

HSIC, A Measure of Independence?

Zoltán Szabó – CMAP, École Polytechnique



Joint work with: Bharath K. Sriperumbudur

LIONS, EPFL
February 28, 2018

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

- ① $I(\mathbb{P}) \geq 0$.

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

- ① $I(\mathbb{P}) \geq 0$.
- ② $I(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \bigotimes_{m=1}^M \mathbb{P}_m$.

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

- ① $I(\mathbb{P}) \geq 0$.
- ② $I(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \bigotimes_{m=1}^M \mathbb{P}_m$.

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

Euclidean Space → Inner Product → Kernel

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

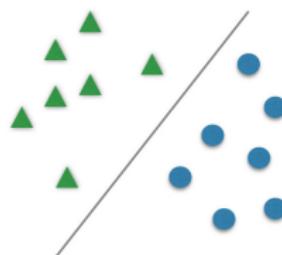
Euclidean Space → Inner Product → Kernel

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

Euclidean Space → Inner Product → Kernel

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

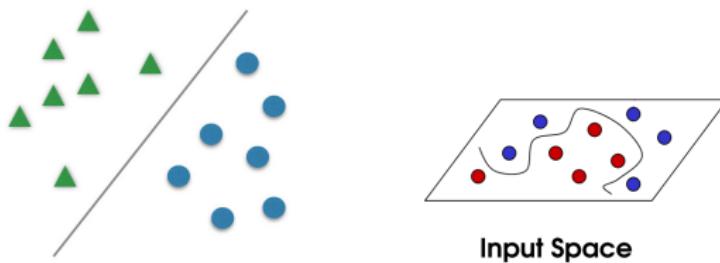
① Classification:



Euclidean Space → Inner Product → Kernel

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

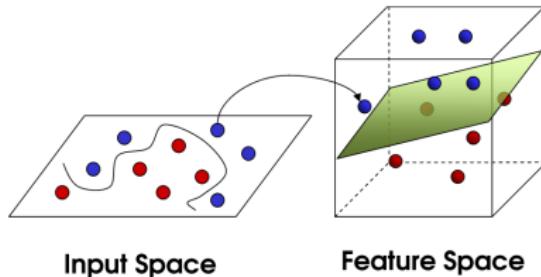
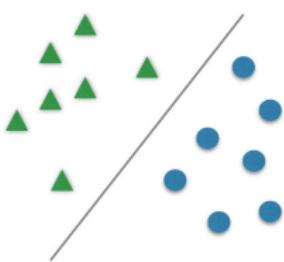
① Classification:



Euclidean Space → Inner Product → Kernel

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

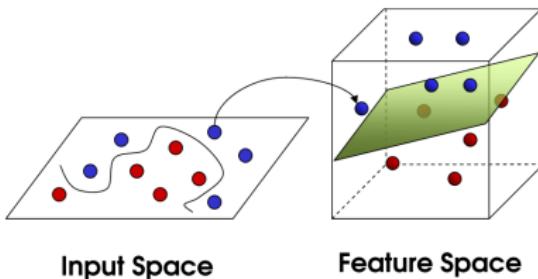
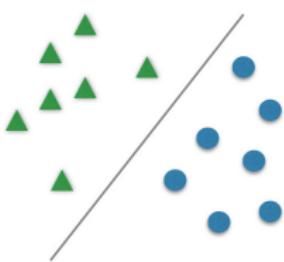
① Classification:



Euclidean Space → Inner Product → Kernel

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

① Classification:



② Representation of distributions:

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

$\varphi(\mathbf{x}) = \mathbf{x}$: mean, $\varphi(\mathbf{x}) = e^{i\langle \cdot, \mathbf{x} \rangle}$: characteristic function.

Diverse Set of Domains, Kernel Examples



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

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

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

Diverse Set of Domains, Kernel Examples



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

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

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

- $\mathcal{X} = \text{strings, texts}$:

- r -spectrum kernel: # of common $\leq r$ -substrings.

Diverse Set of Domains, Kernel Examples



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

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

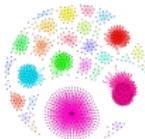
$$k_e(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2}, \quad k_C(\mathbf{x}, \mathbf{y}) = 1 + \frac{1}{\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}.$$

- $\mathcal{X} = \text{strings, texts}$:

- r -spectrum kernel: # of common $\leq r$ -substrings.

- $\mathcal{X} = \text{time-series: dynamic time-warping.}$

Diverse Set of Domains, Kernel Examples



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

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

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

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

Objects of Interest

'KL divergence & mutual information' on kernel-endowed domains.

- Mean embedding:

$$\mu(\mathbb{P}) := \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x)$$

Objects of Interest

'KL divergence & mutual information' on kernel-endowed domains.

- Mean embedding:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot, x)} d\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

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

Objects of Interest

'KL divergence & mutual information' on kernel-endowed domains.

