Measures of (In)dependence Using Positive Definite Kernels

Bharath K. Sriperumbudur

Department of Statistics, Pennsylvania State University

4th Conference of ISNPS
Salerno, Italy
June 13, 2018

(Joint work with Zoltán Szabó, CMAP, École Polytechnique)

Supported by National Science Foundation (DMS-1713011)

Outline

- Motivating example: Comparing distributions
- Hilbert space embedding of measures
  - Mean element
  - Distance on probabilities
  - Characteristic kernels
- Kernel measure of dependence
  - Cross-covariance operator
- Tensor kernels and joint independence

Motivating Example: Coin Toss

- Toss 1: $T\ H\ H\ H\ T\ T\ H\ T\ T\ H\ H\ T\ H$
- Toss 2: $H\ T\ T\ H\ T\ H\ T\ T\ H\ H\ H\ T\ T$

Are the coins/tosses statistically similar?

Toss 1 is a sample from $\mathbb{P}$:=Bernoulli(p) and Toss 2 is a sample from $\mathbb{Q}$:=Bernoulli(q).

Is $p = q$ or not?, i.e., compare

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\{0,1\}} x\, d\mathbb{P}(x) \quad \text{and} \quad \mathbb{E}_{\mathbb{Q}}[X] = \int_{\{0,1\}} x\, d\mathbb{Q}(x).$$

Motivating Example: Coin Toss

- Toss 1: $T\ H\ H\ H\ T\ T\ H\ T\ T\ H\ H\ T\ H$
- Toss 2: $H\ T\ T\ H\ T\ H\ T\ T\ H\ H\ H\ T\ T$

Are the coins/tosses statistically similar?

Toss 1 is a sample from $\mathbb{P}$:=Bernoulli(p) and Toss 2 is a sample from $\mathbb{Q}$:=Bernoulli(q).

Is $p = q$ or not?, i.e., compare

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\{0,1\}} x \, d\mathbb{P}(x) \qquad \text{and} \qquad \mathbb{E}_{\mathbb{Q}}[X] = \int_{\{0,1\}} x \, d\mathbb{Q}(x).$$

Coin Toss Example

In other words, we compare

$$\int_{\mathbb{R}} \Phi(x)\, d\mathbb{P}(x) \quad \text{and} \quad \int_{\mathbb{R}} \Phi(x)\, d\mathbb{Q}(x)$$

where Φ is an identity map,

$$\Phi(x) = x.$$

A positive definite kernel corresponding to Φ is

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_2 = xy,$$

which is called a linear kernel. Therefore, comparing two Bernoulli is equivalent to

$$\int_{\{0,1\}} k(y, x)\, d\mathbb{P}(x) \stackrel{?}{=} \int_{\{0,1\}} k(y, x)\, d\mathbb{Q}(x)$$

for all $y \in \{0, 1\}$, i.e., compare the expectations of the kernel.

Comparing two Gaussians

$$\mathbb{P} = N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad \mathbb{Q} = N(\mu_2, \sigma_2^2)$$

Comparing $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to comparing μ_1, μ_2 and σ_1^2, σ_2^2, i.e.,

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X]$$

and

$$\mathbb{E}_{\mathbb{P}}[X^2] = \int_{\mathbb{R}} x^2 \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x^2 \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X^2].$$

Concisely

$$\int_{\mathbb{R}} \Phi(x) \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} \Phi(x) \, d\mathbb{Q}(x)$$

where

$$\Phi(x) = (x, x^2).$$

Compare the first moment of the feature map

Comparing two Gaussians

$$\mathbb{P} = N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad \mathbb{Q} = N(\mu_2, \sigma_2^2)$$

Comparing $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to comparing μ_1, μ_2 and σ_1^2, σ_2^2, i.e.,

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X]$$

and

$$\mathbb{E}_{\mathbb{P}}[X^2] = \int_{\mathbb{R}} x^2 \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x^2 \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X^2].$$

Concisely

$$\int_{\mathbb{R}} \Phi(x) \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} \Phi(x) \, d\mathbb{Q}(x)$$

where

$$\Phi(x) = (x, x^2).$$

Compare the first moment of the feature map

Comparing two Gaussians

Using the map Φ, we can construct a positive definite kernel as

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathbb{R}^2} = xy + x^2 y^2$$

which is called a polynomial kernel of order 2.

