The Power of Cumulants in Reproducing Kernel Hilbert Spaces

Zoltán Szabó @ Department of Statistics, LSE

Joint work with: Patric Bonnier, Harald Oberhauser

- @ Mathematical Institute, University of Oxford.


Computer Science Colloquium, University of Warwick
Feb. 12, 2024

Moments on $\mathbb{R} \ni X \sim \gamma$

- Moments $\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i \in \mathbb{N}}$:

$$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}, \qquad \mu^{(0)}(\gamma) := 1.$$

Moments and cumulants [McCullagh, 2018] on $\mathbb{R} \ni X \sim \gamma$

- Moments $\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i \in \mathbb{N}}$:

$$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}, \qquad \mu^{(0)}(\gamma) := 1.$$

- Cumulants $\kappa(\gamma) = \left(\kappa^{(i)}(\gamma)\right)_{i \in \mathbb{N}}$: from the moment-generating function

$$\sum_{i \in \mathbb{N}} \kappa^{(i)}(\gamma) \frac{\theta^i}{i!} = \log\left(\sum_{i \in \mathbb{N}} \mu^{(i)}(\gamma) \frac{\theta^i}{i!}\right).$$

Moments and cumulants [McCullagh, 2018] on $\mathbb{R} \ni X \sim \gamma$

- Moments $\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$:

$$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}, \qquad \mu^{(0)}(\gamma) := 1.$$

- Cumulants $\kappa(\gamma) = \left(\kappa^{(i)}(\gamma)\right)_{i\in\mathbb{N}}$: from the moment-generating function

$$\sum_{i\in\mathbb{N}} \kappa^{(i)}(\gamma)\frac{\theta^i}{i!} = \log\left(\sum_{i\in\mathbb{N}} \mu^{(i)}(\gamma)\frac{\theta^i}{i!}\right).$$

$\kappa^{(1)}(\gamma) = \mathbb{E}(X)$	mean
$\kappa^{(2)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^2$	variance
$\kappa^{(3)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^3$	3rd central moment
$\kappa^{(4)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^4 - 3\left[\mathbb{E}(X - \mathbb{E}X)^2\right]^2$	
$\kappa^{(5)}(\gamma) = \mathbb{E}(X - \mathbb{E}X)^5 - 10\mathbb{E}(X - \mathbb{E}X)^3\mathbb{E}(X - \mathbb{E}X)^2$	

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description [Speed, 1983, Speed, 1984]

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description [Speed, 1983, Speed, 1984]

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \mathbb{E}^2(X)$$

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description [Speed, 1983, Speed, 1984]

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \overbrace{\mathbb{E}^2(X)}^{\mathbb{E}(XX')}, \qquad \{\{1,2\}\}, \{\{1\},\{2\}\}$$

where $X, X' \sim \gamma$, independent.

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description [Speed, 1983, Speed, 1984]

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \overbrace{\mathbb{E}^2(X)}^{\mathbb{E}(XX')}, \qquad \{\{1,2\}\}, \{\{1\},\{2\}\}$$

$$\kappa^{(3)}(\gamma) = \mathbb{E}\left(X^3\right) - 3\mathbb{E}\left(X^2\right)\mathbb{E}(X) + 2\mathbb{E}^3(X)$$

where $X, X' \sim \gamma$, independent.

Unzipping cumulants on $\mathbb{R}$: (known) combinatorial description [Speed, 1983, Speed, 1984]

$$\kappa^{(1)}(\gamma) = \mathbb{E}(X), \qquad \{\{1\}\}$$

$$\kappa^{(2)}(\gamma) = \mathbb{E}\left(X^2\right) - \overbrace{\mathbb{E}^2(X)}^{\mathbb{E}(XX')}, \qquad \{\{1,2\}\}, \{\{1\},\{2\}\}$$

$$\kappa^{(3)}(\gamma) = \mathbb{E}\left(X^3\right) - \overbrace{3\mathbb{E}\left(X^2\right)\mathbb{E}(X)}^{\mathbb{E}(XXX')+\mathbb{E}(XX'X)+\mathbb{E}(X'XX)} + 2\mathbb{E}^3(X), \qquad \{\{1,2,3\}\}, \{\{1,2\},\{3\}\}, \{\{1,3\},\{2\}\}, \{\{2,3\},\{1\}\}, \{\{1\},\{2\},\{3\}\},$$

$\ldots$

where $X, X' \sim \gamma$, independent.

Question

What are the weights in front of the moments?

Unzipping cumulants on $\mathbb{R}$

m	elements of $\pi \in P(m)$	$\lvert\pi\rvert$
1	$\{1\}$	1
2	$\{1,2\}$	1
	$\{1\},\{2\}$	2
3	$\{1,2,3\}$	1
	$\{1,2\}, \{3\}$	2
	$\{1,3\}, \{2\}$	2
	$\{2,3\}, \{1\}$	2
	$\{1\}, \{2\}, \{3\}$	3

with $P(m) :=$ all partitions of $[m]$, $[m] := \{1, \dots, m\}$

Unzipping cumulants on $\mathbb{R}$: the weights

m	elements of $\pi \in P(m)$	$\lvert\pi\rvert$	c_π
1	$\{1\}$	1	1
2	$\{1,2\}$	1	1
	$\{1\},\{2\}$	2	-1
3	$\{1,2,3\}$	1	1
	$\{1,2\}, \{3\}$	2	-1
	$\{1,3\}, \{2\}$	2	-1
	$\{2,3\}, \{1\}$	2	-1
	$\{1\}, \{2\}, \{3\}$	3	2

with $P(m) :=$ all partitions of $[m]$, $[m] := \{1, \ldots, m\}$,

$$c_\pi = (-1)^{|\pi|-1}(|\pi| - 1)!$$

Motivation, i.e. one reason why one likes cumulants

Moment and cumulants on $\mathbb{R}^d$

Change $\mathbb{E}\left(X^i\right) \in \mathbb{R}$ to $\mathbb{E}\left(X_1^{i_1} \cdots X_d^{i_d}\right) \in \mathbb{R}$ ($\mathbf{i} \in \mathbb{N}^d$). log, $P(m)$: ✓

Known theorem [Billingsley, 2012]

Let γ be a probability measure on a bounded subset of $\mathbb{R}^d$ with cumulants $\kappa(\gamma)$ and let $(X_1, \ldots, X_d) \sim \gamma$. Then

1. $\gamma \mapsto \kappa(\gamma)$ is injective.
2. $X_1, \ldots, X_d$ are independent $\Leftrightarrow \kappa^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}_+^d$.

Motivation, i.e. one reason why one likes cumulants

Moment and cumulants on $\mathbb{R}^d$

Change $\mathbb{E}\left(X^i\right) \in \mathbb{R}$ to $\mathbb{E}\left(X_1^{i_1} \cdots X_d^{i_d}\right) \in \mathbb{R}$ ($\mathbf{i} \in \mathbb{N}^d$). log, $P(m)$: ✓

Known theorem [Billingsley, 2012]

Let γ be a probability measure on a bounded subset of $\mathbb{R}^d$ with cumulants $\kappa(\gamma)$ and let $(X_1, \ldots, X_d) \sim \gamma$. Then

1. $\gamma \mapsto \kappa(\gamma)$ is injective.
2. $X_1, \ldots, X_d$ are independent $\Leftrightarrow \kappa^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}_+^d$.

Motivation

1. Various data types, nonlinear features: kernels.
2. Linear: not even characteristic (see MMD and HSIC).
3. Computable estimators.

Idea

Lifting

$$(X_1, \ldots, X_d) \in \times_{j=1}^{d} \mathcal{X}_j \to (\Phi_1(X_1), \ldots, \Phi_d(X_d)) \in \times_{j=1}^{d} \mathcal{H}_{k_j}.$$

Idea

Lifting

$(X_1,\ldots,X_d) \in \times_{j=1}^d \mathcal{X}_j \to (\Phi_1(X_1),\ldots,\Phi_d(X_d)) \in \times_{j=1}^d \mathcal{H}_{k_j}.$

Ingredients:

1. Moments: swap out $\mathbb{E}\left(X_1^{i_1}\cdots X_d^{i_d}\right) \in \mathbb{R}$ to

$$\mathbb{E}\left([\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}\right) \in \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}.$$

Idea

Lifting

$$(X_1,\dots,X_d)\in \times_{j=1}^d \mathcal{X}_j \to (\Phi_1(X_1),\dots,\Phi_d(X_d))\in \times_{j=1}^d \mathcal{H}_{k_j}.$$

Ingredients:

1. Moments: swap out $\mathbb{E}\left(X_1^{i_1}\cdots X_d^{i_d}\right)\in\mathbb{R}$ to

$$\mathbb{E}\left([\Phi_1(X_1)]^{\otimes i_1}\otimes\cdots\otimes[\Phi_d(X_d)]^{\otimes i_d}\right)\in \mathcal{H}_{k_1}^{\otimes i_1}\otimes\cdots\otimes\mathcal{H}_{k_d}^{\otimes i_d}.$$

2. From moments to cumulants:
   - log on tensor algebras, or
   - combinatorial description of cumulants ($\leftarrow$ a bit simpler, but $\Leftrightarrow$).

Idea

Lifting

$$(X_1, \dots, X_d) \in \times_{j=1}^{d} \mathcal{X}_j \to (\Phi_1(X_1), \dots, \Phi_d(X_d)) \in \times_{j=1}^{d} \mathcal{H}_{k_j}.$$

Ingredients:

1. Moments: swap out $\mathbb{E}\left(X_1^{i_1} \cdots X_d^{i_d}\right) \in \mathbb{R}$ to

$$\mathbb{E}\left([\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}\right) \in \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}.$$

2. From moments to cumulants:
   - log on tensor algebras, or
   - combinatorial description of cumulants ($\leftarrow$ a bit simpler, but $\Leftrightarrow$).
3. Computation: by the 'expected kernel trick' (V-statistics).

Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

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

- Def-2 (reproducing kernel):

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

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

Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

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

- Def-2 (reproducing kernel):

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

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

Notes

- $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k = \overline{\mathrm{Span}}(k(\cdot, x) : x \in \mathcal{X})$: Fourier analysis, approximation with polynomials, splines, ...
- Examples ($c \ge 0$, $p \in \mathbb{N}$): $k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^p$, $k_G(\mathbf{x}, \mathbf{y}) = e^{-c\|\mathbf{x}-\mathbf{y}\|_2^2}$.
- 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, Balanca and Herbin, 2012, Fellmann et al., 2023], 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, Schulz et al., 2022, Nikolentzos and Vazirgiannis, 2023].

Mean embedding

- Mean embedding (integral; [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]):

$$\mu_k(\gamma) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k,\text{ a la }\Phi(x)=e^{i\langle\cdot,x\rangle}} \mathrm{d}\gamma(x) \in \mathcal{H}_k.$$

Mean embedding, MMD

- Mean embedding (integral; [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]):

$$\mu_k(\gamma) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k,\ \text{a la } \Phi(x)=e^{i\langle\cdot,x\rangle}} \mathrm{d}\gamma(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\gamma, \eta) := \|\mu_k(\gamma) - \mu_k(\eta)\|_{\mathcal{H}_k}.$$

Mean embedding, MMD, HSIC

- Mean embedding (integral; [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]):

$$\mu_k(\gamma) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k,\ \text{a la } \Phi(x)=e^{i\langle\cdot,x\rangle}} \mathrm{d}\gamma(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\gamma, \eta) := \|\mu_k(\gamma) - \mu_k(\eta)\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion [Gretton et al., 2005] ($d = 2$), [Quadrianto et al., 2009, Sejdinovic et al., 2013a] ($d \geq 3$), $k := \otimes_{j=1}^d k_j$:

$$\mathrm{HSIC}_k(\gamma) := \mathrm{MMD}_k\left(\gamma, \otimes_{j=1}^d \gamma|_{\mathcal{X}_j}\right)$$

Mean embedding, MMD, HSIC

- Mean embedding (integral; [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]):

$$\mu_k(\gamma) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k,\ \text{a la}\ \Phi(x)=e^{i\langle\cdot,x\rangle}} \mathrm{d}\gamma(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\gamma,\eta) := \|\mu_k(\gamma) - \mu_k(\eta)\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion [Gretton et al., 2005] ($d = 2$), [Quadrianto et al., 2009, Sejdinovic et al., 2013a] ($d \geq 3$), $k := \otimes_{j=1}^d k_j$:

$$\begin{aligned}
\mathrm{HSIC}_k(\gamma) &:= \mathrm{MMD}_k\left(\gamma, \otimes_{j=1}^d \gamma|_{\mathcal{X}_j}\right), \\
&= \Big\| \underbrace{\mu_{\otimes_{j=1}^d k_j}(\gamma) - \otimes_{j=1}^d \mu_{k_j}\left(\gamma|_{\mathcal{X}_j}\right)}_{\text{cross-covariance operator}} \Big\|_{\mathcal{H}_k}.
\end{aligned}$$

Mean embedding, MMD, HSIC

- Mean embedding (integral; [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]):

$$\mu_k(\gamma) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k,\text{ a la }\Phi(x)=e^{i\langle\cdot,x\rangle}} \mathrm{d}\gamma(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\gamma, \eta) := \left\|\mu_k(\gamma) - \mu_k(\eta)\right\|_{\mathcal{H}_k}.$$

- Hilbert-Schmidt independence criterion [Gretton et al., 2005] ($d = 2$), [Quadrianto et al., 2009, Sejdinovic et al., 2013a] ($d \geq 3$), $k := \otimes_{j=1}^d k_j$:

$$\begin{aligned}\mathrm{HSIC}_k(\gamma) &:= \mathrm{MMD}_k\left(\gamma, \otimes_{j=1}^d \gamma|_{\mathcal{X}_j}\right)\\ &= \Bigg\| \underbrace{\mu_{\otimes_{j=1}^d k_j}(\gamma) - \otimes_{j=1}^d \mu_{k_j}\left(\gamma|_{\mathcal{X}_j}\right)}_{\text{cross-covariance operator}} \Bigg\|_{\mathcal{H}_k}.\end{aligned}$$

Notes before clarification of what $\otimes_{j=1}^d k_j$ and $\otimes_{j=1}^d \mu_{k_j}\left(\gamma|_{\mathcal{X}_j}\right)$ are.

- M MD:

$$\mathrm{MMD}_k(\gamma,\eta) = \|\mu_k(\gamma) - \mu_k(\eta)\|_{\mathcal{H}_k} = \sup_{f \in B_k} \underbrace{\langle f, \mu_k(\gamma) - \mu_k(\eta)\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\gamma} f(x) - \mathbb{E}_{x\sim\eta} f(x)}$$

MMD, HSIC: information theoretical & statistical links

- M MD:

$$\mathrm{MMD}_k(\gamma,\eta) = \|\mu_k(\gamma) - \mu_k(\eta)\|_{\mathcal{H}_k} = \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\gamma) - \mu_k(\eta)\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\gamma} f(x) - \mathbb{E}_{x\sim\eta} f(x)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],

- M MD:

$$\mathrm{MMD}_k(\gamma,\eta) = \|\mu_k(\gamma) - \mu_k(\eta)\|_{\mathcal{H}_k} = \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\gamma) - \mu_k(\eta)\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\gamma} f(x) - \mathbb{E}_{x\sim\eta} f(x)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],
  - $\overset{\dagger}{\Leftrightarrow}$ energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005], a.k.a. N-distance [Zinger et al., 1992, Klebanov, 2005].

† [Sejdinovic et al., 2013b].

- M MD:

$$\mathrm{MMD}_k(\gamma,\eta) = \|\mu_k(\gamma) - \mu_k(\eta)\|_{\mathcal{H}_k} = \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\gamma) - \mu_k(\eta)\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\gamma} f(x) - \mathbb{E}_{x\sim\eta} f(x)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],
  - $\stackrel{\dagger}{\Leftrightarrow}$ energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005], a.k.a. N-distance [Zinger et al., 1992, Klebanov, 2005].
- HSIC ($d = 2$) $\stackrel{\dagger}{\Leftrightarrow}$ distance covariance [Székely et al., 2007, Székely and Rizzo, 2009, Lyons, 2013].

† [Sejdinovic et al., 2013b].

Interaction measures: $d = 3$, $(X_1, X_2, X_3) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$

$(X_1, X_2, X_3) \sim \gamma$, Lancaster interaction measure [Lancaster, 1969]

$$\begin{aligned} L(\gamma) := \gamma &- \gamma|_{\mathcal{X}_1\mathcal{X}_2} \otimes \gamma|_{\mathcal{X}_3} - \gamma|_{\mathcal{X}_2\mathcal{X}_3} \otimes \gamma|_{\mathcal{X}_1} - \gamma|_{\mathcal{X}_1\mathcal{X}_3} \otimes \gamma|_{\mathcal{X}_2} \\ &+ 2\gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2} \otimes \gamma|_{\mathcal{X}_3}. \end{aligned}$$

Interaction measures: $d = 3$, $(X_1, X_2, X_3) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$

$(X_1, X_2, X_3) \sim \gamma$, Lancaster interaction measure [Lancaster, 1969]

$$L(\gamma) := \gamma - \gamma|_{\mathcal{X}_1\mathcal{X}_2} \otimes \gamma|_{\mathcal{X}_3} - \gamma|_{\mathcal{X}_2\mathcal{X}_3} \otimes \gamma|_{\mathcal{X}_1} - \gamma|_{\mathcal{X}_1\mathcal{X}_3} \otimes \gamma|_{\mathcal{X}_2} + 2\gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2} \otimes \gamma|_{\mathcal{X}_3}.$$

In case of some factorization ($\not\Leftarrow$):

$$(X_1, X_2) \perp\!\!\!\perp X_3 \vee (X_1, X_3) \perp\!\!\!\perp X_2 \vee (X_2, X_3) \perp\!\!\!\perp X_1 \Rightarrow L(\gamma) = 0.$$

Interaction measures: $d = 3$, $(X_1, X_2, X_3) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$

$(X_1, X_2, X_3) \sim \gamma$, Lancaster interaction measure [Lancaster, 1969]

$$L(\gamma) := \gamma - \gamma|_{\mathcal{X}_1\mathcal{X}_2} \otimes \gamma|_{\mathcal{X}_3} - \gamma|_{\mathcal{X}_2\mathcal{X}_3} \otimes \gamma|_{\mathcal{X}_1} - \gamma|_{\mathcal{X}_1\mathcal{X}_3} \otimes \gamma|_{\mathcal{X}_2} + 2\gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2} \otimes \gamma|_{\mathcal{X}_3}.$$

In case of some factorization ($\not\Leftarrow$):

$$(X_1, X_2) \perp\!\!\!\perp X_3 \vee (X_1, X_3) \perp\!\!\!\perp X_2 \vee (X_2, X_3) \perp\!\!\!\perp X_1 \Rightarrow L(\gamma) = 0.$$

Idea [Sejdinovic et al., 2013a]: $k = k_1 \otimes k_2 \otimes k_3$, $k_m : \mathcal{X}_m \otimes \mathcal{X}_m \to \mathbb{R}$

$\|\mu_k(L(\gamma))\|^2_{\mathcal{H}_k} \overset{?}{>} 0 \Rightarrow$ no factorization.

Interaction measures: $d \geq 2$, $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$

- Partition measure (partitioning): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

Interaction measures: $d \geq 2$, $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$

- Partition measure (partitioning): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

- Weights: $c_\pi = (-1)^{|\pi|-1}(|\pi|-1)!$.
- Streitberg interaction [Streitberg, 1990], $X \sim \gamma$:

$$S(\gamma) = \sum_{\pi \in P(d)} c_\pi \gamma_\pi$$

Interaction measures: $d \geq 2$, $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$

- Partition measure (partitioning): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

- Weights: $c_\pi = (-1)^{|\pi|-1}(|\pi|-1)!$.
- Streitberg interaction [Streitberg, 1990], $X \sim \gamma$:

$$S(\gamma) = \sum_{\pi \in P(d)} c_\pi \gamma_\pi \xrightarrow{\text{spec.: } d=2} S(\gamma) = \gamma - \gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2}.$$

Interaction measures: $d \geq 2$, $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$

- Partition measure (partitioning): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

- Weights: $c_\pi = (-1)^{|\pi|-1}(|\pi|-1)!$.
- Streitberg interaction [Streitberg, 1990], $X \sim \gamma$:

$$S(\gamma) = \sum_{\pi \in P(d)} c_\pi \gamma_\pi \xrightarrow{\text{spec.: } d=2} S(\gamma) = \gamma - \gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2}.$$

- One could kernelize it (analogously to Lancaster interaction, KLI):

$$\|\mu_k(S(\gamma))\|_{\mathcal{H}_k}^2, \; k = \otimes_{j=1}^d k_j.$$

We now return to the meaning of

$\otimes_{j=1}^d k_j$ and $\otimes_{j=1}^d \mu_{k_j}\left(\gamma|_{\mathcal{X}_j}\right)$ in HSIC.