- Mean embedding:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot, x)} d\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

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

- Hilbert-Schmidt independence criterion, $k = \otimes_{m=1}^M k_m$:

$$\text{HSIC}_k(\mathbb{P}) := \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right).$$

Objects of Interest

- Mean embedding:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{\varphi(x)}_{k(\cdot, x)} d\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy:

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

- Hilbert-Schmidt independence criterion, $k = \bigotimes_{m=1}^M k_m$:

$$\text{HSIC}_k(\mathbb{P}) := \text{MMD}_k\left(\mathbb{P}, \bigotimes_{m=1}^M \mathbb{P}_m\right).$$

When is HSIC an independence measure? Conditions on k_m -s?

Ingredients

Ingredients: Domain of the Distributions (\mathcal{X})

- HSIC $\Rightarrow \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$: product space.
- \mathcal{X}_m : different modalities \rightarrow images, texts, audio, ...



Ingredients: Domain of the Distributions (\mathcal{X})

- HSIC $\Rightarrow \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$: product space.
- \mathcal{X}_m : different modalities \rightarrow images, texts, audio, ...



Assumption

\mathcal{X}_m : kernel-enriched domains.

Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

Given: \mathcal{X} set. \mathcal{H} (ilbert space).

- Kernel:

$$k(\textcolor{red}{a}, \textcolor{red}{b}) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

Given: \mathcal{X} set. \mathcal{H} (ilbert space).

- Kernel:

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Reproducing kernel of a $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$:

$$\underbrace{k(\cdot, b)}_{\text{replicating function}} \in \mathcal{H}, \quad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

Given: \mathcal{X} set. \mathcal{H} (ilbert space).

- Kernel:

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Reproducing kernel of a $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$:

$$\underbrace{k(\cdot, b)}_{\text{reproducing property}} \in \mathcal{H}, \quad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



$$\xrightarrow{\text{spec.}} k(a, b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}.$$

Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

Given: \mathcal{X} set. \mathcal{H} (ilbert space).

- Kernel:

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Reproducing kernel of a $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$:

$$\underbrace{k(\cdot, b)}_{\text{a blue bell curve}} \in \mathcal{H}, \quad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



$$\xrightarrow{\text{spec.}} k(a, b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}. \quad \mathcal{H}_k = \overline{\{\sum_{i=1}^n \alpha_i k(\cdot, x_i)\}}.$$

Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

Given: \mathcal{X} set. \mathcal{H} (ilbert space).

- Kernel:

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}.$$

- Reproducing kernel of a $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$:

$$\underbrace{k(\cdot, b)}_{\text{replicates } f(b)} \in \mathcal{H}, \quad \underbrace{f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}}_{\text{reproducing property}}.$$



$$\xrightarrow{\text{spec.}} k(a, b) = \langle k(\cdot, a), k(\cdot, b) \rangle_{\mathcal{H}}. \quad \mathcal{H}_k = \overline{\{\sum_{i=1}^n \alpha_i k(\cdot, x_i)\}}.$$

Equivalent definitions. We represent distributions in an RKHS...

Mean Embedding

- Dirac measure: $\delta_x \mapsto k(\cdot, x)$.

Mean Embedding

- Dirac measure: $\delta_x \mapsto k(\cdot, x)$. Generally:

$$\mu_{\mathbb{P}} := \underbrace{\int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x)}_{\text{Bochner integral}} \in \mathcal{H}_k.$$

Mean Embedding

- Dirac measure: $\delta_x \mapsto k(\cdot, x)$. Generally:

$$\mu_{\mathbb{P}} := \underbrace{\int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x)}_{\text{Bochner integral}} \in \mathcal{H}_k.$$

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

Mean Embedding

- Dirac measure: $\delta_x \mapsto k(\cdot, x)$. Generally:

$$\mu_{\mathbb{P}} := \underbrace{\int_{\mathcal{X}} k(\cdot, x) d\mathbb{P}(x)}_{\text{Bochner integral}} \in \mathcal{H}_k.$$

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

Mean Embedding, MMD: Applications & Review

- Applications:
 - two-sample testing [Borgwardt et al., 2006, Gretton et al., 2012],
 - domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017],
 - kernel Bayesian inference [Song et al., 2011, Fukumizu et al., 2013]
 - approximate Bayesian computation [Park et al., 2016], probabilistic programming [Schölkopf et al., 2015],
 - model criticism [Lloyd et al., 2014, Kim et al., 2016],
 - distribution classification [Muandet et al., 2011, Zaheer et al., 2017], distribution regression [Szabó et al., 2016, Law et al., 2018],
 - topological data analysis [Kusano et al., 2016].
- Review [Muandet et al., 2017].

Let us switch to HSIC.

