Independence Measures and Testing

Zoltán Szabó

March 24, 2021

Contents

- Examples.
- Independence measures based on
  - copula,
  - maximum correlation,
  - distance,
  - kernel.

High-level idea

- $X = (X_1, \ldots, X_M) \in \times_{m\in[M]} \mathcal{X}_m$. X_m: m^{th} 'coordinate' of X. $X \sim \mathbb{P}$.
- $(X_m)_{m\in[M]}$ are independent iff $\boxed{\mathbb{P} = \otimes_{m\in[M]} \mathbb{P}_m}$.

High-level idea

- $X = (X_1, \ldots, X_M) \in \times_{m\in[M]} \mathcal{X}_m$. X_m: m^{th} 'coordinate' of X. $X \sim \mathbb{P}$.
- $(X_m)_{m\in[M]}$ are independent iff $\boxed{\mathbb{P} = \otimes_{m\in[M]} \mathbb{P}_m}$.

1. Copula: $\mathcal{X}_m = \mathbb{R}$, C = copula of X, Π = independence copula.

$$\boxed{\left\| C - \Pi \right\|_{L^2\left([0,1]^M\right)}}.$$

High-level idea

- $X = (X_1, \ldots, X_M) \in \times_{m \in [M]} \mathcal{X}_m$. X_m: m^{th} 'coordinate' of X. $X \sim \mathbb{P}$.
- $(X_m)_{m \in [M]}$ are independent iff $\mathbb{P} = \otimes_{m \in [M]} \mathbb{P}_m$.

1. Copula: $\mathcal{X}_m = \mathbb{R}$, C = copula of X, Π = independence copula.

$$\|C - \Pi\|_{L^2([0,1]^M)}.$$

2. Maximum correlation ($M = 2$):

$$\sup_{f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2} \mathrm{corr}(f_1(X_1), f_2(X_2)).$$

High-level idea

- $X = (X_1, \ldots, X_M) \in \times_{m\in[M]}\mathcal{X}_m$. X_m: m^{th} 'coordinate' of X. $X \sim \mathbb{P}$.
- $(X_m)_{m\in[M]}$ are independent iff $\mathbb{P} = \otimes_{m\in[M]}\mathbb{P}_m$.

1. Copula: $\mathcal{X}_m = \mathbb{R}$, C = copula of X, Π = independence copula.

$$\|C - \Pi\|_{L^2([0,1]^M)}.$$

2. Maximum correlation ($M = 2$):

$$\sup_{f_1\in\mathcal{H}_1, f_2\in\mathcal{H}_2} \mathrm{corr}(f_1(X_1), f_2(X_2)).$$

3. Distance: ϕ = characteristic function of X, ϕ_m = that of X_m.

$$\left\|\phi - \prod_{m\in[M]} \phi_m\right\|_{L^2(w)}.$$

High-level idea

- $X = (X_1, \ldots, X_M) \in \times_{m\in[M]}\mathcal{X}_m$. X_m: m^{th} 'coordinate' of X. $X \sim \mathbb{P}$.
- $(X_m)_{m\in[M]}$ are independent iff $\boxed{\mathbb{P} = \otimes_{m\in[M]}\mathbb{P}_m}$.

1. Copula: $\mathcal{X}_m = \mathbb{R}$, C = copula of X, Π = independence copula.
$$\boxed{\|C - \Pi\|_{L^2([0,1]^M)}}.$$
2. Maximum correlation ($M = 2$):
$$\boxed{\sup_{f_1\in\mathcal{H}_1, f_2\in\mathcal{H}_2} \operatorname{corr}(f_1(X_1), f_2(X_2))}.$$
3. Distance: ϕ = characteristic function of X, ϕ_m = that of X_m.
$$\boxed{\left\|\phi - \prod_{m\in[M]} \phi_m\right\|_{L^2(w)}}.$$
4. Kernel: C = covariance of X under some feature φ.
$$\boxed{\|C\|_{\mathrm{HS}} = D(\mathbb{P}, \otimes_{m\in[M]}\mathbb{P}_m)}.$$

Our questions

1. Validness.
2. (Statistical) properties.
3. Estimation.
4. Applications.

Code (Python)

- Information theoretical estimators (ITE) toolbox:
  - 53 entropy, independence, divergence, association measures and kernels of probability distributions,
  - (by at least an order) the largest package in the domain,
  - ~ 80 successful projects worldwide.
  - `https://bitbucket.org/szzoli/ite-in-python/`

Code (Python)

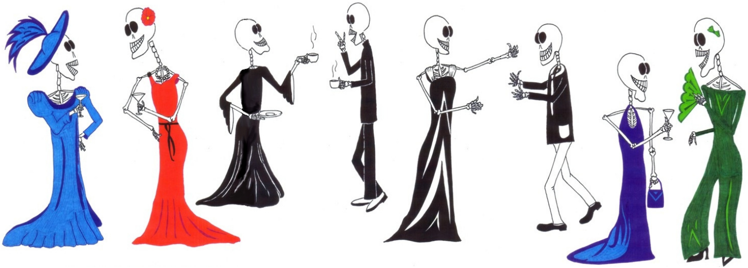
- Information theoretical estimators (ITE) toolbox:
  - 53 entropy, independence, divergence, association measures and kernels of probability distributions,
  - (by at least an order) the largest package in the domain,
  - ~ 80 successful projects worldwide.
  - `https://bitbucket.org/szzoli/ite-in-python/`
- Independence testing toolbox:

  `https://github.com/wittawatj/fsic-test`

Examples

Cocktail party problem:

- independent groups of people / music bands,
- observation = mixed sources.



ISA equations

Observation:

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right].$$

Goal: $\hat{\mathbf{s}}$ from $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. Assumptions:

- independent groups: $I\left(\mathbf{s}^1, \ldots, \mathbf{s}^M\right) = 0$,
- $\mathbf{s}^m$-s: non-Gaussian,
- $\mathbf{A}$: invertible.

ISA solution

Find $\mathbf{W}$ which makes the estimated components independent:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \left[\mathbf{y}^1; \ldots; \mathbf{y}^M\right],$$

$$J(\mathbf{W}) = I\left(\mathbf{y}^1, \ldots, \mathbf{y}^M\right) \to \min_{\mathbf{W}}.$$

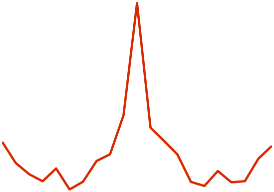
Given two images:

Goal: find the transformation which takes the right one to the left.

Outlier-robust image registration [Kybic, 2004, Neemuchwala et al., 2007]

Given two images:





Goal: find the transformation which takes the right one to the left.

Outlier-robust image registration: equations

- Reference image: $\mathbf{y}_{\text{ref}}$,
- test image: $\mathbf{y}_{\text{test}}$,
- possible transformations: $\boldsymbol{\Theta}$.

Objective:

$$J(\boldsymbol{\theta}) = \underbrace{I(\mathbf{y}_{\text{ref}}, \mathbf{y}_{\text{test}}(\boldsymbol{\theta}))}_{\text{similarity}} \rightarrow \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}},$$

Feature selection

- **Goal**: find
  - the feature subset (# of rooms, criminal rate, local taxes)
  - most relevant for house price prediction (y).

Feature selection: equations

- Features: $x^1, \ldots, x^F$. Subset: $S \subseteq \{1, \ldots, F\}$.
- MaxRelevance - MinRedundancy principle [Peng et al., 2005]:

$$J(S) = \frac{1}{|S|} \sum_{i \in S} I\left(x^i, y\right) - \frac{1}{|S|^2} \sum_{i,j \in S} I\left(x^i, x^j\right) \to \max_{S \subseteq \{1,\ldots,F\}}.$$

Independence testing: translation

- How do we detect dependency? (paired samples)

x_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

x_2: No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

...

y_1: Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat et concerne l'aide financiére qu'on a annoncée pour les agriculteurs. La plupart des agriculteurs n'ont encore rien reu de cet argent.

y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

...

Independence testing: translation

- How do we detect dependency? (paired samples)

x_1: Honourable senators, I have a question for the Leader of the Government in the Senate with regard to the support funding to farmers that has been announced. Most farmers have not received any money yet.

x_2: No doubt there is great pressure on provincial and municipal governments in relation to the issue of child care, but the reality is that there have been no cuts to child care funding from the federal government to the provinces. In fact, we have increased federal investments for early childhood development.

...

y_1: Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat et concerne l'aide financiére qu'on a annoncée pour les agriculteurs. La plupart des agriculteurs n'ont encore rien reu de cet argent.

y_2: Il est évident que les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions en ce qui concerne les services de garde, mais le gouvernement n'a pas réduit le financement qu'il verse aux provinces pour les services de garde. Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes enfants.

...

Are the French paragraphs translations of the English ones, or have nothing to do with it, i.e. $\mathbb{P}_{XY} = \mathbb{P}_X \otimes \mathbb{P}_Y$?

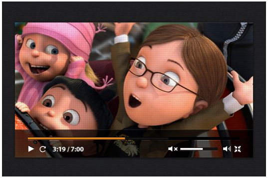
Dependency testing of media annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

Dependency testing of media annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

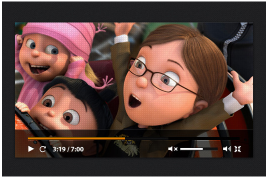
  - (video, caption) pairs



Dependency testing of media annotations

- We are given paired samples. Task: test independence.
- Examples:
  - (song, year of release) pairs

  - (video, caption) pairs



- $\{(\mathbf{x}_n, y_n)\}_{n\in[N]} \stackrel{?}{\to} H_0 : \mathbb{P}_{XY} = \mathbb{P}_X \otimes \mathbb{P}_Y,\ H_1 : \mathbb{P}_{XY} \neq \mathbb{P}_X \otimes \mathbb{P}_Y.$

Copula

Setting, Sklar's theorem [Nelsen, 2006]

- Setting: $\mathbf{X} = [X_m]_{m \in [M]} \in \mathbb{R}^M$.
- Cdf of $\mathbf{X}$ and X_m-s:

$$F(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}), \qquad F_m(x) = \mathbb{P}(X_m \leq x).$$

Assumption: F_m-s are continuous.

Setting, Sklar's theorem [Nelsen, 2006]

- Setting: $\mathbf{X} = [X_m]_{m \in [M]} \in \mathbb{R}^M$.
- Cdf of $\mathbf{X}$ and X_m-s:

$$F(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}), \qquad F_m(x) = \mathbb{P}(X_m \leq x).$$

  Assumption: F_m-s are continuous.
- Sklar's theorem: $\exists!$ a function (copula) $C : [0,1]^M \to [0,1]$ s.t.

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_M(x_M)) \quad \forall \mathbf{x} \in \mathbb{R}^M.$$

Setting, Sklar's theorem [Nelsen, 2006]

- Setting: $\mathbf{X} = [X_m]_{m\in[M]} \in \mathbb{R}^M$.
- Cdf of $\mathbf{X}$ and X_m-s:

$$F(\mathbf{x}) = \mathbb{P}(\mathbf{X} \le \mathbf{x}), \qquad F_m(x) = \mathbb{P}(X_m \le x).$$

  Assumption: F_m-s are continuous.
- Sklar's theorem: $\exists!$ a function (copula) $C : [0,1]^M \to [0,1]$ s.t.

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_M(x_M)) \quad \forall \mathbf{x} \in \mathbb{R}^M.$$

- C: It represents the joint distribution of $U_m = F_m(X_m)$

$$C(\mathbf{u}) = \mathbb{P}(\mathbf{U} \le \mathbf{u}), \qquad \mathbf{U} = [F_m(X_m)]_{m\in[M]}, \qquad \mathbf{u} \in [0,1]^M.$$

Copula bounds

For any copula C

$$W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u}), \quad \forall \mathbf{u} \in [0,1]^M,$$

Copula bounds

For any copula C

$$W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u}), \quad \forall \mathbf{u} \in [0,1]^M,$$

$$W(\mathbf{u}) = \max\left(\sum_{m \in [M]} u_m - (M-1), 0\right),$$

- W: Fréchet-Hoeffding lower bound. Copula for $M = 2$ only.

Copula bounds

For any copula C

$$W(\mathbf{u}) \le C(\mathbf{u}) \le M(\mathbf{u}), \quad \forall \mathbf{u} \in [0,1]^M,$$
$$W(\mathbf{u}) = \max\left(\sum_{m\in[M]} u_m - (M-1), 0\right),$$
$$M(\mathbf{u}) = \min_{m\in[M]} u_m.$$

- W: Fréchet-Hoeffding lower bound. Copula for $M=2$ only.
- M:
  - Fréchet-Hoeffding upper bound (comonotonicity copula).
  - Strictly increasing functional relation between X_i and X_j.

Copula: independence

- $[X_m]_{m\in[M]}$ **independent** iff.

$$\boxed{C} = \underbrace{\boxed{\Pi}}_{\text{product copula}}, \qquad \Pi(u_1,\ldots,u_M) = \prod_{m\in[M]} u_m.$$

Copula: independence

- $[X_m]_{m\in[M]}$ independent iff.

$$\boxed{C} = \underbrace{\boxed{\Pi}}_{\text{product copula}}, \qquad \Pi(u_1,\ldots,u_M) = \prod_{m\in[M]} u_m.$$

- L^p distance of C and Π:

$$I_p(C) = [h_p(M)]^{\frac{1}{p}} \left\| C - \Pi \right\|_{L^p\left([0,1]^M\right)}, \quad p \in [1,\infty].$$

Normalizing constant $h_p(M)$: to ensure $I_p(C) \in [0,1]$.

- $p = 2$: Hoeffding's Φ [Gaißer et al., 2010, Schmid et al., 2010],

$$\Phi^2(C) = h_2(M) \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u},$$

$$[h_2(M)]^{-1} = \|M - \Pi\|^2_{L^2([0,1]^M)}$$

.

