From Distance Covariance to Hilbert-Schmidt Independence Criterion

Zoltán Szabó – CMAP, École Polytechnique

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

Statistical Seminar in Rennes
Oct. 26, 2018

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

1. $I(\mathbb{P}) \geqslant 0$.

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

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

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

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

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

Kernels on $\mathbb{R}^d$: generalization of $\mathbf{x}^T\mathbf{y}$

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

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

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

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

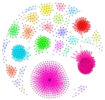
Kernels on $\mathbb{R}^d$: generalization of $\mathbf{x}^T\mathbf{y}$

$\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$: $\boxed{k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}}$

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

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

Kernels exist on various domains!



Some kernel-enriched domains: $(\mathcal{X}, k)$

- Trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002], time series [Cuturi, 2011], strings [Lodhi et al., 2002],
- mixture models, hidden Markov models or linear dynamical systems [Jebara et al., 2004],
- sets [Haussler, 1999, Gärtner et al., 2002], fuzzy domains [Guevara et al., 2017], distributions [Hein and Bousquet, 2005, Martins et al., 2009, Muandet et al., 2011],
- groups [Cuturi et al., 2005] $\xrightarrow{\text{spec.}}$ permutations [Jiao and Vert, 2016],
- graphs [Vishwanathan et al., 2010, Kondor and Pan, 2016].

Kernel, RKHS: intuition

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

- Kernel:

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

Kernel, RKHS: intuition

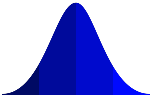
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

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

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

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



Kernel, RKHS: intuition

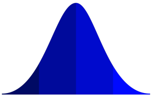
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

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

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

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



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

Kernel, RKHS: intuition

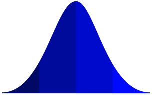
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

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

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

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



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

- Def-1 (feature space):

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

- Def-2 (reproducing kernel, constructive):

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

Kernels: $+2$ definitions

- Def-1 (feature space):

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

- Def-2 (reproducing kernel, constructive):

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

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \in \mathbb{R}^{n \times n} \succeq 0$.

Kernels: $+2$ definitions

- Def-1 (feature space):

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

- Def-2 (reproducing kernel, constructive):

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

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \in \mathbb{R}^{n \times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

Kernels: +2 definitions

- Def-1 (feature space):

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

- Def-2 (reproducing kernel, constructive):

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

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^{n} \in \mathbb{R}^{n \times n} \succeq 0$.
- Def-4 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

- All these definitions are equivalent, $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k$.
- We represent distributions in RKHSs: $\mu_{\mathbb{P}} \in \mathcal{H}_k$.

Distribution representation

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

Distribution representation

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

- Cdf:

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

Distribution representation

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

- Cdf:

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

- Characteristic function:

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

Distribution representation

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

- Cdf:

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

- Characteristic function:

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

- Moment generating function:

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

Distribution representation

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

- Cdf:

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

- Characteristic function:

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

- Moment generating function:

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

Trick

φ: on any kernel-endowed domain!

Mean embedding:

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

Distribution representation: on kernel-enriched domains

Mean embedding:

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

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

Mean embedding:

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

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

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

Mean embedding:

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

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

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

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

- Mean embedding:

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

Mean embedding → MMD, HSIC

'KL divergence & mutual information' 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:

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

Mean embedding → MMD, HSIC

'KL divergence & mutual information' 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:

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

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

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

Mean embedding → MMD, HSIC

'KL divergence & mutual information' 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:

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

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

$$\mathrm{HSIC}_k(\mathbb{P}) = \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right),$$
$$= \Big\| \underbrace{\mu_{\otimes_{m=1}^{M} k_m}(\mathbb{P}) - \otimes_{m=1}^{M} \mu_{k_m}(\mathbb{P}_m)}_{\text{cross-covariance operator}} \Big\|_{\otimes_{m=1}^{M} \mathcal{H}_{k_m}}.$$

MMD, HSIC: easy to estimate!