MMD with $\mathbf{k} = \otimes_{m=1}^M k_m$:

$$\mathbf{k}(x, x') := \prod_{m=1}^M k_m(x_m, x'_m),$$

$$\text{HSIC}_{\mathbf{k}}(\mathbb{P}) := \text{MMD}_{\mathbf{k}}\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right).$$

MMD with $k = \otimes_{m=1}^M k_m$:

$$k(x, x') := \prod_{m=1}^M k_m(x_m, x'_m),$$

$$\text{HSIC}_k(\mathbb{P}) := \text{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right).$$

Applications:

- blind source separation [Gretton et al., 2005],
- feature selection [Song et al., 2012], post selection inference [Yamada et al., 2016],
- independence testing [Gretton et al., 2008], causal inference [Mooij et al., 2016, Pfister et al., 2017, Strobl et al., 2017].

Central in Applications: Characteristic Property

- MMD: k is called **characteristic** [Fukumizu et al., 2008] if

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Central in Applications: Characteristic Property

- MMD: k is called **characteristic** [Fukumizu et al., 2008] if

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

- HSIC: $k = \otimes_{m=1}^M k_m$ will be called **\mathcal{I} -characteristic** if

$$\text{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^M \mathbb{P}_m.$$

Central in Applications: Characteristic Property

- MMD: k is called **characteristic** [Fukumizu et al., 2008] if

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

- HSIC: $k = \otimes_{m=1}^M k_m$ will be called **\mathcal{I} -characteristic** if

$$\text{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^M \mathbb{P}_m.$$

- $\otimes_{m=1}^M k_m$: characteristic $\Rightarrow \mathcal{I}$ -characteristic.

Central in Applications: Characteristic Property

- MMD: k is called **characteristic** [Fukumizu et al., 2008] if

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

- HSIC: $k = \otimes_{m=1}^M k_m$ will be called **\mathcal{I} -characteristic** if

$$\text{HSIC}_k(\mathbb{P}) = 0 \Leftrightarrow \mathbb{P} = \otimes_{m=1}^M \mathbb{P}_m.$$

- $\otimes_{m=1}^M k_m$: characteristic $\Rightarrow \mathcal{I}$ -characteristic.

Wanted

- $\otimes_{m=1}^M k_m$ is **\mathcal{I} -characteristic**: conditions in terms of k_m -s?
- $\otimes_{m=1}^M k_m$ is **characteristic**: relation?

Characteristic Property: Description on \mathbb{R}^d

For continuous bounded shift-invariant kernels on \mathbb{R}^d :

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') \stackrel{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x} - \mathbf{x}', \boldsymbol{\omega} \rangle} d\Lambda(\boldsymbol{\omega})$$

(*): Bochner's theorem.

Characteristic Property: Description on \mathbb{R}^d

For continuous bounded shift-invariant kernels on \mathbb{R}^d :

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') \stackrel{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x} - \mathbf{x}', \boldsymbol{\omega} \rangle} d\Lambda(\boldsymbol{\omega}) \Rightarrow \\ \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} = \|c_{\mathbb{P}} - c_{\mathbb{Q}}\|_{L^2(\Lambda)}.$$

(*): Bochner's theorem, $c_{\mathbb{P}}$: characteristic function of \mathbb{P} .

Characteristic Property: Description on \mathbb{R}^d

For continuous bounded shift-invariant kernels on \mathbb{R}^d :

$$k(\mathbf{x}, \mathbf{x}') = k_0(\mathbf{x} - \mathbf{x}') \stackrel{(*)}{=} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x} - \mathbf{x}', \boldsymbol{\omega} \rangle} d\Lambda(\boldsymbol{\omega}) \Rightarrow \\ \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k} = \|c_{\mathbb{P}} - c_{\mathbb{Q}}\|_{L^2(\Lambda)}.$$

(*): Bochner's theorem, $c_{\mathbb{P}}$: characteristic function of \mathbb{P} .

Theorem ([Sriperumbudur et al., 2010])

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

Examples on \mathbb{R} ; Similarly \mathbb{R}^d

kernel name k_0	$\hat{k}_0(\omega)$	$supp(\hat{k}_0)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2 \omega^2}{2}}$
Laplacian	$e^{-\sigma x }$	$\sqrt{\frac{2}{\pi}} \frac{\sigma}{\sigma^2 + \omega^2}$
B_{2n+1} -spline	$*^{2n+2} \chi_{[-\frac{1}{2}, \frac{1}{2}]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}} \chi_{[-\sigma, \sigma]}(\omega)$
Fejér	$\frac{1}{n+1} \frac{\sin^2 \frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi} \sum_{j=-n}^n \left(1 - \frac{ j }{n+1}\right) \delta(\omega - j)$
		$\{0, \pm 1, \pm 2, \dots, \pm n\}$

k is called **characteristic** (recall!)/**universal** if

$$\mu_k : \underbrace{\mathcal{M}_1^+(\mathcal{X})}_{\text{probability measures on } \mathcal{X}} \mapsto \mathcal{H}_k, \quad \mu_k : \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{bounded signed measures on } \mathcal{X}} \mapsto \mathcal{H}_k$$

is injective.

k is called **characteristic** (recall!)/**universal** if

$$\mu_k : \underbrace{\mathcal{M}_1^+(\mathcal{X})}_{\text{probability measures on } \mathcal{X}} \mapsto \mathcal{H}_k, \quad \mu_k : \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{bounded signed measures on } \mathcal{X}} \mapsto \mathcal{H}_k$$

is injective.