Therefore, comparing two Gaussians is equivalent to

$$\int_{\mathbb{R}} k(y, x)\, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} k(y, x)\, d\mathbb{Q}(x)$$

for all $y \in \mathbb{R}$, i.e., compare the expectations of the kernel.

Comparing general $\mathbb{P}$ and $\mathbb{Q}$

Moment generating function is defined as

$$M_{\mathbb{P}}(y) = \int_{\mathbb{R}} e^{xy}\, d\mathbb{P}(x)$$

and (if it exists) captures the information about a distribution, i.e.,

$$M_{\mathbb{P}} = M_{\mathbb{Q}} \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Choosing

$$\Phi(x) = \left(1, x, \frac{x^2}{\sqrt{2!}}, \ldots, \frac{x^i}{\sqrt{i!}}, \ldots\right) \in \ell_2(\mathbb{N}), \forall\, x \in \mathbb{R}$$

it is easy to verify that

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\ell_2(\mathbb{N})} = e^{xy}$$

and so

$$\int_{\mathbb{R}} k(x,y)\, d\mathbb{P}(x) = \int_{\mathbb{R}} k(x,y)\, d\mathbb{Q}(x), \forall\, y \in \mathbb{R} \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Generalization

Based on the above, we can compare $\mathbb{P}$ and $\mathbb{Q}$ defined on any measurable space $\mathcal{X}$ as

$$\int_{\mathcal{X}} k(x, y)\, d\mathbb{P}(x) = \int_{\mathcal{X}} k(x, y)\, d\mathbb{Q}(x), \forall\, y \in \mathcal{X}$$

using a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle.$$

In other words, we consider the map

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} := \underbrace{\int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)}_{\mathbb{E}_{X \sim \mathbb{P}} k(\cdot, X)}$$

and compare $\mathbb{P}$ and $\mathbb{Q}$ through $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ (Mean element).

Hilbert Space Embedding of Measures

Reproducing Kernel Hilbert Space

A Hilbert space, $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is called a reproducing kernel Hilbert space (RKHS) if the evaluational functional

$$\delta_x : \mathcal{H} \to \mathbb{R}, \qquad f \mapsto f(x)$$

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

- Riesz representation: $\forall\, x \in \mathcal{X}$, $\exists$ unique $k_x \in \mathcal{H}$ such that

$$\delta_x(f) = f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \;\; \forall\, f \in \mathcal{H}.$$

- Reproducing property, symmetry and positive-definiteness:

$$k(y,x) := k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = k_x(y) = k(x,y), \;\; x, y \in \mathcal{X}.$$

  $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel.

- Moore-Aronszajn Theorem: For every positive definite kernel, k on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS, $\mathcal{H}$ with k as its r.k.

Properties of RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) \,:\, x \in \mathcal{X}\}}$
- Norm convergence implies pointwise convergence!!
- k is bounded if and only every $f \in \mathcal{H}$ is bounded.
- If $\sqrt{k(x,x)}$ is p-integrable, then $\mathcal{H}$ consists of p-integrable functions.
- Every $f \in \mathcal{H}$ is continuous if and only if $k(\cdot, x)$ is continuous for all $x \in \mathcal{X}$.
- Every $f \in \mathcal{H}$ is m-times continuously differentiable if k is m-times continuously differentiable.

Explicit Realization of RKHS

- $\mathcal{X} = \mathbb{R}^d$ and $k(x, y) = \psi(x - y)$ where ψ is a positive definite function.
- Let $\psi \in L^1(\mathcal{X})$. Then

$$\mathcal{H} = \left\{ f \in L^2(\mathcal{X}) \cap C_b(\mathcal{X}) \,\middle|\, \int \frac{|\hat{f}(\omega)|^2}{\hat{\psi}(\omega)}\, d\omega < \infty \right\}$$

endowed with

$$\langle f, g \rangle_{\mathcal{H}} = (2\pi)^{-d/2} \int \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\psi}(\omega)}\, d\omega$$

is an RKHS with k as the r.k, where $\hat{\psi}$ is the Fourier transform of ψ.