For $p \in \{1, 2, \infty\}$

- $p = 2$: Hoeffding's Φ [Gaißer et al., 2010, Schmid et al., 2010],

$$\Phi^2(C) = h_2(M) \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u},$$

$$[h_2(M)]^{-1} = \|M - \Pi\|^2_{L^2([0,1]^M)}$$

$$= \frac{2}{(M+1)(M+2)} - \frac{1}{2^M} \frac{M!}{\prod_{i=0}^{M} \left(i + \frac{1}{2}\right)} + \frac{1}{3^M}.$$

For $p \in \{1, 2, \infty\}$

- $p = 2$: Hoeffding's Φ [Gaißer et al., 2010, Schmid et al., 2010],

$$\Phi^2(C) = h_2(M) \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u},$$

$$[h_2(M)]^{-1} = \|M - \Pi\|^2_{L^2([0,1]^M)}$$

$$= \frac{2}{(M+1)(M+2)} - \frac{1}{2^M} \frac{M!}{\prod_{i=0}^{M}\left(i + \frac{1}{2}\right)} + \frac{1}{3^M}.$$

- For $p = 1$ and $p = \infty$: Schweizer-Wolff's σ and κ [Schweizer and Wolff, 1981] ($M = 2$).

Properties

1. Normalization:

$$\begin{aligned}
\Phi^2(C) &\in [0,1], \quad \forall C, \\
\Phi^2(C) &= 0 \Leftrightarrow C = \Pi, \\
\Phi^2(C) &= 1 \Leftrightarrow C = M \qquad\qquad\qquad (M \geq 3), \\
&\quad\ \Leftrightarrow C = M \text{ or } C = W \quad (M = 2).
\end{aligned}$$

Properties

1. Normalization:

$$\begin{aligned}
\Phi^2(C) &\in [0,1], \quad \forall C,\\
\Phi^2(C) = 0 &\Leftrightarrow C = \Pi,\\
\Phi^2(C) = 1 &\Leftrightarrow C = M && (M \geq 3),\\
&\Leftrightarrow C = M \text{ or } C = W && (M = 2).
\end{aligned}$$

2. Invariance w.r.t. permutations: for any π permutation

$$\Phi^2(\mathbf{X}) = \Phi^2(\pi(\mathbf{X})).$$

Properties – continued

3. Monotonicity: For any C_1, C_2, C_3, C_4 copula for which

$$W \preccurlyeq C_1 \preccurlyeq C_2 \preccurlyeq \boxed{\Pi} \preccurlyeq C_3 \preccurlyeq C_4 \preccurlyeq M,$$

3. Monotonicity: For any C_1, C_2, C_3, C_4 copula for which

$$W \preccurlyeq C_1 \preccurlyeq C_2 \preccurlyeq \boxed{\Pi} \preccurlyeq C_3 \preccurlyeq C_4 \preccurlyeq M,$$
$$\Phi^2(C_1) \geq \Phi^2(C_2), \quad \Phi^2(C_3) \leq \Phi^2(C_4).$$

Approaching Π, the value of Φ^2 is decreasing.

Properties – continued

3. Monotonicity: For any C_1, C_2, C_3, C_4 copula for which

$$\color{blue}{W} \preccurlyeq C_1 \preccurlyeq C_2 \preccurlyeq \boxed{\Pi} \preccurlyeq C_3 \preccurlyeq C_4 \preccurlyeq \color{red}{M},$$
$$\Phi^2(C_1) \geq \Phi^2(C_2), \quad \Phi^2(C_3) \leq \Phi^2(C_4).$$

   Approaching Π, the value of Φ^2 is decreasing.

4. Invariance under strictly monotone transformations:
   - $M \geq 2$: strictly increasing transformation of the coordinates.
   - $M = 2$: strictly decreasing transformation of one/both coordinates.

3. Monotonicity: For any C_1, C_2, C_3, C_4 copula for which

$$\color{blue}{W} \preccurlyeq C_1 \preccurlyeq C_2 \preccurlyeq \boxed{\Pi} \preccurlyeq C_3 \preccurlyeq C_4 \preccurlyeq \color{red}{M},$$
$$\Phi^2(C_1) \geq \Phi^2(C_2), \quad \Phi^2(C_3) \leq \Phi^2(C_4).$$

   Approaching Π, the value of Φ^2 is decreasing.

4. Invariance under strictly monotone transformations:
   - $M \geq 2$: strictly increasing transformation of the coordinates.
   - $M = 2$: strictly decreasing transformation of one/both coordinates.

5. Continuity:

$$(C_n)_{n\in\mathbb{N}} \xrightarrow{n\to\infty} C \text{ pointwise } \Rightarrow \Phi^2(C_n) \xrightarrow{n\to\infty} \Phi^2(C).$$

Estimation of Φ^2

- Goal: given $(\mathbf{X}_n)_{n\in[N]}$ estimate $\Phi^2(\mathbf{X})$.

Estimation of Φ^2

- Goal: given $(\mathbf{X}_n)_{n\in[N]}$ estimate $\Phi^2(\mathbf{X})$.
- Empirical cdf-s, ranks:

$$\hat{F}_m(x) = \frac{1}{N}\sum_{n\in[N]} \mathbb{I}_{\{X_{m,n}\le x\}},$$

$$\hat{U}_{m,n} = \hat{F}_m(X_{m,n}) = \frac{1}{N}(\text{rank of } X_{m,n} \text{ in } X_{m,1},\ldots,X_{m,N}).$$

Estimation of Φ^2

- Goal: given $(\mathbf{X}_n)_{n\in[N]}$ estimate $\Phi^2(\mathbf{X})$.
- Empirical cdf-s, ranks:

$$\hat{F}_m(x) = \frac{1}{N}\sum_{n\in[N]} \mathbb{I}_{\{X_{m,n}\le x\}},$$

$$\hat{U}_{m,n} = \hat{F}_m(X_{m,n}) = \frac{1}{N}(\text{rank of } X_{m,n} \text{ in } X_{m,1},\ldots,X_{m,N}).$$

- Empirical copula:

$$\hat{C}_N(\mathbf{u}) = \frac{1}{N}\sum_{n\in[N]}\prod_{m\in[M]} \mathbb{I}_{\{\hat{U}_{m,n}\le u_m\}},\ \mathbf{u}\in[0,1]^M.$$

Estimation of Φ^2 – continued

- Recall:

$$\Phi^2(C) = h_2(M) \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u},$$

$$[h_2(M)]^{-1} = \frac{2}{(M+1)(M+2)} - \frac{1}{2^M} \frac{M!}{\prod_{i=0}^{M}\left(i + \frac{1}{2}\right)} + \frac{1}{3^M}.$$

Estimation of Φ^2 – continued

- Recall:

$$\Phi^2(C) = h_2(M)\int_{[0,1]^M}[C(\mathbf{u}) - \Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u},$$

$$[h_2(M)]^{-1} = \frac{2}{(M+1)(M+2)} - \frac{1}{2^M}\frac{M!}{\prod_{i=0}^{M}\left(i+\frac{1}{2}\right)} + \frac{1}{3^M}.$$

- Estimator:

$$\widehat{\Phi_N^2}(C) = \Phi^2\left(\hat{C}_N\right) = h_2(M)\underbrace{\int_{[0,1]^M}\left[\hat{C}_N(\mathbf{u}) - \Pi(\mathbf{u})\right]^2 \mathrm{d}\mathbf{u}}_{=:(*)}.$$

Estimation of Φ^2 – continued

- Recall:

$$\Phi^2(C) = h_2(M) \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u},$$

$$[h_2(M)]^{-1} = \frac{2}{(M+1)(M+2)} - \frac{1}{2^M} \frac{M!}{\prod_{i=0}^{M}\left(i+\frac{1}{2}\right)} + \frac{1}{3^M}.$$

- Estimator:

$$\widehat{\Phi^2_N}(C) = \Phi^2\left(\hat{C}_N\right) = h_2(M) \underbrace{\int_{[0,1]^M} \left[\hat{C}_N(\mathbf{u}) - \Pi(\mathbf{u})\right]^2 \mathrm{d}\mathbf{u}}_{=:(*)}.$$

Good news

$(*)$ can also be computed analytically!

Computation of $(*)$

$$\int_{[0,1]^M}\left[\hat{C}_N(\mathbf{u})-\prod_m u_m\right]^2 \mathrm{d}\mathbf{u}=\int_{[0,1]^M}\left[\frac{1}{N}\sum_n\prod_m \mathbb{I}_{\{\hat{U}_{m,n}\le u_m\}}-\prod_m u_m\right]^2 \mathrm{d}\mathbf{u}$$

(i): $\frac{1}{N}\sum_{n\in[N]} a = a$

Computation of $(*)$

$$\int_{[0,1]^M}\left[\hat{C}_N(\mathbf{u})-\prod_m u_m\right]^2 \mathrm{d}\mathbf{u}=\int_{[0,1]^M}\left[\frac{1}{N}\sum_n\prod_m \mathbb{I}_{\{\hat{U}_{m,n}\le u_m\}}-\prod_m u_m\right]^2 \mathrm{d}\mathbf{u}$$

$$\overset{(i)}{=}\int_{[0,1]^M}\left[\frac{1}{N}\sum_n \underbrace{\left(\prod_m \mathbb{I}_{\{\hat{U}_{m,n}\le u_m\}}-\prod_m u_m\right)}_{b_n}\right]^2 \mathrm{d}\mathbf{u}$$

(i): $\frac{1}{N}\sum_{n\in[N]} a = a$, (ii) $\left(\frac{1}{N}\sum_{n\in[N]} b_n\right)^2=\frac{1}{N^2}\sum_{j,k\in[N]} b_j b_k$, $\int \leftrightarrow \sum$.

Computation of $(*)$

$$\int_{[0,1]^M}\left[\hat{C}_N(\mathbf{u})-\prod_m u_m\right]^2 \mathrm{d}\mathbf{u}=\int_{[0,1]^M}\left[\frac{1}{N}\sum_n\prod_m \mathbb{I}_{\{\hat{U}_{m,n}\le u_m\}}-\prod_m u_m\right]^2\mathrm{d}\mathbf{u}$$

$$\overset{(i)}{=}\int_{[0,1]^M}\left[\frac{1}{N}\sum_n\underbrace{\left(\prod_m \mathbb{I}_{\{\hat{U}_{m,n}\le u_m\}}-\prod_m u_m\right)}_{b_n}\right]^2\mathrm{d}\mathbf{u}$$

$$\overset{(ii)}{=}\frac{1}{N^2}\sum_{j,k\in[N]}\int_{[0,1]^M}\underbrace{\left(\prod_m \mathbb{I}_{\{\hat{U}_{m,j}\le u_m\}}-\prod_m u_m\right)}_{b_j}\underbrace{\left(\prod_m \mathbb{I}_{\{\hat{U}_{m,k}\le u_m\}}-\prod_m u_m\right)}_{b_k}\mathrm{d}\mathbf{u}.$$

(i): $\frac{1}{N}\sum_{n\in[N]} a = a$, (ii) $\left(\frac{1}{N}\sum_{n\in[N]} b_n\right)^2=\frac{1}{N^2}\sum_{j,k\in[N]} b_j b_k$, $\int \leftrightarrow \sum$.

Computation of $(*)$ – continued

$$\frac{1}{N^2}\sum_{j,k\in[N]}\int_{[0,1]^M}\left(\prod_m \mathbb{I}_{\left\{\hat{U}_{m,j}\le u_m\right\}} - \prod_m u_m\right)\left(\prod_m \mathbb{I}_{\left\{\hat{U}_{m,k}\le u_m\right\}} - \prod_m u_m\right)$$

(i) $(a-b)(c-d) = ac + bd - ad - bc$

Computation of $(*)$ – continued

$$\frac{1}{N^2}\sum_{j,k\in[N]}\int_{[0,1]^M}\left(\prod_m \mathbb{I}_{\left\{\hat{U}_{m,j}\le u_m\right\}}-\prod_m u_m\right)\left(\prod_m \mathbb{I}_{\left\{\hat{U}_{m,k}\le u_m\right\}}-\prod_m u_m\right)$$

$$\stackrel{(i)}{=}\frac{1}{N^2}\sum_{j,k\in[N]}\int_{[0,1]^M}\left[\left(\prod_m \mathbb{I}_{\left\{\max(\hat{U}_{m,j},\hat{U}_{m,k})\le u_m\right\}}+\prod_{m\in[M]} {u_m}^2\right)\right.$$

$$\left.-\prod_m u_m \mathbb{I}_{\left\{\hat{U}_{m,j}\le u_m\right\}}-\prod_m u_m \mathbb{I}_{\left\{\hat{U}_{m,k}\le u_m\right\}}\right]\mathrm{d}\mathbf{u}.$$

(i) $(a-b)(c-d)=ac+bd-ad-bc$

$$\int_{[0,1]^M} \prod_m \mathbb{I}_{\left\{\max(\hat{U}_{m,j},\hat{U}_{m,k})\le u_m\right\}} \mathrm{d}\mathbf{u} \stackrel{\text{(i)}}{=} \prod_m \int_{[0,1]} \mathbb{I}_{\left\{\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\le u_m\right\}} \mathrm{d}u_m$$

(i) $\int_{[0,1]^M} \prod_m f_m(u_m) \mathrm{d}\mathbf{u} = \prod_m \int_{[0,1]} f_m(u_m) \mathrm{d}u_m$

Computation of $(*)$ – continued

$$\int_{[0,1]^M} \prod_m \mathbb{I}_{\left\{\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\le u_m\right\}} \mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} \mathbb{I}_{\left\{\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\le u_m\right\}} \mathrm{d}u_m$$

$$\overset{(ii)}{=} \prod_m \left[1 - \max\left(\hat{U}_{m,j}, \hat{U}_{m,k}\right)\right],$$