- Example: $\mathcal{M}_b(\mathcal{X}) \ni \mathbb{P} = \bigotimes_{m=1}^M \mathbb{P}_m$.

k is called **characteristic** (recall!)/**universal** if

$$\mu_k : \underbrace{\mathcal{M}_1^+(\mathcal{X})}_{\text{probability measures on } \mathcal{X}} \mapsto \mathcal{H}_k, \quad \mu_k : \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{bounded signed measures on } \mathcal{X}} \mapsto \mathcal{H}_k$$

is injective.

- Example: $\mathcal{M}_b(\mathcal{X}) \ni \mathbb{P} = \bigotimes_{m=1}^M \mathbb{P}_m$.
- **Universal** \Rightarrow characteristic $\Rightarrow \mathcal{I}$ -characteristic.

k is called **characteristic** (recall!)/**universal** if

$$\mu_k : \underbrace{\mathcal{M}_1^+(\mathcal{X})}_{\text{probability measures on } \mathcal{X}} \mapsto \mathcal{H}_k, \quad \mu_k : \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{bounded signed measures on } \mathcal{X}} \mapsto \mathcal{H}_k$$

is injective.

- Example: $\mathcal{M}_b(\mathcal{X}) \ni \mathbb{P} = \bigotimes_{m=1}^M \mathbb{P}_m$.
- **Universal** \Rightarrow characteristic $\Rightarrow \mathcal{I}$ -characteristic.

Challenge

Characteristic/ \mathcal{I} -characteristic/universality of $\bigotimes_{m=1}^M k_m$ in terms of k_m -s!

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:
 $k_1 \& k_2$: universal $\Rightarrow k_1 \otimes k_2$: universal ($\Rightarrow \mathcal{I}$ -characteristic).

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:
 $k_1 \& k_2$: universal $\Rightarrow k_1 \otimes k_2$: universal ($\Rightarrow \mathcal{I}$ -characteristic).
- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013]:
 $k_1 \& k_2$: characteristic $\Leftrightarrow k_1 \otimes k_2$: \mathcal{I} -characteristic.

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:
 $k_1 \& k_2$: universal $\Rightarrow k_1 \otimes k_2$: universal ($\Rightarrow \mathcal{I}$ -characteristic).
- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013]:
 $k_1 \& k_2$: characteristic $\Leftrightarrow k_1 \otimes k_2$: \mathcal{I} -characteristic.

Goal

Extension to $M \geq 2$.

Existing Results, $M = 2$

- [Blanchard et al., 2011, Waegeman et al., 2012, Gretton, 2015]:
 $k_1 \& k_2$: universal $\Rightarrow k_1 \otimes k_2$: universal ($\Rightarrow \mathcal{I}$ -characteristic).
- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013]:
 $k_1 \& k_2$: characteristic $\Leftrightarrow k_1 \otimes k_2$: \mathcal{I} -characteristic.

Goal

Extension to $M \geq 2$.

Main Challenge

' $\otimes k_m$: \mathcal{I} -characteristic $\Leftrightarrow k_m$: characteristic ($\forall m$)' does NOT hold.

Idea: Characteristic Property as pd

- Characteristic property:

$$\mathbb{F} = \mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_{\mathbb{F}} \neq 0.$$

Idea: Characteristic Property as pd

- Characteristic property:

$$\mathbb{F} = \mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_{\mathbb{F}} \neq 0.$$

Here: $\mathbb{F} \in \mathcal{M}_b(\mathcal{X})$, $\mathbb{F}(\mathcal{X}) = \underbrace{\mathbb{P}_1(\mathcal{X})}_{1} - \underbrace{\mathbb{P}_2(\mathcal{X})}_{1} = 0$.

Idea: Characteristic Property as pd

- Characteristic property:

$$\mathbb{F} = \mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_{\mathbb{F}} \neq 0.$$

Here: $\mathbb{F} \in \mathcal{M}_b(\mathcal{X})$, $\mathbb{F}(\mathcal{X}) = \underbrace{\mathbb{P}_1(\mathcal{X})}_1 - \underbrace{\mathbb{P}_2(\mathcal{X})}_1 = 0$.

- Observation [Sriperumbudur et al., 2010]: k is characteristic iff.

$$\|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 > 0, \quad \forall \underbrace{\mathbb{F} \in \mathcal{M}_b(\mathcal{X}) \setminus \{0\} \quad \mathbb{F}(\mathcal{X}) = 0}_{\mathcal{F}_1}.$$

Idea: Characteristic Property as pd

- Characteristic property:

$$\mathbb{F} = \mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_{\mathbb{F}} \neq 0.$$

Here: $\mathbb{F} \in \mathcal{M}_b(\mathcal{X})$, $\mathbb{F}(\mathcal{X}) = \underbrace{\mathbb{P}_1(\mathcal{X})}_{1} - \underbrace{\mathbb{P}_2(\mathcal{X})}_{1} = 0$.