Hilbert Space Embedding of Measures

- Canonical feature map:

$$\Phi(x) = k(\cdot, x) \in \mathcal{H}, \qquad x \in \mathcal{X}$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

- Therefore

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} := \int_{\mathcal{X}} \Phi(x)\, d\mathbb{P}(x) = \underbrace{\int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)}_{\mathbb{E}_{X \sim \mathbb{P}} k(\cdot, X)} \in \mathcal{H}$$

  (Smola et al., ALT 2007)

- Generalizes
  - characteristic function: $k(\cdot, x) = e^{-\sqrt{-1}\langle x, \cdot \rangle_2}$
  - Weierstrass transform: $k(\cdot, x) = e^{-\sigma \|x - \cdot\|_2^2}$, $\sigma > 0$.

Kernel Distance

- It is natural to consider

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_\infty$$

when comparing $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$.

- Since $\|\cdot\|_{\mathcal{H}}$ dominates $\|\cdot\|_\infty$, we use

$$\begin{aligned}
\rho_k(\mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \\
&= \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}}, \\
&= \int \int k(x, y)\, d(\mathbb{P} - \mathbb{Q})\, d(\mathbb{P} - \mathbb{Q})
\end{aligned}$$

called the kernel distance between $\mathbb{P}$ and $\mathbb{Q}$.

Interpretation of ρ_k (S et al., JMLR 2010)

Suppose $k(x,y)=\psi(x-y)$, $x,y\in\mathbb{R}^d$ where ψ is a positive definite function. By Bochner's theorem,

$$\psi(x)=\int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle x,\omega\rangle_2}\,d\Lambda(\omega),$$

where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.

- $$\rho_k(\mathbb{P},\mathbb{Q})=\|\phi_{\mathbb{P}}-\phi_{\mathbb{Q}}\|_{L^2(\Lambda)}$$

  where $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ are the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$.

- If $\psi\to\delta$, then $\rho_k(\mathbb{P},\mathbb{Q})\to\|p-q\|_{L^2(\mathbb{R}^d)}$.

Maximum Mean Discrepancy

(Gretton et al., NIPS 2007; S et al., JMLR 2010)

$$\rho_k(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int f(x)\, d\mathbb{P}(x) - \int f(x)\, d\mathbb{Q}(x) \right|,$$

where

$$\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}.$$

- The choice of $\mathcal{H}$ determines the power of ρ_k to distinguish between $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernel

k is said to be characteristic if

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

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x, y) = c > 0, \forall\, x, y \in \mathcal{X}$, then

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $\rho_k(\mathbb{P}, \mathbb{Q}) = 0, \forall\, \mathbb{P}, \mathbb{Q}$.

- Example: Let $k(x, y) = xy,\ x, y \in \mathbb{R}$. Then

$$\rho_k(\mathbb{P}, \mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x, y) = (1 + xy)^2,\ x, y \in \mathbb{R}$. Then

$$\rho_k^2(\mathbb{P}, \mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernel

k is said to be characteristic if

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

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x, y) = c > 0, \forall\, x, y \in \mathcal{X}$, then

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $\rho_k(\mathbb{P}, \mathbb{Q}) = 0, \forall\, \mathbb{P}, \mathbb{Q}$.

- Example: Let $k(x, y) = xy$, $x, y \in \mathbb{R}$. Then

$$\rho_k(\mathbb{P}, \mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x, y) = (1 + xy)^2$, $x, y \in \mathbb{R}$. Then

$$\rho_k^2(\mathbb{P}, \mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernel

k is said to be characteristic if

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

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x, y) = c > 0, \forall x, y \in \mathcal{X}$, then

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $\rho_k(\mathbb{P}, \mathbb{Q}) = 0, \forall \mathbb{P}, \mathbb{Q}$.

- Example: Let $k(x, y) = xy$, $x, y \in \mathbb{R}$. Then

$$\rho_k(\mathbb{P}, \mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x, y) = (1 + xy)^2$, $x, y \in \mathbb{R}$. Then

$$\rho_k^2(\mathbb{P}, \mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernel

k is said to be characteristic if

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

for any $\mathbb{P}$ and $\mathbb{Q}$.