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], 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 $\xrightarrow{\text{spec.}}$ 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 $k = \otimes_{m=1}^{M} k_m$:

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

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

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].

Central in applications: characteristic property

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

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

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

Central in applications: characteristic property

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

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

  Injectivitity 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

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

Central in applications: characteristic property

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

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

  Injectivitity 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.

Central in applications: characteristic property

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

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

  Injectivitity 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?

Known: description of characteristic property on $\mathbb{R}^d$

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

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

$(*)$: Bochner's theorem.

Known: description of characteristic property on $\mathbb{R}^d$

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

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

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

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

Known: description of characteristic property on $\mathbb{R}^d$

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

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

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

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

Theorem ([Sriperumbudur et al., 2010])

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

Examples on $\mathbb{R}$; similarly $\mathbb{R}^d$

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

Well-known: $M = 2$

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

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

Well-known: $M = 2$

- [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.

Well-known: $M = 2$

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

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

- Distance covariance [Lyons, 2013, Sejdinovic et al., 2013]:

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

Goal

Extension to $M \geqslant 2$.

Discrete case: 'easy', e.g. k_1, k_2: char $\not\Rightarrow$ $k_1 \otimes k_2$: 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 $\not\Rightarrow$ $k_1 \otimes k_2$: 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.

$$\forall \mathbb{F} \in \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{finite signed measures on } \mathcal{X}} \backslash \{0\} \;\&\; \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \underbrace{\|\mu_{\mathbb{F}}\|^2_{\mathcal{H}_k}}_{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\mathrm{d}\mathbb{F}(x)\mathrm{d}\mathbb{F}(x')} > 0.$$

Discrete case: 'easy', e.g. k_1, k_2: char $\not\Rightarrow$ $k_1 \otimes k_2$: 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.

$$\forall \mathbb{F} \in \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{finite signed measures on } \mathcal{X}} \backslash\{0\} \ \&\ \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \underbrace{\|\mu_{\mathbb{F}}\|^2_{\mathcal{H}_k}}_{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\mathrm{d}\mathbb{F}(x)\mathrm{d}\mathbb{F}(x')} > 0.$$

- Witness construction:

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

- 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.

$$\forall \mathbb{F} \in \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{finite signed measures on } \mathcal{X}} \backslash \{0\} \;\&\; \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \underbrace{\|\mu_{\mathbb{F}}\|^2_{\mathcal{H}_k}}_{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\mathrm{d}\mathbb{F}(x)\mathrm{d}\mathbb{F}(x')} > 0.$$

- Witness construction:

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

Discrete case: 'easy', e.g. k_1, k_2: char $\not\Rightarrow$ $k_1 \otimes k_2$: 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.

$$\forall \mathbb{F} \in \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{finite signed measures on } \mathcal{X}} \backslash \{0\} \ \& \ \mathbb{F}(\mathcal{X}) = 0 \Rightarrow \underbrace{\|\mu_{\mathbb{F}}\|^2_{\mathcal{H}_k}}_{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\mathrm{d}\mathbb{F}(x)\mathrm{d}\mathbb{F}(x')} > 0.$$

- Witness construction:

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

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

Non-$\mathcal{I}$-characteristicity: analytical solution

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

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

Non-$\mathcal{I}$-characteristicity: analytical solution

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

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

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

Non-$\mathcal{I}$-characteristicity: analytical solution

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

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

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

k_1, k_2: universal, k_3: char $\not\Rightarrow$ $\otimes_{m=1}^3 k_m$: $\mathcal{I}$-characteristic

Example

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

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

Results [Szabó and Sriperumbudur, 2018]

Proposition (characteristic property)

- $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow (k_m)_{m=1}^{M}$ *are characteristic.*
- $\boxed{\not\Leftarrow}$ *[*$|\mathcal{X}_m| = 2,\ k_m(x,x') = 2\delta_{x,x'} - 1$*]*

Results [Szabó and Sriperumbudur, 2018]