- Observation [Sriperumbudur et al., 2010]: k is characteristic iff.

$$\|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 > 0, \quad \forall \underbrace{\mathbb{F} \in \mathcal{M}_b(\mathcal{X}) \setminus \{0\}}_{\mathcal{F}_1} \quad \mathbb{F}(\mathcal{X}) = 0.$$

- We saw: k is universal iff.

$$\|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 > 0, \quad \forall \underbrace{\mathbb{F} \in \mathcal{M}_b(\mathcal{X}) \setminus \{0\}}_{\mathcal{F}_2}.$$

Examples

\mathcal{F}	$\mathcal{F}\text{-pd } k$
$\mathcal{M}_b(\mathcal{X})$	universal
$[\mathcal{M}_b(\mathcal{X})]^0$	characteristic

$$[\mathcal{M}_b(\mathcal{X})]^0 \subseteq \mathcal{M}_b(\mathcal{X}).$$

$$\text{characteristic} \Leftarrow \text{universal}.$$

Examples

\mathcal{F}	$\mathcal{F}\text{-pd } k$
$\mathcal{M}_b(\mathcal{X})$	universal
$[\mathcal{M}_b(\mathcal{X})]^0$	characteristic
$\mathcal{I} := \{\mathbb{P} - \otimes_{m=1}^M \mathbb{P}_m\}$	\mathcal{I} -characteristic

$$\mathcal{I} \subseteq [\mathcal{M}_b(\mathcal{X})]^0 \subseteq \mathcal{M}_b(\mathcal{X}).$$

$$\mathcal{I}\text{-characteristic} \Leftarrow \text{characteristic} \Leftarrow \text{universal}.$$

$\otimes_{m=1}^M k_m :$

\mathcal{I} -char \longleftrightarrow char \longleftrightarrow universal



$(k_m)_{m=1}^M :$

char $\xrightarrow{\text{[Sriperumbudur et al., 2011]}}$ -universal
 $\xleftarrow{\text{[Sriperumbudur et al., 2011]}}$

Results

Characteristic Property of $\otimes_{m=1}^M k_m$

Proposition

- (i) $\otimes_{m=1}^M k_m$: characteristic $\Rightarrow (k_m)_{m=1}^M$ are characteristic.
- (ii) \Leftarrow

Characteristic Property of $\otimes_{m=1}^M k_m$

Proposition

- (i) $\otimes_{m=1}^M k_m$: characteristic $\Rightarrow (k_m)_{m=1}^M$ are characteristic.
- (ii) \Leftarrow

Example

- $\mathcal{X}_m = \{1, 2\}$, $\tau_{\mathcal{X}_m} = \mathcal{P}(\{1, 2\})$, $k_m(x, x') = 2\delta_{x,x'} - 1$, $M = 2$.
- $k_1 = k_2$: characteristic, but $k_1 \otimes k_2$ is not characteristic.
- $k_1 \otimes k_2$ is \mathcal{I} -characteristic.

Characteristic Property of $\otimes_{m=1}^M k_m$

Proposition

- (i) $\otimes_{m=1}^M k_m$: characteristic $\Rightarrow (k_m)_{m=1}^M$ are characteristic.
- (ii) \Leftarrow

Example

- $\mathcal{X}_m = \{1, 2\}$, $\tau_{\mathcal{X}_m} = \mathcal{P}(\{1, 2\})$, $k_m(x, x') = 2\delta_{x,x'} - 1$, $M = 2$.
- $k_1 = k_2$: characteristic, but $k_1 \otimes k_2$ is not characteristic.
- $k_1 \otimes k_2$ is \mathcal{I} -characteristic.

Witness: $0 \neq \mathbb{F}_{\mathbf{A}} \in \mathcal{M}_b(\mathcal{X})$, $\mathbf{A} \in \mathbb{R}^{2 \times 2}$,

$$0 = \mathbb{F}(\mathcal{X}), \quad 0 = \|\mu_k(\mathbb{F})\|_{\mathcal{H}_k}^2.$$

\mathcal{I} -characteristic Property

In the previous example:

$$k_1, k_2: \text{characteristic} \Rightarrow k_1 \otimes k_2: \mathcal{I}\text{-characteristic}.$$

In fact:

- this holds for any bounded kernel,
- +converse for any $M \geqslant 2$!

k_1, k_2, k_3 : characteristic $\Rightarrow \otimes_{m=1}^3 k_m$: \mathcal{I} -characteristic

Example

- $\mathcal{X}_m = \{1, 2\}$, $\tau_{\mathcal{X}_m} = \mathcal{P}(\{1, 2\})$, $k_m(x, x') = 2\delta_{x,x'} - 1$, $M = 3$.
- Then
 - $(k_m)_{m=1}^3$: characteristic.
 - $\otimes_{m=1}^3 k_m$: is **not** \mathcal{I} -characteristic. Witness:

$$p_{1,1,1} = \frac{1}{5}, \quad p_{1,1,2} = \frac{1}{10}, \quad p_{1,2,1} = \frac{1}{10}, \quad p_{1,2,2} = \frac{1}{10},$$
$$p_{2,1,1} = \frac{1}{5}, \quad p_{2,1,2} = \frac{1}{10}, \quad p_{2,2,1} = \frac{1}{10}, \quad p_{2,2,2} = \frac{1}{10}.$$

Non- \mathcal{I} -characteristicity: Analytical Solution

Parameter: $\mathbf{z} = (z_0, z_1, \dots, z_5) \in [0, 1]^6$.

