Characteristic Tensor Kernels

Zoltán Szabó – CMAP, École Polytechnique

Joint work with: Bharath K. Sriperumbudur (PSU)

CREST Statistics Seminar, ENSAE
October 9, 2017

Objects of Interest

Divergence & independence measures on kernel-endowed domains.

- Mean embedding:

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

Objects of Interest

Divergence & independence measures on kernel-endowed domains.

- Mean embedding:

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

- Maximum mean discrepancy (MMD):

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

Objects of Interest

Divergence & independence measures on kernel-endowed domains.

- Mean embedding:

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

- Maximum mean discrepancy (MMD):

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

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

$$\mathrm{HSIC}_k\left(\mathbb{P}\right) := \left\|\mu_k\left(\mathbb{P}\right) - \mu_k\left(\otimes_{m=1}^{M} \mathbb{P}_m\right)\right\|_{\mathcal{H}_k}.$$

Objects of Interest

- Mean embedding:

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

- Maximum mean discrepancy (MMD):

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

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

$$\mathrm{HSIC}_k\left(\mathbb{P}\right) := \left\|\mu_k\left(\mathbb{P}\right) - \mu_k\left(\otimes_{m=1}^{M} \mathbb{P}_m\right)\right\|_{\mathcal{H}_k}.$$

Question

Conditions on k_m-s so that MMD and HSIC are characteristic?

Intuition: Distribution Representation via Functions

- Cumulative density function:

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

Intuition: Distribution Representation via Functions

- Cumulative density function:

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

Intuition: Distribution Representation via Functions

- Cumulative density function:

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

- Characteristic function:

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

Intuition: Distribution Representation via Functions

- Cumulative density function:

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

- Characteristic function:

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

- Moment generating function:

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

Intuition: Distribution Representation via Functions

- Cumulative density function:

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

- Characteristic function:

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

- Moment generating function:

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

Pattern

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

Ingredients

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

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

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

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

Assumption

$\mathcal{X}_m$: kernel-endowed domains ($\Rightarrow$ inner product).

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

Given: $\mathcal{X}$ set.

- Kernel:

$$k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad \mathcal{H} : \text{some Hilbert space.}$$

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

Given: $\mathcal{X}$ set.

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad \mathcal{H} : \text{some Hilbert space.}$$

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

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

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

Given: $\mathcal{X}$ set.

- Kernel:

$$k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{H}}, \quad \mathcal{H} : \text{some Hilbert space.}$$

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

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

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

Ingredients: Kernel, RKHS – continued

Alternatives:

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric is positive definite if

$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \succeq 0 \quad \forall n \in \mathbb{Z}^+, \forall \{x_i\}_{i=1}^{n}.$$

Ingredients: Kernel, RKHS – continued

Alternatives:

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric is positive definite if

$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \succeq 0 \quad \forall n \in \mathbb{Z}^+, \forall \{x_i\}_{i=1}^{n}.$$

- $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ Hilbert space,

$$\delta_x : f \in \mathcal{H} \mapsto f(x) \in \mathbb{R}$$

  is continuous for all $x \in \mathcal{X}$.

Ingredients: Kernel, RKHS – continued

Alternatives:

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric is positive definite if

$$\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \succeq 0 \quad \forall n \in \mathbb{Z}^+, \forall \{x_i\}_{i=1}^{n}.$$

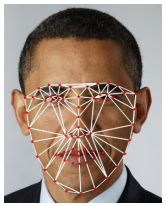
- $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ Hilbert space,

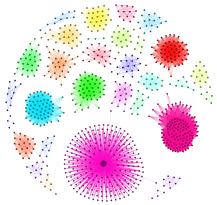
$$\delta_x : f \in \mathcal{H} \mapsto f(x) \in \mathbb{R}$$

  is continuous for all $x \in \mathcal{X}$.

All these definitions are equivalent.

Kernel Examples



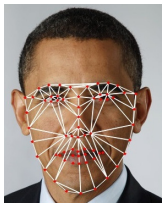


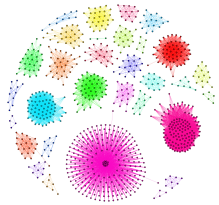
- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

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

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

Kernel Examples





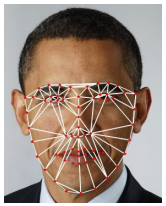
- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

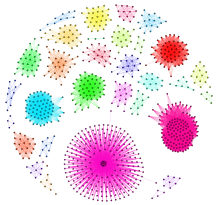
$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

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

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

Kernel Examples





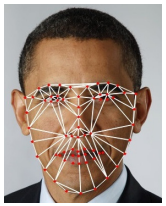
- $\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

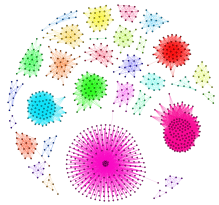
$$k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, \qquad k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2},$$

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

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

Kernel Examples





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

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

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

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

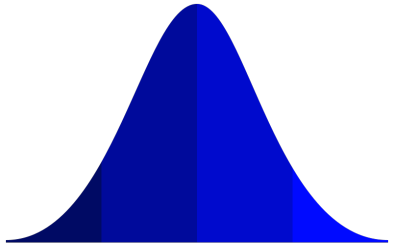
RKHS: Constructively

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\overset{1:1}{\longleftrightarrow}$ $\mathcal{H}_k$ RKHS.

RKHS: Constructively

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\overset{1:1}{\longleftrightarrow}$ $\mathcal{H}_k$ RKHS.

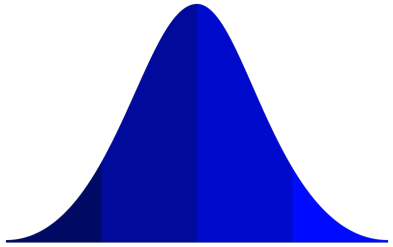
- Elements of $\mathcal{H}_k$: $\overline{\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}\}}$.



RKHS: Constructively

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\overset{1:1}{\longleftrightarrow}$ $\mathcal{H}_k$ RKHS.

- Elements of $\mathcal{H}_k$: $\overline{\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}\}}$.

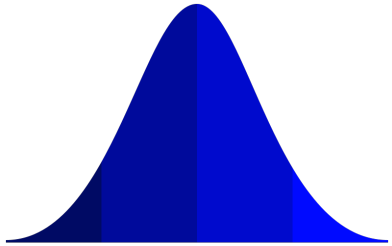


- Inner product:
  - $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} := k(x, y)$.
  - Extension: by linearity & limit.

RKHS: Constructively

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ kernel, $\overset{1:1}{\longleftrightarrow}$ $\mathcal{H}_k$ RKHS.