Not all kernels are characteristic.

- Example: If $k(x, y) = c > 0, \forall x, y \in \mathcal{X}$, then

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) = c, \quad \mu_{\mathbb{Q}} = c$$

  and $\rho_k(\mathbb{P}, \mathbb{Q}) = 0, \forall \mathbb{P}, \mathbb{Q}$.

- Example: Let $k(x, y) = xy$, $x, y \in \mathbb{R}$. Then

$$\rho_k(\mathbb{P}, \mathbb{Q}) = |\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X]|.$$

  Characteristic for Bernoulli's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

- Example: Let $k(x, y) = (1 + xy)^2$, $x, y \in \mathbb{R}$. Then

$$\rho_k^2(\mathbb{P}, \mathbb{Q}) = 2(\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[X])^2 + (\mathbb{E}_{\mathbb{P}}[X^2] - \mathbb{E}_{\mathbb{Q}}[X^2]).$$

  Characteristic for Gaussian's but not for all $\mathbb{P}$ and $\mathbb{Q}$.

Characteristic Kernels on $\mathbb{R}^d$

- Translation invariant kernel: $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$; bounded and continuous.

k is characteristic $\Leftrightarrow$ $\text{supp}(\Lambda) = \mathbb{R}^d$. (S et al., COLT 2008; JMLR, 2010)

- Corollary: Compactly supported ψ are characteristic (S et al., COLT 2008; JMLR, 2010).

- Extensions: Locally compact Abelian groups, compact non-Abelian groups, Semigroup $\mathbb{R}^n_+$ (Fukumizu et al., NIPS 2009)

- Richness of $\mathcal{H}$ and distinguishability of $\mathbb{P}$ and $\mathbb{Q}$ (Gretton et al., NIPS 2007; Fukumizu et al., NIPS 2009; S et al., JMLR 2011)

Examples

$\mathcal{X} = \mathbb{R}^d$:

- Gaussian kernel:

$$k(x,y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}, \ \sigma > 0$$

- Matérn kernel:

$$k(x,y) = \frac{2^{1-s}}{\Gamma(s)} \|x-y\|_2^{s-\frac{d}{2}} \mathfrak{K}_{\frac{d}{2}-s}(\|x-y\|_2), \ s > \frac{d}{2}$$

- Inverse multiquadrics kernel:

$$k(x,y) = \left(1 + \frac{\|x-y\|_2^2}{c^2}\right)^{-t}, \ t > 0, \ c \in (0,\infty)$$

Metrization of weak*-topology (S, Bernoulli 2016)

Suppose $\mathcal{X}$ is a Polish and locally compact Hausdorff space. Let k satisfies the following:

- k is bounded on $\mathcal{X} \times \mathcal{X}$
- $k(\cdot, x) \in C_0(\mathcal{X})$ for all $x \in \mathcal{X}$
- $x \mapsto k(x, x)$ is continuous
- $\int \int_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y) > 0,\ \forall\, \mu \in M_b(\mathcal{X}) \backslash \{0\}$ (Universality)

Then

$$\rho_k(\mathbb{P}_{(n)}, \mathbb{P}) \to 0 \Leftrightarrow \mathbb{P}_{(n)} \xrightarrow{w} \mathbb{P}$$

as $n \to \infty$.

Universality:

$$\mu \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x)$$

is one-to-one.

Metrization of weak*-topology (S, Bernoulli 2016)

Suppose $\mathcal{X}$ is a Polish and locally compact Hausdorff space. Let k satisfies the following:

- k is bounded on $\mathcal{X} \times \mathcal{X}$
- $k(\cdot, x) \in C_0(\mathcal{X})$ for all $x \in \mathcal{X}$
- $x \mapsto k(x, x)$ is continuous
- $\int \int_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y) > 0, \ \forall\, \mu \in M_b(\mathcal{X}) \backslash \{0\}$ (Universality)

Then

$$\rho_k(\mathbb{P}_{(n)}, \mathbb{P}) \to 0 \Leftrightarrow \mathbb{P}_{(n)} \xrightarrow{w} \mathbb{P}$$

as $n \to \infty$.

