

Characteristic Tensor Product Kernels

Zoltán Szabó – CMAP, École Polytechnique



Joint work with: Bharath K. Sriperumbudur
(Department of Statistics, PSU)

ISNPS, Salerno, Italy
June 14, 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:

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

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:

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

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:

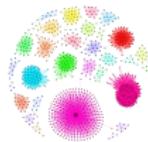
$$I(\mathbb{P}) = KL\left(\mathbb{P}, \bigotimes_{m=1}^M \mathbb{P}_m\right).$$

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

From \mathbb{R}^d to Diverse Set of Domains, Kernel Examples



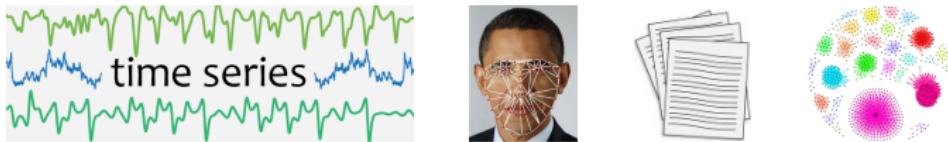
- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$$

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

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

From \mathbb{R}^d to Diverse Set of Domains, Kernel Examples



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

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$$

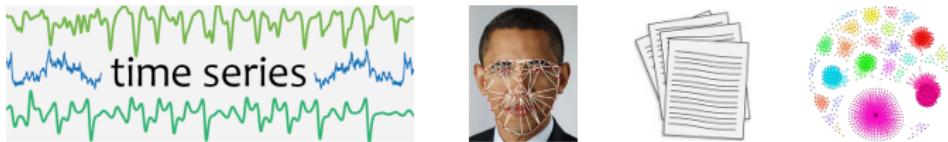
$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, \quad k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

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

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

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

From \mathbb{R}^d to Diverse Set of Domains, Kernel Examples



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

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$$

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

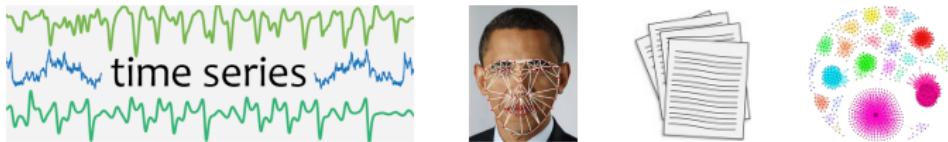
$$k_e(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2}, \quad k_C(\mathbf{x}, \mathbf{y}) = 1 + \frac{1}{\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}.$$

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

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

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

From \mathbb{R}^d to Diverse Set of Domains, Kernel Examples



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

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$$

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

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

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

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

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

- $\mathcal{X} = \text{graphs, sets, permutations, ...}$

Distribution Representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x).$$

Distribution Representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x).$$

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(z) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, z)}(x).$$

Distribution Representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x).$$

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(z) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, z)}(x).$$

- Characteristic function:

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

Distribution Representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x).$$

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(z) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, z)}(x).$$

- Characteristic function:

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

- Moment generating function:

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

Distribution Representation

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x).$$

- Cdf:

$$\mathbb{P} \mapsto F_{\mathbb{P}}(z) = \mathbb{E}_{x \sim \mathbb{P}} \chi_{(-\infty, z)}(x).$$

- Characteristic function:

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

- Moment generating function:

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

Trick

φ : on any kernel-endowed domain!

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

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

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

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

- Kernel:

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

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

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

- Kernel:

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

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

$$\underbrace{k(\cdot, b)}_{\text{a function}} \in \mathcal{H},$$



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

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

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

- Kernel:

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

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

$$\underbrace{k(\cdot, b)}_{\text{a function}} \in \mathcal{H},$$



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

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

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

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

- Kernel:

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

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

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


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

- 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], goodness-of-fit [Balasubramanian et al., 2017],
 - distribution classification [Muandet et al., 2011, Lopez-Paz et al., 2015], [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].

Switching to HSIC ...

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

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

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

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., 2018],
- independence testing [Gretton et al., 2008], causal inference [Mooij et al., 2016, Pfister et al., 2017, Strobl et al., 2017].

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

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

Injectivity of $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ on finite signed measures: **universality** [Steinwart, 2001].

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

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

Injectivity of $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ on finite signed measures: **universality** [Steinwart, 2001].

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

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

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

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

Injectivity of $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ on finite signed measures: **universality** [Steinwart, 2001].

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

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

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

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

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

Injectivity of $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ on finite signed measures: **universality** [Steinwart, 2001].

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

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

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

Wanted

- Characteristic properties of $\otimes_{m=1}^M k_m$ **in terms of k_m -s?**

Theorem ([Sriperumbudur et al., 2010])

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

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

Theorem ([Sriperumbudur et al., 2010])

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

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

Example on \mathbb{R} :

kernel name	k_0	$\hat{k}_0(\omega)$	$\text{supp}(\hat{k}_0)$
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}
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}} \chi_{[-\sigma, \sigma]}(\omega)$	$[-\sigma, \sigma]$

- [Blanchard et al., 2011, Gretton, 2015]:

$k_1 \& k_2$: universal $\Rightarrow k_1 \otimes k_2$: universal ($\Rightarrow \mathcal{I}$ -characteristic).

- [Blanchard et al., 2011, 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.

- [Blanchard et al., 2011, 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 $M \geq 2$.



Discrete case: 'easy', e.g. k_1, k_2 : char $\Rightarrow k_1 \otimes k_2$: char.

- Characteristic property:

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

Discrete case: 'easy', e.g. k_1, k_2 : char $\Rightarrow k_1 \otimes k_2$: char.