- Elements of $\mathcal{H}_k$: $\overline{\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}\}}$.



- Inner product:
  - $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} := k(x, y)$.
  - Extension: by linearity & limit.

We represent distributions in an RKHS.

Mean embedding: kernel trick $\longrightarrow$ mean trick

- Kernel: $k(x, x') = \left\langle k(\cdot, x), k(\cdot, x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot, x)$, $\mathbb{P} = \delta_x$.

Mean embedding: kernel trick $\longrightarrow$ mean trick

- Kernel: $k(x, x') = \left\langle k(\cdot, x), k(\cdot, x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot, x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

Mean embedding: kernel trick $\longrightarrow$ mean trick

- Kernel: $k(x, x') = \left\langle k(\cdot, x), k(\cdot, x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot, x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mu_{\mathbb{P}} := \sum_{i=1}^{N} w_i k(\cdot, x_i) \in \mathcal{H}_k, \qquad \mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

Mean embedding: kernel trick $\longrightarrow$ mean trick

- Kernel: $k(x,x') = \left\langle k(\cdot,x), k(\cdot,x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot,x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mu_{\mathbb{P}} := \sum_{i=1}^{N} w_i k(\cdot,x_i) \in \mathcal{H}_k, \quad \mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

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

Mean embedding: kernel trick $\longrightarrow$ mean trick

- Kernel: $k(x,x') = \left\langle k(\cdot,x), k(\cdot,x') \right\rangle_{\mathcal{H}_k}$, $\mu_{\mathbb{P}} := k(\cdot,x)$, $\mathbb{P} = \delta_x$.
- Mean embedding (feature of $\mathbb{P}$):

$$\mu_{\mathbb{P}} := \sum_{i=1}^{N} w_i k(\cdot,x_i) \in \mathcal{H}_k, \quad \mathbb{P} = \sum_{i=1}^{N} w_i \delta_{x_i},$$

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

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

Mean Embedding

- Applications:
  - two-sample testing [Gretton et al., 2012], domain adaptation [Zhang et al., 2013],
  - kernel belief propagation [Song et al., 2011], kernel Bayes' rule [Fukumizu et al., 2013], model criticism [Lloyd et al., 2014],
  - approximate Bayesian computation [Park et al., 2016], probabilistic programming [Schölkopf et al., 2015],
  - distribution classification [Muandet et al., 2011], distribution regression [Szabó et al., 2016], topological data analysis [Kusano et al., 2016].
- Review [Muandet et al., 2017].

Mean Embedding $\rightarrow$ MMD

- k is called characteristic if

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

is injective [Fukumizu et al., 2008, Sriperumbudur et al., 2010].

Mean Embedding $\rightarrow$ MMD

- k is called characteristic if

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

  is injective [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- In this case

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

  is a metric.

Maximum Mean Discrepancy → HSIC

- $\mathcal{X} := \times_{m=1}^{M} \mathcal{X}_m$. Tensor product of $(k_m)_{m=1}^{M}$ kernels:

$$\left(\otimes_{m=1}^{M} k_m\right)\left(x, x'\right) = \prod_{m=1}^{M} k_m\left(x_m, x'_m\right), \quad x, x' \in \mathcal{X}$$

  is a kernel, $\mathcal{H}_{\otimes_{m=1}^{M} k_m} = \otimes_{m=1}^{M} \mathcal{H}_{k_m}$ [Berlinet and Thomas-Agnan, 2004].

Maximum Mean Discrepancy → HSIC

- $\mathcal{X} := \times_{m=1}^{M} \mathcal{X}_m$. Tensor product of $(k_m)_{m=1}^{M}$ kernels:

$$\left(\otimes_{m=1}^{M} k_m\right)\left(x, x'\right) = \prod_{m=1}^{M} k_m\left(x_m, x'_m\right), \quad x, x' \in \mathcal{X}$$

  is a kernel, $\mathcal{H}_{\otimes_{m=1}^{M} k_m} = \otimes_{m=1}^{M} \mathcal{H}_{k_m}$ [Berlinet and Thomas-Agnan, 2004].
- Choosing $k := \otimes_{m=1}^{M} k_m$, $\mathbb{Q} := \otimes_{m=1}^{M} \mathbb{P}_m$ in MMD:

$$\mathrm{HSIC}_k\left(\mathbb{P}\right) := \left\| \mu_k\left(\mathbb{P}\right) - \mu_k\left(\otimes_{m=1}^{M} \mathbb{P}_m\right) \right\|_{\mathcal{H}_k}.$$

HSIC Applications

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

MMD: Easy to Estimate

Using $\{x_i\}_{i=1}^{N_x} \sim \mathbb{P}, \{y_j\}_{j=1}^{N_y} \sim \mathbb{Q}$,

$$\widehat{MMD^2}(\mathbb{P}, \mathbb{Q}) = MMD^2\left(\hat{\mathbb{P}}, \hat{\mathbb{Q}}\right)$$

MMD: Easy to Estimate

Using $\{x_i\}_{i=1}^{N_x} \sim \mathbb{P}, \{y_j\}_{j=1}^{N_y} \sim \mathbb{Q}$,

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

MMD: Easy to Estimate

Using $\{x_i\}_{i=1}^{N_x} \sim \mathbb{P}, \{y_j\}_{j=1}^{N_y} \sim \mathbb{Q}$,

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

$$= \left\| \frac{1}{N_x}\sum_{i=1}^{N_x} k(\cdot, x_i) - \frac{1}{N_y}\sum_{j=1}^{N_y} k(\cdot, y_j) \right\|_{\mathcal{H}_k}^2$$

MMD: Easy to Estimate

Using $\{x_i\}_{i=1}^{N_x} \sim \mathbb{P}, \{y_j\}_{j=1}^{N_y} \sim \mathbb{Q}$,

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

$$= \left\| \frac{1}{N_x}\sum_{i=1}^{N_x} k(\cdot,x_i) - \frac{1}{N_y}\sum_{j=1}^{N_y} k(\cdot,y_j) \right\|_{\mathcal{H}_k}^2$$

$$= \underbrace{\frac{1}{N_x^2}\sum_{i,j=1}^{N_x} k(x_i,x_j)}_{\overline{\mathbf{G}_{\mathbb{P},\mathbb{P}}}} + \underbrace{\frac{1}{N_y^2}\sum_{i,j=1}^{N_y} k(y_i,y_j)}_{\overline{\mathbf{G}_{\mathbb{Q},\mathbb{Q}}}} - 2\underbrace{\frac{1}{N_xN_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} k(x_i,y_j)}_{\overline{\mathbf{G}_{\mathbb{P},\mathbb{Q}}}}.$$