Universality:

$$\mu \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mu(x)$$

is one-to-one.

Relation to Other Distances

- Total variation, Kullback-Leibler and Hellinger: (S et al., JMLR 2010)
  - Kernel distance is weaker but computationally efficient
- Wasserstein distance, bounded Lipschitz metric: (S et al., EJS 2012)
  - Topologically similar to kernel distance but computationally expensive
- Energy distance: (Sejdinovic et al., AoS 2013)
  - Special case of kernel distance

Measure of (In)dependence

$$\rho_k(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y)$$

Measuring (In)dependence

- Let X and Y be Gaussian random variables on $\mathbb{R}$. Then

  $$X \text{ and } Y \text{ are independent } \Leftrightarrow \mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$$

- In general, $\mathrm{Cov}(X, Y) = 0 \not\Rightarrow X \perp Y$.
- Covariance captures the linear relationship between X and Y.
- Feature space view point: How about $\mathrm{Cov}(\Phi(X), \Psi(Y))$?
- Suppose

  $$\Phi(X) = (1, X, X^2) \text{ and } \Psi(Y) = (1, Y, Y^2, Y^3).$$

  Then $\mathrm{Cov}(\Phi(X), \Phi(Y))$ captures $\mathrm{Cov}(X^i, Y^j)$ for $i \in \{0, 1, 2\}$ and $j \in \{0, 1, 2, 3\}$.

Measuring (In)dependence

- Let X and Y be Gaussian random variables on $\mathbb{R}$. Then

  X and Y are independent $\Leftrightarrow \mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$

- In general, $\mathrm{Cov}(X,Y) = 0 \not\Rightarrow X \perp Y$.
- Covariance captures the linear relationship between X and Y.
- Feature space view point: How about $\mathrm{Cov}(\Phi(X), \Psi(Y))$?
- Suppose

  $$\Phi(X) = (1, X, X^2) \text{ and } \Psi(Y) = (1, Y, Y^2, Y^3).$$

  Then $\mathrm{Cov}(\Phi(X), \Phi(Y))$ captures $\mathrm{Cov}(X^i, Y^j)$ for $i \in \{0,1,2\}$ and $j \in \{0,1,2,3\}$.

Measuring (In)Dependence

- Characterization of independence:

$$X \perp Y \Leftrightarrow \mathrm{Cov}(f(X), g(Y)) = 0, \forall \text{ measurable functions } f \text{ and } g.$$

- Dependence measure:

$$\sup_{f,g} |\mathrm{Cov}(f(X), g(Y))| = \sup_{f,g} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|$$

Similar to the MMD between $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$.

- Restricting functions in RKHS: (constrained covariance)

$$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|.$$

(Gretton et al., AISTATS 2005, JMLR 2005)

Measuring (In)Dependence

- Characterization of independence:

$$X \perp Y \Leftrightarrow \mathrm{Cov}(f(X), g(Y)) = 0, \forall \text{ measurable functions } f \text{ and } g.$$

- Dependence measure:

$$\sup_{f,g} |\mathrm{Cov}(f(X), g(Y))| = \sup_{f,g} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|$$

Similar to the MMD between $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$.

- Restricting functions in RKHS: (constrained covariance)

$$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} |\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]|.$$

(Gretton et al., AISTATS 2005, JMLR 2005)

Covariance Operator

- Assuming $\mathbb{E}\sqrt{k_X(X,X)k_Y(Y,Y)} < \infty$, we obtain

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \langle f, \mathbb{E}[k_X(\cdot,X)\otimes k_Y(\cdot,Y)]g\rangle_{\mathcal{H}_X} \\ &= \langle g, \mathbb{E}[k_Y(\cdot,Y)\otimes k_X(\cdot,X)]f\rangle_{\mathcal{H}_Y}\end{aligned}$$

-

$$\mathrm{Cov}(f(X),g(Y)) = \langle f, C_{XY}g\rangle_{\mathcal{H}_X} = \langle g, C_{YX}f\rangle_{\mathcal{H}_Y}$$

where

$$C_{XY} := \mathbb{E}[k_X(\cdot,X)\otimes k_Y(\cdot,Y)] - \mu_{\mathbb{P}_X}\otimes\mu_{\mathbb{P}_Y}$$

is a cross-covariance operator from $\mathcal{H}_Y$ to $\mathcal{H}_X$ and $C_{YX} = C^*_{XY}$.