Tensor product: $\bigotimes_{j=1}^{d} a_j$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{a}\mathbf{b}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

$\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.

Tensor product: $\otimes_{j=1}^{d} a_j$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{a}\mathbf{b}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.
- For $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$ Hilbert spaces, i.e. for $d = 2$:

$$(a \otimes b)(v, w) := \langle a, v \rangle_{\mathcal{H}_1} \langle b, w \rangle_{\mathcal{H}_2}.$$

Tensor product: $\otimes_{j=1}^{d} a_j$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{ab}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{ab}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.
- For $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$ Hilbert spaces, i.e. for $d = 2$:

$$(a \otimes b)(v, w) := \langle a, v \rangle_{\mathcal{H}_1} \langle b, w \rangle_{\mathcal{H}_2}.$$

- For $d \geq 2$ and $a_j \in \mathcal{H}_j$,

$$\left(\otimes_{j=1}^{d} a_j\right)(b_1, \ldots, b_d) := \prod_{j=1}^{d} \langle a_j, b_j \rangle_{\mathcal{H}_j}.$$

Tensor product: $\otimes_{j=1}^{d}\mathcal{H}_j$

$$\otimes_{j=1}^{d}\mathcal{H}_j := \overline{\mathrm{Span}}(\otimes_{j=1}^{d} a_j \,:\, a_j \in \mathcal{H}_j).$$

Tensor product: $\otimes_{j=1}^d \mathcal{H}_j$

$$\otimes_{j=1}^d \mathcal{H}_j := \overline{\operatorname{Span}}(\otimes_{j=1}^d a_j \,:\, a_j \in \mathcal{H}_j).$$

$\xrightarrow{\text{spec.}}$ The tensor product of RKHSs is an RKHS

[Berlinet and Thomas-Agnan, 2004]

$$\mathcal{H}_k = \otimes_{j=1}^d \mathcal{H}_{k_j},$$

$$k(x,x') := (\otimes_{j=1}^d k_j)(x,x') := \prod_{j=1}^d \underbrace{k_j(x_j,x_j')}_{\text{coordinate-wise similarity}}.$$

Validness of MMD & HSIC, their estimation

Validness:

- $\mathrm{MMD}_k(\gamma, \eta) = 0 \Leftrightarrow \gamma = \eta$: k is characteristic (description on $\mathbb{R}^d$) [Fukumizu et al., 2008, Sriperumbudur et al., 2010].

Validness of MMD & HSIC, their estimation

Validness:

- $\mathrm{MMD}_k(\gamma,\eta)=0 \Leftrightarrow \gamma=\eta$: k is characteristic (description on $\mathbb{R}^d$) [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- $\mathrm{HSIC}_k(\gamma)=0 \Leftrightarrow \gamma=\otimes_{j=1}^{d}\gamma|_{\mathcal{X}_j} \xLeftarrow{\text{[Szabó and Sriperumbudur, 2018]}}$ k_j-s are universal [Steinwart, 2001, Micchelli et al., 2006].

Validness of MMD & HSIC, their estimation

Validness:

- $\mathrm{MMD}_k(\gamma, \eta) = 0 \Leftrightarrow \gamma = \eta$: k is characteristic (description on $\mathbb{R}^d$) [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- $\mathrm{HSIC}_k(\gamma) = 0 \Leftrightarrow \gamma = \otimes_{j=1}^{d} \gamma|_{\mathcal{X}_j} \xLeftarrow{\text{[Szabó and Sriperumbudur, 2018]}}$ k_j-s are universal [Steinwart, 2001, Micchelli et al., 2006].

Properties:

1. Injectivity of μ_k on probability / finite signed measures, so

universal $\Rightarrow$ characteristic.

Validness of MMD & HSIC, their estimation

Validness:

- $\text{MMD}_k(\gamma, \eta) = 0 \Leftrightarrow \gamma = \eta$: k is characteristic (description on $\mathbb{R}^d$) [Fukumizu et al., 2008, Sriperumbudur et al., 2010].
- $\text{HSIC}_k(\gamma) = 0 \Leftrightarrow \gamma = \otimes_{j=1}^{d} \gamma|_{\mathcal{X}_j} \xLeftarrow{\text{[Szabó and Sriperumbudur, 2018]}}$ k_j-s are universal [Steinwart, 2001, Micchelli et al., 2006].

Properties:

1. Injectivity of μ_k on probability / finite signed measures, so

   universal $\Rightarrow$ characteristic.

2. Easy-to-estimate: expected kernel trick

$$\langle \mu_k(\gamma), \mu_k(\eta) \rangle_{\mathcal{H}_k} = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \mathrm{d}\gamma(x) \mathrm{d}\eta(y).$$

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, Schrab et al., 2022, Hagrass et al., 2022, Zhang et al., 2022], and its differential private variant [Raj et al., 2019]; independence [Gretton et al., 2008, Pfister et al., 2018, Jitkrittum et al., 2017a, Albert et al., 2022] and goodness-of-fit testing [Jitkrittum et al., 2017b, Balasubramanian et al., 2021, Baum et al., 2022], causal discovery [Mooij et al., 2016, Pfister et al., 2018, Chakraborty and Zhang, 2019, Schölkopf et al., 2021], learning fair representations [Deka and Sutherland, 2023],
- feature selection [Camps-Valls et al., 2010, Song et al., 2012, Wang et al., 2022] $\xrightarrow{\text{app.}}$ biomarker detection [Climente-González et al., 2019], wind power prediction [Bouche et al., 2023], clustering [Song et al., 2007, Climente-González et al., 2019],
- domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2021], change-point detection [Harchaoui and Cappé, 2007, Kalinke et al., 2023], post selection inference [Yamada et al., 2018], sensitivity analysis [da Veiga, 2021, Fellmann et al., 2023],
- 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], Bayesian optimization [Buathong et al., 2020],
- topological data analysis [Kusano et al., 2016], learning on distributions [Sutherland, 2016]: classification [Muandet et al., 2011, Lopez-Paz et al., 2015, Zaheer et al., 2017], regression [Szabó et al., 2016, Sutherland et al., 2016, Law et al., 2018, Fang et al., 2020, Mücke, 2021],
- generative adversarial networks [Dziugaite et al., 2015, Li et al., 2015, Sutherland et al., 2017, Binkowski et al., 2018], understanding the dynamics of complex dynamical systems [Klus et al., 2019, Klus et al., 2020],
- functional data analysis [Wynne and Nagy, 2021, Wynne and Duncan, 2022], ...

Kernelized moments – towards kernelized cumulants

- From now:
  - $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$, $X \sim \gamma$,
  - kernels $k_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$, $j \in [d]$,
  - lifting $\Phi(X) = (\Phi_j(X_j))_{j=1}^d$ with $\Phi_j(x_j) := k_j(\cdot, x_j)$,
  - RKHS $\mathcal{H}^{\otimes \mathbf{i}} := \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}$ with kernel $k^{\otimes \mathbf{i}} := k_1^{\otimes i_1} \otimes \cdots \otimes k_d^{\otimes i_d}$, and feature

$$\Phi^{\otimes \mathbf{i}}(X) := [\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}.$$

Kernelized moments – towards kernelized cumulants

- From now:
  - $X = (X_j)_{j=1}^d \in \times_{j=1}^d \mathcal{X}_j$, $X \sim \gamma$,
  - kernels $k_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$, $j \in [d]$,
  - lifting $\Phi(X) = (\Phi_j(X_j))_{j=1}^d$ with $\Phi_j(x_j) := k_j(\cdot, x_j)$,
  - RKHS $\mathcal{H}^{\otimes \mathbf{i}} := \mathcal{H}_{k_1}^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_{k_d}^{\otimes i_d}$ with kernel $k^{\otimes \mathbf{i}} := k_1^{\otimes i_1} \otimes \cdots \otimes k_d^{\otimes i_d}$, and feature

$$\Phi^{\otimes \mathbf{i}}(X) := [\Phi_1(X_1)]^{\otimes i_1} \otimes \cdots \otimes [\Phi_d(X_d)]^{\otimes i_d}.$$

- Moment sequence:

$$\mu(\gamma) = \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d}, \qquad \mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[\Phi^{\otimes \mathbf{i}}(X)\right] \in \mathcal{H}^{\otimes \mathbf{i}}.$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X \sim \gamma$,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}\left[\Phi(X)\right]$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X \sim \gamma$,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}\left[\Phi(X)\right],$$

$$\kappa_k^{(2)}(\gamma) = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right] - \mathbb{E}\left[\Phi(X)\right] \otimes \mathbb{E}\left[\Phi(X)\right]$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$
\begin{aligned}
\kappa_k^{(1)}(\gamma) &= \mathbb{E}\left[\Phi(X)\right], \\
\kappa_k^{(2)}(\gamma) &= \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right] - \mathbb{E}\left[\Phi(X)\right] \otimes \mathbb{E}\left[\Phi(X)\right], \\
\kappa_k^{(3)}(\gamma) &= \mathbb{E}\left[\Phi^{\otimes 3}(X)\ \right] - \mathbb{E}\left[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')\right] \\
&\quad - \mathbb{E}\left[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)\right] - \mathbb{E}\left[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)\right] \\
&\quad + 2\mathbb{E}^{\otimes 3}\left[\Phi(X)\right].
\end{aligned}
$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$\begin{aligned}
\kappa_k^{(1)}(\gamma) &= \mathbb{E}\big[\Phi(X)\big], \\
\kappa_k^{(2)}(\gamma) &= \mathbb{E}\big[\Phi(X) \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X)\big] \otimes \mathbb{E}\big[\Phi(X)\big], \\
\kappa_k^{(3)}(\gamma) &= \mathbb{E}\Big[\Phi^{\otimes 3}(X)\ \Big] - \mathbb{E}\big[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')\big] \\
&\quad - \mathbb{E}\big[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)\big] \\
&\quad + 2\mathbb{E}^{\otimes 3}\big[\Phi(X)\big].
\end{aligned}$$

