

Tensor Product Kernels: Characteristic Property and Beyond

Zoltán Szabó – CMAP, École Polytechnique



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

Advanced Methods Group
Cubist Systematic Strategies, LLC
November 28, 2017

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

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

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

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

Motivation: 'Classical' Information Theory

- Kullback-Leibler divergence:

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

- Mutual information:

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

Properties:

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

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

Euclidean Space $\rightarrow$ Inner Product $\rightarrow$ Kernel

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

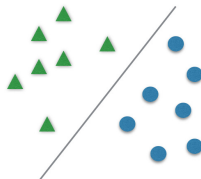
Euclidean Space $\rightarrow$ Inner Product $\rightarrow$ Kernel

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

Euclidean Space $\rightarrow$ Inner Product $\rightarrow$ Kernel

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

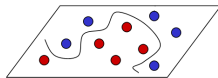
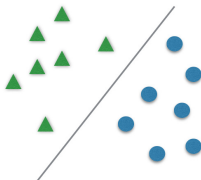
① Classification:



Euclidean Space $\rightarrow$ Inner Product $\rightarrow$ Kernel

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

1 Classification:

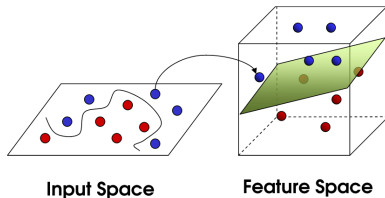
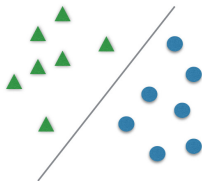


Input Space

Euclidean Space $\rightarrow$ Inner Product $\rightarrow$ Kernel

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

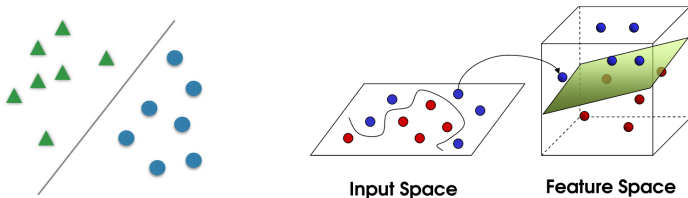
1 Classification:



Euclidean Space $\rightarrow$ Inner Product $\rightarrow$ Kernel

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

1 Classification:

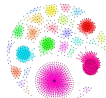


2 Representation of distributions (this talk):

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

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

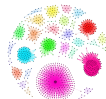
Diverse Set of Domains, Kernel Examples



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

$$\begin{aligned} k_p(\mathbf{x}, \mathbf{y}) &= (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, & k_G(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}, \\ k_e(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2}, & k_C(\mathbf{x}, \mathbf{y}) &= 1 + \frac{1}{\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}. \end{aligned}$$

Diverse Set of Domains, Kernel Examples

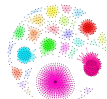


- $\mathcal{X} = \mathbb{R}^d, \gamma > 0$:

$$\begin{aligned} k_p(\mathbf{x}, \mathbf{y}) &= (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, & k_G(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}, \\ k_e(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2}, & k_C(\mathbf{x}, \mathbf{y}) &= 1 + \frac{1}{\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}. \end{aligned}$$

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

Diverse Set of Domains, Kernel Examples

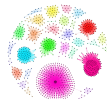


- $\mathcal{X} = \mathbb{R}^d, \gamma > 0$:

$$\begin{aligned} k_p(\mathbf{x}, \mathbf{y}) &= (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, & k_G(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}, \\ k_e(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2}, & k_C(\mathbf{x}, \mathbf{y}) &= 1 + \frac{1}{\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}. \end{aligned}$$

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

Diverse Set of Domains, Kernel Examples



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

$$\begin{aligned} k_p(\mathbf{x}, \mathbf{y}) &= (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p, & k_G(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}, \\ k_e(\mathbf{x}, \mathbf{y}) &= e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2}, & k_C(\mathbf{x}, \mathbf{y}) &= 1 + \frac{1}{\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}. \end{aligned}$$

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

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

- Mean embedding:

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

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

- Mean embedding:

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

- Maximum mean discrepancy:

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

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

- Mean embedding:

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

- Maximum mean discrepancy:

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

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

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

- Mean embedding:

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

- Maximum mean discrepancy:

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

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

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

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

Ingredients

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

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



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

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



Assumption

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

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

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

- **Kernel:**

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

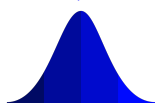
Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

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

- **Kernel:**

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

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

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


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

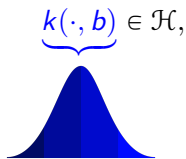
Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

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

- **Kernel:**

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

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



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

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

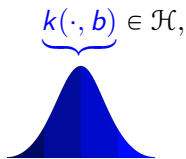
Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m, k := k_m$)

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

- **Kernel:**

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

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



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

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

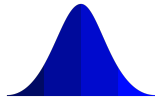
Ingredients: Kernel, RKHS ($\mathcal{X} := \mathcal{X}_m$, $k := k_m$)

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

- **Kernel:**

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

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

$$\underbrace{k(\cdot, b)}_{\text{blue bell curve}} \in \mathcal{H},$$


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

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

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

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

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

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

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

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

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

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

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

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

- Applications:
 - two-sample testing [Gretton et al., 2012], domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017],
 - interpretable machine learning [Kim et al., 2016],
 - 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].

Let us switch to HSIC.

MMD $\xrightarrow{\text{spec.}}$ HSIC

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

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

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

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

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

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

Applications:

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

Central in Applications: Characteristic Property

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

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

Central in Applications: Characteristic Property

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

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

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

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

Central in Applications: Characteristic Property

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

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

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

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

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

Central in Applications: Characteristic Property

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

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

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

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

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

Wanted

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

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

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

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

(*) : Bochner's theorem.

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

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

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

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

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

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

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

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

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

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

Theorem ([Sriperumbudur et al., 2010])

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

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

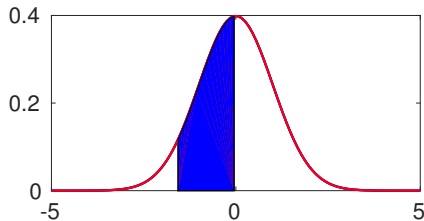
kernel name	k_0	$\hat{k}_0(\omega)$	$\text{supp}(\hat{k}_0)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2 \omega^2}{2}}$	$\mathbb{R}$
Laplacian	$e^{-\sigma x }$	$\sqrt{\frac{2}{\pi}} \frac{\sigma}{\sigma^2 + \omega^2}$	$\mathbb{R}$
B_{2n+1} -spline	$*^{2n+2} \chi_{[-\frac{1}{2}, \frac{1}{2}]}(x)$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

Measures: Intuition

- Probability measures. Idea:

$$\int_{\mathcal{X}} p(x) dx = 1, \quad p \geq 0.$$

Usage: $\mathbb{P}(H) = \int_H p(x) dx$.

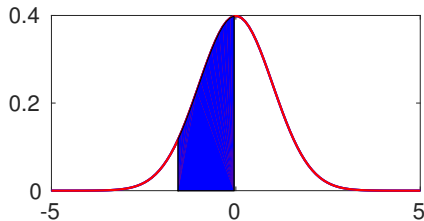


Measures: Intuition

- Probability measures. Idea:

$$\int_{\mathcal{X}} p(x) dx \stackrel{\text{NO}}{=} 1, \quad p \geq 0.$$

Usage: $\mathbb{P}(H) = \int_H p(x) dx$.



- Non-negative (finite) measures:

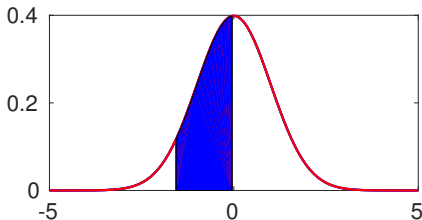
$$\int p(x) dx < \infty, \quad p \geq 0.$$

Measures: Intuition

- Probability measures. Idea:

$$\int_{\mathcal{X}} p(x) dx \stackrel{\text{NO}}{=} 1, \quad p \stackrel{\text{NO}}{\geq} 0.$$

Usage: $\mathbb{P}(H) = \int_H p(x) dx$.