(i) $\int_{[0,1]^M} \prod_m f_m(u_m) \mathrm{d}\mathbf{u} = \prod_m \int_{[0,1]} f_m(u_m) \mathrm{d}u_m$, (ii) $\int_{[0,1]} \mathbb{I}_{\{a\le u\}} \mathrm{d}u = 1 - a$

Computation of $(*)$ – continued

$$\int_{[0,1]^M} \prod_m \mathbb{I}_{\left\{\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\le u_m\right\}} \mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} \mathbb{I}_{\left\{\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\le u_m\right\}} \mathrm{d}u_m \overset{(ii)}{=} \prod_m \left[1-\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\right],$$

$$\int_{[0,1]^M} \prod_m u_m^2 \mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} u_m^2 \mathrm{d}\mathbf{u} = \prod_m \left[\frac{u_m^3}{3}\right]_0^1 = \frac{1}{3^M},$$

(i) $\int_{[0,1]^M} \prod_m f_m(u_m) \mathrm{d}\mathbf{u} = \prod_m \int_{[0,1]} f_m(u_m) \mathrm{d}u_m$, (ii) $\int_{[0,1]} \mathbb{I}_{\{a\le u\}} \mathrm{d}u = 1-a$

Computation of $(*)$ – continued

$$\int_{[0,1]^M}\prod_m \mathbb{I}_{\{\max(\hat{U}_{m,j},\hat{U}_{m,k})\le u_m\}}\mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} \mathbb{I}_{\{\max(\hat{U}_{m,j},\hat{U}_{m,k})\le u_m\}}\mathrm{d}u_m$$
$$\overset{(ii)}{=} \prod_m \left[1-\max(\hat{U}_{m,j},\hat{U}_{m,k})\right],$$

$$\int_{[0,1]^M}\prod_m u_m^2\mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} u_m^2\mathrm{d}\mathbf{u} = \prod_m \left[\frac{u_m^3}{3}\right]_0^1 = \frac{1}{3^M},$$

$$\int_{[0,1]^M}\prod_m u_m\mathbb{I}_{\{\hat{U}_{m,j}\le u_m\}}\mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} u_m\mathbb{I}_{\{\hat{U}_{m,j}\le u_m\}}\mathrm{d}u_m$$

(i) $\int_{[0,1]^M}\prod_m f_m(u_m)\mathrm{d}\mathbf{u} = \prod_m \int_{[0,1]} f_m(u_m)\mathrm{d}u_m$, (ii) $\int_{[0,1]}\mathbb{I}_{\{a\le u\}}\mathrm{d}u = 1-a$

$$\int_{[0,1]^M} \prod_m \mathbb{I}_{\left\{\max(\hat{U}_{m,j},\hat{U}_{m,k})\le u_m\right\}} \mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} \mathbb{I}_{\left\{\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\le u_m\right\}} \mathrm{d}u_m$$

$$\overset{(ii)}{=} \prod_m \left[1 - \max(\hat{U}_{m,j}, \hat{U}_{m,k})\right],$$

$$\int_{[0,1]^M} \prod_m u_m^2 \mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} u_m^2 \mathrm{d}\mathbf{u} = \prod_m \left[\frac{u_m^3}{3}\right]_0^1 = \frac{1}{3^M},$$

$$\int_{[0,1]^M} \prod_m u_m \mathbb{I}_{\{\hat{U}_{m,j}\le u_m\}} \mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} u_m \mathbb{I}_{\{\hat{U}_{m,j}\le u_m\}} \mathrm{d}u_m$$

$$\overset{(iii)}{=}$$

(i) $\int_{[0,1]^M} \prod_m f_m(u_m) \mathrm{d}\mathbf{u} = \prod_m \int_{[0,1]} f_m(u_m) \mathrm{d}u_m$, (ii) $\int_{[0,1]} \mathbb{I}_{\{a\le u\}} \mathrm{d}u = 1 - a$,

(iii) $\int_{[0,1]} u \mathbb{I}_{\{b\le u\}} \mathrm{d}u = \int_{[b,1]} u \mathrm{d}u = \left[\frac{u^2}{2}\right]_b^1 = \frac{1}{2} - \frac{b^2}{2} = \frac{1}{2}\left(1 - b^2\right)$.

Computation of $(*)$ – continued

$$\int_{[0,1]^M}\prod_m \mathbb{I}_{\left\{\max(\hat{U}_{m,j},\hat{U}_{m,k})\le u_m\right\}}\mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} \mathbb{I}_{\left\{\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\le u_m\right\}}\mathrm{d}u_m \overset{(ii)}{=} \prod_m \left[1-\max(\hat{U}_{m,j},\hat{U}_{m,k})\right],$$

$$\int_{[0,1]^M}\prod_m u_m^2 \mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} u_m^2 \mathrm{d}\mathbf{u} = \prod_m \left[\frac{u_m^3}{3}\right]_0^1 = \frac{1}{3^M},$$

$$\int_{[0,1]^M}\prod_m u_m \mathbb{I}_{\{\hat{U}_{m,j}\le u_m\}}\mathrm{d}\mathbf{u} \overset{(i)}{=} \prod_m \int_{[0,1]} u_m \mathbb{I}_{\{\hat{U}_{m,j}\le u_m\}}\mathrm{d}u_m \overset{(iii)}{=} \frac{1}{2^M}\prod_m \left(1-\hat{U}_{m,j}^2\right).$$

(i) $\int_{[0,1]^M}\prod_m f_m(u_m)\mathrm{d}\mathbf{u} = \prod_m \int_{[0,1]} f_m(u_m)\mathrm{d}u_m$, (ii) $\int_{[0,1]} \mathbb{I}_{\{a\le u\}}\mathrm{d}u = 1-a$,

(iii) $\int_{[0,1]} u\mathbb{I}_{\{b\le u\}}\mathrm{d}u = \int_{[b,1]} u\mathrm{d}u = \left[\frac{u^2}{2}\right]_b^1 = \frac{1}{2}-\frac{b^2}{2} = \frac{1}{2}\left(1-b^2\right)$.

(*): collecting the terms

$$(*) = \frac{1}{N^2} \sum_{j,k \in [N]} \prod_m \left[1 - \max\left(\hat{U}_{m,j}, \hat{U}_{m,k}\right)\right] + \frac{1}{3^M} - \frac{2}{N}\frac{1}{2^M} \sum_{j \in [N]} \prod_m \left(1 - \hat{U}_{m,j}^2\right).$$

Easiness came from

$\int_{[0,1]} q_m(u_m) \mathrm{d}u_m$, with quadratic q_m.

(*): collecting the terms

$$(*) = \frac{1}{N^2}\sum_{j,k\in[N]}\prod_m\left[1-\max\left(\hat{U}_{m,j},\hat{U}_{m,k}\right)\right] + \frac{1}{3^M} - \frac{2}{N}\frac{1}{2^M}\sum_{j\in[N]}\prod_m\left(1-\hat{U}_{m,j}^2\right).$$

Easiness came from

$\int_{[0,1]} q_m(u_m)\mathrm{d}u_m$, with quadratic q_m.

Next step

application in independence testing.

Independent testing

- Given $\prod_{m\in[M]}\mathcal{X}_m \ni (\mathbf{X}_n)_{n\in[N]} \sim \mathbb{P}$ samples, level $\alpha \in (0,1)$.
- Goal: to check if

$$H_0 : \mathbb{P} = \otimes_{m\in[M]}\mathbb{P}_m, \qquad H_1 : \mathbb{P} \neq \otimes_{m\in[M]}\mathbb{P}_m.$$

Independent testing

- Given $\prod_{m\in[M]} \mathcal{X}_m \ni (\mathbf{X}_n)_{n\in[N]} \sim \mathbb{P}$ samples, level $\alpha \in (0,1)$.
- Goal: to check if

$$H_0 : \mathbb{P} = \otimes_{m\in[M]} \mathbb{P}_m, \qquad H_1 : \mathbb{P} \neq \otimes_{m\in[M]} \mathbb{P}_m.$$

- Role of α: wanted $\mathbb{P}(\text{reject } H_0 | H_0) \leq \alpha$.

Independent testing

- Given $\prod_{m\in[M]} \mathcal{X}_m \ni (\mathbf{X}_n)_{n\in[N]} \sim \mathbb{P}$ samples, level $\alpha \in (0,1)$.
- Goal: to check if

$$H_0 : \mathbb{P} = \otimes_{m\in[M]} \mathbb{P}_m, \qquad H_1 : \mathbb{P} \neq \otimes_{m\in[M]} \mathbb{P}_m.$$

- Role of α: wanted $\mathbb{P}(\text{reject } H_0 | H_0) \leq \alpha$.
- Decision:
  - compute a test statistic: $T(\mathbf{X}_1, \ldots, \mathbf{X}_N)$. Example: $\hat{\Phi}_N^2$.
  - $F_T :=$ cdf of $T(\mathbf{X}_1, \ldots, \mathbf{X}_N)$ under H_0: null distribution.
  - $q :=$ its $(1-\alpha)$-quantile.
  - reject H_0 if $T(\mathbf{X}_1, \ldots, \mathbf{X}_N) > q$.

Independent testing

- Given $\prod_{m\in[M]} \mathcal{X}_m \ni (\mathbf{X}_n)_{n\in[N]} \sim \mathbb{P}$ samples, level $\alpha \in (0,1)$.
- Goal: to check if

$$H_0 : \mathbb{P} = \otimes_{m\in[M]} \mathbb{P}_m, \qquad H_1 : \mathbb{P} \neq \otimes_{m\in[M]} \mathbb{P}_m.$$

- Role of α: wanted $\mathbb{P}(\text{reject } H_0 | H_0) \le \alpha$.
- Decision:
  - compute a test statistic: $T(\mathbf{X}_1,\ldots,\mathbf{X}_N)$. Example: $\hat{\Phi}_N^2$.
  - $F_T :=$ cdf of $T(\mathbf{X}_1,\ldots,\mathbf{X}_N)$ under H_0: null distribution.
  - $q :=$ its $(1-\alpha)$-quantile.
  - reject H_0 if $T(\mathbf{X}_1,\ldots,\mathbf{X}_N) > q$.
- Good to have F_T, and the distribution of $T(\mathbf{X}_1,\ldots,\mathbf{X}_N)$ under the alternative (power).

Towards the null distribution for Φ^2

Warm-up: recall that $C \in L^\infty\left([0,1]^M\right)$. $\Rightarrow$ $\hat{C}_N$: $L^\infty\left([0,1]^M\right)$-valued r.v.

Towards the null distribution for Φ^2

Warm-up: recall that $C \in L^\infty\left([0,1]^M\right)$. $\Rightarrow \hat{C}_N$: $L^\infty\left([0,1]^M\right)$-valued r.v.

Theorem (Asymptotic behaviour of the empirical copula process)

Assume that $\partial^m C$ continuous for all $m \in [M]$. Then

$$\sqrt{N}\left(\hat{C}_N - C\right) \xrightarrow{w} \mathbb{G}_C,$$

$$\mathbb{G}_C(\mathbf{u}) = \mathbb{B}_C(\mathbf{u}) - \sum_{m\in[M]} \partial^m C(\mathbf{u})\mathbb{B}_C\left(\mathbf{u}^{(m)}\right),$$

$$\mathbf{u}^{(m)} = [1;\ldots;1;u_m;1;\ldots;1],$$

where $\mathbb{B}_C$ is a (tight) centered GP on $[0,1]^M$ with covariance function, $\wedge$ acts coordinate-wise,

$$\mathbb{E}\left[\mathbb{B}_C(\mathbf{u})\mathbb{B}_C(\mathbf{v})\right] = C(\mathbf{u}\wedge\mathbf{v}) - C(\mathbf{u})C(\mathbf{v}).$$

$\mathbb{B}_C$: is the so-called M-dimensional Brownian bridge.

Using the previus result

- If $C \neq \Pi$ (see H_1):

$$\sqrt{N}\left(\hat{\Phi}_N^2 - \Phi^2\right) \xrightarrow{w} N\left(0, \sigma^2\right),$$

$$\sigma^2 = [2h_2(M)]^2 \int_{[0,1]^M} \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})] \times \mathbb{E}[\mathbb{G}_C(\mathbf{u})\mathbb{G}_C(\mathbf{v})][C(\mathbf{v}) - \Pi(\mathbf{v})] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}.$$

Note: $C \neq \Pi$ guarantees that $\sigma^2 \neq 0$ (non-degenerate normal).

Using the previus result

- If $C \neq \Pi$ (see H_1):

$$\sqrt{N}\left(\hat{\Phi}_N^2 - \Phi^2\right) \xrightarrow{w} N\left(0, \sigma^2\right),$$

$$\sigma^2 = [2h_2(M)]^2 \int_{[0,1]^M} \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})] \times \mathbb{E}[\mathbb{G}_C(\mathbf{u})\mathbb{G}_C(\mathbf{v})][C(\mathbf{v}) - \Pi(\mathbf{v})] \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}.$$

  Note: $C \neq \Pi$ guarantees that $\sigma^2 \neq 0$ (non-degenerate normal).
- If $C = \Pi$ (see H_0):

$$\sqrt{N}(\hat{C}_N - \Pi) \xrightarrow{w} \mathbb{G}_\Pi, \quad N\hat{\Phi}_N^2 = h_2(M) \int_{[0,1]^M} N[\hat{C}_N(u) - \Pi(u)]^2 \mathrm{d}\mathbf{u}$$

Using the previus result

- If $C \neq \Pi$ (see H_1):

$$\sqrt{N}\left(\hat{\Phi}_N^2 - \Phi^2\right) \xrightarrow{w} N\left(0, \sigma^2\right),$$

$$\sigma^2 = [2h_2(M)]^2 \int_{[0,1]^M} \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})] \times \mathbb{E}[\mathbb{G}_C(\mathbf{u})\mathbb{G}_C(\mathbf{v})][C(\mathbf{v}) - \Pi(\mathbf{v})]\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}.$$

  Note: $C \neq \Pi$ guarantees that $\sigma^2 \neq 0$ (non-degenerate normal).