- $d = 2$, $m = 2$: $(X_1, X_2) \sim \gamma$,

$$\kappa_{k_1,k_2}^{(2,0)}(\gamma) = \mathbb{E}\left[\Phi_1^{\otimes 2}(X_1)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_1(X_1)\right],$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$\kappa_k^{(1)}(\gamma) = \mathbb{E}\big[\Phi(X)\big],$$
$$\kappa_k^{(2)}(\gamma) = \mathbb{E}\big[\Phi(X) \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X)\big] \otimes \mathbb{E}\big[\Phi(X)\big],$$
$$\begin{aligned}\kappa_k^{(3)}(\gamma) = {} & \mathbb{E}\Big[\Phi^{\otimes 3}(X)\ \Big] - \mathbb{E}\big[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')\big] \\ & - \mathbb{E}\big[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)\big] - \mathbb{E}\big[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)\big] \\ & + 2\mathbb{E}^{\otimes 3}\big[\Phi(X)\big].\end{aligned}$$

- $d = 2$, $m = 2$: $(X_1, X_2) \sim \gamma$,

$$\kappa_{k_1,k_2}^{(2,0)}(\gamma) = \mathbb{E}\left[\Phi_1^{\otimes 2}(X_1)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_1(X_1)\right],$$
$$\kappa_{k_1,k_2}^{(1,1)}(\gamma) = \mathbb{E}\left[\Phi_1(X_1) \otimes \Phi_2(X_2)\right] - \mathbb{E}\left[\Phi_1(X_1)\right] \otimes \mathbb{E}\left[\Phi_2(X_2)\right]$$

Kernelized cumulants: examples first, analogous to $\mathbb{R}$

- $d = 1$, $m \in [3]$: $X, X' \sim \gamma$, independent,

$$
\begin{aligned}
\kappa_k^{(1)}(\gamma) &= \mathbb{E}[\Phi(X)],\\
\kappa_k^{(2)}(\gamma) &= \mathbb{E}[\Phi(X) \otimes \Phi(X)] - \mathbb{E}[\Phi(X)] \otimes \mathbb{E}[\Phi(X)],\\
\kappa_k^{(3)}(\gamma) &= \mathbb{E}\left[\Phi^{\otimes 3}(X)\ \right] - \mathbb{E}[\Phi(X) \otimes \Phi(X) \otimes \Phi(X')]\\
&\quad - \mathbb{E}[\Phi(X) \otimes \Phi(X') \otimes \Phi(X)] - \mathbb{E}[\Phi(X') \otimes \Phi(X) \otimes \Phi(X)]\\
&\quad + 2\mathbb{E}^{\otimes 3}[\Phi(X)].
\end{aligned}
$$

- $d = 2$, $m = 2$: $(X_1, X_2) \sim \gamma$,

$$
\begin{aligned}
\kappa_{k_1,k_2}^{(2,0)}(\gamma) &= \mathbb{E}\left[\Phi_1^{\otimes 2}(X_1)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_1(X_1)\right],\\
\kappa_{k_1,k_2}^{(1,1)}(\gamma) &= \mathbb{E}\left[\Phi_1(X_1) \otimes \Phi_2(X_2)\right] - \mathbb{E}\left[\Phi_1(X_1)\right] \otimes \mathbb{E}\left[\Phi_2(X_2)\right],\\
\kappa_{k_1,k_2}^{(0,2)}(\gamma) &= \mathbb{E}\left[\Phi_2^{\otimes 2}(X_2)\right] - \mathbb{E}^{\otimes 2}\left[\Phi_2(X_2)\right].
\end{aligned}
$$

Wanted: repetition and partitioning. Weights: as before (c_π).

Kernelized cumulants: $X \sim \gamma$ prob. measure on $\times_{j=1}^{d} \mathcal{X}_j$

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1, \ldots, X_1}_{i_1 \text{ times}}, \underbrace{X_2, \ldots, X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d, \ldots, X_d}_{i_d \text{ times}}).$$

Kernelized cumulants: $X \sim \gamma$ prob. measure on $\times_{j=1}^{d} \mathcal{X}_j$

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1, \ldots, X_1}_{i_1 \text{ times}}, \underbrace{X_2, \ldots, X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d, \ldots, X_d}_{i_d \text{ times}}).$$

- Partitioning (partition measure, as for KLI): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

Kernelized cumulants: $X \sim \gamma$ prob. measure on $\times_{j=1}^d \mathcal{X}_j$

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1,\ldots,X_1}_{i_1 \text{ times}}, \underbrace{X_2,\ldots,X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d,\ldots,X_d}_{i_d \text{ times}}).$$

- Partitioning (partition measure, as for KLI): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

- Kernelized cumulants: $m = \deg(\mathbf{i}) := \sum_{j=1}^d i_j \xRightarrow{\text{OK}} \gamma_\pi^{\mathbf{i}} = \left(\gamma^{\mathbf{i}}\right)_\pi$,

$$\kappa_{k_1,\ldots,k_d}(\gamma) := \left(\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\right)_{\mathbf{i}\in\mathbb{N}^d},$$
$$\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) := \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} k^{\otimes \mathbf{i}}(\cdot,(X_1,\ldots,X_m)).$$

Kernelized cumulants: $X \sim \gamma$ prob. measure on $\times_{j=1}^d \mathcal{X}_j$

- Repetition (diagonal measure): $\mathbf{i} \in \mathbb{N}^d$,

$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1, \ldots, X_1}_{i_1 \text{ times}}, \underbrace{X_2, \ldots, X_2}_{i_2 \text{ times}}, \ldots, \underbrace{X_d, \ldots, X_d}_{i_d \text{ times}}).$$

- Partitioning (partition measure, as for KLI): $\pi \in P(d)$, $b = |\pi|$,

$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}} \otimes \cdots \otimes \gamma|_{\mathcal{X}_{\pi_b}}.$$

- Kernelized cumulants: $m = \deg(\mathbf{i}) := \sum_{j=1}^d i_j \overset{\text{OK}}{\Longrightarrow} \gamma_\pi^{\mathbf{i}} = \left(\gamma^{\mathbf{i}}\right)_\pi$,

$$\kappa_{k_1,\ldots,k_d}(\gamma) := \left(\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\right)_{\mathbf{i}\in\mathbb{N}^d},$$
$$\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) := \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} k^{\otimes \mathbf{i}}(\cdot, (X_1, \ldots, X_m)).$$

$\Rightarrow$ expected kernel trick is applicable

Cumulants characterize distributions

Point-separating k := injectivity of $\Phi \Leftarrow$ characteristic $k \Leftarrow$ universal k.

Cumulants characterize distributions

Point-separating k := injectivity of $\Phi \Leftarrow$ characteristic $k \Leftarrow$ universal k.

Theorem

- Assume:
  - γ, η: probability measures on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).

Cumulants characterize distributions

Point-separating k := injectivity of $\Phi \Leftarrow$ characteristic $k \Leftarrow$ universal k.

Theorem

- Assume:
  - γ, η: probability measures on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \eta \Leftrightarrow \kappa_{k_1,\ldots,k_d}(\gamma) = \kappa_{k_1,\ldots,k_d}(\eta)$

Cumulants characterize distributions

Point-separating k $:=$ injectivity of $\Phi \Leftarrow$ characteristic $k \Leftarrow$ universal k.

Theorem

- Assume:
  - γ, η: probability measures on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \eta \Leftrightarrow \kappa_{k_1,\dots,k_d}(\gamma) = \kappa_{k_1,\dots,k_d}(\eta)$, and

$$
\begin{aligned}
d^{\mathbf{i}}(\gamma,\eta) &:= \|\kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\gamma) - \kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\eta)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} \\
&= \sum_{\pi,\tau\in P(m)} c_\pi c_\tau \Big[\mathbb{E}_{\gamma^{\mathbf{i}}_\pi\otimes\gamma^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m)) \\
&\qquad + \mathbb{E}_{\eta^{\mathbf{i}}_\pi\otimes\eta^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m)) \\
&\qquad - 2\mathbb{E}_{\gamma^{\mathbf{i}}_\pi\otimes\eta^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m))\Big].
\end{aligned}
$$

Cumulants characterize independence

Theorem

- Assume:
  - γ: probability measure on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \gamma|_{\mathcal{X}_1} \otimes \cdots \otimes \gamma|_{\mathcal{X}_d} \Leftrightarrow \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$ for every $\mathbf{i} \in \mathbb{N}_+^d$

Cumulants characterize independence

Theorem

- Assume:
  - γ: probability measure on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \gamma|_{\mathcal{X}_1} \otimes \cdots \otimes \gamma|_{\mathcal{X}_d} \Leftrightarrow \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$ for every $\mathbf{i} \in \mathbb{N}_+^d$, and

$$\|\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} = \sum_{\pi,\tau \in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \gamma^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}\left((X_j)_{j=1}^m, (Y_j)_{j=1}^m\right),$$

where $m = \deg(\mathbf{i})$.

Cumulants characterize independence

Theorem

- Assume:
  - γ: probability measure on $\times_{j=1}^{d}\mathcal{X}_j$,
  - $(\mathcal{X}_j)_{j=1}^{d}$ are Polish spaces,
  - k_j: bounded, continuous, point-separating kernel ($j \in [d]$).
- Then, $\gamma = \gamma|_{\mathcal{X}_1} \otimes \cdots \otimes \gamma|_{\mathcal{X}_d} \Leftrightarrow \kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma) = 0$ for every $\mathbf{i} \in \mathbb{N}_+^d$, and

$$\|\kappa^{\mathbf{i}}_{k_1,\ldots,k_d}(\gamma)\|^2_{\mathcal{H}^{\otimes \mathbf{i}}} = \sum_{\pi,\tau \in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma^{\mathbf{i}}_\pi \otimes \gamma^{\mathbf{i}}_\tau} k^{\otimes \mathbf{i}}\left((X_j)_{j=1}^m, (Y_j)_{j=1}^m\right),$$

where $m = \deg(\mathbf{i})$.

