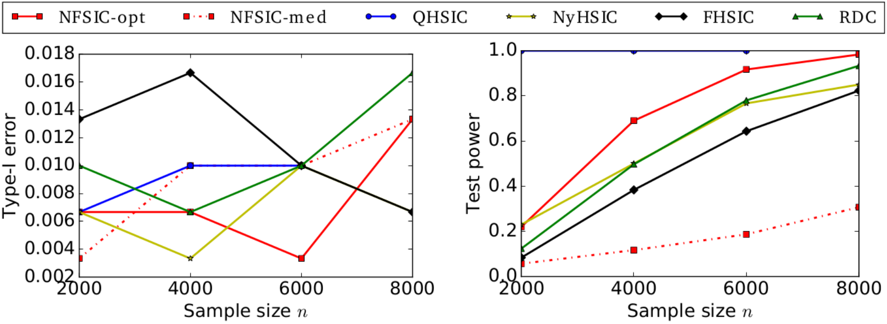
Demo-2: videos and captions

Youtube video (x) vs. caption (y).

- VideoStory46K [Habibian et al., 2014]
- $x \in \mathbb{R}^{2000=d_x}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $y \in \mathbb{R}^{1878=d_y}$: bag of words. TF.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.

Youtube video (x) vs. caption (y).

- VideoStory46K [Habibian et al., 2014]
- $x \in \mathbb{R}^{2000=d_x}$: Fisher vector encoding of motion boundary histograms [Wang and Schmid, 2013].
- $y \in \mathbb{R}^{1878=d_y}$: bag of words. TF.
- Left: break (x, y) pairs, i.e. H_0; right: H_1 is true.



Summary

- Focus: independence testing.
- HSIC: accurate but expensive.

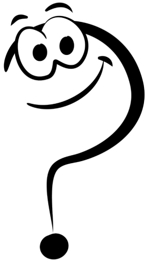
Summary

- Focus: independence testing.
- HSIC: accurate but expensive.
- We suggested 2 linear-time alternatives (FSIC, NFSIC).

Summary

- Focus: independence testing.
- HSIC: accurate but expensive.
- We suggested 2 linear-time alternatives (FSIC, NFSIC).
- NFSIC:
  - χ^2-test.
  - Adaptive.
  - Applications: song-year, video-caption.

Thank you for the attention!



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

HSIC versus FSIC

Question

Which one to choose?

- *HSIC* $= \|u\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$.
- *FSIC* $= \|u\|_{L^2(\mathcal{V})}$.

Question

Which one to choose?

- $HSIC = \|u\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$.
  - When $p_{xy} - p_x p_y$ is diffuse, close to flat.
- $FSIC = \|u\|_{L^2(\mathcal{V})}$.

HSIC versus FSIC

Question

Which one to choose?

- $HSIC = \|u\|_{\mathcal{H}_k \otimes \mathcal{H}_\ell}$.
    - When $p_{xy} - p_x p_y$ is diffuse, close to flat.
- $FSIC = \|u\|_{L^2(\mathcal{V})}$.
    - When $p_{xy} - p_x p_y$ is local, with many peaks.

Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011).
The million song dataset.
In *International Conference on Music Information Retrieval (ISMIR)*.

Gretton, A. (2015).
A simpler condition for consistency of a kernel independence test.
Technical report, University College London.
(`https://arxiv.org/abs/1501.06103`).

Habibian, A., Mensink, T., and Snoek, C. G. (2014).
Videostory: A new multimedia embedding for few-example recognition and translation of events.

In *ACM International Conference on Multimedia*, pages 17–26.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017).
An adaptive test of independence with analytic kernel embeddings.
In *International Conference on Machine Learning (ICML)*, Sydney, Australia.
(accepted).

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., and Lanckriet, G. R. G. (2010).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.

Wang, H. and Schmid, C. (2013).
Action recognition with improved trajectories.
In *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558.