- If $C = \Pi$ (see H_0): by the continuous mapping theorem

$$\sqrt{N}(\hat{C}_N - \Pi) \xrightarrow{w} \mathbb{G}_\Pi, \quad N\hat{\Phi}_N^2 = h_2(M) \int_{[0,1]^M} N[\hat{C}_N(u) - \Pi(u)]^2 \mathrm{d}\mathbf{u}$$

$$\xrightarrow{w} \underbrace{h_2(M) \int_{[0,1]^M} [\mathbb{G}_\Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u}}_{\text{asymptotic null distribution (simulation)}}.$$

Using the previus result

- If $C \neq \Pi$ (see H_1):

$$\sqrt{N}\left(\hat{\Phi}_N^2 - \Phi^2\right) \xrightarrow{w} N\left(0, \sigma^2\right),$$

$$\sigma^2 = [2h_2(M)]^2 \int_{[0,1]^M} \int_{[0,1]^M} [C(\mathbf{u}) - \Pi(\mathbf{u})] \times \mathbb{E}[\mathbb{G}_C(\mathbf{u})\mathbb{G}_C(\mathbf{v})][C(\mathbf{v}) - \Pi(\mathbf{v})]\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}.$$

  Note: $C \neq \Pi$ guarantees that $\sigma^2 \neq 0$ (non-degenerate normal).

- If $C = \Pi$ (see H_0): by the continuous mapping theorem

$$\sqrt{N}(\hat{C}_N - \Pi) \xrightarrow{w} \mathbb{G}_\Pi, \quad N\hat{\Phi}_N^2 = h_2(M) \int_{[0,1]^M} N[\hat{C}_N(u) - \Pi(u)]^2 \mathrm{d}\mathbf{u}$$

$$\xrightarrow{w} \underbrace{h_2(M) \int_{[0,1]^M} [\mathbb{G}_\Pi(\mathbf{u})]^2 \mathrm{d}\mathbf{u}}_{\text{asymptotic null distribution (simulation)}}.$$

How was the $C \neq \Pi$ case obtained?

The delta method: idea

- We know $\sqrt{N}[\hat{C}_N - C] \xrightarrow{w} T := \mathbb{G}_C$.

The delta method: idea

- We know $\sqrt{N}[\hat{C}_N - C] \xrightarrow{w} T := \mathbb{G}_C$.
- Let $\varphi(C) := \Phi^2(C)$. One expects that

$$\begin{aligned}\sqrt{N}[\varphi(\hat{C}_N) - \varphi(C)] &\approx \sqrt{N}\varphi'_C(\hat{C}_N - C)\\ &\xrightarrow{w} \varphi'_C(T).\end{aligned}$$

The delta method: idea

- We know $\sqrt{N}[\hat{C}_N - C] \xrightarrow{w} T := \mathbb{G}_C$.
- Let $\varphi(C) := \Phi^2(C)$. One expects that

$$\begin{aligned}\sqrt{N}[\varphi(\hat{C}_N) - \varphi(C)] &\approx \sqrt{N}\varphi'_C(\hat{C}_N - C)\\ &\xrightarrow{w} \varphi'_C(T).\end{aligned}$$

Good news

This heuristic goes through with the right notion of differentiability [van der Vaart, 1998, Chapter 18, 20].

The delta method: idea

- We know $\sqrt{N}[\hat{C}_N - C] \xrightarrow{w} T := \mathbb{G}_C$.
- Let $\varphi(C) := \Phi^2(C)$. One expects that

$$\begin{aligned}\sqrt{N}[\varphi(\hat{C}_N) - \varphi(C)] &\approx \sqrt{N}\varphi'_C(\hat{C}_N - C)\\ &\xrightarrow{w} \varphi'_C(T).\end{aligned}$$

Good news

This heuristic goes through with the right notion of differentiability [van der Vaart, 1998, Chapter 18, 20].

- $\varphi : L^\infty\left([0,1]^M\right) \to \mathbb{R}$. Gateaux (directional) / Hadamard / Fréchet (classical). On $\mathbb{R}^d$: Hadamard = Fréchet.

Hadamard differentiability

- $\varphi : L^{\infty}\left([0,1]^M\right) \to \mathbb{R}$.
- Its derivative at copula $C \in L^{\infty}\left([0,1]^M\right)$, shortly φ'_C, is a continuous linear functional for which

$$\left|\frac{\varphi(C + t_n D_n) - \varphi(C)}{t_n} - \varphi'_C(D)\right| \xrightarrow{n\to\infty} 0$$

for all $t_n \xrightarrow{n\to\infty} 0$, $D_n \xrightarrow{n\to\infty} D$. Note: direction D_n might change with n, but eventually converge.

Computing the Hadamard derivative

$$\frac{\varphi(C+t_nD_n)-\varphi(C)}{t_n} = \frac{h_2(M)\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})+t_nD_n(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}$$
$$-\frac{h_2(M)\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}$$

(i) $(a+b)^2-a^2=2ab+b^2$.

Computing the Hadamard derivative

$$\frac{\varphi(C+t_nD_n)-\varphi(C)}{t_n} = \frac{h_2(M)\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})+t_nD_n(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}$$

$$- \frac{h_2(M)\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}$$

$$\overset{(i)}{=} \frac{2h_2(M)t_n\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})]D_n(\mathbf{u})\mathrm{d}\mathbf{u}}{t_n} + \underbrace{\frac{t_n^2\int_{[0,1]^M}[D_n(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}}_{\to 0}$$

(i) $(a+b)^2 - a^2 = 2ab + b^2$.

$$\frac{\varphi(C+t_nD_n)-\varphi(C)}{t_n} = \frac{h_2(M)\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})+t_nD_n(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}$$

$$-\ \frac{h_2(M)\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}$$

$$\overset{(i)}{=} \frac{2h_2(M)t_n\int_{[0,1]^M}[C(\mathbf{u})-\Pi(\mathbf{u})]D_n(\mathbf{u})\mathrm{d}\mathbf{u}}{t_n} + \underbrace{\frac{t_n^2\int_{[0,1]^M}[D_n(\mathbf{u})]^2\mathrm{d}\mathbf{u}}{t_n}}_{\to 0}$$

$$\to \int_{[0,1]^M}\underbrace{2h_2(M)[C(\mathbf{u})-\Pi(\mathbf{u})]D(\mathbf{u})}_{f_{C,D}(\mathbf{u})}\mathrm{d}\mathbf{u} = \boxed{\varphi'_C(D)}.$$

(i) $(a+b)^2 - a^2 = 2ab + b^2$.

Hadamard derivative: wrap up

We got that

$$\sqrt{N}[\varphi(\hat{C}_N) - \varphi(C)] \xrightarrow{w} \varphi'_C(\mathbb{G}_C)$$

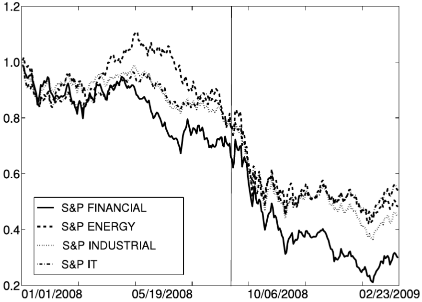
Using that $\mathbb{G}_C$ is a tight GP, with Lemma 3.9.8 [van der Vaart and Wellner, 2000] implies

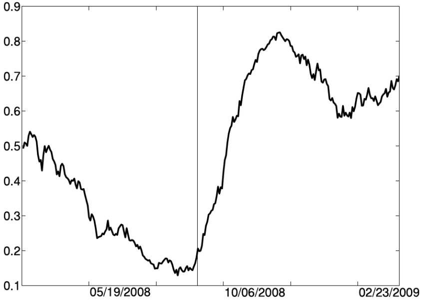
$$\varphi'_C(\mathbb{G}_C) = N\left(0, \sigma^2\right),$$

$$\sigma^2 = \int_{[0,1]^M} \int_{[0,1]^M} \mathbb{E}[f_{C,\mathbb{G}_C}(\mathbf{u}) f_{C,\mathbb{G}_C}(\mathbf{v})] \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v}$$

as claimed.

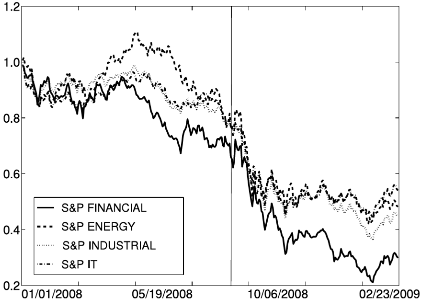
- We consider 4 global sector indices.

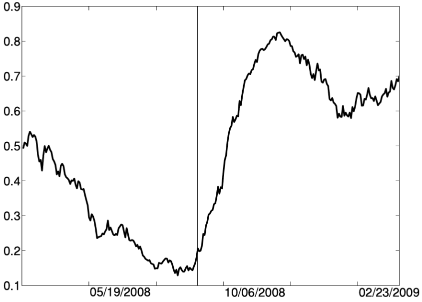




A financial application [Gaißer et al., 2010]

- We consider 4 global sector indices.

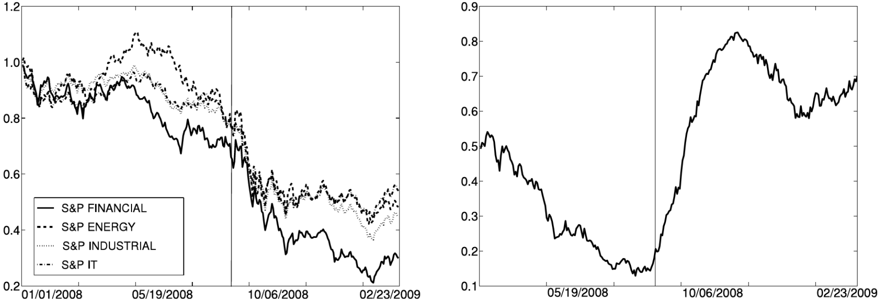




- Sept. 15, 2008: bankrupcy of Lehman Brothers Inc.

A financial application [Gaißer et al., 2010]

- We consider 4 global sector indices.



- Sept. 15, 2008: bankrupcy of Lehman Brothers Inc.
- Left: steep decay. Right: high dependency captured by $\hat{\Phi}^2_N$.

Hoeffding Φ^2: summary

- Valid independence measure: $\mathcal{X}_m = \mathbb{R}$, $M \geq 2$.
- Various favorable properties.
- Estimator: plug-in, analytic formula.
- Null distribution: continuous mapping theorem (Brownian bridge, simulation),
- Alternative: normal (delta method, Hadamard derivative).

Maximum correlation

Independence measures

- Given: random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim \mathbb{P}_{XY}$.
- Goal: measure the dependence of X and Y.

Independence measures

- Given: random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim \mathbb{P}_{XY}$.
- Goal: measure the dependence of X and Y.
- Desiderata for a $Q(\mathbb{P}_{XY})$ independence measure [Rényi, 1959]:

  1. $Q(\mathbb{P}_{XY})$ is well-defined,
  2. $Q(\mathbb{P}_{XY}) \in [0, 1]$,
  3. $Q(\mathbb{P}_{XY}) = 0$ iff. $X \perp Y$.
  4. $Q(\mathbb{P}_{XY}) = 1$ iff. $Y = f(X)$ or $X = g(Y)$.

Independence measures

- He showed:

$$Q(\mathbb{P}_{XY}) = \sup_{f,g:\ \text{measurable}} \mathrm{corr}(f(X), g(Y)),$$

  satisfies 1-4.
- Too ambitious:
  - computationally intractable.
  - many measurable functions.

Independence measures: measurable $\rightarrow$ continuous

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \text{ metric} \rightarrow \mathbb{R}, \text{ bounded continuous}\}$ would also work.
- Still too large!

Independence measures: measurable $\rightarrow$ continuous

- $C_b(\mathcal{X}) = \{f : \mathcal{X}$ metric $\rightarrow \mathbb{R}$, bounded continuous$\}$ would also work.
- Still too large!
- Idea: to take function spaces
  - dense in $C_b(\mathcal{X})$,
  - computionally tractable.

Independence measures: measurable $\rightarrow$ continuous

- $C_b(\mathcal{X}) = \{f : \mathcal{X} \text{ metric} \rightarrow \mathbb{R}, \text{ bounded continuous}\}$ would also work.
- Still too large!
- Idea: to take function spaces
  - dense in $C_b(\mathcal{X})$,
  - computionally tractable.

Key: Balance

denseness $\rightarrow$ universality, computation $\rightarrow$ RKHS.

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

Constructively, $\mathcal{H}_k = \overline{\left\{\sum_{i=1}^n \alpha_i k(\cdot, x_i)\right\}}$.

Kernel, RKHS: generalized inner product, -linear methods

- Def-1 (feature space):

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

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

Kernel, RKHS: generalized inner product, -linear methods

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

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

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad x, y \in \mathcal{X}.$$

- Def-2 (reproducing kernel):

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

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

- All these definitions are equivalent, $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k$.
- Examples on $\mathbb{R}^d$ ($\gamma > 0$, $p \in \mathbb{Z}^+$): $k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$, $k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}$, $k_e(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2}$.

Kernels [Steinwart and Christmann, 2008, Saitoh and Sawano, 2016]: various data types

- 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],
- 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, Draief et al., 2018, Bai et al., 2020, Borgwardt et al., 2020].

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated:
  - feature maps $\varphi(x) = k(\cdot, x)$, $\psi(y) = \ell(\cdot, y)$,
  - RKHS-s $\mathcal{H}_k$, $\mathcal{H}_\ell$.