Estimation in both cases

$\mathbb{E}k^{\otimes \mathbf{i}}((X_1,\ldots,X_m),(Y_1,\ldots,Y_m)) \Rightarrow$ V-statistics ✓

Distance between kernel variance embeddings

- By our theorem: if $\gamma = \eta$, then $d^{(2)}(\gamma, \eta) = 0$.
- V-statistic estimator of $d^{(2)}(\gamma, \eta)$:

$$\frac{1}{N^2}\mathrm{Tr}\Big[(\mathbf{K}_x \mathbf{J}_N)^2\Big] + \frac{1}{M^2}\mathrm{Tr}\Big[(\mathbf{K}_y \mathbf{J}_M)^2\Big] - \frac{2}{NM}\mathrm{Tr}\Big[\mathbf{K}_{xy}\mathbf{J}_M \mathbf{K}_{xy}^\top \mathbf{J}_N\Big],$$

with $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \gamma$, $(y_m)_{m=1}^M \overset{\text{i.i.d.}}{\sim} \eta$, $\mathbf{K}_x = [k(x_i, x_j)]_{i,j=1}^N$, $\mathbf{K}_y = [k(y_i, y_j)]_{i,j=1}^M$, $\mathbf{K}_{x,y} = [k(x_i, y_j)]_{i,j=1}^{N,M}$, $\mathbf{J}_n = \mathbf{I}_n - \mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^\top$.

Distance between kernel variance/skewness embeddings

- By our theorem: if $\gamma = \eta$, then $d^{(2)}(\gamma, \eta) = 0$.
- V-statistic estimator of $d^{(2)}(\gamma, \eta)$:

$$\frac{1}{N^2}\mathrm{Tr}\left[(\mathbf{K}_x\mathbf{J}_N)^2\right] + \frac{1}{M^2}\mathrm{Tr}\left[(\mathbf{K}_y\mathbf{J}_M)^2\right] - \frac{2}{NM}\mathrm{Tr}\left[\mathbf{K}_{xy}\mathbf{J}_M\mathbf{K}_{xy}^\top\mathbf{J}_N\right],$$

with $(x_n)_{n=1}^N \stackrel{\text{i.i.d.}}{\sim} \gamma$, $(y_m)_{m=1}^M \stackrel{\text{i.i.d.}}{\sim} \eta$, $\mathbf{K}_x = [k(x_i, x_j)]_{i,j=1}^N$, $\mathbf{K}_y = [k(y_i, y_j)]_{i,j=1}^M$, $\mathbf{K}_{x,y} = [k(x_i, y_j)]_{i,j=1}^{N,M}$, $\mathbf{J}_n = \mathbf{I}_n - \mathbf{H}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$.

Time complexity

Quadratic as MMD.

- $d^{(3)}(\gamma, \eta)$: similarly; quadratic time.

Cross-skewness independence criterion (CSIC)

- By our theorem: if $\gamma = \gamma|_{\mathcal{X}_1} \otimes \gamma|_{\mathcal{X}_2}$, then $\kappa^{(2,1)}_{k,\ell}(\gamma) = 0$ and $\kappa^{(1,2)}_{k,\ell}(\gamma) = 0$.
- V-statistic estimator of $\|\kappa^{(1,2)}_{k,\ell}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1} \otimes \mathcal{H}_\ell^{\otimes 2}}$:

$$\frac{1}{N^2}\Bigg\langle \mathbf{K} \circ \mathbf{K} \circ \mathbf{L} - 4\mathbf{K} \circ \mathbf{KH} \circ \mathbf{L} - 2\mathbf{K} \circ \mathbf{K} \circ \mathbf{LH} + 4\mathbf{KH} \circ \mathbf{K} \circ \mathbf{LH}$$
$$+2\mathbf{K} \circ \mathbf{L}\left\langle \frac{\mathbf{K}}{N^2} \right\rangle + 2\mathbf{KH} \circ \mathbf{HK} \circ \mathbf{L} + 4\mathbf{K} \circ \mathbf{HK} \circ \mathbf{LH} + \mathbf{K} \circ \mathbf{K}\left\langle \frac{\mathbf{L}}{N^2} \right\rangle$$
$$-8\mathbf{K} \circ \mathbf{LH}\left\langle \frac{\mathbf{K}}{N^2} \right\rangle - 4\mathbf{K} \circ \mathbf{HK}\left\langle \frac{\mathbf{L}}{N^2} \right\rangle + 4\left\langle \frac{\mathbf{K}}{N^2} \right\rangle^2 \mathbf{L}\Bigg\rangle,$$

  with kernels $k : \mathcal{X}_1^2 \to \mathbb{R}$, $\ell : \mathcal{X}_2^2 \to \mathbb{R}$, $\mathbf{K} := \mathbf{K}_x, \mathbf{L} := \mathbf{L}_y$, $\langle \mathbf{A} \rangle := \sum_{i,j} A_{i,j}$.
- Time complexity: quadratic.

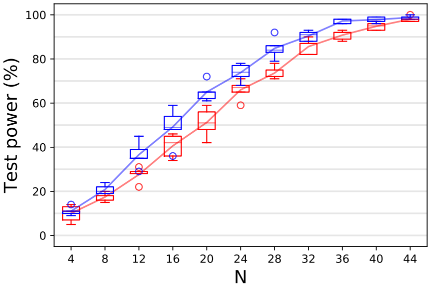
Numerical illustrations ($\alpha = 0.05$): improved power

1. Seoul bicycle rental data [E et al., 2020]:
   - features: temperature, humidity, wind speed, visibility, rainfall, snowfall, ...

Numerical illustrations ($\alpha = 0.05$): improved power

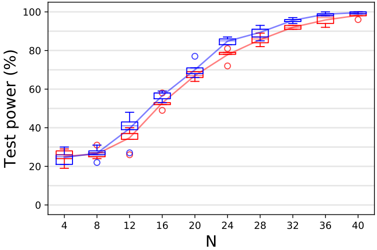
1. Seoul bicycle rental data [E et al., 2020]:
   - features: temperature, humidity, wind speed, visibility, rainfall, snowfall, ...

   - two-sample test (MMD, $d^{(2)}$): winter vs fall, $d = 11$,



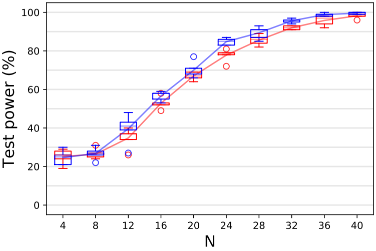
Numerical illustrations ($\alpha = 0.05$): improved power

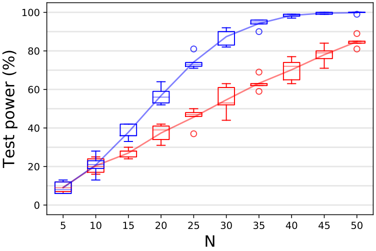
2. Brazilian traffic data [Ferreira, 2016]:
   - independence test (HSIC, CSIC); (blockage, fire, flood, ...) vs slowness of traffic; $d_1 = 16$, $d_2 = 1$; l.h.s.



Numerical illustrations ($\alpha = 0.05$): improved power

2. Brazilian traffic data [Ferreira, 2016]:
   - independence test (HSIC, CSIC); (blockage, fire, flood, ...) vs slowness of traffic; $d_1 = 16$, $d_2 = 1$; l.h.s.,
   - two-sample test (MMD, $d^{(3)}$): slow vs fast moving traffic, $d = 16$; r.h.s.





Summary

- We developed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).

Summary

- We developed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).
- Unified umbrella (ITE toolkit: `https://bitbucket.org/szzoli/ite/`):
  - MMD $\xleftarrow{\text{m}=\text{d}=1}$ k-cumulants $\xrightarrow{\mathbf{i}=\mathbf{1}_2}$ HSIC ($d = 2$).
  - k-Lancaster interaction $\xleftarrow{d=3}$ k-Streitberg interaction $\xleftarrow{\mathbf{i}=\mathbf{1}_d}$ k-cumulants.

Summary

- We developed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).
- Unified umbrella (ITE toolkit: `https://bitbucket.org/szzoli/ite/`):
  - MMD $\xleftarrow{\mathrm{m}=\mathrm{d}=1}$ k-cumulants $\xrightarrow{\mathbf{i}=\mathbf{1}_2}$ HSIC ($d = 2$).
  - k-Lancaster interaction $\xleftarrow{d=3}$ k-Streitberg interaction $\xleftarrow{\mathbf{i}=\mathbf{1}_d}$ k-cumulants.
- Relaxed kernel assumptions: point-separating.
- Higher-order cumulants: potential to improve power.
- Paper (presented at NeurIPS-2023), code.

Summary

- We developed a kernelized extension of cumulants,
- leveraging a combinatorial route (and tensor algebras).
- Unified umbrella (ITE toolkit: `https://bitbucket.org/szzoli/ite/`):
  - MMD $\xleftarrow{\mathrm{m=d=1}}$ k-cumulants $\xrightarrow{\mathbf{i}=\mathbf{1}_2}$ HSIC ($d = 2$).
  - k-Lancaster interaction $\xleftarrow{d=3}$ k-Streitberg interaction $\xleftarrow{\mathbf{i}=\mathbf{1}_d}$ k-cumulants.
- Relaxed kernel assumptions: point-separating.
- Higher-order cumulants: potential to improve power.
- Paper (presented at NeurIPS-2023), code.

Supplement

- Bell numbers.
- Characteristic kernels.
- Universal kernels:
  - equivalent definitions, Hahn-Banach theorem.
  - properties, examples.
- Moments and cumulants on $\mathbb{R}^d$.
- Estimator for $d^{(3)}(\gamma, \eta)$.
- Bochner integral.
- Mean embedding: expected kernel trick.