Non- \mathcal{I} -characteristicity: Analytical Solution

Parameter: $\mathbf{z} = (z_0, z_1, \dots, z_5) \in [0, 1]^6$. Example: $p_{1,1,1} =$

$$\frac{z_2 + z_1 + z_4 + z_5 - 3z_2z_1 - 4z_2z_4 - 4z_1z_4 - z_2z_3 - 2z_2z_0 - 2z_1z_3 - 3z_2z_5 - 2z_4z_3 - z_1z_0 - 3z_1z_5 - 2z_4z_0 - 4z_4z_5 - z_3z_0 - z_3z_5 - z_0z_5 + 2z_2z_1^2 + 2z_2^2z_1 + 4z_2z_4^2 + 2z_2^2z_4 + 4z_1z_4^2 + 2z_1^2z_4 + 2z_2^2z_0 + 2z_1^2z_3 + 2z_2z_5^2 + 2z_2^2z_5 + 2z_4^2z_3 + 2z_1z_5^2 + 2z_1^2z_5 + 2z_4z_0^2 + 2z_4z_5^2 + 4z_4^2z_5 - z_2^2 - z_1^2 - 3z_4^2 + 2z_4^3 - z_5^2 + 6z_2z_1z_4 + 2z_2z_1z_3 + 2z_2z_4z_3 + 2z_2z_1z_0 + 4z_2z_1z_5 + 4z_2z_4z_0 + 4z_1z_4z_3 + 6z_2z_4z_5 + 2z_1z_4z_0 + 6z_1z_4z_5 + 2z_2z_3z_0 + 2z_2z_3z_5 + 2z_1z_3z_0 + 2z_2z_0z_5 + 2z_1z_3z_5 + 2z_4z_3z_0 + 2z_4z_3z_5 + 2z_1z_0z_5 + 2z_4z_0z_5}{2z_2z_1 - z_1 - 2z_4 - z_3 - z_0 - 2z_5 - z_2 + 2z_2z_4 + 2z_1z_4 + 2z_2z_0 + 2z_1z_3 + 2z_2z_5 + 2z_4z_3 + 2z_1z_5 + 2z_4z_0 + 4z_4z_5 + 2z_3z_0 + 2z_3z_5 + 2z_0z_5 + 2z_4^2 + 2z_5^2}.$$

Non- \mathcal{I} -characteristicity: Analytical Solution

Parameter: $\mathbf{z} = (z_0, z_1, \dots, z_5) \in [0, 1]^6$. Example: $p_{1,1,1} =$

$$\frac{z_2 + z_1 + z_4 + z_5 - 3z_2z_1 - 4z_2z_4 - 4z_1z_4 - z_2z_3 - 2z_2z_0 - 2z_1z_3 - 3z_2z_5 - 2z_4z_3 - z_1z_0 - 3z_1z_5 - 2z_4z_0 - 4z_4z_5 - z_3z_0 - z_3z_5 - z_0z_5 + 2z_2z_1^2 + 2z_2^2z_1 + 4z_2z_4^2 + 2z_2^2z_4 + 4z_1z_4^2 + 2z_1^2z_4 + 2z_2^2z_0 + 2z_1^2z_3 + 2z_2z_5^2 + 2z_2^2z_5 + 2z_4^2z_3 + 2z_1z_5^2 + 2z_1^2z_5 + 2z_4^2z_0 + 2z_4z_5^2 + 4z_4^2z_5 - z_2^2 - z_1^2 - 3z_4^2 + 2z_4^3 - z_5^2 + 6z_2z_1z_4 + 2z_2z_1z_3 + 2z_2z_4z_3 + 2z_2z_1z_0 + 4z_2z_1z_5 + 4z_2z_4z_0 + 4z_1z_4z_3 + 6z_2z_4z_5 + 2z_1z_4z_0 + 6z_1z_4z_5 + 2z_2z_3z_0 + 2z_2z_3z_5 + 2z_1z_3z_0 + 2z_2z_0z_5 + 2z_1z_3z_5 + 2z_4z_3z_0 + 2z_4z_3z_5 + 2z_1z_0z_5 + 2z_4z_0z_5}{2z_2z_1 - z_1 - 2z_4 - z_3 - z_0 - 2z_5 - z_2 + 2z_2z_4 + 2z_1z_4 + 2z_2z_0 + 2z_1z_3 + 2z_2z_5 + 2z_4z_3 + 2z_1z_5 + 2z_4z_0 + 4z_4z_5 + 2z_3z_0 + 2z_3z_5 + 2z_0z_5 + 2z_4^2 + 2z_5^2}.$$

We chose: $\mathbf{z} = \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right)$.