MMD: Easy to Estimate

Using $\{x_i\}_{i=1}^{N_x} \sim \mathbb{P}, \{y_j\}_{j=1}^{N_y} \sim \mathbb{Q}$,

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

$$= \left\|\frac{1}{N_x}\sum_{i=1}^{N_x} k(\cdot,x_i) - \frac{1}{N_y}\sum_{j=1}^{N_y} k(\cdot,y_j)\right\|_{\mathcal{H}_k}^2$$

$$= \underbrace{\frac{1}{N_x^2}\sum_{i,j=1}^{N_x} k(x_i,x_j)}_{\overline{\mathbf{G}_{\mathbb{P},\mathbb{P}}}} + \underbrace{\frac{1}{N_y^2}\sum_{i,j=1}^{N_y} k(y_i,y_j)}_{\overline{\mathbf{G}_{\mathbb{Q},\mathbb{Q}}}} - 2\underbrace{\frac{1}{N_xN_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} k(x_i,y_j)}_{\overline{\mathbf{G}_{\mathbb{P},\mathbb{Q}}}}.$$

A bit biased: $\overline{\mathbf{G}_{\mathbb{P},\mathbb{P}}} \leftarrow \frac{1}{N_x^2}\sum_{i=1}^{N_x}\sum_{j=1,j\neq i}^{N_x} k(x_i,x_j)$, $\overline{\mathbf{G}_{\mathbb{Q},\mathbb{Q}}} \leftarrow \ldots$, if needed.

HSIC: Easy to Estimate

- Given: $\{(x_i, y_i)\}_{i=1}^N$ paired samples.
- Estimate:

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

HSIC: Easy to Estimate

- Given: $\{(x_i, y_i)\}_{i=1}^N$ paired samples.
- Estimate:

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

$$\mathbf{G}_x = [k_1(x_i, x_j)]_{i,j=1}^n$$

HSIC: Easy to Estimate

- Given: $\{(x_i, y_i)\}_{i=1}^N$ paired samples.
- Estimate:

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

$$\mathbf{G}_x = [k_1(x_i, x_j)]_{i,j=1}^n, \ \tilde{\mathbf{G}}_x = \mathbf{H}\mathbf{G}_x\mathbf{H}, \ \mathbf{H} = \mathbf{I} - \frac{\mathbf{E}}{N}.$$

- Plug-in estimator; similarly easy to debias.

Central in Applications: Characteristic Property

- MMD: 'k is characteristic' means

$$MMD_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Central in Applications: Characteristic Property

- MMD: 'k is characteristic' means

$$MMD_k(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

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

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

Central in Applications: Characteristic Property

- MMD: 'k is characteristic' means

$$MMD_k(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

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

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

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

Central in Applications: Characteristic Property

- MMD: 'k is characteristic' means

$$MMD_k(\mathbb{P},\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

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

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

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

Wanted

- Precise relation between characteristic and $\mathcal{I}$-characteristic?
- Conditions in terms of k_m-s?

Characteristic property

Well-understood for

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

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

Characteristic property

Well-understood for

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

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

- 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 $\mathbb{R}^d$:

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

- 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	$\widehat{k_0}(\omega)$	$supp(\widehat{k_0})$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
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	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
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	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
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	$\widehat{k_0}(\omega)$	$supp(\widehat{k_0})$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
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	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
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	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

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

Universal kernel

Let $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$.

Definition

Assume:

- $\mathcal{X}$: compact metric space.
- k: continuous kernel on $\mathcal{X}$.

k is called *(c)-universal* [Steinwart, 2001] if $\mathcal{H}_k$ is dense in $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Universal $\Rightarrow$ Characteristic

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ_k is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.

Universal $\Rightarrow$ Characteristic

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ_k is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.
- If $\mathcal{X} =$ LCP: c_0-universality [Sriperumbudur et al., 2010]
  - $\mathcal{H}_k$ dense in $C_0(\mathcal{X})$, equivalently
  - $\mu_k : \mathcal{M}_b(\mathcal{X}) \mapsto \mathcal{H}_k$ is injective.

Universal $\Rightarrow$ Characteristic

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ_k is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.
- If $\mathcal{X} =$ LCP: c_0-universality [Sriperumbudur et al., 2010]
  - $\mathcal{H}_k$ dense in $C_0(\mathcal{X})$, equivalently
  - $\mu_k : \mathcal{M}_b(\mathcal{X}) \mapsto \mathcal{H}_k$ is injective.

  LCP examples: $\mathbb{R}^d$, countable discrete.

Universal $\Rightarrow$ Characteristic

- [Micchelli et al., 2006]: k is c-universal $\Leftrightarrow$ μ_k is injective on $\mathcal{M}_b(\mathcal{X})$, the set of finite signed Borel measures on $\mathcal{X}$.
- If $\mathcal{X} =$ LCP: c_0-universality [Sriperumbudur et al., 2010]
  - $\mathcal{H}_k$ dense in $C_0(\mathcal{X})$, equivalently
  - $\mu_k : \mathcal{M}_b(\mathcal{X}) \mapsto \mathcal{H}_k$ is injective.

  LCP examples: $\mathbb{R}^d$, countable discrete.
- c_0-universality $\Rightarrow$ characteristic.

Local Summary

- Setup: $\mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$, $(\mathcal{X}_m, k_m)$: kernel-enriched domains.
- Mean embedding, MMD, HSIC, $\otimes_{m=1}^{M} k_m$.
- Characteristic, c_0-universality.

Local Summary

- Setup: $\mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$, $(\mathcal{X}_m, k_m)$: kernel-enriched domains.
- Mean embedding, MMD, HSIC, $\otimes_{m=1}^{M} k_m$.
- Characteristic, c_0-universality.

Little is known about the

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

Known '$\mathcal{I}$-characteristic' Results, $M = 2$

- [Waegeman et al., 2012, Gretton, 2015]:

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

Known '$\mathcal{I}$-characteristic' Results, $M = 2$

- [Waegeman et al., 2012, Gretton, 2015]:

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

- Stronger: combinining [Lyons, 2013] (DCov) and [Sejdinovic et al., 2013] (DCov $\Leftrightarrow$ HSIC)

  $k_1 \& k_2$: characteristic $\Leftrightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-characteristic.

Known '$\mathcal{I}$-characteristic' Results, $M = 2$

- [Waegeman et al., 2012, Gretton, 2015]:

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

- Stronger: combinining [Lyons, 2013] (DCov) and [Sejdinovic et al., 2013] (DCov $\Leftrightarrow$ HSIC)

  $k_1 \& k_2$: characteristic $\Leftrightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-characteristic.

Question

Extension to $M \geqslant 2$?