Bell numbers

- $B(m) :=$ number of elements in $P(m)$.
- $B_0 = B_1 = 1,\ B_2 = 2,\ B_3 = 5,\ B_4 = 15,\ B_5 = 52,\ B_6 = 203,$
  $B_7 = 877,\ B_8 = 4140,\ \ldots$

Bell numbers

- $B(m) :=$ number of elements in $P(m)$.
- $B_0 = B_1 = 1,\ B_2 = 2,\ B_3 = 5,\ B_4 = 15,\ B_5 = 52,\ B_6 = 203,$ $B_7 = 877,\ B_8 = 4140,\ \ldots$
- Recursion:

$$B_{m+1} = |P(m+1)| = \sum_{k=0}^{m} \binom{m}{k} B_k.$$

Bell numbers – continued

- Easy computation by the Bell triangle (like Pascal triangle for $\binom{n}{k}$)

$$
\begin{array}{lllll}
1 & & & & \\
1 & 2 & & & \\
2 & 3 & 5 & & \\
5 & 7 & 10 & 15 & \\
15 & 20 & 27 & 37 & 52 \\
52 & \dots & & &
\end{array}
$$

Bell numbers – continued

- Easy computation by the Bell triangle (like Pascal triangle for $\binom{n}{k}$)

$$\begin{array}{ccccc}
1 & & & & \\
1 & 2 & & & \\
2 & 3 & 5 & & \\
5 & 7 & 10 & 15 & \\
15 & 20 & 27 & 37 & 52 \\
52 & \dots & & &
\end{array}$$

- Asymptotics [de Bruijn, 1981, Lovász, 1993]:

$$\frac{\ln B_m}{m} = \ln m - \ln\ln m - 1 + \frac{\ln\ln m}{\ln m} + \frac{1}{\ln m} + \frac{1}{2}\left(\frac{\ln\ln m}{\ln m}\right)^2 + \mathcal{O}\left(\frac{\ln\ln m}{\ln^2 m}\right)$$

as $m \to \infty$.

Contents

Description of characteristic kernels 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.

Description of characteristic kernels on $\mathbb{R}^d$

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

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

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

$(*)$: Bochner's theorem, c_γ: characteristic function of γ.

Description of characteristic kernels on $\mathbb{R}^d$

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

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

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

$(*)$: Bochner's theorem, c_γ: characteristic function of γ.

Theorem ([Sriperumbudur et al., 2010])

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

Examples on $\mathbb{R}$; similarly $\mathbb{R}^d$ [Sriperumbudur et al., 2010]

For Poisson kernel: $\sigma \in (0,1)$.

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	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$	$\mathbb{R}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

Examples on $\mathbb{R}$; similarly $\mathbb{R}^d$ [Sriperumbudur et al., 2010]

For Poisson kernel: $\sigma \in (0,1)$.

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	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\|j\|}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\|j\|}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\pm2,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

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

Contents, MMD validness

Universal kernel

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

Definition

Assume:

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

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

Universal kernel

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

Definition

Assume:

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

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

$\mathcal{X}$ assumption $\Rightarrow$

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

Denseness in $C(\mathcal{X}) \Leftrightarrow$ injectivity of μ_k on $\mathcal{M}_b(\mathcal{X})$

- k universal means that $\mathcal{H}_k$ is dense in $C(\mathcal{X})$.

Denseness in $C(\mathcal{X}) \Leftrightarrow$ injectivity of μ_k on $\mathcal{M}_b(\mathcal{X})$

- k universal means that $\mathcal{H}_k$ is dense in $C(\mathcal{X})$.
- Hahn-Banach theorem [Rudin, 1991]: Let H be a subspace of a normed space C. H is dense in C iff.

$$\{0\} = H^{\perp} := \{F \in C' : \forall f \in H, F(f) = 0\}.$$

Denseness in $C(\mathcal{X}) \Leftrightarrow$ injectivity of μ_k on $\mathcal{M}_b(\mathcal{X})$

- k universal means that $\mathcal{H}_k$ is dense in $C(\mathcal{X})$.
- Hahn-Banach theorem [Rudin, 1991]: Let H be a subspace of a normed space C. H is dense in C iff.

$$\{0\} = H^{\perp} := \{F \in C' : \forall f \in H, F(f) = 0\}.$$

- Denseness $\Leftrightarrow$

$$\{0\} = \mathcal{H}_k^{\perp} = \Big\{\mathbb{F} \in \underbrace{C(\mathcal{X})'}_{=\mathcal{M}_b(\mathcal{X})} : \forall f \in \mathcal{H}_k, 0 = T_{\mathbb{F}}(f) = \underbrace{\int_{\mathcal{X}} f \,\mathrm{d}\mathbb{F}}_{\langle f, \mu_k(\mathbb{F})\rangle_{\mathcal{H}_k}} \Big\}$$

Denseness in $C(\mathcal{X}) \Leftrightarrow$ injectivity of μ_k on $\mathcal{M}_b(\mathcal{X})$

- k universal means that $\mathcal{H}_k$ is dense in $C(\mathcal{X})$.
- Hahn-Banach theorem [Rudin, 1991]: Let H be a subspace of a normed space C. H is dense in C iff.

$$\{0\} = H^{\perp} := \{F \in C' : \forall f \in H, F(f) = 0\}.$$

- Denseness $\Leftrightarrow$

$$\{0\} = \mathcal{H}_k^{\perp} = \left\{\mathbb{F} \in \underbrace{C(\mathcal{X})'}_{=\mathcal{M}_b(\mathcal{X})} : \forall f \in \mathcal{H}_k, 0 = T_{\mathbb{F}}(f) = \underbrace{\int_{\mathcal{X}} f \mathrm{d}\mathbb{F}}_{\langle f, \mu_k(\mathbb{F})\rangle_{\mathcal{H}_k}} \right\}$$

$$= \{\mathbb{F} \in \mathcal{M}_b(\mathcal{X}) : \mu_k(\mathbb{F}) = 0\}.$$

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x, x) > 0$ for all $x \in \mathcal{X}$.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x, x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x, x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\Phi(x) = k(\cdot, x)$ is injective, i.e.

$$\rho_k(x, y) = \|\Phi(x) - \Phi(y)\|_{\mathcal{H}_k}$$

  is a metric.

Properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

- $k(x,x) > 0$ for all $x \in \mathcal{X}$.
- Every restriction of k to an $\mathcal{X}' \subseteq \mathcal{X}$ compact set is universal.
- $\Phi(x) = k(\cdot, x)$ is injective, i.e.

$$\rho_k(x,y) = \|\Phi(x) - \Phi(y)\|_{\mathcal{H}_k}$$

  is a metric.
- The normalized kernel (like corr)

$$\tilde{k}(x,y) := \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

  is universal.

Universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^\infty \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r), \; r \in (0, \infty].$$

Universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

- For an $C^\infty \ni f : (-r, r) \to \mathbb{R}$

$$f(t) = \sum_{n=0}^{\infty} a_n t^n \quad t \in (-r, r),\ r \in (0, \infty].$$

- If $a_n > 0\ \forall n$, then

$$k(\mathbf{x}, \mathbf{y}) = f(\langle \mathbf{x}, \mathbf{y} \rangle)$$

is universal on $\mathcal{X} := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \sqrt{r} \right\}$.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x}, \mathbf{y}) = e^{\alpha \langle \mathbf{x}, \mathbf{y} \rangle}$: previous result with $f(t) = e^{\alpha t} \Rightarrow a_n = \frac{\alpha^n}{n!}$.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x},\mathbf{y}) = e^{\alpha\langle\mathbf{x},\mathbf{y}\rangle}$: previous result with $f(t) = e^{\alpha t} \Rightarrow a_n = \frac{\alpha^n}{n!}$.
- $k(\mathbf{x},\mathbf{y}) = e^{-\alpha\|\mathbf{x}-\mathbf{y}\|_2^2}$: exp. kernel & normalization.

Universal kernels on compact subsets of $\mathbb{R}^d$, $\alpha > 0$

- $k(\mathbf{x},\mathbf{y}) = (1 - \langle \mathbf{x},\mathbf{y}\rangle)^{-\alpha}$ binomial kernel
  - on $\mathcal{X}$ compact $\subset \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 < 1\}$.
  - $f(t) = (1-t)^{-\alpha} = \sum_{n=0}^{\infty} \underbrace{\binom{-\alpha}{n}(-1)^n}_{>0} t^n \quad (|t| < 1)$,

where $\binom{b}{n} = \sum_{i=1}^{n} \frac{b-i+1}{i}$.

Contents

Moments and cumulants on $\mathbb{R}^d \ni X \sim \gamma$, $\mathbf{i} \in \mathbb{N}^d$

	$d = 1$	$d \geq 1$
moment sequence	$\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i \in \mathbb{N}}$	$\mu(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d}$
moments	$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}$	$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$

Moments and cumulants on $\mathbb{R}^d \ni X \sim \gamma$, $\mathbf{i} \in \mathbb{N}^d$

	$d = 1$	$d \geq 1$
moment sequence	$\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i \in \mathbb{N}}$	$\mu(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d}$
moments	$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}$	$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$
m-th moment	$\mu^{(m)}(\gamma)$	$\mu^m(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\deg(\mathbf{i})=m}$

where $\deg(\mathbf{i}) := i_1 + \cdots + i_d$, $\mu^0(\gamma) = 1$

Moments and cumulants on $\mathbb{R}^d \ni X \sim \gamma$, $\mathbf{i} \in \mathbb{N}^d$

	$d = 1$	$d \geq 1$
moment sequence	$\mu(\gamma) := \left(\mu^{(i)}(\gamma)\right)_{i \in \mathbb{N}}$	$\mu(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\mathbf{i} \in \mathbb{N}^d}$
moments	$\mu^{(i)}(\gamma) := \mathbb{E}\left(X^i\right) \in \mathbb{R}$	$\mu^{\mathbf{i}}(\gamma) := \mathbb{E}\left[X_1^{i_1} \cdots X_d^{i_d}\right] \in \mathbb{R}$
m-th moment	$\mu^{(m)}(\gamma)$	$\mu^m(\gamma) := \left(\mu^{\mathbf{i}}(\gamma)\right)_{\deg(\mathbf{i})=m}$

and cumulants $\kappa(\gamma) = (\kappa^{\mathbf{i}}(\gamma))_{\mathbf{i} \in \mathbb{N}^d}$

$$\sum_{\mathbf{i} \in \mathbb{N}^d} \kappa^{\mathbf{i}}(\gamma) \frac{\boldsymbol{\theta}^{\mathbf{i}}}{\mathbf{i}!} = \log \left(\sum_{\mathbf{i} \in \mathbb{N}^d} \mu^{\mathbf{i}}(\gamma) \frac{\boldsymbol{\theta}^{\mathbf{i}}}{\mathbf{i}!} \right), \quad \boldsymbol{\theta} \in \mathbb{R}^d,$$

where $\deg(\mathbf{i}) := i_1 + \cdots + i_d$, $\mu^0(\gamma) = 1$, $\mathbf{i}! = i_1! \cdots i_d!$, $\boldsymbol{\theta}^{\mathbf{i}} = \theta_1^{i_1} \cdots \theta_d^{i_d}$.