KCCA: definition

- Given: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Associated:
  - feature maps $\varphi(x) = k(\cdot, x)$, $\psi(y) = \ell(\cdot, y)$,
  - RKHS-s $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- KCCA measure of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$

$$\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \mathrm{corr}(f(X), g(Y)),$$

$$\mathrm{corr}(f(X), g(Y)) = \frac{\mathrm{cov}(f(X), g(Y))}{\sqrt{\mathrm{var}[f(X)]\,\mathrm{var}[g(Y)]}}.$$

KCCA: notes

- Optimization domain: $\mathcal{H}_k \times \mathcal{H}_\ell \ni (f, g)$.
- By reproducing property: we will get a finite-D task.
- k,ℓ linear: traditional CCA.
- In practice: we have $\{(x_n, y_n)\}_{n=1}^N$ samples from (X, Y).

KCCA: notes

- Optimization domain: $\mathcal{H}_k \times \mathcal{H}_\ell \ni (f, g)$.
- By reproducing property: we will get a finite-D task.
- k,ℓ linear: traditional CCA.
- In practice: we have $\{(x_n, y_n)\}_{n=1}^N$ samples from (X, Y).

Recall the reproducing property

$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} \quad \forall f \in \mathcal{H}_k, x \in \mathcal{X}.$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i) \right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i) \right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i) \right\rangle_{\mathcal{H}_\ell}}$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i) \right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i) \right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i) \right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i) \right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

KCCA: empirical estimate

$$\widehat{\operatorname{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)\right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\operatorname{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]^2$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)\right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\mathrm{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]^2 = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}^2,$$

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]}_{\left\langle f, \varphi(x_n) - \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)\right\rangle_{\mathcal{H}_k}} \underbrace{\left[g(y_n) - \frac{1}{N}\sum_{i=1}^{N} g(y_i)\right]}_{\left\langle g, \psi(y_n) - \frac{1}{N}\sum_{i=1}^{N}\psi(y_i)\right\rangle_{\mathcal{H}_\ell}}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

Similarly:

$$\widehat{\mathrm{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N}\left[f(x_n) - \frac{1}{N}\sum_{i=1}^{N} f(x_i)\right]^2 = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}^2,$$

$$\widehat{\mathrm{var}}[g(Y)] = \frac{1}{N}\sum_{n=1}^{N} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}^2.$$

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ component of f $\perp$

$$span\left(\{\tilde{\varphi}(x_n)\}_{n=1}^N\right) = \left\{\sum_{n=1}^N c_n\tilde{\varphi}(x_n), \mathbf{c} = [c_n] \in \mathbb{R}^N\right\}$$

has no affect in the objective.

KCCA: empirical estimate

- f: appears only as $\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k}$ [similarly: g in $\langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell}$]. $\Rightarrow$
- $\forall$ component of f $\perp$

$$span\left(\{\tilde{\varphi}(x_n)\}_{n=1}^N\right) = \left\{\sum_{n=1}^N c_n\tilde{\varphi}(x_n), \mathbf{c} = [c_n] \in \mathbb{R}^N\right\}$$

has no affect in the objective.

Key idea

Enough to consider $f = \sum_{i=1}^N c_i\tilde{\varphi}(x_i)$, $g = \sum_{i=1}^N d_i\tilde{\psi}(y_i)$.

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k}$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \tilde{k}(x_i, x_n)$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^N c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^N d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^N c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^N c_i \tilde{k}(x_i, x_n) = (\mathbf{c}^T \tilde{\mathbf{G}}_X)_n,$$

KCCA: empirical estimate

Using that $f = \sum_{i=1}^{N} c_i \tilde{\varphi}(x_i)$, $g = \sum_{i=1}^{N} d_i \tilde{\psi}(y_i)$:

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{N} c_i \tilde{k}(x_i, x_n) = (\mathbf{c}^T \tilde{\mathbf{G}}_X)_n,$$

$$\langle g, \tilde{\psi}(y_n) \rangle_{\mathcal{H}_\ell} = (\mathbf{d}^T \tilde{\mathbf{G}}_Y)_n,$$

with the centered kernels ($\tilde{k}$, $\tilde{\ell}$) and Gram matrices ($\tilde{\mathbf{G}}_X$, $\tilde{\mathbf{G}}_Y$).

Until now

All the objective terms can be expressed by $\mathbf{c}$, $\mathbf{d}$, $\tilde{\mathbf{G}}_X$, $\tilde{\mathbf{G}}_Y$.

KCCA: empirical estimate

$$\widehat{\mathrm{cov}}(f(X), g(Y)) = \frac{1}{N} \sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n) \rangle_{\mathcal{H}_\ell},$$

$$\widehat{\mathrm{var}}[f(X)] = \frac{1}{N} \sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k}^2, \widehat{\mathrm{var}}[g(Y)] = \frac{1}{N} \sum_{n=1}^{N} \langle g, \tilde{\psi}(y_n) \rangle_{\mathcal{H}_\ell}^2,$$

and we have

$$\langle f, \tilde{\varphi}(x_n) \rangle_{\mathcal{H}_k} = (\mathbf{c}^T \tilde{\mathbf{G}}_X)_n, \qquad \langle g, \tilde{\psi}(y_n) \rangle_{\mathcal{H}_\ell} = (\mathbf{d}^T \tilde{\mathbf{G}}_Y)_n.$$

KCCA: empirical estimate

$$\widehat{\text{cov}}(f(X), g(Y)) = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell},$$

$$\widehat{\text{var}}[f(X)] = \frac{1}{N}\sum_{n=1}^{N} \langle f, \tilde{\varphi}(x_n)\rangle^2_{\mathcal{H}_k}, \widehat{\text{var}}[g(Y)] = \frac{1}{N}\sum_{n=1}^{N} \langle g, \tilde{\psi}(y_n)\rangle^2_{\mathcal{H}_\ell},$$

and we have

$$\langle f, \tilde{\varphi}(x_n)\rangle_{\mathcal{H}_k} = (\mathbf{c}^T \tilde{\mathbf{G}}_X)_n, \qquad \langle g, \tilde{\psi}(y_n)\rangle_{\mathcal{H}_\ell} = (\mathbf{d}^T \tilde{\mathbf{G}}_Y)_n.$$

Thus,

$$\widehat{\text{cov}}(f(X), g(Y)) = \frac{1}{N}\mathbf{c}^T \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d},$$

$$\widehat{\text{var}}[f(X)] = \frac{1}{N}\mathbf{c}^T (\tilde{\mathbf{G}}_X)^2 \mathbf{c}, \quad \widehat{\text{var}}[g(Y)] = \frac{1}{N}\mathbf{d}^T (\tilde{\mathbf{G}}_Y)^2 \mathbf{d}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\text{KCCA}}}^{\text{temp}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N, \mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}}{\sqrt{\mathbf{c}^T (\tilde{\mathbf{G}}_X)^2 \mathbf{c}} \sqrt{\mathbf{d}^T (\tilde{\mathbf{G}}_Y)^2 \mathbf{d}}}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{\text{KCCA}}}^{\text{temp}}(X,Y;\mathcal{H}_k,\mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N,\mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d}}{\sqrt{\mathbf{c}^T(\tilde{\mathbf{G}}_X)^2\mathbf{c}}\sqrt{\mathbf{d}^T(\tilde{\mathbf{G}}_Y)^2\mathbf{d}}}.$$

In practice ($\kappa > 0$):

$$\widehat{\rho_{\text{KCCA}}}(X,Y) := \widehat{\rho_{\text{KCCA}}}(X,Y;\mathcal{H}_k,\mathcal{H}_\ell,\kappa) = \sup_{\mathbf{c}\in\mathbb{R}^N,\mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d}}{\sqrt{\mathbf{c}^T\left(\tilde{\mathbf{G}}_X+\kappa\mathbf{I}_N\right)^2\mathbf{c}}\sqrt{\mathbf{d}^T\left(\tilde{\mathbf{G}}_Y+\kappa\mathbf{I}_N\right)^2\mathbf{d}}}.$$

KCCA: finite-D form

Empirical estimate of KCCA:

$$\widehat{\rho_{KCCA}}^{\text{temp}}(X,Y;\mathcal{H}_k,\mathcal{H}_\ell) = \sup_{\mathbf{c}\in\mathbb{R}^N,\mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d}}{\sqrt{\mathbf{c}^T(\tilde{\mathbf{G}}_X)^2\mathbf{c}}\sqrt{\mathbf{d}^T(\tilde{\mathbf{G}}_Y)^2\mathbf{d}}}.$$

In practice ($\kappa > 0$):

$$\widehat{\rho_{KCCA}}(X,Y) := \widehat{\rho_{KCCA}}(X,Y;\mathcal{H}_k,\mathcal{H}_\ell,\kappa)$$

$$= \sup_{\mathbf{c}\in\mathbb{R}^N,\mathbf{d}\in\mathbb{R}^N} \frac{\mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d}}{\sqrt{\mathbf{c}^T\left(\tilde{\mathbf{G}}_X+\kappa\mathbf{I}_N\right)^2\mathbf{c}}\sqrt{\mathbf{d}^T\left(\tilde{\mathbf{G}}_Y+\kappa\mathbf{I}_N\right)^2\mathbf{d}}}.$$

Question

How do we solve it?

KCCA: solution

Stationary points of $\widehat{\rho_{\text{KCCA}}}(X, Y)$:

$$\mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(X, Y)}{\partial \mathbf{c}}, \qquad \mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(X, Y)}{\partial \mathbf{d}},$$

which simplifies to

$$\tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d} = \frac{(\mathbf{c}^T \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d})(\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \mathbf{c}}{\mathbf{c}^T (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \mathbf{c}}, \quad \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X \mathbf{c} = \frac{(\mathbf{d}^T \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X \mathbf{c})(\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N)^2 \mathbf{d}}{\mathbf{d}^T (\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N)^2 \mathbf{d}}$$

KCCA: solution

Stationary points of $\widehat{\rho_{\text{KCCA}}}(X,Y)$:

$$\mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(X,Y)}{\partial \mathbf{c}}, \qquad \mathbf{0} = \frac{\partial \widehat{\rho_{\text{KCCA}}}(X,Y)}{\partial \mathbf{d}},$$

which simplifies to

$$\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d} = \frac{(\mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d})(\tilde{\mathbf{G}}_X+\kappa\mathbf{I}_N)^2\mathbf{c}}{\mathbf{c}^T(\tilde{\mathbf{G}}_X+\kappa\mathbf{I}_N)^2\mathbf{c}}, \; \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X\mathbf{c} = \frac{(\mathbf{d}^T\tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X\mathbf{c})(\tilde{\mathbf{G}}_Y+\kappa\mathbf{I}_N)^2\mathbf{d}}{\mathbf{d}^T(\tilde{\mathbf{G}}_Y+\kappa\mathbf{I}_N)^2\mathbf{d}}$$

Normalization:

- $(\mathbf{c},\mathbf{d})$: solution $\Rightarrow (a\mathbf{c}, b\mathbf{d})$: solution $a,b \in \mathbb{R}, \neq 0$.
- denominators $:= 1$.

KCCA: final task

Find the maximal eigenvalue, $\lambda := \mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d}$, of the generalized eigenvalue problem:

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_Y + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$$\mathbf{A}\mathbf{z} = \lambda\mathbf{B}\mathbf{z}.$$

KCCA: final task

Find the maximal eigenvalue, $\lambda := \mathbf{c}^T \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d}$, of the generalized eigenvalue problem:

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^T \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_Y + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$$\mathbf{A}\mathbf{z} = \lambda \mathbf{B}\mathbf{z}.$$

Questions

1. Is KCCA an independence measure? ($\Leftarrow$ universality)
2. Meaning/handling of the regularization (κ).
3. $M \geq 2$ components.
4. Computation of $\tilde{\mathbf{G}}_X$, $\tilde{\mathbf{G}}_Y$.

Q1 (indep. measure) $\Leftarrow$ universal k, ℓ

If $X \perp Y$, then $\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005b].

Q1 (indep. measure) $\Leftarrow$ universal k, ℓ

If $X \perp Y$, then $\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005b].
- Enough: universal kernel on a compact metric domain.

Q1 (indep. measure) $\Leftarrow$ universal k, ℓ

If $X \perp Y$, then $\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = 0$. Opposite direction:

- For 'rich' $\mathcal{H}_k$, $\mathcal{H}_\ell$
  [Bach and Jordan, 2002, Gretton et al., 2005b].
- Enough: universal kernel on a compact metric domain.
- Example ($\gamma > 0$):
  - Gaussian: $k(\mathbf{x}, \mathbf{x}') = e^{-\gamma \left\|\mathbf{x}-\mathbf{x}'\right\|_2^2}$.
  - Laplacian kernel: $k(\mathbf{x}, \mathbf{x}') = e^{-\gamma \left\|\mathbf{x}-\mathbf{x}'\right\|_2}$.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

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 $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

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 $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow C(\mathcal{X}) = C_b(\mathcal{X})$.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

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 $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow C(\mathcal{X}) = C_b(\mathcal{X})$.
- k: continuous, $\mathcal{X}$: compact $\Rightarrow k$: bounded.

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

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 $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow C(\mathcal{X}) = C_b(\mathcal{X})$.
- k: continuous, $\mathcal{X}$: compact $\Rightarrow k$: bounded.
- k: continuous, bounded $\Rightarrow \mathcal{H}_k \subset C(\mathcal{X})$ [Steinwart and Christmann, 2008].

Q1: universal kernel, $C(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \text{ continuous}\}$

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 $(C(\mathcal{X}), \|\cdot\|_\infty)$.

Assumptions

- $\mathcal{X}$ assumption $\Rightarrow C(\mathcal{X}) = C_b(\mathcal{X})$.
- k: continuous, $\mathcal{X}$: compact $\Rightarrow$ k: bounded.
- k: continuous, bounded $\Rightarrow \mathcal{H}_k \subset C(\mathcal{X})$ [Steinwart and Christmann, 2008].
- Extensions of c-universality to non-compact spaces:
  - c_0-universality, cc-universality, …[Carmeli et al., 2010, Sriperumbudur et al., 2010b, Simon-Gabriel and Schölkopf, 2018].