Known '$\mathcal{I}$-characteristic' Results, $M = 2$

- [Waegeman et al., 2012, Gretton, 2015]:

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

- Stronger: combinining [Lyons, 2013] (DCov) and [Sejdinovic et al., 2013] (DCov $\Leftrightarrow$ HSIC)

  $k_1 \& k_2$: characteristic $\Leftrightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-characteristic.

Question

Extension to $M \geqslant 2$?

Main Challenge

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

Idea: Characteristic Property as Ispd

- Characteristic property:

$$\underbrace{\left\|\mu_{\mathbb{P}_1-\mathbb{P}_2}\right\|_k^2}_{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x'),\,\mathbb{F}:=\mathbb{P}_1-\mathbb{P}_2\neq 0,\,\mathbb{F}(\mathcal{X})=0} > 0, \quad \forall \mathbb{P}_1,\mathbb{P}_2 \in \mathcal{M}_1^+(\mathcal{X}), \mathbb{P}_1 \neq \mathbb{P}_2.$$

Idea: Characteristic Property as Ispd

- Characteristic property:

$$\underbrace{\left\|\mu_{\mathbb{P}_1-\mathbb{P}_2}\right\|_k^2}_{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x'),\,\mathbb{F}:=\mathbb{P}_1-\mathbb{P}_2\neq 0,\,\mathbb{F}(\mathcal{X})=0} > 0, \quad \forall \mathbb{P}_1,\mathbb{P}_2 \in \mathcal{M}_1^+(\mathcal{X}), \mathbb{P}_1 \neq \mathbb{P}_2.$$

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

$$\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x') > 0, \ \forall \underbrace{\mathbb{F} \in \mathcal{M}_b(\mathcal{X})\backslash\{0\} \ \ \mathbb{F}(\mathcal{X}) = 0}_{\mathcal{F}_1}.$$

Idea: Characteristic Property as Ispd

- Characteristic property:

$$\underbrace{\|\mu_{\mathbb{P}_1-\mathbb{P}_2}\|_k^2}_{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x'),\,\mathbb{F}:=\mathbb{P}_1-\mathbb{P}_2\neq 0,\,\mathbb{F}(\mathcal{X})=0} > 0, \quad \forall \mathbb{P}_1,\mathbb{P}_2 \in \mathcal{M}_1^+(\mathcal{X}), \mathbb{P}_1 \neq \mathbb{P}_2.$$

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

$$\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x') > 0, \ \forall \underbrace{\mathbb{F} \in \mathcal{M}_b(\mathcal{X})\backslash\{0\} \ \ \mathbb{F}(\mathcal{X}) = 0}_{\mathcal{F}_1}.$$

- We have also seen: k is c_0-universal iff.

$$\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x') > 0, \ \forall \underbrace{\mathbb{F} \in \mathcal{M}_b(\mathcal{X})\backslash\{0\}}_{\mathcal{F}_2}.$$

$\mathcal{F}$-ispd Tensor Product Kernels

From now on: $\mathcal{X} = \otimes_{m=1}^{M} \mathcal{X}_m$.

Definition

Let $\mathcal{F} \subseteq \mathcal{M}_b(\mathcal{X})$, $0 \in \mathcal{F}$. $k = \otimes_{m=1}^{M} k_m$ is called $\mathcal{F}$-ispd if

$$\mu_k(\mathbb{F}) = 0 \Rightarrow \mathbb{F} = 0 \quad (\mathbb{F} \in \mathcal{F}), \text{ equivalently}$$

$$\|\mu_k(\mathbb{F})\|_{\mathcal{H}_k}^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} k\left(x, x'\right) \, \mathrm{d}\mathbb{F}(x) \, \mathrm{d}\mathbb{F}(x') > 0, \quad \forall \, \mathbb{F} \in \mathcal{F} \backslash \{0\}.$$

Examples

$\mathcal{F}$	$\mathcal{F}$-ispd k
$\mathcal{M}_b(\mathcal{X})$	c_0-universal
$[\mathcal{M}_b(\mathcal{X})]^0$	characteristic

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

$$\Leftarrow \quad \Leftarrow \quad \text{characteristic} \Leftarrow c_0\text{-universal.}$$
$$\Downarrow$$

Examples

$\mathcal{F}$	$\mathcal{F}$-ispd k
$\mathcal{M}_b(\mathcal{X})$	c_0-universal
$[\mathcal{M}_b(\mathcal{X})]^0$	characteristic
$\mathcal{I} := \left\{\mathbb{P} - \otimes_{m=1}^{M} \mathbb{P}_m : \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})\right\}$	$\mathcal{I}$-characteristic

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

$$\begin{array}{c} \cup| \\ \mathcal{I} \end{array}$$

$$\Leftarrow \qquad \Leftarrow \quad \text{characteristic} \Leftarrow \quad c_0\text{-universal.}$$

$$\Downarrow$$

$$\mathcal{I}\text{-characteristic}$$

Examples

$\mathcal{F}$	$\mathcal{F}$-ispd k
$\mathcal{M}_b(\mathcal{X})$	c_0-universal
$[\mathcal{M}_b(\mathcal{X})]^0$	characteristic
$\mathcal{I} := \left\{\mathbb{P} - \otimes_{m=1}^M \mathbb{P}_m : \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})\right\}$	$\mathcal{I}$-characteristic
$\left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0$	$\otimes$-characteristic

$$\subseteq \left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0 \subseteq \quad [\mathcal{M}_b(\mathcal{X})]^0 \subseteq \quad \mathcal{M}_b(\mathcal{X}).$$

$$\begin{array}{c} \cup| \\ \mathcal{I} \end{array}$$

$$\Leftarrow \otimes\text{-characteristic} \Leftarrow \quad \text{characteristic} \Leftarrow \quad c_0\text{-universal.}$$

