Tensor Product Kernels for Independence

Zoltán Szabó – Dept. of Statistics, LSE

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

DataSig Seminar,
Mathematical Institute, University of Oxford
May 26, 2022

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

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

Motivation: 'classical' information theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

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:

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

- Mutual information:

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

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 $\mathbf{x}^T\mathbf{y}$

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

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

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

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

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

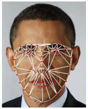
$\mathcal{X} = \mathbb{R}^d$, $\gamma > 0$:

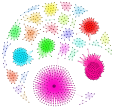
$$\boxed{k(\mathbf{x},\mathbf{y}) = \langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle_{\mathcal{H}}}$$

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

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

Kernels exist on various domains!





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

- Strings
  [Watkins, 1999, Lodhi et al., 2002, Leslie et al., 2002, Kuang et al., 2004, Leslie and Kuang, 2004, Saigo et al., 2004, Cuturi and Vert, 2005],
- time series [Rüping, 2001, Cuturi et al., 2007, Cuturi, 2011, Király and Oberhauser, 2019],
- trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002],
- groups and specifically rankings [Cuturi et al., 2005, Jiao and Vert, 2016],
- sets [Haussler, 1999, Gärtner et al., 2002], probability distributions [Berlinet and Thomas-Agnan, 2004, Hein and Bousquet, 2005, Smola et al., 2007, Sriperumbudur et al., 2010],
- various generative models [Jaakkola and Haussler, 1999, Tsuda et al., 2002, Seeger, 2002, Jebara et al., 2004],
- fuzzy domains [Guevara et al., 2017], or
- graphs [Kondor and Lafferty, 2002, Gärtner et al., 2003, Kashima et al., 2003, Borgwardt and Kriegel, 2005, Shervashidze et al., 2009, Vishwanathan et al., 2010, Kondor and Pan, 2016, Bai et al., 2020, Borgwardt et al., 2020].

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

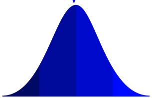
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- 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

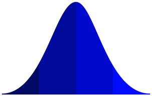
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

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

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

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



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

Kernel, RKHS: intuition

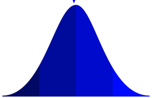
Given: $\mathcal{X}$ set. $\mathcal{H}$(ilbert space).

- Kernel:

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

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

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



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

Kernels: $+2$ definitions

- Def-1 (feature space):

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

- Def-2 (reproducing kernel, constructive):

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

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:

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

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:

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

Trick

φ: on any kernel-endowed domain!

Distribution representation: on kernel-enriched domains

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:

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

Distribution representation: on kernel-enriched domains

Mean embedding:

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

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

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

Mean embedding:

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

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

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

Mean embedding → MMD, HSIC

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

- Mean embedding:

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

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

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

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

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

MMD, HSIC: easy to estimate!

Mean embedding, MMD, HSIC: a few applications

- two-sample testing [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Borgwardt et al., 2006, Harchaoui et al., 2007, Gretton et al., 2012, Jitkrittum et al., 2016], and its differential private variant [Raj et al., 2019]; independence [Gretton et al., 2008, Pfister et al., 2018, Jitkrittum et al., 2017a] and goodness-of-fit testing [Jitkrittum et al., 2017b, Balasubramanian et al., 2021], causal discovery [Mooij et al., 2016, Pfister et al., 2018],
- domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017], change-point detection [Harchaoui and Cappé, 2007], post selection inference [Yamada et al., 2018],
- 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],
- topological data analysis [Kusano et al., 2016],
- 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],
- generative adversarial networks [Dziugaite et al., 2015, Li et al., 2015, Binkowski et al., 2018], understanding the dynamics of complex dynamical systems [Klus et al., 2018, Klus et al., 2019], ...

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 $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ on finite signed measures: universality [Steinwart, 2001].

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 $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ on finite signed measures: universality [Steinwart, 2001].

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

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

Central in applications: characteristic property

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

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

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

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

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

- $\otimes_{m=1}^{M} k_m$: 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 $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ 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 $\mathbb{R}^d$:

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

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

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

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	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\|j\|}{n+1}\right)\delta(\omega-j)$	$\{0,\pm 1,\pm 2,\dots,\pm n\}$

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]:

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

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

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

Goal

Extension to $M \geqslant 2$.