Q1: properties of universal kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

If k is universal, then

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

Q1: 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.

Q1: 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.
- $\varphi(x) = k(\cdot, x)$ is injective, i.e.

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

is a metric.

Q1: 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.
- $\varphi(x) = k(\cdot, x)$ is injective, i.e.

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

  is a metric.
- The normalized kernel (recall: corr)

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

  is universal.

Q1: universal Taylor kernels [Steinwart, 2001, Steinwart and Christmann, 2008]

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

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

Q1: 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 $\mathcal{X} := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le \sqrt{r} \right\}$.

Q1: 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 $f(t) = e^{\alpha t} \Rightarrow a_n = \frac{\alpha^n}{n!}$.

- $k(\mathbf{x}, \mathbf{y}) = e^{\alpha \langle \mathbf{x}, \mathbf{y} \rangle}$: 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.

- $k(\mathbf{x}, \mathbf{y}) = (1 - \langle \mathbf{x}, \mathbf{y} \rangle)^{-\alpha}$ 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 $\binom{b}{n} = \sum_{i=1}^{n} \frac{b-i+1}{i}$.

Q2 (κ)

In fact, we estimated

$$\rho_{\text{KCCA}}(X, Y; \mathcal{H}_k, \mathcal{H}_\ell, \kappa) = \sup_{f \in \mathcal{H}_k, g \in \mathcal{H}_\ell} \operatorname{corr}(f(X), g(Y); \kappa),$$

$$\operatorname{corr}(f(X), g(Y); \kappa) = \frac{\operatorname{cov}(f(X), g(Y))}{\sqrt{\operatorname{var}[f(X)] + \kappa \|f\|^2_{\mathcal{H}_k}} \sqrt{\operatorname{var}[g(Y)] + \kappa \|g\|^2_{\mathcal{H}_\ell}}}.$$

In fact, we estimated

$$\rho_{\mathrm{KCCA}}(X,Y;\mathcal{H}_k,\mathcal{H}_\ell,\kappa)=\sup_{f\in\mathcal{H}_k,g\in\mathcal{H}_\ell}\operatorname{corr}(f(X),g(Y);\kappa),$$

$$\operatorname{corr}(f(X),g(Y);\kappa)=\frac{\operatorname{cov}(f(X),g(Y))}{\sqrt{\operatorname{var}[f(X)]+\kappa\left\|f\right\|_{\mathcal{H}_k}^2}\sqrt{\operatorname{var}[g(Y)]+\kappa\left\|g\right\|_{\mathcal{H}_\ell}^2}}.$$

For consistent KCCA estimate:

- $\kappa_N\to 0$ [Leurgans et al., 1993](spline-RKHS), [Fukumizu et al., 2007] (general RKHS).
- analysis: covariance operators.

Q3 ($M \geq 2$): symmetry, other form

For

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_Y + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$([\mathbf{c}, \mathbf{d}], \lambda)$ solution $\Rightarrow$ $([-\mathbf{c}; \mathbf{d}], -\lambda)$: solution. Thus, eigenvalues:

$$\{\lambda_1, -\lambda_1, \ldots, \lambda_N, -\lambda_N\}.$$

Q3 ($M \geq 2$): symmetry, other form

For

$$\begin{bmatrix} \mathbf{0} & \tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \mathbf{c}^T\tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y\mathbf{d} \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_Y + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

$([\mathbf{c}, \mathbf{d}], \lambda)$ solution $\Rightarrow$ $([-\mathbf{c}; \mathbf{d}], -\lambda)$: solution. Thus, eigenvalues:

$$\{\lambda_1, -\lambda_1, \ldots, \lambda_N, -\lambda_N\}.$$

Adding the r.h.s. to both sides:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 & \tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X & (\tilde{\mathbf{G}}_\mathbf{Y} + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1+\lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

with eigenvalues $\{1+\lambda_1, 1-\lambda_1, \ldots, 1+\lambda_N, 1-\lambda_N\}$.

Q3 ($M \ge 2$)

2-variables $[(X, Y)]$:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_{\mathbf{X}} + \kappa \mathbf{I}_N)^2 & \tilde{\mathbf{G}}_X \tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y \tilde{\mathbf{G}}_X & (\tilde{\mathbf{G}}_{\mathbf{Y}} + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1+\lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_X + \kappa \mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

Q3 ($M \geq 2$)

2-variables $[(X, Y)]$:

$$\begin{bmatrix} (\tilde{\mathbf{G}}_{\mathbf{X}} + \kappa\mathbf{I}_N)^2 & \tilde{\mathbf{G}}_X\tilde{\mathbf{G}}_Y \\ \tilde{\mathbf{G}}_Y\tilde{\mathbf{G}}_X & (\tilde{\mathbf{G}}_{\mathbf{Y}} + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = (1+\lambda) \begin{bmatrix} (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_X + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

For M-variables (pairwise dependence):

$$\begin{bmatrix} (\tilde{\mathbf{G}}_1 + \kappa\mathbf{I}_N)^2 & \tilde{\mathbf{G}}_1\tilde{\mathbf{G}}_2 & \dots & \tilde{\mathbf{G}}_1\tilde{\mathbf{G}}_M \\ \tilde{\mathbf{G}}_2\tilde{\mathbf{G}}_1 & (\tilde{\mathbf{G}}_2 + \kappa\mathbf{I}_N)^2 & \dots & \tilde{\mathbf{G}}_2\tilde{\mathbf{G}}_M \\ \vdots & \vdots & & \vdots \\ \tilde{\mathbf{G}}_M\tilde{\mathbf{G}}_1 & \tilde{\mathbf{G}}_M\tilde{\mathbf{G}}_2 & \dots & (\tilde{\mathbf{G}}_M + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_M \end{bmatrix} =$$

$$\gamma \begin{bmatrix} (\tilde{\mathbf{G}}_1 + \kappa\mathbf{I}_N)^2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{G}}_2 + \kappa\mathbf{I}_N)^2 & \dots & \mathbf{0} \\ \vdots & \vdots & & \\ \mathbf{0} & \mathbf{0} & \dots & (\tilde{\mathbf{G}}_M + \kappa\mathbf{I}_N)^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_M \end{bmatrix}.$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$(\tilde{\mathbf{G}}_X)_{ij} = \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k}$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N \times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k}
\end{aligned}
$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_X)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_X)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_X)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_X)_{nm}
\end{aligned}$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j)\rangle_{\mathcal{H}_k}\\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N}\varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N}\varphi(x_m)\right\rangle_{\mathcal{H}_k}\\
&= (\mathbf{G}_X)_{ij} - \frac{1}{N}\sum_{m=1}^{N}(\mathbf{G}_X)_{im} - \frac{1}{N}\sum_{n=1}^{N}(\mathbf{G}_X)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N}(\mathbf{G}_X)_{nm}\\
&= \left(\mathbf{G}_X - \mathbf{G}_X\frac{\mathbf{E}_N}{N} - \frac{\mathbf{E}_N}{N}\mathbf{G}_X - \frac{\mathbf{E}_N}{N}\mathbf{G}_X\frac{\mathbf{E}_N}{N}\right)_{ij},
\end{aligned}
$$

Centered Gram matrix

In short

$\tilde{\mathbf{G}}_X = \mathbf{H}\mathbf{G}_X\mathbf{H}$ with $\mathbf{H} = \mathbf{I}_N - \frac{\mathbf{E}_N}{N}$; $\mathbf{H}; \mathbf{E}_N \in \mathbb{R}^{N\times N}$.

$$
\begin{aligned}
(\tilde{\mathbf{G}}_X)_{ij} &= \tilde{k}(x_i, x_j) = \langle \tilde{\varphi}(x_i), \tilde{\varphi}(x_j) \rangle_{\mathcal{H}_k} \\
&= \left\langle \varphi(x_i) - \frac{1}{N}\sum_{n=1}^{N} \varphi(x_n), \varphi(x_j) - \frac{1}{N}\sum_{m=1}^{N} \varphi(x_m) \right\rangle_{\mathcal{H}_k} \\
&= (\mathbf{G}_X)_{ij} - \frac{1}{N}\sum_{m=1}^{N} (\mathbf{G}_X)_{im} - \frac{1}{N}\sum_{n=1}^{N} (\mathbf{G}_X)_{ni} - \frac{1}{N^2}\sum_{n,m=1}^{N} (\mathbf{G}_X)_{nm} \\
&= \left(\mathbf{G}_X - \mathbf{G}_X\frac{\mathbf{E}_N}{N} - \frac{\mathbf{E}_N}{N}\mathbf{G}_X - \frac{\mathbf{E}_N}{N}\mathbf{G}_X\frac{\mathbf{E}_N}{N} \right)_{ij}, \\
&= (\mathbf{H}\mathbf{G}_X\mathbf{H})_{ij},
\end{aligned}
$$

$\mathbf{H}$: symmetric ($\mathbf{H} = \mathbf{H}^T$), idempotent ($\mathbf{H}^2 = \mathbf{H}$).

Maximal correlation: summary

- Independence measure ($M = 2$): with universal kernels.
- $M \geq 2$: pairwise independence.
- Universal kernels: various examples & constructions.
- Consistent estimation: ✓
- Computation: generalized eigenvalue task (almost closed-form).
- Image registration: it was KCCA.

Maximal correlation: summary

- Independence measure ($M = 2$): with universal kernels.
- $M \geq 2$: pairwise independence.
- Universal kernels: various examples & constructions.
- Consistent estimation: ✓
- Computation: generalized eigenvalue task (almost closed-form).
- Image registration: it was KCCA.

Questions

1. Analytic estimators (distance / kernel evaluations)?
2. Other usage of covariance?

Distances

Setting

- In this part: $X \in \mathbb{R}^{d_1}$, $Y \in \mathbb{R}^{d_2}$ $(M = 2)$.
- Characteristic function ($\exists$, $X \xrightarrow{1:1} \phi_X$):

$$\phi_X(\mathbf{t}) = \mathbb{E}\left[e^{i\langle \mathbf{t}, X\rangle}\right], \qquad \phi_Y(\mathbf{s}) = \mathbb{E}\left[e^{i\langle \mathbf{s}, Y\rangle}\right],$$

$$\phi_{XY}(\mathbf{t}, \mathbf{s}) = \mathbb{E}\left[e^{i(\langle \mathbf{t}, X\rangle + \langle \mathbf{s}, Y\rangle)}\right].$$

Setting

- In this part: $X \in \mathbb{R}^{d_1}$, $Y \in \mathbb{R}^{d_2}$ $(M = 2)$.
- Characteristic function ($\exists$, $X \xrightarrow{1:1} \phi_X$):

$$\phi_X(\mathbf{t}) = \mathbb{E}\left[e^{i\langle \mathbf{t}, X\rangle}\right], \qquad \phi_Y(\mathbf{s}) = \mathbb{E}\left[e^{i\langle \mathbf{s}, Y\rangle}\right],$$

$$\phi_{XY}(\mathbf{t}, \mathbf{s}) = \mathbb{E}\left[e^{i(\langle \mathbf{t}, X\rangle + \langle \mathbf{s}, Y\rangle)}\right].$$

- X and Y are independent iff.

$$\boxed{\phi_{XY}(\mathbf{t}, \mathbf{s}) = \phi_X(\mathbf{t})\phi_Y(\mathbf{s})} \quad \forall \mathbf{s} \in \mathbb{R}^{d_1}, \mathbf{t} \in \mathbb{R}^{d_2}.$$

dCov [Székely et al., 2007, Székely and Rizzo, 2009]

- Idea:

$$
\begin{aligned}
\mathrm{dCov}^2(X,Y) &= \left\|\phi_{XY} - \phi_X\phi_Y\right\|^2_{L^2(w)} \\
&= \int_{\mathbb{R}^{d_1+d_2}} \left[\phi_{XY}(\mathbf{s},\mathbf{t}) - \phi_X(\mathbf{s})\phi_Y(\mathbf{t})\right]^2 w(\mathbf{s},\mathbf{t})\mathrm{d}\mathbf{s}\mathrm{d}\mathbf{t},
\end{aligned}
$$

dCov [Székely et al., 2007, Székely and Rizzo, 2009]

- Idea:

$$\begin{aligned}
\mathrm{dCov}^2(X,Y) &= \|\phi_{XY} - \phi_X\phi_Y\|^2_{L^2(w)} \\
&= \int_{\mathbb{R}^{d_1+d_2}} [\phi_{XY}(\mathbf{s},\mathbf{t}) - \phi_X(\mathbf{s})\phi_Y(\mathbf{t})]^2 w(\mathbf{s},\mathbf{t})\mathrm{d}\mathbf{s}\mathrm{d}\mathbf{t}, \\
w(\mathbf{t},\mathbf{s}) &= \frac{1}{c(d_1)c(d_2)\|\mathbf{t}\|_2^{d_1+1}\|\mathbf{s}\|_2^{d_2+1}},
\end{aligned}$$

- Idea:

$$\begin{aligned}
\mathrm{dCov}^2(X, Y) &= \|\phi_{XY} - \phi_X\phi_Y\|^2_{L^2(w)} \\
&= \int_{\mathbb{R}^{d_1+d_2}} [\phi_{XY}(\mathbf{s}, \mathbf{t}) - \phi_X(\mathbf{s})\phi_Y(\mathbf{t})]^2 w(\mathbf{s}, \mathbf{t}) \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{t}, \\
w(\mathbf{t}, \mathbf{s}) &= \frac{1}{c(d_1)c(d_2)\|\mathbf{t}\|_2^{d_1+1}\|\mathbf{s}\|_2^{d_2+1}}, \\
c(d) &= \frac{\pi^{\frac{1+d}{2}}}{\Gamma\left(\frac{1+d}{2}\right)}.
\end{aligned}$$

- Idea:

$$\begin{aligned}
\mathrm{dCov}^2(X,Y) &= \|\phi_{XY} - \phi_X\phi_Y\|^2_{L^2(w)} \\
&= \int_{\mathbb{R}^{d_1+d_2}} [\phi_{XY}(\mathbf{s},\mathbf{t}) - \phi_X(\mathbf{s})\phi_Y(\mathbf{t})]^2 w(\mathbf{s},\mathbf{t}) \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{t}, \\
w(\mathbf{t},\mathbf{s}) &= \frac{1}{c(d_1)c(d_2)\|\mathbf{t}\|_2^{d_1+1}\|\mathbf{s}\|_2^{d_2+1}}, \\
c(d) &= \frac{\pi^{\frac{1+d}{2}}}{\Gamma\left(\frac{1+d}{2}\right)}.
\end{aligned}$$

- Enough for its finiteness: $\mathbb{E}\|X\|_2 < \infty$, $\mathbb{E}\|Y\|_2 < \infty$ (finite 1st moments).

dCov [Székely et al., 2007, Székely and Rizzo, 2009]

- Idea:

$$\mathrm{dCov}^2(X,Y) = \|\phi_{XY} - \phi_X\phi_Y\|^2_{L^2(w)}$$
$$= \int_{\mathbb{R}^{d_1+d_2}} [\phi_{XY}(\mathbf{s},\mathbf{t}) - \phi_X(\mathbf{s})\phi_Y(\mathbf{t})]^2 w(\mathbf{s},\mathbf{t})\mathrm{d}\mathbf{s}\mathrm{d}\mathbf{t},$$
$$w(\mathbf{t},\mathbf{s}) = \frac{1}{c(d_1)c(d_2)\|\mathbf{t}\|_2^{d_1+1}\|\mathbf{s}\|_2^{d_2+1}},$$
$$c(d) = \frac{\pi^{\frac{1+d}{2}}}{\Gamma\left(\frac{1+d}{2}\right)}.$$

- Enough for its finiteness: $\mathbb{E}\|X\|_2 < \infty$, $\mathbb{E}\|Y\|_2 < \infty$ (finite 1st moments).