Contents, moments and cumulants on $\mathbb{R}$, motivation of cumulants

Estimator for $d^{(3)}(\gamma,\eta) = \|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$, $N = M$

$$d^{(3)}(\gamma,\eta) = \|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} + \|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}} - 2\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}}$$

Estimator for $d^{(3)}(\gamma,\eta) = \|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$, $N = M$

$$d^{(3)}(\gamma,\eta) = \|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} + \|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}} - 2\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}}$$

$$\begin{aligned}\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}} \approx \frac{1}{N^2}\Bigg\langle & \mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{K}_{xy} - 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \\ & - 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} + 6\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \circ \mathbf{H}\mathbf{K}_{xy} \\ & + 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \left\langle \frac{\mathbf{K}_{xy}}{N^2} \right\rangle + 2\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \\ & + 2\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \circ \mathbf{K}_{xy}\mathbf{H} - 6\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \left\langle \frac{\mathbf{K}_{xy}}{N^2} \right\rangle \\ & - 6\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \left\langle \frac{\mathbf{K}_{xy}}{N^2} \right\rangle + 4\left\langle \frac{\mathbf{K}}{N^2} \right\rangle^2 \mathbf{K}_{xy} \Bigg\rangle.\end{aligned}$$

Note: Matrix multiplication takes precedence over the Hadamard one.

Estimator for $d^{(3)}(\gamma,\eta)$ – continued

$$\|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} \approx \frac{1}{N^2}\Big\langle \mathbf{K}_x \circ \mathbf{K}_x \circ \mathbf{K}_x - 6\mathbf{K}_x \circ \mathbf{K}_x\mathbf{H} \circ \mathbf{K}_x$$
$$+ 4\mathbf{K}_x\mathbf{H} \circ \mathbf{K}_x \circ \mathbf{K}_x\mathbf{H} + 3\mathbf{K}_x \circ \mathbf{K}_x \left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle$$
$$+ 6\mathbf{K}_x\mathbf{H} \circ \mathbf{H}\mathbf{K}_x \circ \mathbf{K}_x - 12\mathbf{K}_x \circ \mathbf{H}\mathbf{K}_x \left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle$$
$$+ 4\left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle^2 \mathbf{K}_x \Big\rangle.$$

$\|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$: similarly (change $\mathbf{K}_x$ to $\mathbf{K}_y$).

Contents, $d^{(2)}(\gamma,\eta)$ estimation

Bochner integral [Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \gamma)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to \mathcal{H}$-valued function (note: Banach-valued f ✓).

Bochner integral [Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \gamma)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to \mathcal{H}$-valued function (note: Banach-valued f ✓).
- For $f = \sum_{i=1}^{n} c_i \chi_{A_i}$ $(A_i \in \mathcal{A}, c_i \in \mathcal{H})$ step functions

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

Bochner integral [Diestel and Uhl, 1977, Dinculeanu, 2000, Steinwart and Christmann, 2008]

- Given:
  - $(\mathcal{X}, \mathcal{A}, \gamma)$: σ-finite measure space,
  - $f : (\mathcal{X}, \mathcal{A}) \to \mathcal{H}$-valued function (note: Banach-valued f ✓).
- For $f = \sum_{i=1}^{n} c_i \chi_{A_i}$ $(A_i \in \mathcal{A}, c_i \in \mathcal{H})$ step functions

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

- f measurable function is Bochner γ-integrable if
  - $\exists$ $(f_n)_{n\in\mathbb{N}}$ step functions: $\lim_{n\to\infty} \int_{\mathcal{X}} \|f - f_n\|_{\mathcal{H}} \mathrm{d}\gamma = 0$.
  - In this case $\lim_{n\to\infty} \int_{\mathcal{X}} f_n \mathrm{d}\gamma$ exists, $=: \int_{\mathcal{X}} f \mathrm{d}\gamma$.

Bochner integral: properties

- $f : \mathcal{X} \to \mathcal{H}$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \, \mathrm{d}\gamma < \infty$.

Bochner integral: properties

- $f : \mathcal{X} \to \mathcal{H}$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \, \mathrm{d}\gamma < \infty$.
- In this case $\|\int_{\mathcal{X}} f \mathrm{d}\gamma\|_{\mathcal{H}} \leq \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \, \mathrm{d}\gamma$. ('Jensen inequality')

Bochner integral: properties

- $f: \mathcal{X} \to \mathcal{H}$ is Bochner integrable $\Leftrightarrow \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \mathrm{d}\gamma < \infty$.
- In this case $\|\int_{\mathcal{X}} f \mathrm{d}\gamma\|_{\mathcal{H}} \leq \int_{\mathcal{X}} \|f\|_{\mathcal{H}} \mathrm{d}\gamma$. ('Jensen inequality')
- In our context: $\mathcal{H} = \mathcal{H}_k$,

$$\mu_k(\gamma) \text{ exists iff. } \int_{\mathcal{X}} \underbrace{\|k(\cdot, x)\|_{\mathcal{H}_k}}_{\sqrt{k(x,x)}} \mathrm{d}\gamma(x) < \infty.$$

Specifically: for bounded kernel $(\sup_{x,x' \in \mathcal{X}} k(x, x') < \infty)$ ✓.

Bochner integral: properties – continued

- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

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

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

Bochner integral: properties – continued

- If
  - $S : B \to B_2$: bounded linear operator,
  - $f : X \to B$: Bochner integrable, then

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

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

In short

$\left|\int f \mathrm{d}\gamma\right| \leq \int |f| \mathrm{d}\gamma$ and $c \int f \mathrm{d}\gamma = \int c f \mathrm{d}\gamma$ generalize nicely.

Contents, mean embedding and friends

Mean embedding: expected kernel trick

$$\langle \mu_k(\gamma), \mu_k(\eta) \rangle_{\mathcal{H}_k} \overset{(a)}{=} \langle \mu_k(\gamma), \int_{\mathcal{X}} k(\cdot, y) \mathrm{d}\eta(y) \rangle_{\mathcal{H}_k}$$

(a): μ_k definition

Mean embedding: expected kernel trick

$$\begin{aligned}\langle \mu_k(\gamma), \mu_k(\eta)\rangle_{\mathcal{H}_k} &\overset{(a)}{=} \left\langle \mu_k(\gamma), \int_{\mathcal{X}} k(\cdot, y)\mathrm{d}\eta(y)\right\rangle_{\mathcal{H}_k} \\ &\overset{(b)}{=} \int_{\mathcal{X}} \langle \mu_k(\gamma), k(\cdot, y)\rangle_{\mathcal{H}_k}\mathrm{d}\eta(y)\end{aligned}$$

(a): μ_k definition, (b): $S\left(\int_{\mathcal{X}} f\mathrm{d}\mu\right) = \int_{\mathcal{X}} Sf\mathrm{d}\mu$, $S(z) = \langle \mu_k(\gamma), z\rangle_{\mathcal{H}_k}$ [$\Leftarrow$ Hille's theorem]

Mean embedding: expected kernel trick

$$\begin{aligned}
\langle \mu_k(\gamma), \mu_k(\eta) \rangle_{\mathcal{H}_k} &\overset{(a)}{=} \left\langle \mu_k(\gamma), \int_{\mathcal{X}} k(\cdot, y) \mathrm{d}\eta(y) \right\rangle_{\mathcal{H}_k} \\
&\overset{(b)}{=} \int_{\mathcal{X}} \langle \mu_k(\gamma), k(\cdot, y) \rangle_{\mathcal{H}_k} \mathrm{d}\eta(y) \\
&\overset{(c)}{=} \int_{\mathcal{X}} \left\langle \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\gamma(x), k(\cdot, y) \right\rangle_{\mathcal{H}_k} \mathrm{d}\eta(y)
\end{aligned}$$

(a): μ_k definition, (b): $S\left(\int_{\mathcal{X}} f \mathrm{d}\mu\right) = \int_{\mathcal{X}} Sf \mathrm{d}\mu$, $S(z) = \langle \mu_k(\gamma), z \rangle_{\mathcal{H}_k}$ [$\Leftarrow$ Hille's theorem], (c): μ_k definition

Mean embedding: expected kernel trick

$$
\begin{aligned}
\langle \mu_k(\gamma), \mu_k(\eta) \rangle_{\mathcal{H}_k} &\overset{(a)}{=} \left\langle \mu_k(\gamma), \int_{\mathcal{X}} k(\cdot, y) \mathrm{d}\eta(y) \right\rangle_{\mathcal{H}_k} \\
&\overset{(b)}{=} \int_{\mathcal{X}} \langle \mu_k(\gamma), k(\cdot, y) \rangle_{\mathcal{H}_k} \mathrm{d}\eta(y) \\
&\overset{(c)}{=} \int_{\mathcal{X}} \left\langle \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\gamma(x), k(\cdot, y) \right\rangle_{\mathcal{H}_k} \mathrm{d}\eta(y) \\
&\overset{(d)}{=} \int_{\mathcal{X}} \int_{\mathcal{X}} \langle k(\cdot, x) k(\cdot, y) \rangle_{\mathcal{H}_k} \mathrm{d}\gamma(x) \mathrm{d}\eta(y)
\end{aligned}
$$