- Non-negative (finite) measures:

$$\int p(x) dx < \infty, \quad p \geq 0.$$

- Finite signed measures:

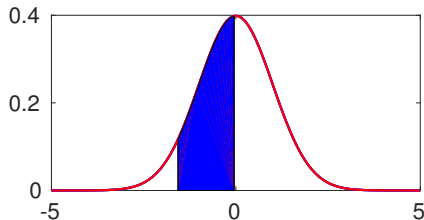
$$\int p(x) dx < \infty.$$

Measures: Intuition

- Probability measures. Idea:

$$\int_{\mathcal{X}} p(x) dx \stackrel{\text{NO}}{=} 1, \quad p \stackrel{\text{NO}}{\geq} 0, \quad \sum_i p_i = 1, \quad p_i \geq 0 \ (\forall i).$$

Usage: $\mathbb{P}(H) = \int_H p(x) dx$.



- Non-negative (finite) measures:

$$\int p(x) dx < \infty, \quad p \geq 0.$$

- Finite signed measures:

$$\int p(x) dx < \infty.$$

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

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

$$\mu_k : \underbrace{\mathcal{M}_b(\mathcal{X})}_{\text{bounded signed measures on } \mathcal{X}} \mapsto \mathcal{H}_k$$

is injective.

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

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

is injective.

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

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

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

is injective.

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

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

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

is injective.

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

Challenge

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

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

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

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

Goal

Extension to $M \geq 2$.

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

Goal

Extension to $M \geq 2$.

Main Challenge

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

Idea: Characteristic Property as lspd

- Characteristic property:

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

Idea: Characteristic Property as Ispd

- Characteristic property:

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

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

Idea: Characteristic Property as Ispd

- Characteristic property:

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

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

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

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

Idea: Characteristic Property as Ispd

- Characteristic property:

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

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

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

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

Idea: Characteristic Property as Ispd

- Characteristic property:

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

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

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

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

- We saw: k is **universal** iff.

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

From now on: $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$. Let $\mathcal{F} \subseteq \mathcal{M}_b(\mathcal{X})$, $0 \in \mathcal{F}$.

Definition

$k = \otimes_{m=1}^M k_m$ is called $\mathcal{F}$ -ispd if

$$\|\mu_k(\mathbb{F})\|_{\mathcal{H}_k}^2 > 0, \quad \forall \mathbb{F} \in \mathcal{F} \setminus \{0\}$$

From now on: $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$. Let $\mathcal{F} \subseteq \mathcal{M}_b(\mathcal{X})$, $0 \in \mathcal{F}$.

Definition

$k = \otimes_{m=1}^M k_m$ is called $\mathcal{F}$ -ispd if

$$\begin{aligned} \|\mu_k(\mathbb{F})\|_{\mathcal{H}_k}^2 &> 0, \quad \forall \mathbb{F} \in \mathcal{F} \setminus \{0\}, \text{ equivalently} \\ \mu_k(\mathbb{F}) &= 0 \Rightarrow \mathbb{F} = 0 \quad (\mathbb{F} \in \mathcal{F}). \end{aligned}$$

Examples

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

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

UI

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

$\Downarrow$

Examples

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

$$\begin{array}{ccc}
 \subseteq & & \subseteq \quad [\mathcal{M}_b(\mathcal{X})]^0 \subseteq \mathcal{M}_b(\mathcal{X}) . \\
 & & \quad \cup \\
 & & \quad \mathcal{I} \\
 \Leftarrow & & \Leftarrow \text{characteristic} \Leftarrow \text{universal} . \\
 & & \quad \Downarrow \\
 & & \mathcal{I}\text{-characteristic}
 \end{array}$$

Examples

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

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

UI

$\mathcal{I}$

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

$\Downarrow$

$\mathcal{I}$ -characteristic

Examples

$\mathcal{F}$	$\mathcal{F}$ -ispd k
$\mathcal{M}_b(\mathcal{X})$	universal
$[\mathcal{M}_b(\mathcal{X})]^0$	characteristic
$\mathcal{I} := \{\mathbb{P} - \bigotimes_{m=1}^M \mathbb{P}_m\}$	$\mathcal{I}$ -characteristic
$[\bigotimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)]^0$	$\bigotimes$ -characteristic
$\bigotimes_{m=1}^M \mathcal{M}_b^0(\mathcal{X}_m)$	$\bigotimes_0$ -characteristic