Proposition (characteristic property)

- $\otimes_{m=1}^{M} k_m$: *characteristic* $\Rightarrow$ $(k_m)_{m=1}^{M}$ *are characteristic.*
- $\not\Leftarrow$ $[|\mathcal{X}_m| = 2,\ k_m(x,x') = 2\delta_{x,x'} - 1]$

Proposition ($\mathcal{I}$-characteristic property)

- k_1, k_2: *characteristic* $\Rightarrow$ $k_1 \otimes k_2$: $\mathcal{I}$-*characteristic.*
- $\Leftarrow$: *for* $\forall M \geqslant 2$.
- k_1, k_2, k_3: *characteristic* $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-*characteristic [Ex].*
- k_1, k_2: *universal,* k_3: *char* $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-*characteristic [Ex].*

Results: continued [Szabó and Sriperumbudur, 2018]

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

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.

Results: continued [Szabó and Sriperumbudur, 2018]

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

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

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,

$$supp\left(\Lambda_{\otimes_{m=1}^M k_m}\right) = \times_{m=1}^M supp(\Lambda_{k_m}).$$

Results: continued [Szabó and Sriperumbudur, 2018]

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

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,

$$supp\left(\Lambda_{\otimes_{m=1}^M k_m}\right) = \times_{m=1}^M supp(\Lambda_{k_m}).$$

Proposition (Universality)

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

$$\mu_{\mathbb{F}} = 0 \overset{?}{\Rightarrow} \mathbb{F} = 0.$$

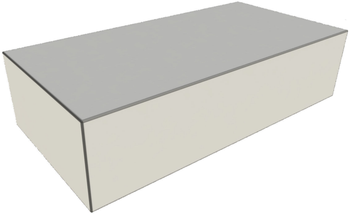
The tricky direction: if $(k_m)_{m=1}^M$ are universal ...

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

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

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

Proof idea

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

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

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

Proof idea

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

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

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

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

Summary

We studied the validness of HSIC.

- HSIC $\Rightarrow$ product structure:
  - Space: $\mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$. Kernel: $k = \otimes_{m=1}^{M} k_m$.
  - $=\text{MMD}(\mathbb{P}, \otimes_m \mathbb{P}_m) = \|\text{cross-cov. op.}\|_{\mathcal{H}_k}$.
- 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: $k = \otimes_{m=1}^{M} k_m$.
  - $=\text{MMD}(\mathbb{P}, \otimes_m \mathbb{P}_m) = \|\text{cross-cov. op.}\|_{\mathcal{H}_k}$.
- Complete answer in terms of k_m-s.
- ITE toolkit, JMLR:

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

Z. Szabó, B. K. Sriperumbudur. Characteristic and Universal Tensor Product Kernels. JMLR 18(233):1-29, 2018.

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.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Neural Information Processing Systems (NIPS)*, pages 625–632.

Cuturi, M. (2011).
Fast global alignment kernels.
In *International Conference on Machine Learning (ICML)*, pages 929–936.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:1169–1198.

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.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

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.

Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.
In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, Department of Computer Science, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hein, M. and Bousquet, O. (2005).
Hilbertian metrics and positive definite kernels on probability measures.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Jiao, Y. and Vert, J.-P. (2016).
The Kendall and Mallows kernels for permutations.

In *International Conference on Machine Learning (ICML; PMLR)*, volume 37, pages 2982–2990.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*, pages 291–298.

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.

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.
In *Neural Information Processing Systems (NIPS)*, pages 2982–2990.

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.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).

Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

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.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2009).
Nonextensive information theoretic kernels on measures.
The Journal of Machine Learning Research, 10:935–975.

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.

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.

Szabó, Z. and Sriperumbudur, B. K. (2018).
Characteristic and universal tensor product kernels.
Journal of Machine Learning Research, 18(233):1–29.

Vishwanathan, S. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010).
Graph kernels.
Journal of Machine Learning Research, 11:1201–1242.

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.