(a): μ_k definition, (b): $S\left(\int_{\mathcal{X}} f \mathrm{d}\mu\right) = \int_{\mathcal{X}} S f \mathrm{d}\mu$, $S(z) = \langle \mu_k(\gamma), z \rangle_{\mathcal{H}_k}$ [$\Leftarrow$ Hille's theorem], (c): μ_k definition, (d): (b) with $S_y(z) = \langle z, k(\cdot, y) \rangle_{\mathcal{H}_k}$

Mean embedding: expected kernel trick

$$
\begin{aligned}
\langle \mu_k(\gamma), \mu_k(\eta) \rangle_{\mathcal{H}_k} &\overset{(a)}{=} \left\langle \mu_k(\gamma), \int_{\mathcal{X}} k(\cdot, y) \mathrm{d}\eta(y) \right\rangle_{\mathcal{H}_k} \\
&\overset{(b)}{=} \int_{\mathcal{X}} \langle \mu_k(\gamma), k(\cdot, y) \rangle_{\mathcal{H}_k} \mathrm{d}\eta(y) \\
&\overset{(c)}{=} \int_{\mathcal{X}} \left\langle \int_{\mathcal{X}} k(\cdot, x) \mathrm{d}\gamma(x), k(\cdot, y) \right\rangle_{\mathcal{H}_k} \mathrm{d}\eta(y) \\
&\overset{(d)}{=} \int_{\mathcal{X}} \int_{\mathcal{X}} \langle k(\cdot, x) k(\cdot, y) \rangle_{\mathcal{H}_k} \mathrm{d}\gamma(x) \mathrm{d}\eta(y) \\
&\overset{(e)}{=} \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) \mathrm{d}\gamma(x) \mathrm{d}\eta(y),
\end{aligned}
$$

(a): μ_k definition, (b): $S\left(\int_{\mathcal{X}} f \mathrm{d}\mu\right) = \int_{\mathcal{X}} S f \mathrm{d}\mu$, $S(z) = \langle \mu_k(\gamma), z \rangle_{\mathcal{H}_k}$ [$\Leftarrow$ Hille's theorem], (c): μ_k definition, (d): (b) with $S_y(z) = \langle z, k(\cdot, y) \rangle_{\mathcal{H}_k}$, (e): reproducing property.

Contents, MMD estimation

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022).
Adaptive test of independence based on HSIC measures.
The Annals of Statistics, 50(2):858–879.

Aronszajn, N. (1950).
Theory of reproducing kernels.
Transactions of the American Mathematical Society, 68:337–404.

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.

Balanca, P. and Herbin, E. (2012).
A set-indexed Ornstein-Uhlenbeck process.
Electronic Communications in Probability, 17:1–14.

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.

Baum, J., Kanagawa, H., and Gretton, A. (2022).
A kernel Stein test of goodness of fit for sequential models.
Technical report.
(https://arxiv.org/abs/2210.10741).

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

Billingsley, P. (2012).
Probability and Measure.
John Wiley and Sons.

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. (2021).
Domain generalization by marginal transfer learning.
Journal of Machine Learning Research, 22:1–55.

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.

Bouche, D., Flamary, R., d'Alché Buc, F., Plougonven, R., Clausel, M., Badosa, J., and Drobinski, P. (2023).
Wind power predictions from nowcasts to 4-hour forecasts: a learning approach with variable selection.
Renewable Energy.

Buathong, P., Ginsbourger, D., and Krityakierne, T. (2020).
Kernels over sets of finite sets using RKHS embeddings, with application to Bayesian (combinatorial) optimization.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2731–2741.

Camps-Valls, G., Mooij, J. M., and Schölkopf, B. (2010).
Remote sensing feature selection by kernel dependence measures.
IEEE Geoscience and Remote Sensing Letters, 7(3):587–591.

Chakraborty, S. and Zhang, X. (2019).
Distance metrics for measuring joint dependence with application to causal inference.
Journal of the American Statistical Association, 114(528):1638–1650.

Climente-González, H., Azencott, C.-A., Kaski, S., and Yamada, M. (2019).
Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data.
Bioinformatics, 35(14):i427–i435.

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.

da Veiga, S. (2021).
Kernel-based ANOVA decomposition and Shapley effects – application to global sensitivity analysis.
Technical report.
(https://arxiv.org/abs/2101.05487).

de Bruijn, N. G. (1981).
Asymptotic Methods in Analysis.
Dover.

Deka, N. and Sutherland, D. J. (2023).
MMD-B-Fair: Learning fair representations with statistical testing.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 9564–9576.

Diestel, J. and Uhl, J. J. (1977).
Vector Measures.
American Mathematical Society. Providence.

Dinculeanu, N. (2000).

Vector Integration and Stochastic Integration in Banach Spaces.
Wiley.

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.

E, S. V., Park, J., and Cho, Y. (2020).
Using data mining techniques for bike sharing demand prediction in metropolitan city.
Computer Communications, 153:353–366.

Fang, Z., Guo, Z.-C., and Zhou, D.-X. (2020).
Optimal learning rates for distribution regression.
Journal of Complexity, page 101426.

Fellmann, N., Blanchet-Scalliet, C., Helbert, C., Spagnol, A., and Sinoquet, D. (2023).

Kernel-based sensivity analysis for (excursion) sets.
Technical report.
(https://arxiv.org/abs/2305.09268).

Ferreira, R. P. (2016).
Combination of artificial intelligence techniques for prediction the behavior of urban vehicular traffic in the city of São Paulo.
In *Anais do 10. Congresso Brasileiro de Inteligência Computacional*, pages 1–5.

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., 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., 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. (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.

Hagrass, O., Sriperumbudur, B. K., and Li, B. (2022).
Spectral regularized kernel two-sample tests.
Technical report.
(https://arxiv.org/abs/2212.09201).

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.

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.

Kalinke, F., Heyden, M., Fouché, E., and Böhm, K. (2023).
Maximum mean discrepancy on exponential windows for online change detection.
Technical report.
(https://arxiv.org/abs/2205.12706).

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.

Klebanov, L. (2005).
N-Distances and Their Applications.
Charles University, Prague.

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. (2020).
Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces.
Journal of Nonlinear Science, 30:283—315.

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.

Lancaster, H. O. (1969).
The Chi-Squared Distribution.
Wiley, London.

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.

Lovász, L. (1993).
Combinatorial Problems and Exercise.
2nd ed. Amsterdam, Netherlands: North-Holland.

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

McCullagh, P. (2018).
Tensor Methods in Statistics.

Courier Dover Publications.

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

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.

Mücke, N. (2021).
Stochastic gradient descent meets distribution regression.

In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2143–2151.

Müller, A. (1997).
Integral probability metrics and their generating classes of functions.
Advances in Applied Probability, 29:429–443.

Nikolentzos, G. and Vazirgiannis, M. (2023).
Graph alignment kernels using Weisfeiler and Leman hierarchies.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2019–2034.

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.

Quadrianto, N., Song, L., and Smola, A. (2009).
Kernelized sorting.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296.

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.

Rudin, W. (1991).
Functional Analysis.
McGraw-Hill, USA.

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., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021).
Toward causal representation learning.
Proceedings of the IEEE, 109(5):612–634.

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.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022).
Efficient aggregated kernel tests using incomplete U-statistics.
In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18793–18807.

Schulz, T. H., Welke, P., and Wrobel, S. (2022).
Graph filtration kernels.
In *AAAI Conference on Artifical Intelligence (AAAI)*, pages 8196–8203.

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

Sejdinovic, D., Gretton, A., and Bergsma, W. (2013a).
A kernel test for three-variable interactions.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013b).
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.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012).
Feature selection via dependence maximization.
Journal of Machine Learning Research, 13(1):1393–1434.

Song, L., Smola, A. J., Gretton, A., and Borgwardt, K. M. (2007).
A dependence maximization view of clustering.
In *International Conference on Machine Learning (ICML)*, pages 815–822.

Speed, T. P. (1983).
Cumulants and partition lattices.
Australian Journal of Statistics, 25(2):378–388.

Speed, T. P. (1984).
Cumulants and partition lattices II.
Australian Journal of Statistics, 4(1):34–53.

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.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer.

Streitberg, B. (1990).
Lancaster interactions revisited.
Annals of Statistics, 18(4):1878–1885.

Sutherland, D. (2016).

Scalable, Flexible and Active Learning on Distributions.
PhD thesis.

Sutherland, D., Oliva, J., Póczos, B., and Schneider, J. (2016).
Linear-time learning on distributions with approximate kernel embeddings.
In *AAAI Conference on Artifical Intelligence (AAAI)*, pages 2073–2079.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017).
Generative models and model criticism via optimized maximum mean discrepancy.
In *International Conference on Learning Representations (ICLR)*.

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.

Székely, G. J. and Rizzo, M. L. (2009).
Brownian distance covariance.
The Annals of Applied Statistics, 3:1236–1265.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007).
Measuring and testing dependence by correlation of distances.
The Annals of Statistics, 35:2769–2794.

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.

Wang, A., Du, J., Zhang, X., and Shi, J. (2022).
Ranking features to promote diversity: An approach based on sparse distance correlation.
Technometrics, 64(3):384–395.

Watkins, C. (1999).
Dynamic alignment kernels.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.

Wynne, G. and Duncan, A. B. (2022).

A kernel two-sample test for functional data.
Journal of Machine Learning Research, 23:1–51.

Wynne, G. and Nagy, S. (2021).
Statistical depth meets machine learning: Kernel mean embeddings and depth in functional data analysis.
Technical report.
(https://arxiv.org/abs/2105.12778).

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.

Zhang, Q., Wild, V., Filippi, S., Flaxman, S., and Sejdinovic, D. (2022).
Bayesian kernel two-sample testing.
Journal of Computational and Graphical Statistics, 31(4):1164–1176.

Zinger, A., Kakosyan, A., and Klebanov, L. (1992).
A characterization of distributions by mean values of statistics and certain probabilistic metrics.
Journal of Soviet Mathematics.

Zolotarev, V. (1983).
Probability metrics.
Theory of Probability and its Applications, 28:278–302.