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) = \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} |\langle f, C_{XY} g\rangle_{\mathcal{H}_X}|$$
  $$= \|C_{XY}\|_{\text{op}} = \|C_{YX}\|_{\text{op}},$$

  which is the maximum singular value of C_{XY}.
- Choosing $k_X(\cdot, X) = \langle \cdot, X\rangle_2$ and $k_Y(\cdot, Y) = \langle \cdot, Y\rangle_2$, for Gaussian distributions,
  $$X \perp Y \Leftrightarrow C_{YX} = 0$$
- In general,
  $$X \perp Y \overset{?}{\Leftrightarrow} C_{YX} = 0.$$

Dependence Measures

- $$
\begin{aligned}
COCO(\mathbb{P}_{XY};\mathcal{H}_X,\mathcal{H}_Y) &= \sup_{\substack{\|f\|_{\mathcal{H}_X}=1\\ \|g\|_{\mathcal{H}_Y}=1}} |\langle f, C_{XY}g\rangle_{\mathcal{H}_X}| \\
&= \|C_{XY}\|_{\mathrm{op}} = \|C_{YX}\|_{\mathrm{op}},
\end{aligned}
$$

  which is the maximum singular value of C_{XY}.

- Choosing $k_X(\cdot, X) = \langle \cdot, X\rangle_2$ and $k_Y(\cdot, Y) = \langle \cdot, Y\rangle_2$, for Gaussian distributions,

  $$X \perp Y \Leftrightarrow C_{YX} = 0$$

- In general,

  $$X \perp Y \overset{?}{\Leftrightarrow} C_{YX} = 0.$$

Dependence Measures

- How about we consider other singular values?
- How about $\|C_{YX}\|_{HS}^2$, which is the sum of squared singular values of C_{YX}?

Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., ALT 2005, JMLR 2005)

- $\|C_{YX}\|_{\text{op}} \leq \|C_{YX}\|_{HS}$

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} \left| \mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)] \right|.$$

- How about we use different constraint, i.e., $\|f \otimes g\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} \leq 1$?

$$\begin{aligned}
\sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \mathrm{Cov}(f(X), g(Y)) &= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f, C_{XY} g\rangle_{\mathcal{H}_X} \\
&= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f \otimes g, C_{XY}\rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} = \|C_{XY}\|_{HS}
\end{aligned}$$

- $$\begin{aligned}
\|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} &= \|\mathbb{E}[k_X(\cdot, X) \otimes k_Y(\cdot, Y)] - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_X}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \left\| \int k_X(\cdot, X) \otimes k_Y(\cdot, Y)\, d(\mathbb{P}_{XY} - \mathbb{P}_X \times \mathbb{P}_Y) \right\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \rho_{k_X\otimes k_Y}(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y)
\end{aligned}$$

Dependence Measures

- $$COCO(\mathbb{P}_{XY};\mathcal{H}_X,\mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1\\ \|g\|_{\mathcal{H}_Y}=1}} \left|\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]\right|.$$

- How about we use different constraint, i.e., $\|f \otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \le 1$?

$$\begin{aligned}
\sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\le 1} \mathrm{Cov}(f(X), g(Y)) &= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\le 1} \langle f, C_{XY} g\rangle_{\mathcal{H}_X} \\
&= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\le 1} \langle f \otimes g, C_{XY}\rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} = \|C_{XY}\|_{HS}
\end{aligned}$$

- $$\begin{aligned}
\|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} &= \|\mathbb{E}[k_X(\cdot,X)\otimes k_Y(\cdot,Y)] - \mu_{\mathbb{P}_X}\otimes\mu_{\mathbb{P}_X}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \left\| \int k_X(\cdot,X)\otimes k_Y(\cdot,Y)\, d(\mathbb{P}_{XY} - \mathbb{P}_X\times\mathbb{P}_Y)\right\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \rho_{k_X\otimes k_Y}(\mathbb{P}_{XY}, \mathbb{P}_X\times\mathbb{P}_Y)
\end{aligned}$$