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

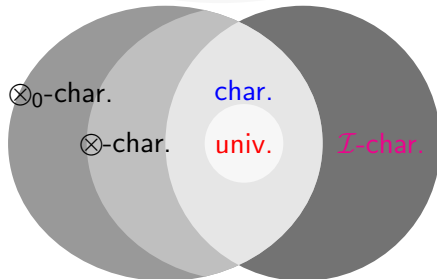
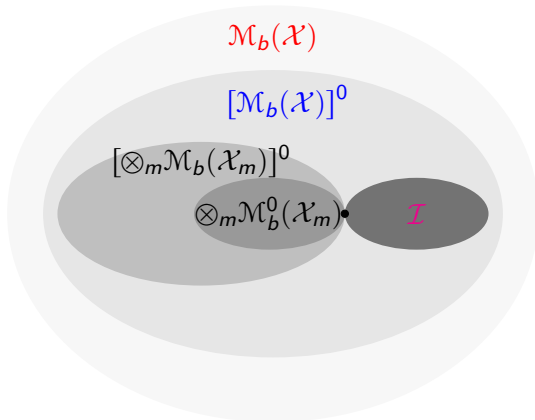
UI

$\mathcal{I}$

$$\bigotimes_0\text{-characteristic} \Leftarrow \bigotimes\text{-characteristic} \Leftarrow \text{characteristic} \Leftarrow \text{universal}.$$



$\mathcal{I}$ -characteristic



$\otimes_0$ -char $\longleftrightarrow$ $\otimes$ -char $\longleftrightarrow$ char $\longleftrightarrow$ universal



$\mathcal{I}$ -char

$(k_m)_{m=1}^M$ char $\xrightarrow{\text{[Sriperumbudur et al., 2011]}}$ $(k_m)_{m=1}^M$ -universal
 $\xleftarrow{\text{[Sriperumbudur et al., 2011]}}$

Results

Various Characteristic Properties of $\bigotimes_{m=1}^M k_m$

Proposition

- (i) $\bigotimes_{m=1}^M k_m$: *characteristic* $\Rightarrow$ $\bigotimes$ -*characteristic*.
- (ii) $\bigotimes_{m=1}^M k_m$: $\bigotimes$ -*characteristic* $\Rightarrow$ $\bigotimes_0$ -*characteristic*.
- (iii) $\bigotimes_{m=1}^M k_m$: $\bigotimes_0$ -*characteristic* $\Leftrightarrow (k_m)_{m=1}^M$ are *characteristic*.

Various Characteristic Properties of $\bigotimes_{m=1}^M k_m$

Proposition

- (i) $\bigotimes_{m=1}^M k_m$: *characteristic* $\Rightarrow$ $\bigotimes$ -*characteristic*.
- (ii) $\bigotimes_{m=1}^M k_m$: $\bigotimes$ -*characteristic* $\Rightarrow$ $\bigotimes_0$ -*characteristic*.
- (iii) $\bigotimes_{m=1}^M k_m$: $\bigotimes_0$ -*characteristic* $\Leftrightarrow (k_m)_{m=1}^M$ are *characteristic*.

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

$$\underbrace{\|\mu_k(\mathbb{F})\|_{\mathcal{H}_k}^2}_{>0} = \underbrace{\prod_{m=1}^M}_{\forall} \underbrace{\|\mu_{k_m}(\mathbb{F}_m)\|_{\mathcal{H}_{k_m}}^2}_{>0},$$

$\otimes_0$ -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

Finite signed measures on $\mathcal{X}_m = \{1, 2\}$:

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

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

Finite signed measures on $\mathcal{X}_m = \{1, 2\}$:

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

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

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

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

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

Finite signed measures on $\mathcal{X}_m = \{1, 2\}$:

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

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

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

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

This gives

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

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

Finite signed measures on $\mathcal{X}_m = \{1, 2\}$:

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

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.