Proposition

Assume $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \rightarrow \mathbb{R}$ are continuous, translation-invariant kernels. Then the followings are equivalent:

- (i) $(k_m)_{m=1}^M$ -s are characteristic.
- (ii) $\otimes_{m=1}^M k_m$: \mathcal{I} -characteristic.
- (iii) $\otimes_{m=1}^M k_m$: characteristic.

Proposition

Assume $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \rightarrow \mathbb{R}$ are continuous, translation-invariant kernels. Then the followings are equivalent:

- (i) $(k_m)_{m=1}^M$ -s are characteristic.
- (ii) $\otimes_{m=1}^M k_m$: \mathcal{I} -characteristic.
- (iii) $\otimes_{m=1}^M k_m$: characteristic.

We already know

$$(iii) \Rightarrow (ii) \Rightarrow (i).$$

Proposition

Assume $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \rightarrow \mathbb{R}$ are continuous, translation-invariant kernels. Then the followings are equivalent:

- (i) $(k_m)_{m=1}^M$ -s are characteristic.
- (ii) $\otimes_{m=1}^M k_m$: \mathcal{I} -characteristic.
- (iii) $\otimes_{m=1}^M k_m$: characteristic.

We already know

$$(iii) \Rightarrow (ii) \Rightarrow (i).$$

Remains: $(iii) \Leftarrow (i)$. Proof: Bochner theorem.

Universality of $\otimes_{m=1}^M k_m$

We saw: for $M \geq 3$

$(k_m)_{m=1}^M$ are characteristic $\Rightarrow \otimes_{m=1}^M k_m$: \mathcal{I} -characteristic.

Proposition

$\otimes_{m=1}^M k_m$: universal $\Leftrightarrow (k_m)_{m=1}^M$ are universal.

The Tricky Direction: If $(k_m)_{m=1}^M$ are Universal . . .

Goal: injectivity of $\mu = \mu_{\otimes_{m=1}^M k_m}$ on $\mathcal{M}_b(\mathcal{X})$, i.e.

$$\mu(\mathbb{F}) = 0 \stackrel{?}{\Rightarrow} \mathbb{F} = 0.$$

The Tricky Direction: If $(k_m)_{m=1}^M$ are Universal . . .

Goal: injectivity of $\mu = \mu_{\otimes_{m=1}^M k_m}$ on $\mathcal{M}_b(\mathcal{X})$, i.e.

$$\mu(\mathbb{F}) = 0 \stackrel{?}{\Rightarrow} \mathbb{F} = 0.$$

Enough:

$$\mathbb{F}\left(\times_{m=1}^M B_m\right) = 0, \quad \forall B_m.$$



Proof Idea

$$0 = \mu(\mathbb{F}) = \int_{\mathcal{X}} \otimes_{m=1}^M k_m(\cdot, x_m) d\mathbb{F}(x),$$

$$0 = \mathbb{F}\left(\times_{m=1}^M B_m\right) = \int_{\mathcal{X}} \times_{m=1}^M \chi_{B_m}(x_m) d\mathbb{F}(x), \quad \forall B_m.$$

Proof Idea

$$0 = \mu(\mathbb{F}) = \int_{\mathcal{X}} \otimes_{m=1}^M k_m(\cdot, x_m) d\mathbb{F}(x),$$

$$0 = \int_{\mathcal{X}} \prod_{m=1}^{\textcolor{red}{J}} \chi_{B_m}(x_m) \otimes_{m=\textcolor{red}{J}+1}^M k_m(\cdot, x_m) d\mathbb{F}(x), \quad \forall B_m,$$

$$0 = \mathbb{F}\left(\times_{m=1}^M B_m\right) = \int_{\mathcal{X}} \times_{m=1}^M \chi_{B_m}(x_m) d\mathbb{F}(x), \quad \forall B_m.$$