Dependence Measures

- $$COCO(\mathbb{P}_{XY}; \mathcal{H}_X, \mathcal{H}_Y) := \sup_{\substack{\|f\|_{\mathcal{H}_X}=1 \\ \|g\|_{\mathcal{H}_Y}=1}} \left| \mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)] \right|.$$
- How about we use different constraint, i.e., $\|f \otimes g\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} \leq 1$?

$$\begin{aligned}
\sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \mathrm{Cov}(f(X), g(Y)) &= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f, C_{XY} g \rangle_{\mathcal{H}_X} \\
&= \sup_{\|f\otimes g\|_{\mathcal{H}_X\otimes\mathcal{H}_Y}\leq 1} \langle f \otimes g, C_{XY} \rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} = \|C_{XY}\|_{HS}
\end{aligned}$$

- $$\begin{aligned}
\|C_{XY}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} &= \|\mathbb{E}[k_X(\cdot, X) \otimes k_Y(\cdot, Y)] - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_X}\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \left\| \int k_X(\cdot, X) \otimes k_Y(\cdot, Y)\, d(\mathbb{P}_{XY} - \mathbb{P}_X \times \mathbb{P}_Y) \right\|_{\mathcal{H}_X\otimes\mathcal{H}_Y} \\
&= \rho_{k_X\otimes k_Y}(\mathbb{P}_{XY}, \mathbb{P}_X \times \mathbb{P}_Y)
\end{aligned}$$

Tensor Product Kernels

$$k((x,y),(x',y')) = k_X(x,x')k_Y(y,y').$$

Question

Suppose $k = \otimes_{m=1}^{M} k_m$, i.e.,

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

Define

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

We define k to be $\mathcal{I}$-characteristic if

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

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

Characteristic properties of $\otimes_{m=1}^{M} k_m$ in terms of k_m-s?

Known Results ($M = 2$)

- (Blanchard et al., NIPS 2011; Gretton, 2015):

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

- (Lyons, AoP 2013; Sejdinovic et al., AoS 2013):

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

Goal: Extension to $M \geq 3$

Results (Szabó and S, 2017)

- k_1 & k_2: characteristic $\not\Rightarrow$ $k_1 \otimes k_2$: characteristic
- $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-characteristic $\Rightarrow$ (k_m)-s: characteristic
- $\otimes_{m=1}^{M} k_m$: characteristic $\Rightarrow$ (k_m)-s: characteristic
- k_1, k_2 & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!)
- k_1 k_2: universal & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!!)
- (k_m)-s: universal $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic
- If k_m-s are translation-invariant and characteristic on $\mathbb{R}^d$, then all notions are equivalent

Results (Szabó and S, 2017)

- k_1 & k_2: characteristic $\not\Rightarrow$ $k_1 \otimes k_2$: characteristic
- $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-characteristic $\Rightarrow$ (k_m)-s: characteristic
- $\otimes_{m=1}^{M} k_m$: characteristic $\Rightarrow$ (k_m)-s: characteristic
- k_1, k_2 & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!)
- k_1 k_2: universal & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!!)
- (k_m)-s: universal $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic
- If k_m-s are translation-invariant and characteristic on $\mathbb{R}^d$, then all notions are equivalent

Results (Szabó and S, 2017)

- k_1 & k_2: characteristic $\not\Rightarrow$ $k_1 \otimes k_2$: characteristic
- $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-characteristic $\Rightarrow$ (k_m)-s: characteristic
- $\otimes_{m=1}^{M} k_m$: characteristic $\Rightarrow$ (k_m)-s: characteristic
- k_1, k_2 & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!)
- k_1 k_2: universal & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!!)
- (k_m)-s: universal $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic
- If k_m-s are translation-invariant and characteristic on $\mathbb{R}^d$, then all notions are equivalent

Results (Szabó and S, 2017)