By construction

X and Y are independent iff. $\boxed{\mathrm{dCov}(X,Y) = 0}$.

Estimator

- Given: $\{(X_n, Y_n)\}_{n\in[N]}$ samples.

Estimator

- Given: $\{(X_n, Y_n)\}_{n\in[N]}$ samples.
- Empirical characteristic functions:

$$\phi_X^N(\mathbf{t}) = \frac{1}{N}\sum_{n\in[N]} e^{i\langle \mathbf{t}, X_n\rangle}, \qquad \phi_Y^N(\mathbf{s}) = \frac{1}{N}\sum_{n\in[N]} e^{i\langle \mathbf{s}, Y_n\rangle},$$

$$\phi_{XY}^N(\mathbf{t}, \mathbf{s}) = \frac{1}{N}\sum_{n\in[N]} e^{i(\langle \mathbf{t}, X_n\rangle + \langle \mathbf{s}, Y_n\rangle)}.$$

Estimator

- Given: $\{(X_n, Y_n)\}_{n\in[N]}$ samples.
- Empirical characteristic functions:

$$\phi_X^N(\mathbf{t}) = \frac{1}{N}\sum_{n\in[N]} e^{i\langle \mathbf{t}, X_n\rangle}, \qquad \phi_Y^N(\mathbf{s}) = \frac{1}{N}\sum_{n\in[N]} e^{i\langle \mathbf{s}, Y_n\rangle},$$

$$\phi_{XY}^N(\mathbf{t},\mathbf{s}) = \frac{1}{N}\sum_{n\in[N]} e^{i(\langle \mathbf{t}, X_n\rangle + \langle \mathbf{s}, Y_n\rangle)}.$$

- Estimator (plug-in):

$$\widehat{\mathrm{dCov}}_N^2(X, Y) = \left\| \phi_{XY}^N - \phi_X^N \phi_Y^N \right\|_{L^2(w)}^2.$$

Estimator – continued

$$\left\|\phi_{XY}^N - \phi_X^N\phi_Y^N\right\|_{L^2(w)}^2 = \frac{1}{N^2}\sum_{k,l\in[N]} A_{kl}B_{kl},$$

$$a_{kl} = \|X_k - X_l\|_2\,,$$

$$\bar{a}_{k\cdot} = \frac{1}{N}\sum_{l\in[N]} a_{kl},\ \bar{a}_{\cdot l} = \frac{1}{N}\sum_{k\in[N]} a_{kl},$$

$$\bar{a}_{\cdot\cdot} = \frac{1}{N^2}\sum_{k,l\in[N]} a_{kl},$$

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot},$$

B_{kl} is defined similarly from $b_{kl} = \|Y_k - Y_l\|_2$.

The careful choice of w results in a pairwise distance based estimator.

Properties

- $\widehat{\mathrm{dCov}}_N(X, Y) \xrightarrow{N\to\infty} \mathrm{dCov}(X, Y)$ almost surely.

Properties

- $\widehat{\mathrm{dCov}}_N(X,Y) \xrightarrow{N\to\infty} \mathrm{dCov}(X,Y)$ almost surely.
- Alternative form (& explanation for the magic):

$$\begin{aligned}\mathrm{dCov}^2(X,Y) = {} & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|X-X'\|_2\|Y-Y'\|_2 \\ & + \mathbb{E}_{XX'}\left\|X-X'\right\|_2 \mathbb{E}_{YY'}\left\|Y-Y'\right\|_2 \\ & - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\left\|X-X'\right\|_2 \mathbb{E}_{Y'}\left\|Y-Y'\right\|_2\right].\end{aligned}$$

Properties

- $\widehat{\mathrm{dCov}}_N(X,Y) \xrightarrow{N\to\infty} \mathrm{dCov}(X,Y)$ almost surely.
- Alternative form (& explanation for the magic):

$$\begin{aligned}\mathrm{dCov}^2(X,Y) = {} & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|X-X'\|_2\|Y-Y'\|_2 \\ & + \mathbb{E}_{XX'}\|X-X'\|_2\,\mathbb{E}_{YY'}\|Y-Y'\|_2 \\ & - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\|X-X'\|_2\,\mathbb{E}_{Y'}\|Y-Y'\|_2\right].\end{aligned}$$

- Extension [Lyons, 2013]:

$$\begin{aligned}\mathrm{dCov}^2(X,Y) = {} & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_1(X,X')\rho_2(Y,Y') \\ & + \mathbb{E}_{XX'}(X,X')\,\mathbb{E}_{YY'}(Y,Y') \\ & - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_1(X,X')\,\mathbb{E}_{Y'}\rho_2(Y,Y')\right].\end{aligned}$$

Properties

- $\widehat{\mathrm{dCov}}_N(X,Y) \xrightarrow{N\to\infty} \mathrm{dCov}(X,Y)$ almost surely.
- Alternative form (& explanation for the magic):

$$\begin{aligned}\mathrm{dCov}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|X-X'\|_2\|Y-Y'\|_2\\ &+\mathbb{E}_{XX'}\|X-X'\|_2\,\mathbb{E}_{YY'}\|Y-Y'\|_2\\ &-2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\|X-X'\|_2\,\mathbb{E}_{Y'}\|Y-Y'\|_2\right].\end{aligned}$$

- Extension [Lyons, 2013]:

$$\begin{aligned}\mathrm{dCov}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_1(X,X')\rho_2(Y,Y')\\ &+\mathbb{E}_{XX'}(X,X')\,\mathbb{E}_{YY'}(Y,Y')\\ &-2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_1(X,X')\,\mathbb{E}_{Y'}\rho_2(Y,Y')\right].\end{aligned}$$

Questions (answer $\in$ next section)

Asymptotic null distribution? Valid choices of (ρ_1,ρ_2)? $M\geq 2$?

Summary: dCov

- Formulation: factorization of the joint characteristic function.
- dCov: $L^2(w)$-distance.
- Smart choice of $w \Rightarrow$ pairwise distance based estimator.
- Almost sure convergence of the plug-in estimator.
- Extendable to metric spaces.

Setting, mean embedding

- In this part:
  - $X = (X_m)_{m \in [M]} \in \times_{m \in [M]} \mathcal{X}_m$, $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ kernel.

Setting, mean embedding

- In this part:
  - $X = (X_m)_{m\in[M]} \in \times_{m\in[M]}\mathcal{X}_m$, $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ kernel.
- Common trick: feature of $\mathbb{P}$,

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{\text{example: } \mathbb{I}_{(-\infty,\cdot)}(x),\, e^{i\langle\cdot,x\rangle},\, e^{\langle\cdot,x\rangle} \text{ in } \mathbb{R}^d} \mathrm{d}\mathbb{P}(x).$$

Setting, mean embedding

- In this part:
  - $X = (X_m)_{m\in[M]} \in \times_{m\in[M]}\mathcal{X}_m$, $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ kernel.
- Common trick: feature of $\mathbb{P}$,

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \int_{\mathcal{X}} \underbrace{\varphi(x)}_{\text{example: } \mathbb{I}_{(-\infty,\cdot)}(x),\, e^{i\langle\cdot,x\rangle},\, e^{\langle\cdot,x\rangle} \text{ in } \mathbb{R}^d} \mathrm{d}\mathbb{P}(x).$$

- **Mean embedding** [Berlinet and Thomas-Agnan, 2004], [Smola et al., 2007]: $\varphi(x) := k(\cdot, x)$.

Characteristic property, universality

- Characteristic k [Fukumizu et al., 2008, Sriperumbudur et al., 2010a]: if $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ is injective.

Characteristic property, universality

- Characteristic k [Fukumizu et al., 2008, Sriperumbudur et al., 2010a]:

  if $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ is injective.
- Universal k: same but on finite signed measures ($\mathbb{F} = c_1\mathbb{P}_1 - c_2\mathbb{P}_2$, $c_1, c_2 \in \mathbb{R}^{\geq 0}$). Recall:
  - denseness in $C_b(\mathcal{X})$,
  - Taylor construction.

Characteristic property, universality

- **Characteristic k** [Fukumizu et al., 2008, Sriperumbudur et al., 2010a]:

  if $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ is injective.
- **Universal k**: same but on finite signed measures ($\mathbb{F} = c_1\mathbb{P}_1 - c_2\mathbb{P}_2$, $c_1, c_2 \in \mathbb{R}^{\geq 0}$). Recall:
  - denseness in $C_b(\mathcal{X})$,
  - Taylor construction.

Universal $\Rightarrow$ characteristic.

Maximum mean discrepancy

MMD [Gretton et al., 2012]:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} \overset{(*)}{=} \sup_{f \in B_k} \underbrace{\langle f, \mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{X\sim\mathbb{P}} f(X) - \mathbb{E}_{X\sim\mathbb{Q}} f(X)}.$$

Maximum mean discrepancy

M MD [Gretton et al., 2012]:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} \overset{(*)}{=} \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{X\sim\mathbb{P}} f(X) - \mathbb{E}_{X\sim\mathbb{Q}} f(X)}.$$

- MMD_k is metric $\Leftrightarrow$ k: characteristic.
- $(*)$: $\mathrm{MMD}_k \in$ IPM ($\sup_{f\in\mathcal{F}} \mathbb{P}f - \mathbb{Q}f$) [Zolotarev, 1983, Müller, 1997], an easy-to-estimate one!
- Mean trick: $\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k} = \mathbb{E}_{X\sim\mathbb{P}, X'\sim\mathbb{Q}} k(X, X')$.

μ_k, MMD_k

Applications:

- two-sample testing [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Borgwardt et al., 2006, Harchaoui et al., 2007, Gretton et al., 2012, Jitkrittum et al., 2016], and its differential private variant [Raj et al., 2019]; independence [Gretton et al., 2008, Pfister et al., 2018, Jitkrittum et al., 2017a] and goodness-of-fit testing [Jitkrittum et al., 2017b, Balasubramanian et al., 2021], causal discovery [Mooij et al., 2016, Pfister et al., 2018],
- domain adaptation [Zhang et al., 2013], -generalization [Blanchard et al., 2017], change-point detection [Harchaoui and Cappé, 2007], post selection inference [Yamada et al., 2018],
- kernel Bayesian inference [Song et al., 2011, Fukumizu et al., 2013], approximate Bayesian computation [Park et al., 2016], probabilistic programming [Schölkopf et al., 2015], model criticism [Lloyd et al., 2014, Kim et al., 2016],
- topological data analysis [Kusano et al., 2016],
- distribution classification [Muandet et al., 2011, Lopez-Paz et al., 2015, Zaheer et al., 2017], distribution regression [Szabó et al., 2016, Law et al., 2018],
- generative adversarial networks [Dziugaite et al., 2015, Li et al., 2015, Binkowski et al., 2018], understanding the dynamics of complex dynamical systems [Klus et al., 2018, Klus et al., 2019], . . .

Hilbert-Schmidt independence criterion [Gretton et al., 2005a]

$M \geq 2$: [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Pfister et al., 2018, Szabó and Sriperumbudur, 2018]):

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

$$k(x, x') := \prod_{m=1}^{M} k_m(x_m, x'_m), \quad \mathcal{X} = \times_{m\in[M]}\mathcal{X}_m.$$

Shorthand: $k = \otimes_m k_m$.

Hilbert-Schmidt independence criterion [Gretton et al., 2005a]

$M \geq 2$: [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Pfister et al., 2018, Szabó and Sriperumbudur, 2018]):

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

$$k(x, x') := \prod_{m=1}^M k_m(x_m, x'_m), \quad \mathcal{X} = \times_{m \in [M]} \mathcal{X}_m.$$

Shorthand: $k = \otimes_m k_m$.

Alternative view of HSIC (naming from $M = 2$)

$$\text{HSIC}_k(\mathbb{P}) = \|C\|_{\text{HS}},$$

$$C = \mathbb{E}\left[\otimes_{m \in [M]} \varphi_m(X_m)\right] - \otimes_{m \in [M]} \mathbb{E}[\varphi_m(X_m)].$$

Note: $ab^T \leftrightarrow a \otimes b$.