$$\Downarrow$$

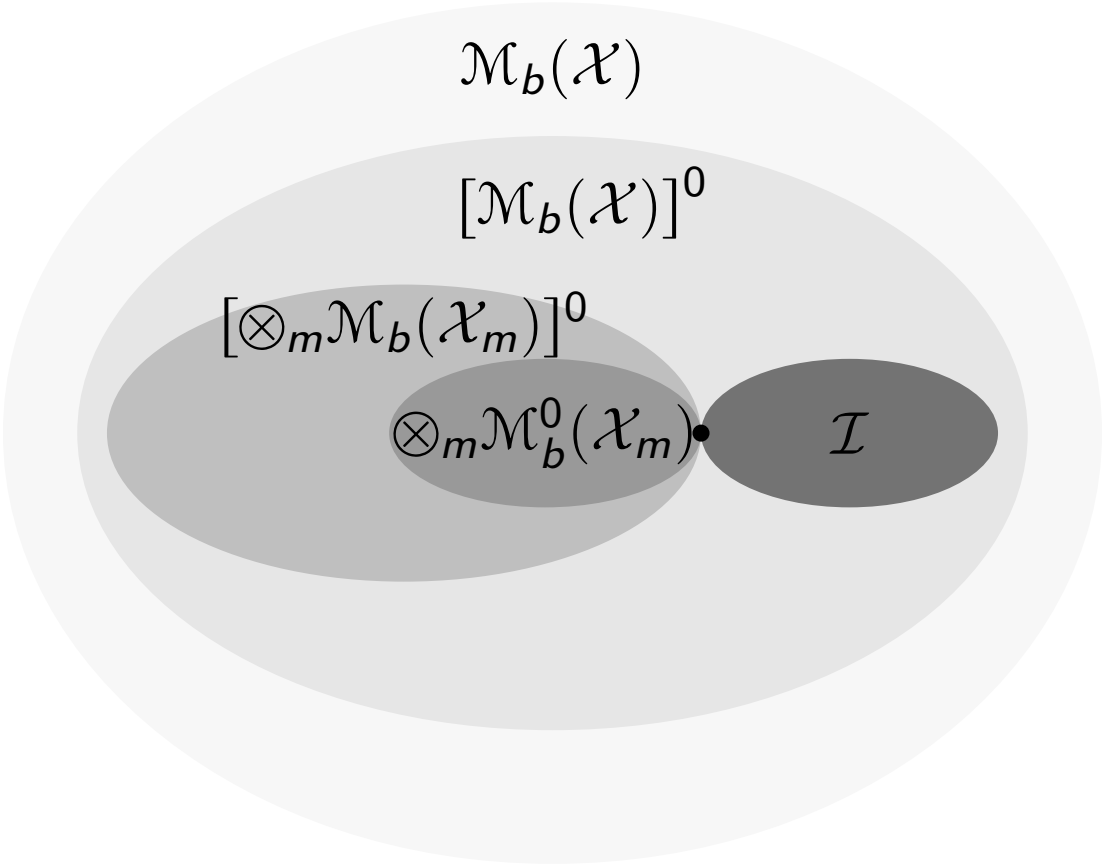
$$\mathcal{I}\text{-characteristic}$$

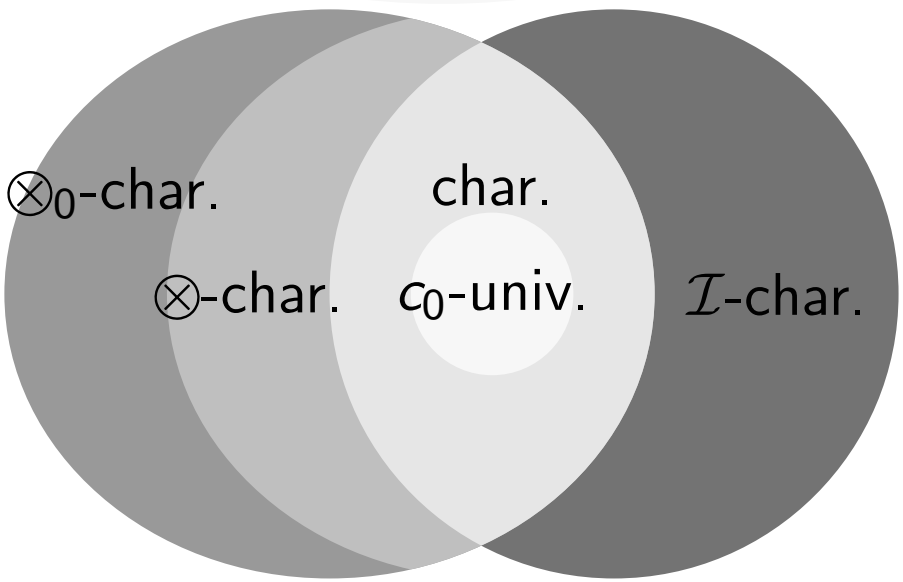
Examples

$\mathcal{F}$	$\mathcal{F}$-ispd k
$\mathcal{M}_b(\mathcal{X})$	c_0-universal
$[\mathcal{M}_b(\mathcal{X})]^0$	characteristic
$\mathcal{I} := \left\{\mathbb{P} - \otimes_{m=1}^M \mathbb{P}_m : \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})\right\}$	$\mathcal{I}$-characteristic
$\left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0$	$\otimes$-characteristic
$\otimes_{m=1}^M \mathcal{M}_b^0(\mathcal{X}_m)$	$\otimes_0$-characteristic

$$\otimes_{m=1}^M \mathcal{M}_b^0(\mathcal{X}_m) \subseteq \left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0 \subseteq [\mathcal{M}_b(\mathcal{X})]^0 \subseteq \mathcal{M}_b(\mathcal{X}).$$

$$\begin{array}{ccccccc} \otimes_0\text{-characteristic} & \Leftarrow & \otimes\text{-characteristic} & \Leftarrow & \text{characteristic} & \Leftarrow & c_0\text{-universal.} \\ & & & & \Downarrow & & \\ & & & & \mathcal{I}\text{-characteristic} & & \end{array}$$


$\mathcal{M}_b(\mathcal{X})$
$[\mathcal{M}_b(\mathcal{X})]^0$
$[\otimes_m \mathcal{M}_b(\mathcal{X}_m)]^0$
$\otimes_m \mathcal{M}_b^0(\mathcal{X}_m)$
$\mathcal{I}$


$\otimes_0$-char.
char.
$\otimes$-char.
c_0-univ.
$\mathcal{I}$-char.

Results

Various characteristic properties of $\otimes_{m=1}^{M} k_m$

Proposition

(i) $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow$ $\otimes$*-characteristic.*

(ii) $\otimes_{m=1}^{M} k_m$: $\otimes$*-characteristic* $\Rightarrow$ $\otimes_0$*-characteristic.*

(iii) $\otimes_{m=1}^{M} k_m$: $\otimes_0$*-characteristic* $\Leftrightarrow$ $(k_m)_{m=1}^{M}$ *are characteristic.*

Various characteristic properties of $\otimes_{m=1}^{M} k_m$

Proposition

(i) $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow$ $\otimes$-*characteristic.*

(ii) $\otimes_{m=1}^{M} k_m$: $\otimes$-*characteristic* $\Rightarrow$ $\otimes_0$-*characteristic.*

(iii) $\otimes_{m=1}^{M} k_m$: $\otimes_0$-*characteristic* $\Leftrightarrow$ $(k_m)_{m=1}^{M}$ *are characteristic.*

(iii) remains. Proof idea: with $k = \otimes_{m=1}^{M} k_m$, $\mathbb{F} = \otimes_{m=1}^{M} \mathbb{F}_m$,

$$\|\mu_k(\mathbb{F})\|_{\mathcal{H}_k}^2 = \prod_{m=1}^{M} \|\mu_{k_m}(\mathbb{F}_m)\|_{\mathcal{H}_{k_m}}^2,$$

$$\mathbb{F} \in \left[\otimes_{m=1}^{M} \mathcal{M}_b^0(\mathcal{X}_m)\right] \backslash \{0\} \Leftrightarrow \forall m : \mathbb{F}_m \in \mathcal{M}_b^0(\mathcal{X}_m) \backslash \{0\}.$$

$\otimes_0$-characteristic $\not\Rightarrow$ even $\otimes$-characteristic

Reverse of (ii) does not hold.