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

- Characteristic property:

$$\mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_k(\mathbb{P}_1 - \mathbb{P}_2) \neq 0.$$

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

- Characteristic property:

$$\mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_k(\mathbb{P}_1 - \mathbb{P}_2) \neq 0.$$

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

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

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

- Characteristic property:

$$\mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_k(\mathbb{P}_1 - \mathbb{P}_2) \neq 0.$$

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

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

- Witness construction:

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

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

- Characteristic property:

$$\mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_k(\mathbb{P}_1 - \mathbb{P}_2) \neq 0.$$

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

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

- Witness construction:

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

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

- Characteristic property:

$$\mathbb{P}_1 - \mathbb{P}_2 \neq 0 \Rightarrow \mu_k(\mathbb{P}_1 - \mathbb{P}_2) \neq 0.$$

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

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

- Witness construction:

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

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

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

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

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?

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.*

$$
\begin{aligned}
p_{1,1,1} &= \frac{1}{5}, \quad p_{1,1,2} = \frac{1}{10}, \quad p_{1,2,1} = \frac{1}{10}, \quad p_{1,2,2} = \frac{1}{10},\\
p_{2,1,1} &= \frac{1}{5}, \quad p_{2,1,2} = \frac{1}{10}, \quad p_{2,2,1} = \frac{1}{10}, \quad p_{2,2,2} = \frac{1}{10}.
\end{aligned}
$$

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.*
- $\boxed{\not\Leftarrow}$ $[|\mathcal{X}_m| = 2,\ k_m(x,x') = 2\delta_{x,x'} - 1]$

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

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

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,

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

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.

$$\mu(\mathbb{F}) = 0 \overset{?}{\Longrightarrow} \mathbb{F} = 0.$$

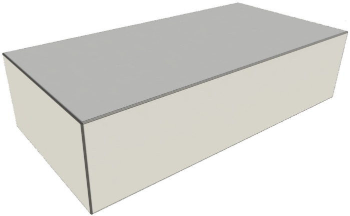
The tricky direction: if $(k_m)_{m=1}^M$ are universal ...

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

$$\mu(\mathbb{F}) = 0 \overset{?}{\Longrightarrow} \mathbb{F} = 0.$$

Enough:

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



Proof idea

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

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

Proof idea

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

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

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

Proof idea

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

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

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

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

Summary

We studied the validness of HSIC.

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

Summary

We studied the validness of HSIC.

- Space: $\mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$. Kernel: $k = \otimes_{m=1}^{M} k_m$.
- $\text{HSIC}_k(\mathbb{P}) = \text{MMD}_k\left(\mathbb{P}, \otimes_m \mathbb{P}_m\right) = \left\| \text{cross-cov. op.} \right\|_{\mathcal{H}_k}$.
- 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!

Bai, L., Cui, L., Rossi, L., Xu, L., Bai, X., and Hancock, E. (2020).
Local-global nested graph kernels using nested complexity traces.
Pattern Recognition Letters, 134:87–95.

Balasubramanian, K., Li, T., and Yuan, M. (2021).
On the optimality of kernel-embedding based goodness-of-fit tests.
Journal of Machine Learning Research, 22(1):1–45.

Baringhaus, L. and Franz, C. (2004).
On a new multivariate two-sample test.
Journal of Multivariate Analysis, 88:190–206.

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

Binkowski, M., Sutherland, D., Arbel, M., and Gretton, A. (2018).
Demystifying MMD GANs.
In *International Conference on Learning Representations (ICLR)*.

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., Ghisu, E., Llinares-López, F., O'Bray, L., and Riec, B. (2020).

Graph kernels: State-of-the-art and future challenges.
Foundations and Trends in Machine Learning, 13(5-6):531–712.

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.

Borgwardt, K. M. and Kriegel, H.-P. (2005).
Shortest-path kernels on graphs.
In *International Conference on Data Mining (ICDM)*, pages 74–81.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Advances 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.

Cuturi, M. and Vert, J.-P. (2005).
The context-tree kernel for strings.
Neural Networks, 18(8):1111–1123.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007).
A kernel for time series based on global alignments.
In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015).
Training generative neural networks via maximum mean discrepancy optimization.

In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 258–267.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.
In *Advances 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., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Gärtner, T., Flach, P., and Wrobel, S. (2003).
On graph kernels: Hardness results and efficient alternatives.
Learning Theory and Kernel Machines, pages 129–143.

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., Rasch, M., Schölkopf, B., and Smola, A. (2012).
A kernel two-sample test.
Journal of Machine Learning Research, 13(25):723–773.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. (2008).
A kernel statistical test of independence.
In *Advances 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 *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Harchaoui, Z., Bach, F., and Moulines, E. (2007).
Testing for homogeneity with kernel Fisher discriminant analysis.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616.