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

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

This gives

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

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

$$\begin{aligned} a_1 + a_2 &= 0, & b_1 &= b_2. \\ a_1 &= a_2, & b_1 + b_2 &= 0. \end{aligned}$$

In the previous example:

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

In fact:

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

Proposition

- (i) k_1, k_2 : *characteristic* $\Rightarrow k_1 \otimes k_2$: $\mathcal{I}$ -*characteristic*.
- (ii) $\otimes_{m=1}^M k_m$: $\mathcal{I}$ -*characteristic* $\Rightarrow (k_m)_{m=1}^M$ are *characteristic*.

Proposition

- (i) k_1, k_2 : *characteristic* $\Rightarrow k_1 \otimes k_2$: $\mathcal{I}$ -*characteristic*.
- (ii) $\otimes_{m=1}^M k_m$: $\mathcal{I}$ -*characteristic* $\Rightarrow (k_m)_{m=1}^M$ are *characteristic*.

Proof idea:

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

Proposition

- (i) k_1, k_2 : *characteristic* $\Rightarrow k_1 \otimes k_2$: $\mathcal{I}$ -*characteristic*.
- (ii) $\otimes_{m=1}^M k_m$: $\mathcal{I}$ -*characteristic* $\Rightarrow (k_m)_{m=1}^M$ are *characteristic*.

Proof idea:

- (i) Induction: see later universality ($\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$).
- (ii) If a k_m is not characteristic, witness can be constructed.

k_1, k_2, k_3 : characteristic $\Rightarrow \bigotimes_{m=1}^3 k_m$: $\mathcal{I}$ -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.
 - $\bigotimes_{m=1}^3 k_m$: is **not** $\mathcal{I}$ -characteristic. Witness:

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

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

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

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

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

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

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

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

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

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

$\mathbb{R}^d$ & Translation-invariance: All Notions Coincide

Proposition

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

- (i) $(k_m)_{m=1}^M$ -s are characteristic.
- (ii) $\otimes_{m=1}^M k_m$: $\otimes_0$ -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 $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \rightarrow \mathbb{R}$ are continuous, translation-invariant kernels. Then the followings are equivalent:

- (i) $(k_m)_{m=1}^M$ -s are characteristic.
- (ii) $\otimes_{m=1}^M k_m$: $\otimes_0$ -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

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

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

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

- Since k_m is characteristic

$$k_m \xrightarrow{\text{Bochner thm}} \Lambda_m, \text{ } \text{supp}(\Lambda_m) = \mathbb{R}^{d_m}.$$

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

- Since k_m is characteristic

$$k_m \xrightarrow{\text{Bochner thm}} \Lambda_m, \text{ } \text{supp}(\Lambda_m) = \mathbb{R}^{d_m}.$$

- Tensor kernel:

$$\bigotimes_{m=1}^M k_m \xrightarrow{\text{Bochner thm}} \Lambda = \bigotimes_{m=1}^M \Lambda_m.$$

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

- Since k_m is characteristic

$$k_m \xrightarrow{\text{Bochner thm}} \Lambda_m, \text{ } \text{supp}(\Lambda_m) = \mathbb{R}^{d_m}.$$

- Tensor kernel:

$$\bigotimes_{m=1}^M k_m \xrightarrow{\text{Bochner thm}} \Lambda = \bigotimes_{m=1}^M \Lambda_m.$$

- $\text{supp}(\Lambda) = \times_{m=1}^M \underbrace{\text{supp}(\Lambda_m)}_{\mathbb{R}^{d_m}} = \mathbb{R}^d.$

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

We saw: for $M \geq 3$

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

Proposition

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

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

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

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

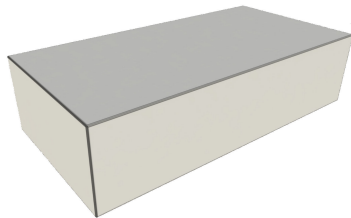
The Tricky Direction: If $(k_m)_{m=1}^M$ are Universal ...