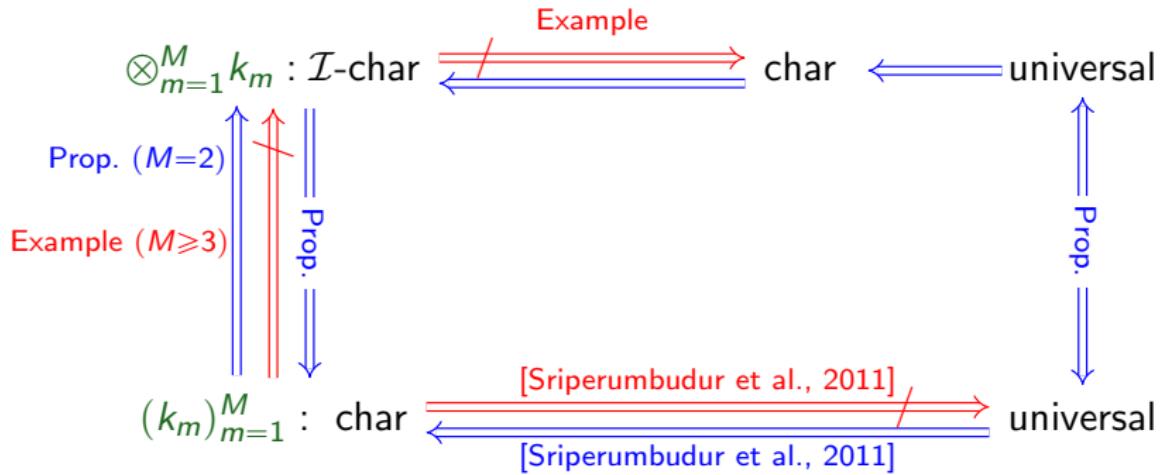
Proof Idea

$$0 = \mu(\mathbb{F}) = \int_{\mathcal{X}} \otimes_{m=1}^M k_m(\cdot, x_m) d\mathbb{F}(x),$$

$$0 = \int_{\mathcal{X}} \prod_{m=1}^{\textcolor{red}{J}} \chi_{B_m}(x_m) \otimes_{m=\textcolor{red}{J}+1}^M k_m(\cdot, x_m) d\mathbb{F}(x), \quad \forall B_m,$$

$$0 = \mathbb{F}\left(\times_{m=1}^M B_m\right) = \int_{\mathcal{X}} \times_{m=1}^M \chi_{B_m}(x_m) d\mathbb{F}(x), \quad \forall B_m.$$

We proceed by induction ($\textcolor{red}{J} = 0, \dots, M$).



We studied the validness of HSIC.

- HSIC \Rightarrow product structure:
 - Space: $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$.
 - Kernel: $k = \otimes_{m=1}^M k_m$.
- \mathcal{F} -pd property \Rightarrow complete answer in terms of k_m -s.

Summary

We studied the validness of HSIC.

- HSIC \Rightarrow product structure:
 - Space: $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$.
 - Kernel: $k = \otimes_{m=1}^M k_m$.
- \mathcal{F} -pd property \Rightarrow complete answer in terms of k_m -s.
- ITE toolkit, preprint (maths \rightarrow JMLR):

<https://bitbucket.org/szzoli/ite/>

<http://arxiv.org/abs/1708.08157>

Thank you for the attention!

Acks: A part of the work was carried out while BKS was visiting ZSz at CMAP, École Polytechnique. BKS is supported by NSF-DMS-1713011.

-  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>.
-  Blanchard, G., Lee, G., and Scott, C. (2011).
Generalizing from several related classification tasks to a new unlabeled sample.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186.
-  Borgwardt, K., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schlkopf, B., and Smola, A. J. (2006).
Integrating structured biological data by kernel maximum mean discrepancy.
Bioinformatics, 22:e49–57.
-  Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.

In *Neural Information Processing Systems (NIPS)*, pages 498–496.



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. (2015).

A simpler condition for consistency of a kernel independence test.

Technical report, University College London.

(<http://arxiv.org/abs/1501.06103>).



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., Bousquet, O., Smola, A., and Schölkopf, B. (2005).

Measuring statistical dependence with Hilbert-Schmidt norms.
In *Algorithmic Learning Theory (ALT)*, pages 63–78.

-  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.
-  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.

 Law, H. C. L., Sutherland, D. J., Sejdinovic, D., and Flaxman, S. (2018).

Bayesian approaches to distribution regression.

In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

 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.

 Lyons, R. (2013).

Distance covariance in metric spaces.

The Annals of Probability, 41:3284–3305.

 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.

Foundations and Trends in Machine Learning, 10(1-2):1–141.

-  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.

-  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).
-  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., Sriperumbudur, B. K., Gretton, A., and Fukumizu, K. (2013).
Equivalence of distance-based and RKHS-based statistics in hypothesis testing.
Annals of Statistics, 41:2263–2291.

-  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.
-  Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012).
Feature selection via dependence maximization.
Journal of Machine Learning Research, 13:1393–1434.
-  Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.
-  Steinwart, I. (2001).
On the influence of the kernel on the consistency of support vector machines.

-  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., Sriperumbudur, B., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40.
-  Waegeman, W., Pahikkala, T., Airola, A., Salakoski, T., Stock, M., and Baets, B. D. (2012). A kernel-based framework for learning graded relations from data. *IEEE Transactions on Fuzzy Systems*, 20:1090–1101.
-  Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2016).

Post selection inference with kernels.

Technical report.

(<https://arxiv.org/abs/1610.03725>).



Zaheer, M., Kottur, S., Ravanbakhsh, S., Póczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017).

Deep sets.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 3394–3404.



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.