- k_1 & k_2: characteristic $\not\Rightarrow$ $k_1 \otimes k_2$: characteristic
- $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-characteristic $\Rightarrow$ (k_m)-s: characteristic
- $\otimes_{m=1}^{M} k_m$: characteristic $\Rightarrow$ (k_m)-s: characteristic
- $k_1,\ k_2$ & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!)
- $k_1\ k_2$: universal & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!!)
- (k_m)-s: universal $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic
- If k_m-s are translation-invariant and characteristic on $\mathbb{R}^d$, then all notions are equivalent

Results (Szabó and S, 2017)

- k_1 & k_2: characteristic $\not\Rightarrow$ $k_1 \otimes k_2$: characteristic
- $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-characteristic $\Rightarrow$ (k_m)-s: characteristic
- $\otimes_{m=1}^{M} k_m$: characteristic $\Rightarrow$ (k_m)-s: characteristic
- k_1, k_2 & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!)
- k_1 k_2: universal & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!!)
- (k_m)-s: universal $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic
- If k_m-s are translation-invariant and characteristic on $\mathbb{R}^d$, then all notions are equivalent

Results (Szabó and S, 2017)

- k_1 & k_2: characteristic $\not\Rightarrow$ $k_1 \otimes k_2$: characteristic
- $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-characteristic $\Rightarrow$ (k_m)-s: characteristic
- $\otimes_{m=1}^{M} k_m$: characteristic $\Rightarrow$ (k_m)-s: characteristic
- k_1, k_2 & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!)
- k_1 k_2: universal & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!!)
- (k_m)-s: universal $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic
- If k_m-s are translation-invariant and characteristic on $\mathbb{R}^d$, then all notions are equivalent

Results (Szabó and S, 2017)

- k_1 & k_2: characteristic $\not\Rightarrow$ $k_1 \otimes k_2$: characteristic
- $\otimes_{m=1}^{M} k_m$: $\mathcal{I}$-characteristic $\Rightarrow$ (k_m)-s: characteristic
- $\otimes_{m=1}^{M} k_m$: characteristic $\Rightarrow$ (k_m)-s: characteristic
- k_1, k_2 & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!)
- k_1 k_2: universal & k_3: characteristic $\not\Rightarrow$ $\otimes_{m=1}^{3} k_m$: $\mathcal{I}$-characteristic (!!)
- (k_m)-s: universal $\Leftrightarrow$ $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic
- If k_m-s are translation-invariant and characteristic on $\mathbb{R}^d$, then all notions are equivalent

Summary

- HSIC as a measure of independence
- Characterization of product kernel in terms of individual kernels

Thank You

References I

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., Sriperumbudur, B. K., Gretton, A., and Schölkopf, B. (2009).
Characteristic kernels on groups and semigroups.
In *Advances in Neural Information Processing Systems 21*, pages 473–480.

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., Schölkopf, B., and Smola, A. (2007).
A kernel method for the two sample problem.
In *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a).
Measuring statistical dependence with Hilbert-Schmidt norms.
In Jain, S., Simon, H. U., and Tomita, E., editors, *Proceedings of Algorithmic Learning Theory*, pages 63–77, Berlin. Springer-Verlag.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b).
Kernel methods for measuring independence.
Journal of Machine Learning Research, 6:2075–2129.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005c).
Kernel constrained covariance for dependence measurement.
In Ghahramani, Z. and Cowell, R., editors, *Proc. 10th International Workshop on Artificial Intelligence and Statistics*, pages 1–8.

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

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(5):2263–2291.

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, Berlin, Germany.

References II

Sriperumbudur, B. K. (2016).
On the optimal estimation of probability measures in weak and strong topologies.
Bernoulli, 22(3):1839–1893.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012).
On the empirical estimation of integral probability metrics.
Electronic Journal of Statistics, 6:1550–1599.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2011).
Universality, characteristic kernels and RKHS embedding of measures.
Journal of Machine Learning Research, 12:2389–2410.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.

Szabó, Z. and Sriperumbudur, B. (2017).
Characteristic and universal tensor product kernels.
Technical report.
(`http://arxiv.org/abs/1708.08157`).