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

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

Enough:

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



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

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

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

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

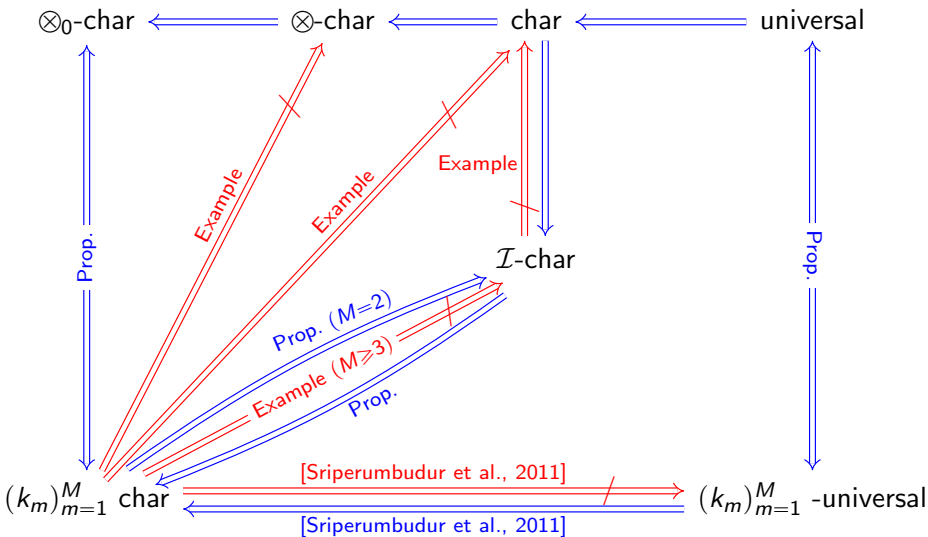
$$0 = \mathbb{F}\left(\times_{m=1}^M B_m\right) = \int_{\mathcal{X}} \times_{m=1}^M \chi_{B_m}(x_m) d\mathbb{F}(x), \quad \forall B_m.$$

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

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

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

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



We studied the validness of HSIC.

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

We studied the validness of HSIC.

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

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

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

Thank you for the attention!

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



Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2017).

Domain generalization by marginal transfer learning.

Technical report.

<https://arxiv.org/abs/1711.07910>.



Blanchard, G., Lee, G., and Scott, C. (2011).

Generalizing from several related classification tasks to a new unlabeled sample.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186.



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.



Gretton, A. (2015).

A simpler condition for consistency of a kernel independence test.

Technical report, University College London.

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



Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).

A kernel two-sample test.

Journal of Machine Learning Research, 13:723–773.



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

Measuring statistical dependence with Hilbert-Schmidt norms.

In *Algorithmic Learning Theory (ALT)*, pages 63–78.



Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008).

A kernel statistical test of independence.

In *Neural Information Processing Systems (NIPS)*, pages 585–592.



Kim, B., Khanna, R., and Koyejo, O. O. (2016).

Examples are not enough, learn to criticize! criticism for interpretability.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 2280–2288.



Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016).

Persistence weighted Gaussian kernel for topological data analysis.

In *International Conference on Machine Learning (ICML)*, pages 2004–2013.



Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2014).

Automatic construction and natural-language description of nonparametric regression models.

In *AAAI Conference on Artificial Intelligence*, pages 1242–1250.



Lyons, R. (2013).

Distance covariance in metric spaces.
The Annals of Probability, 41:3284–3305.



Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016).

Distinguishing cause from effect using observational data:
Methods and benchmarks.
Journal of Machine Learning Research, 17:1–102.



Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2011).

Learning from distributions via support measure machines.
In *Neural Information Processing Systems (NIPS)*, pages 10–18.



Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017).

Kernel mean embedding of distributions: A review and beyond.

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



Park, M., Jitkrittum, W., and Sejdinovic, D. (2016).

K2-ABC: Approximate Bayesian computation with kernel embeddings.

In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 51, pages 51:398–407.



Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2017).

Kernel-based tests for joint independence.

Journal of the Royal Statistical Society: Series B (Statistical Methodology).



Schölkopf, B., Muandet, K., Fukumizu, K., Harmeling, S., and Peters, J. (2015).

Computing functions of random variables via reproducing kernel Hilbert space representations.



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.



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.