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 $\otimes$-characteristic.*
- *$k_1 \otimes k_2$ is $\mathcal{I}$-characteristic.*

Proof idea: $k_1 \otimes k_2$: not $\otimes$-characteristic

Goal: construct a witness $0 \neq \mathbb{F} = \mathbb{F}_1 \otimes \mathbb{F}_2 \in \otimes_{m=1}^{2} \mathcal{M}_b(\mathcal{X}_m)$ s.t.

$$0 = \mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) = \mathbb{F}_1(\mathcal{X}_1)\mathbb{F}_2(\mathcal{X}_2),$$

$$0 = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_2} k_1(x_1, x_1')k_2)(x_2, x_2') \, \mathrm{d}\mathbb{F}(x_1, x_2) \, \mathrm{d}\mathbb{F}(x_1', x_2').$$

Proof idea: $k_1 \otimes k_2$: not $\otimes$-characteristic

Goal: construct a witness $0 \neq \mathbb{F} = \mathbb{F}_1 \otimes \mathbb{F}_2 \in \otimes_{m=1}^{2} \mathcal{M}_b(\mathcal{X}_m)$ s.t.

$$0 = \mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) = \mathbb{F}_1(\mathcal{X}_1)\mathbb{F}_2(\mathcal{X}_2),$$

$$0 = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_2} k_1(x_1, x_1')k_2)(x_2, x_2') \, \mathrm{d}\mathbb{F}(x_1, x_2) \, \mathrm{d}\mathbb{F}(x_1', x_2').$$

Finite signed measures on $\{1, 2\}$:

$$\mathbb{F}_1 = \mathbb{F}_1(\mathbf{a}) = a_1\delta_1 + a_2\delta_2, \qquad \mathbb{F}_2 = \mathbb{F}_2(\mathbf{b}) = b_1\delta_1 + b_2\delta_2.$$

Proof idea: $k_1 \otimes k_2$: not $\otimes$-characteristic

Goal: construct a witness $0 \neq \mathbb{F} = \mathbb{F}_1 \otimes \mathbb{F}_2 \in \otimes_{m=1}^{2} \mathcal{M}_b(\mathcal{X}_m)$ s.t.

$$0 = \mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) = \mathbb{F}_1(\mathcal{X}_1)\mathbb{F}_2(\mathcal{X}_2),$$

$$0 = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_2} k_1(x_1, x_1')k_2)(x_2, x_2') \, \mathrm{d}\mathbb{F}(x_1, x_2) \, \mathrm{d}\mathbb{F}(x_1', x_2').$$

Finite signed measures on $\{1, 2\}$:

$$\mathbb{F}_1 = \mathbb{F}_1(\mathbf{a}) = a_1\delta_1 + a_2\delta_2, \qquad \mathbb{F}_2 = \mathbb{F}_2(\mathbf{b}) = b_1\delta_1 + b_2\delta_2.$$

This gives

$$0 = (a_1 + a_2)(b_1 + b_2), \qquad 0 = (a_1 - a_2)^2(b_1 - b_2)^2.$$

Proof idea: $k_1 \otimes k_2$: not $\otimes$-characteristic

Goal: construct a witness $0 \neq \mathbb{F} = \mathbb{F}_1 \otimes \mathbb{F}_2 \in \otimes_{m=1}^{2} \mathcal{M}_b(\mathcal{X}_m)$ s.t.

$$0 = \mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) = \mathbb{F}_1(\mathcal{X}_1)\mathbb{F}_2(\mathcal{X}_2),$$

$$0 = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_2} k_1(x_1, x_1')k_2)(x_2, x_2') \, \mathrm{d}\mathbb{F}(x_1, x_2) \, \mathrm{d}\mathbb{F}(x_1', x_2').$$

Finite signed measures on $\{1, 2\}$:

$$\mathbb{F}_1 = \mathbb{F}_1(\mathbf{a}) = a_1\delta_1 + a_2\delta_2, \qquad \mathbb{F}_2 = \mathbb{F}_2(\mathbf{b}) = b_1\delta_1 + b_2\delta_2.$$

This gives

$$0 = (a_1 + a_2)(b_1 + b_2), \qquad 0 = (a_1 - a_2)^2(b_1 - b_2)^2.$$

$\Rightarrow$ Two symmetric solutions ($\mathbf{a} \neq \mathbf{0}$, $\mathbf{b} \neq \mathbf{0}$):

$$a_1 + a_2 = 0, \qquad b_1 = b_2$$

$$a_1 = a_2, \qquad b_1 + b_2 = 0.$$

Towards $\mathcal{I}$-characteristicity

In the previous example:

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

In fact:

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

$\mathcal{I}$-characteristic Property

Proposition

(i) *k_1, k_2: characteristic $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-characteristic.*

(ii) *Suppose $\mathcal{X}_m$ is Hausdorff $(\forall m)$. Then*

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

$\mathcal{I}$-characteristic Property

Proposition

(i) *k_1, k_2: characteristic $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-characteristic.*

(ii) *Suppose $\mathcal{X}_m$ is Hausdorff ($\forall m$). Then*

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

Proof idea:

(i) Induction: see later c_0-universality ($\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$).

$\mathcal{I}$-characteristic Property

Proposition

(i) *k_1, k_2: characteristic $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-characteristic.*

(ii) *Suppose $\mathcal{X}_m$ is Hausdorff ($\forall m$). Then*
$\otimes_{m=1}^{M} k_m$*: $\mathcal{I}$-characteristic $\Rightarrow$ $(k_m)_{m=1}^{M}$ are characteristic.*

Proof idea:

(i) Induction: see later c_0-universality ($\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$).

(ii) If a k_m is not characteristic, then we construct a witness for the non-$\mathcal{I}$-characteristic property.

k_1, k_2, k_3: characteristic $\not\Rightarrow \otimes_{m=1}^3 k_m$: $\mathcal{I}$-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: $\mathbf{z} = (z_0, z_1, \ldots, z_5) \in [0,1]^6$.