- Characteristic property:

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

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

$$\forall \mathbb{F} \in \{c(\mathbb{P}_1 - \mathbb{P}_2) : c \in \mathbb{R} \setminus \{0\}\} \text{ & } \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 > 0.$$

Discrete case: 'easy', e.g. k_1, k_2 : char $\Rightarrow k_1 \otimes k_2$: char.

- Characteristic property:

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

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

$$\forall \mathbb{F} \in \{c(\mathbb{P}_1 - \mathbb{P}_2) : c \in \mathbb{R}\} \setminus \{0\} \text{ & } \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 > 0.$$

- Witness construction :

$$\exists \mathbb{F} \in \mathcal{M}_b(\mathcal{X}) \setminus \{0\} \text{ & } \mathbb{F}(\mathcal{X}) = 0 \text{ for which } \|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 = 0.$$

Discrete case: 'easy', e.g. k_1, k_2 : char $\Rightarrow k_1 \otimes k_2$: char.

- Characteristic property:

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

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

$$\forall \mathbb{F} \in \{c(\mathbb{P}_1 - \mathbb{P}_2) : c \in \mathbb{R}\} \setminus \{0\} \text{ & } \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 > 0.$$

- Witness construction :

$$\underbrace{\exists \mathbb{F} \in \mathcal{M}_b(\mathcal{X}) \setminus \{0\} \text{ & } \underbrace{\mathbb{F}(\mathcal{X}) = 0}_{eq_1(\mathbf{A})=0} \text{ for which } \underbrace{\|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 = 0}_{eq_2(\mathbf{A})=0}}_{\mathbf{A}:=(a_{ij})}.$$

Discrete case: 'easy', e.g. k_1, k_2 : char $\Rightarrow k_1 \otimes k_2$: char.

- Characteristic property:

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

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

$$\forall \mathbb{F} \in \{c(\mathbb{P}_1 - \mathbb{P}_2) : c \in \mathbb{R}\} \setminus \{0\} \text{ & } \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 > 0.$$

- Witness construction :

$$\underbrace{\exists \mathbb{F} \in \mathcal{M}_b(\mathcal{X}) \setminus \{0\} \text{ & } \mathbb{F}(\mathcal{X}) = 0}_{\mathbf{A} := (a_{ij})} \text{ for which } \underbrace{\|\mu_{\mathbb{F}}\|_{\mathcal{H}_k}^2 = 0}_{eq_2(\mathbf{A}) = 0}.$$

Example: $\mathcal{X}_m = \{1, 2\}$, $k_m(x, x') = 2\delta_{x,x'} - 1$ (solvable for $\mathbf{A} \neq \mathbf{0}$).

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: $p_{1,1,1} =$

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

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

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: $\mathbf{z} = \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right)$. Universality: helps?

Example

- $\mathcal{X}_m = \{1, 2\}$, $\tau_{\mathcal{X}_m} = \mathcal{P}(\{1, 2\})$, $M = 3$.
- $k_1(x, x') = k_2(x, x') = \delta_{x,x'}$: universal.
- $k_3(x, x') = 2\delta_{x,x'} - 1$: characteristic.
- Different constraints & $P(\mathbf{z})$ solution; same witness: useful.

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

Proposition (characteristic property)

- $\otimes_{m=1}^M k_m$: characteristic $\Rightarrow (k_m)_{m=1}^M$ are characteristic.
- $\Leftrightarrow [|\mathcal{X}_m| = 2, k_m(x, x') = 2\delta_{x,x'} - 1]$

Results [Szabó and Sriperumbudur, 2017]

Proposition (characteristic property)

- $\otimes_{m=1}^M k_m$: characteristic $\Rightarrow (k_m)_{m=1}^M$ are characteristic.
- $\Leftrightarrow |\mathcal{X}_m| = 2, k_m(x, x') = 2\delta_{x,x'} - 1$

Proposition (\mathcal{I} -characteristic property)

- k_1, k_2 : characteristic $\Rightarrow k_1 \otimes k_2$: \mathcal{I} -characteristic.
- \Leftrightarrow : for $\forall M \geq 2$.
- k_1, k_2, k_3 : characteristic $\Rightarrow \otimes_{m=1}^3 k_m$: \mathcal{I} -characteristic [Ex].
- k_1, k_2 : universal, k_3 : char $\Rightarrow \otimes_{m=1}^3 k_m$: \mathcal{I} -characteristic [Ex].

Proposition ($\mathcal{X}_m = \mathbb{R}^{d_m}$, k_m : continuous, shift-invariant, bounded)

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

Proposition ($\mathcal{X}_m = \mathbb{R}^{d_m}$, k_m : continuous, shift-invariant, bounded)

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

Proposition (Universality)

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

We studied the validness of HSIC.

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

We studied the validness of HSIC.

- HSIC \Rightarrow product structure:
 - Space: $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$.
 - Kernel: $k = \otimes_{m=1}^M k_m$.
- Complete answer in terms of k_m -s.
- ITE toolkit, preprint (with minor revision in 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.

-  Balasubramanian, K., Li, T., and Yuan, M. (2017).
On the optimality of kernel-embedding based goodness-of-fit tests.
Technical report.
(<https://arxiv.org/abs/1709.08148>).
-  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., Schölkopf, 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.

-  Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015).
Towards a learning theory of cause-effect inference.
International Conference on Machine Learning (ICML; PMLR), 37:1452–1461.
-  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 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.

Journal of Machine Learning Research, 6(3):67–93.

 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. and Sriperumbudur, B. (2017).
Characteristic and universal tensor product kernels.
Technical report.
(<http://arxiv.org/abs/1708.08157>).
-  Szabó, Z., Sriperumbudur, B., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.
Journal of Machine Learning Research, 17(152):1–40.
-  Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2018).
Post selection inference with kernels.
In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 84, pages 152–160.
-  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.