Harchaoui, Z. and Cappé, O. (2007).
Retrospective multiple change-point estimation with kernels.
In *IEEE/SP Workshop on Statistical Signal Processing*, pages 768–772.

Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, 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.

Jaakkola, T. S. and Haussler, D. (1999).
Exploiting generative models in discriminative classifiers.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–493.

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)*, volume 37, pages 2982–2990.

Jitkrittum, W., Szabó, Z., Chwialkowski, K., and Gretton, A. (2016).

Interpretable distribution features with maximum testing power.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 181–189.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2017a).
An adaptive test of independence with analytic kernel embeddings.
In *International Conference on Machine Learning (ICML)*, volume 70, pages 1742–1751. PMLR.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017b).
A linear-time kernel goodness-of-fit test.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 261–270.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.

In *International Conference on Machine Learning (ICML)*, pages 291–298.

Kashima, H., Tsuda, K., and Inokuchi, A. (2003).
Marginalized kernels between labeled graphs.
In *International Conference on Machine Learning (ICML)*, pages 321–328.

Kim, B., Khanna, R., and Koyejo, O. (2016).
Examples are not enough, learn to criticize! criticism for interpretability.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2280–2288.

Király, F. J. and Oberhauser, H. (2019).
Kernels for sequentially ordered data.
Journal of Machine Learning Research, 20:1–45.

Klus, S., Bittracher, A., Schuster, I., and Schütte, C. (2019).
A kernel-based approach to molecular conformation analysis.
The Journal of Chemical Physics, 149:244109.

Klus, S., Schuster, I., and Muandet, K. (2018).
Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces.
Technical report.
(https://arxiv.org/abs/1712.01572).

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

Kondor, R. I. and Lafferty, J. (2002).
Diffusion kernels on graphs and other discrete input.
In *International Conference on Machine Learning (ICML)*, pages 315–322.

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2004).
Profile-based string kernels for remote homology detection and motif extraction.

Journal of Bioinformatics and Computational Biology, 13(4):527–550.

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., Sejdinovic, D., and Flaxman, S. (2018).
Bayesian approaches to distribution regression.
International Conference on Artificial Intelligence and Statistics (AISTATS), 84:1167–1176.

Leslie, C., Eskin, E., and Noble, W. S. (2002).
The spectrum kernel: A string kernel for SVM protein classification.
Biocomputing, pages 564–575.

Leslie, C. and Kuang, R. (2004).

Fast string kernels using inexact matching for protein sequences.
Journal of Machine Learning Research, 5:1435–1455.

Li, Y., Swersky, K., and Zemel, R. (2015).
Generative moment matching networks.
In *International Conference on Machine Learning (ICML)*, pages 1718–1727.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J., 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), 37:1452–1461.

Lyons, R. (2013).
Distance covariance in metric spaces.
The Annals of Probability, 41:3284–3305.

Mooij, J., 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 *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18.

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)*, volume 51, pages 398–407.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018).
Kernel-based tests for joint independence.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(1):5–31.

Raj, A., Law, H. C. L., Sejdinovic, D., and Park, M. (2019).
A differentially private kernel two-sample test.
In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, pages 697–724.

Rüping, S. (2001).
SVM kernels for time series analysis.
Technical report, University of Dortmund.
(http://www.stefan-rueping.de/publications/rueping-2001-a.pdf).

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).
Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682–1689.

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.

Seeger, M. (2002).
Covariance kernels from Bayesian generative models.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 905–912.

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

Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., and Borgwardt, K. M. (2009).
Efficient graphlet kernels for large graph comparison.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 488–495.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *Algorithmic Learning Theory (ALT)*, pages 13–31.

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.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. (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. and Sriperumbudur, B. K. (2018).
Characteristic and universal tensor product kernels.
Journal of Machine Learning Research, 18(233):1–29.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016).
Learning theory for distribution regression.

Journal of Machine Learning Research, 17(152):1–40.

Székely, G. and Rizzo, M. (2004).
Testing for equal distributions in high dimension.
InterStat, 5:1249–1272.

Székely, G. and Rizzo, M. (2005).
A new test for multivariate normality.
Journal of Multivariate Analysis, 93:58–80.

Tsuda, K., Kin, T., and Asai, K. (2002).
Marginalized kernels for biological sequences.
Bioinformatics, 18:268–275.

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

Watkins, C. (1999).
Dynamic alignment kernels.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.

Yamada, M., Umezu, Y., Fukumizu, K., and Takeuchi, I. (2018).
Post selection inference with kernels.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pages 152–160.

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