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

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

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

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

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

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

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

$\mathbb{R}^d$ & translation-invariance: all notions coincide

Proposition

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

(i) $(k_m)_{m=1}^M$*-s are characteristic.*

(ii) $\otimes_{m=1}^M k_m$*: $\otimes_0$-characteristic.*

(iii) $\otimes_{m=1}^M k_m$*: $\otimes$-characteristic.*

(iv) $\otimes_{m=1}^M k_m$*: $\mathcal{I}$-characteristic.*

(v) $\otimes_{m=1}^M k_m$*: characteristic.*

$\mathbb{R}^d$ & translation-invariance: all notions coincide

Proposition

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

(i) $(k_m)_{m=1}^M$*-s are characteristic.*

(ii) $\otimes_{m=1}^M k_m$*: $\otimes_0$-characteristic.*

(iii) $\otimes_{m=1}^M k_m$*: $\otimes$-characteristic.*

(iv) $\otimes_{m=1}^M k_m$*: $\mathcal{I}$-characteristic.*

(v) $\otimes_{m=1}^M k_m$*: characteristic.*

Proof idea: We already know

$$(v) \Rightarrow (iii) \Rightarrow (ii) \Leftrightarrow (i), \qquad (v) \Rightarrow (iv) \Rightarrow (i).$$

Remains: $(i) \Rightarrow (v)$.

- Bochner theorem and *supp*-characterization:

$$k_m(x_m, x'_m) = \int_{\mathbb{R}^{d_m}} e^{-i\langle \omega_m, x_m - x'_m \rangle} \, \mathrm{d}\Lambda_m(\omega_m), \;\; x_m, x'_m \in \mathbb{R}^{d_m},$$

where $supp(\Lambda_m) = \mathbb{R}^{d_m}$.

$(k_m)_{m=1}^M$: characteristic $\Rightarrow \otimes_{m=1}^M k_m$: characteristic

- Bochner theorem and *supp*-characterization:

$$k_m(x_m, x'_m) = \int_{\mathbb{R}^{d_m}} e^{-i\langle \omega_m, x_m - x'_m \rangle} \, \mathrm{d}\Lambda_m(\omega_m), \;\; x_m, x'_m \in \mathbb{R}^{d_m},$$

  where $supp(\Lambda_m) = \mathbb{R}^{d_m}$.
- Tensor kernel:

$$\left(\otimes_{m=1}^M k_m\right)(x, x') = \int_{\mathbb{R}^d} e^{-i\langle \omega, x - x' \rangle} \, \mathrm{d}\Lambda(\omega), \quad \Lambda := \otimes_{m=1}^M \Lambda_m.$$

$(k_m)_{m=1}^M$: characteristic $\Rightarrow \otimes_{m=1}^M k_m$: characteristic

- Bochner theorem and *supp*-characterization:

$$k_m(x_m, x'_m) = \int_{\mathbb{R}^{d_m}} e^{-i\langle \omega_m, x_m - x'_m \rangle} \mathrm{d}\Lambda_m(\omega_m), \ x_m, x'_m \in \mathbb{R}^{d_m},$$

  where $supp(\Lambda_m) = \mathbb{R}^{d_m}$.
- Tensor kernel:

$$\left(\otimes_{m=1}^M k_m\right)(x, x') = \int_{\mathbb{R}^d} e^{-i\langle \omega, x - x' \rangle} \mathrm{d}\Lambda(\omega), \quad \Lambda := \otimes_{m=1}^M \Lambda_m.$$

- $\mathrm{supp}(\Lambda) = \times_{m=1}^M \mathrm{supp}\left(\Lambda_m\right) = \times_{m=1}^M \mathbb{R}^{d_m} = \mathbb{R}^d$.

c_0-universality of $\otimes_{m=1}^M k_m$

We saw: for $M \geqslant 3$

$$(k_m)_{m=1}^M \text{ characteristic } \not\Rightarrow \otimes_{m=1}^M k_m\text{: } \mathcal{I}\text{-characteristic.}$$

Proposition

Assume $(k_m)_{m=1}^M$ are c_0-kernels on LCP spaces $\mathcal{X}_m$. Then

$$\otimes_{m=1}^M k_m\text{: } c_0\text{-universal} \Leftrightarrow (k_m)_{m=1}^M \text{ are } c_0\text{-universal.}$$

The tricky direction: if $(k_m)_{m=1}^M$ are c_0-universal ...

Checking:

- $\mathcal{X}_m$: LCP $\Rightarrow$ $\times_{m=1}^M \mathcal{X}_m$: LCP.
- k_m: c_0-kernel $\Rightarrow$ $\otimes_{m=1}^M k_m$: c_0-kernel.

The tricky direction: if $(k_m)_{m=1}^M$ are c_0-universal ...

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

$$\underbrace{\mu(\mathbb{F})}_{\int_{\mathcal{X}} \otimes_{m=1}^M k_m(\cdot, x_m) \mathrm{d}\mathbb{F}(x)} = 0 \stackrel{?}{\Rightarrow} \mathbb{F} = 0, \quad \mathbb{F} \in \mathcal{M}_b(\mathcal{X}).$$

The tricky direction: if $(k_m)_{m=1}^M$ are c_0-universal ...

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

$$\underbrace{\mu(\mathbb{F})}_{\int_{\mathcal{X}} \otimes_{m=1}^M k_m(\cdot, x_m) \mathrm{d}\mathbb{F}(x)} = 0 \stackrel{?}{\Rightarrow} \mathbb{F} = 0, \quad \mathbb{F} \in \mathcal{M}_b(\mathcal{X}).$$

To get $\mathbb{F} = 0$ it is enough:

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

Proof idea

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

,

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

Proof idea

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

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

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

Proof idea

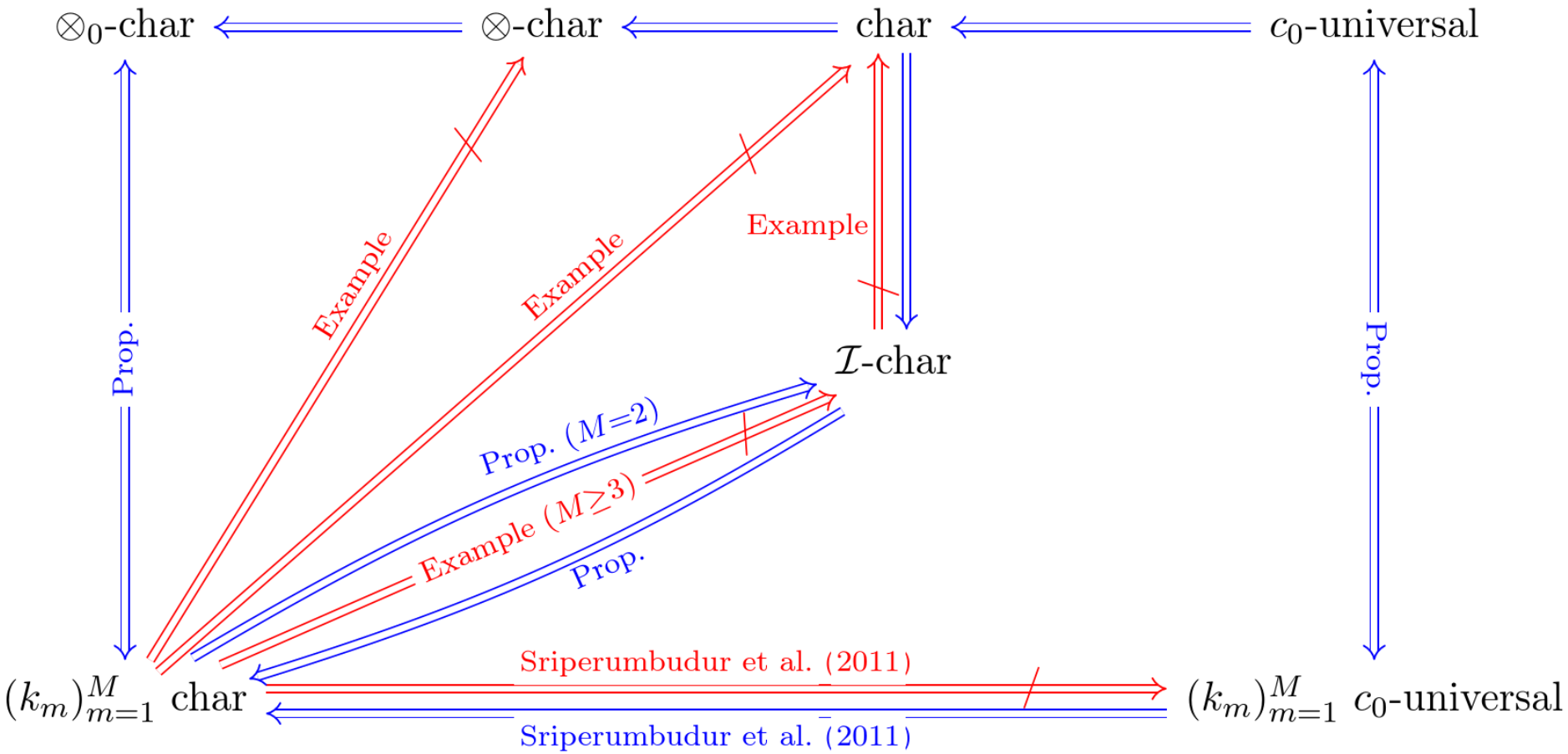
$$\otimes_{m=1}^{M}\mathcal{H}_{k_m} \ni 0 = \int_{\mathcal{X}} \otimes_{m=1}^{M} k_m(\cdot, x_m)\mathrm{d}\mathbb{F}(x),$$

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

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

We proceed by induction ($J = 0, \ldots, M$).

Visual Illustration



Summary

- Divergence & independence measures on kernel-endowed domains:
  - Maximum mean discrepancy,
  - Hilbert Schmidt independence criterion.
- Kernel: $k = \otimes_{m=1}^{M} k_m$.
- $\mathcal{F}$-ispd $\Rightarrow$
  - various characteristic properties.
  - relations & expressed in terms of k_m-s.

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.

Bochner integral: quick summary

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to B$(anach space)-valued measurable function.

Bochner integral: quick summary

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to B$(anach space)-valued measurable function.
- For $f = \sum_{i=1}^{n} c_i \chi_{A_i}$ $(A_i \in \mathcal{A}, c_i \in B)$ measurable step functions

$$\int_{\mathcal{X}} f \mathrm{d}\mu := \sum_{i=1}^{n} c_i \mu(A_i) \in B.$$

Bochner integral: quick summary

- Given:
  - $(\mathcal{X}, \mathcal{A}, \mu)$: measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to B$(anach space)-valued measurable function.
- For $f = \sum_{i=1}^{n} c_i \chi_{A_i}$ $(A_i \in \mathcal{A}, c_i \in B)$ measurable step functions

$$\int_{\mathcal{X}} f \mathrm{d}\mu := \sum_{i=1}^{n} c_i \mu(A_i) \in B.$$

- f measurable function is Bochner μ-integrable if
  - $\exists$ (f_n) measurable step functions: $\lim_{n\to\infty} \int_{\mathcal{X}} \|f - f_n\|_B \mathrm{d}\mu = 0$.
  - In this case $\lim_{n\to\infty} \int_{\mathcal{X}} f_n \mathrm{d}\mu$ exists, $=: \int_{\mathcal{X}} f \mathrm{d}\mu$.

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu < \infty$.

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu < \infty$.
- In this case $\left\|\int_{\mathcal{X}} f \, \mathrm{d}\mu\right\|_B \leqslant \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu$. ('Jensen inequality')

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu < \infty$.
- In this case $\left\|\int_{\mathcal{X}} f \, \mathrm{d}\mu\right\|_B \leqslant \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu$. ('Jensen inequality')
- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

  $S \circ f : X \to B_2$ is Bochner integrable and

$$S\left(\int_{\mathcal{X}} f \, \mathrm{d}\mu\right) = \int_{\mathcal{X}} Sf \, \mathrm{d}\mu.$$

Bochner integral: properties

- $f : \mathcal{X} \to B$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu < \infty$.
- In this case $\left\|\int_{\mathcal{X}} f \, \mathrm{d}\mu\right\|_B \leqslant \int_{\mathcal{X}} \|f\|_B \, \mathrm{d}\mu$. ('Jensen inequality')
- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

  $S \circ f : X \to B_2$ is Bochner integrable and

$$S\left(\int_{\mathcal{X}} f \, \mathrm{d}\mu\right) = \int_{\mathcal{X}} Sf \, \mathrm{d}\mu.$$

In short

$\left|\int f \, \mathrm{d}\mu\right| \leqslant \int |f| \mathrm{d}\mu$ and $c \int f \, \mathrm{d}\mu = \int cf \, \mathrm{d}\mu$ generalize nicely.


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


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.

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.


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.


Micchelli, C. A., Xu, Y., and Zhang, H. (2006).
Universal kernels.
Journal of Machine Learning Research, 7:2651–2667.


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.

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


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`).


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.