Intuition of $a \otimes b$, goal: $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.

Intuition of $a \otimes b$, goal: $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a \left(b^T g\right)$$

Intuition of $a \otimes b$, goal: $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a\left(b^T g\right) = a\langle b, g\rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

Intuition of $a \otimes b$, goal: $a := \varphi(x) \in \mathcal{H}_k$, $b := \psi(y) \in \mathcal{H}_\ell$

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a\left(b^T g\right) = a \langle b, g \rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

- Alternatively

$$\mathbb{R} \ni f^T \left(ab^T\right) g = \left(f^T a\right)\left(b^T g\right)$$

$ab^T : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ bilinear form.

- If $a \in \mathbb{R}^{d_1}$, $b \in \mathbb{R}^{d_2}$, then $ab^T \in \mathbb{R}^{d_1 \times d_2}$.
- For $g \in \mathbb{R}^{d_2}$

$$\left(ab^T\right) g = a \left(b^T g\right) = a \langle b, g \rangle \in \mathbb{R}^{d_1},$$

$ab^T : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$ linear mapping.

- Alternatively

$$\mathbb{R} \ni f^T \left(ab^T\right) g = \left(f^T a\right) \left(b^T g\right) = \langle f, a \rangle \langle g, b \rangle$$

$ab^T : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ bilinear form.

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

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

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

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

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2, n \in \mathbb{Z}^+ \right\}.$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

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

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2, n \in \mathbb{Z}^+ \right\}.$$

- Define inner product on $\mathcal{L}$, and extend by linearity

$$\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle := \langle a_1, a_2 \rangle_{\mathcal{H}_1} \langle b_1, b_2 \rangle_{\mathcal{H}_2} .$$

Definition of $a \otimes b$, $\mathcal{H}_1 \otimes \mathcal{H}_2$

- Given: $\mathcal{H}_1$, $\mathcal{H}_2$ Hilbert spaces.
- $a \otimes b : (f, g) \in \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$ is the bilinear form:

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

- Finite linear combinations of $a \otimes b$-s:

$$\mathcal{L} := \left\{ \sum_{i=1}^{n} c_i (a_i \otimes b_i), c_i \in \mathbb{R}, a_i \in \mathcal{H}_1, b_i \in \mathcal{H}_2, n \in \mathbb{Z}^+ \right\}.$$

- Define inner product on $\mathcal{L}$, and extend by linearity

$$\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle := \langle a_1, a_2 \rangle_{\mathcal{H}_1} \langle b_1, b_2 \rangle_{\mathcal{H}_2}.$$

- $\mathcal{H}_1 \otimes \mathcal{H}_2$: completion of $\mathcal{L}$.

Tensor product of M Hilbert spaces:

$$(a_1 \otimes \ldots \otimes a_M)(h_1, \ldots, h_M) = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m},$$

$a_1 \otimes \ldots \otimes a_M$, $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_M$ would work similarly

Tensor product of M Hilbert spaces:

$$(a_1 \otimes \ldots \otimes a_M)(h_1, \ldots, h_M) = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m},$$

$$\left\langle \otimes_{m=1}^{M} a_m, \otimes_{m=1}^{M} h_m \right\rangle = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m}.$$

$\Rightarrow$ HSIC for M-variables: ✓

$a_1 \otimes \ldots \otimes a_M$, $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_M$ would work similarly

Tensor product of M Hilbert spaces:

$$(a_1 \otimes \ldots \otimes a_M)(h_1, \ldots, h_M) = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m},$$

$$\left\langle \otimes_{m=1}^{M} a_m, \otimes_{m=1}^{M} h_m \right\rangle = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m}.$$

$\Rightarrow$ HSIC for M-variables: $\checkmark$

Also known for RKHS-s [Berlinet and Thomas-Agnan, 2004]

$\mathcal{H}_k = \otimes_{m \in [M]} \mathcal{H}_{k_m}$, $k = \otimes_{m \in [M]} k_m$.

$a_1 \otimes \ldots \otimes a_M$, $\mathcal{H}_1 \otimes \ldots \otimes \mathcal{H}_M$ would work similarly

Tensor product of M Hilbert spaces:

$$(a_1 \otimes \ldots \otimes a_M)(h_1, \ldots, h_M) = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m},$$

$$\left\langle \otimes_{m=1}^{M} a_m, \otimes_{m=1}^{M} h_m \right\rangle = \prod_{m=1}^{M} \langle a_m, h_m \rangle_{\mathcal{H}_m}.$$

$\Rightarrow$ HSIC for M-variables: ✓

Also known for RKHS-s [Berlinet and Thomas-Agnan, 2004]

$\mathcal{H}_k = \otimes_{m \in [M]} \mathcal{H}_{k_m}$, $k = \otimes_{m \in [M]} k_m$.

HS link: $HS(\mathcal{H}_2, \mathcal{H}_1) \cong \mathcal{H}_1 \otimes \mathcal{H}_2$

$h_1 \otimes h_2 \in HS(\mathcal{H}_2, \mathcal{H}_1)$, $\langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle_{HS} = \langle a_1, a_2 \rangle_{\mathcal{H}_1} \langle b_1, b_2 \rangle_{\mathcal{H}_2}$.

How the covariance operator represents covariance?

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $f \in \mathcal{H}_k$, $g \in \mathcal{H}_\ell$.

$$C = \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)] - \underbrace{\mathbb{E}[k(\cdot, X)]}_{=:\mu_X} \otimes \underbrace{\mathbb{E}[\ell(\cdot, Y)]}_{=:\mu_Y},$$

How the covariance operator represents covariance?

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $f \in \mathcal{H}_k$, $g \in \mathcal{H}_\ell$.

$$C = \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)] - \underbrace{\mathbb{E}[k(\cdot, X)]}_{=:\mu_X} \otimes \underbrace{\mathbb{E}[\ell(\cdot, Y)]}_{=:\mu_Y},$$

$$\boxed{\langle f, Cg \rangle_{\mathcal{H}_k}} = \langle f, \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)] g \rangle_{\mathcal{H}_k} - \langle f, (\mu_X \otimes \mu_Y) g \rangle_{\mathcal{H}_k}$$

$$\overset{(i)}{=}$$

(i): $\mathbb{E} \leftrightarrow \langle f, \cdot \rangle_{\mathcal{H}_k}$

How the covariance operator represents covariance?

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $f \in \mathcal{H}_k$, $g \in \mathcal{H}_\ell$.

$$C = \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)] - \underbrace{\mathbb{E}[k(\cdot, X)]}_{=:\mu_X} \otimes \underbrace{\mathbb{E}[\ell(\cdot, Y)]}_{=:\mu_Y},$$

$$\begin{aligned}\boxed{\langle f, Cg \rangle_{\mathcal{H}_k}} &= \langle f, \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)]g \rangle_{\mathcal{H}_k} - \langle f, (\mu_X \otimes \mu_Y)\, g \rangle_{\mathcal{H}_k} \\ &\overset{(i)}{=} \mathbb{E}\langle f, \underbrace{[k(\cdot, X) \otimes \ell(\cdot, Y)]g}_{k(\cdot,X)\langle \ell(\cdot,Y)],g\rangle_{\mathcal{H}_k}} \rangle_{\mathcal{H}_k} - \langle f, \underbrace{(\mu_X \otimes \mu_Y)\, g}_{\mu_X \langle \mu_Y, g\rangle_{\mathcal{H}_\ell}} \rangle_{\mathcal{H}_k}\end{aligned}$$

(i): $\mathbb{E} \leftrightarrow \langle f, \cdot \rangle_{\mathcal{H}_k}$, (ii) $(a \otimes b)c = a \langle b, c \rangle$

How the covariance operator represents covariance?

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $f \in \mathcal{H}_k$, $g \in \mathcal{H}_\ell$.

$$C = \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)] - \underbrace{\mathbb{E}[k(\cdot, X)]}_{=:\mu_X} \otimes \underbrace{\mathbb{E}[\ell(\cdot, Y)]}_{=:\mu_Y},$$

$$\begin{aligned}
\boxed{\langle f, Cg\rangle_{\mathcal{H}_k}} &= \langle f, \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)]g\rangle_{\mathcal{H}_k} - \langle f, (\mu_X \otimes \mu_Y)\, g\rangle_{\mathcal{H}_k} \\
&\overset{(i)}{=} \mathbb{E}\langle f, \underbrace{[k(\cdot, X) \otimes \ell(\cdot, Y)]g}_{k(\cdot,X)\langle \ell(\cdot,Y)],g\rangle_{\mathcal{H}_k}}\rangle_{\mathcal{H}_k} - \langle f, \underbrace{(\mu_X \otimes \mu_Y)\, g}_{\mu_X \langle \mu_Y, g\rangle_{\mathcal{H}_\ell}}\rangle_{\mathcal{H}_k} \\
&\overset{(iii)}{=} \boxed{\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]}.
\end{aligned}$$

(i): $\mathbb{E} \leftrightarrow \langle f, \cdot\rangle_{\mathcal{H}_k}$, (ii) $(a \otimes b)c = a\langle b, c\rangle$,(iii) (mean) reproducing property.

How the covariance operator represents covariance?

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $f \in \mathcal{H}_k$, $g \in \mathcal{H}_\ell$.

$$C = \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)] - \underbrace{\mathbb{E}[k(\cdot, X)]}_{=:\mu_X} \otimes \underbrace{\mathbb{E}[\ell(\cdot, Y)]}_{=:\mu_Y},$$

$$\begin{aligned}
\boxed{\langle f, Cg\rangle_{\mathcal{H}_k}} &= \langle f, \mathbb{E}[k(\cdot, X) \otimes \ell(\cdot, Y)]g\rangle_{\mathcal{H}_k} - \langle f, (\mu_X \otimes \mu_Y)\, g\rangle_{\mathcal{H}_k} \\
&\overset{(i)}{=} \mathbb{E}\langle f, \underbrace{[k(\cdot, X) \otimes \ell(\cdot, Y)]g}_{k(\cdot,X)\langle \ell(\cdot,Y)],g\rangle_{\mathcal{H}_k}} \rangle_{\mathcal{H}_k} - \langle f, \underbrace{(\mu_X \otimes \mu_Y)\, g}_{\mu_X\langle \mu_Y, g\rangle_{\mathcal{H}_\ell}} \rangle_{\mathcal{H}_k} \\
&\overset{(iii)}{=} \boxed{\mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]}.
\end{aligned}$$

(i): $\mathbb{E} \leftrightarrow \langle f, \cdot\rangle_{\mathcal{H}_k}$, (ii) $(a \otimes b)c = a\,\langle b, c\rangle$,(iii) (mean) reproducing property.

Basis of the KCCA consistency proof [Fukumizu et al., 2007].

Questions

1. Meaning of $\mu_k(\mathbb{P}) = \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\boxed{f(x) \in \mathcal{H}_k}} \mathrm{d}\mathbb{P}(x)$?
2. Easy-to-check descriptions of being characteristic?
3. HSIC demo: cocktail party.
4. HSIC estimation.
5. HSIC vs distance covariance?
6. When is HSIC a valid independence measure?
7. Application in hypothesis testing.

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

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

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

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

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

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

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

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

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

Bochner integral: properties

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

Bochner integral: properties

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

Bochner integral: properties

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

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

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

Bochner integral: properties

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

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

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

Next step

When is k characteristic (i.e. MMD_k metric)?

Non-characteristic kernel examples

Polynomial kernels [Sriperumbudur et al., 2010a]:

- $k(x, y) = \langle x, y \rangle$: linear kernel ($L = 1$).

$$\text{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \|m_{\mathbb{P}} - m_{\mathbb{Q}}\|_2^2, \qquad m_{\mathbb{P}} = \int_{\mathcal{X}} x \mathrm{d}\mathbb{P}(x).$$

Non-characteristic kernel examples

Polynomial kernels [Sriperumbudur et al., 2010a]:

- $k(x,y) = \langle x, y\rangle$: linear kernel ($L=1$).

$$\text{MMD}_k^2(\mathbb{P},\mathbb{Q}) = \|m_{\mathbb{P}} - m_{\mathbb{Q}}\|_2^2, \qquad m_{\mathbb{P}} = \int_{\mathcal{X}} x \mathrm{d}\mathbb{P}(x).$$

- $k(x,y) = (\langle x, y\rangle + 1)^2$ ($L=2$):

$$\text{MMD}_k^2(\mathbb{P},\mathbb{Q}) = 2\|m_{\mathbb{P}} - m_{\mathbb{Q}}\|_2^2 + \left\| \Sigma_{\mathbb{P}} - \Sigma_{\mathbb{Q}} + m_{\mathbb{P}} m_{\mathbb{P}}^T - m_{\mathbb{Q}} m_{\mathbb{Q}}^T \right\|_F^2,$$

where $\|\cdot\|_F$: Frobenius norm; $\Sigma_{\mathbb{P}}$: cov. matrix w.r.t. $\mathbb{P}$.

Characteristic property of k

We focus on continuous bounded shift-invariant kernels:

Bochner's theorem [Wendland, 2005]

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

where Λ is a finite Borel measure (w.l.o.g. probability).

We expect it to be encoded in Λ! First, examples.

Shift-invariant kernels on $\mathbb{R}$ [Sriperumbudur et al., 2010a]

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

kernel name	k_0	$\widehat{k_0}(\omega)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$
Laplacian	$e^{-\sigma\|x\|}$	$\sqrt{\frac{2}{\pi}} \frac{\sigma}{\sigma^2+\omega^2}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}} \frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\|j\|}\delta(\omega-j)$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\delta(\omega-j)$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\|j\|}{n+1}\right)\delta(\omega-j)$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$

Shift-invariant kernels on $\mathbb{R}$ [Sriperumbudur et al., 2010a]

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

kernel name	k_0	$\widehat{k_0}(\omega)$
Gaussian	$e^{-\frac{x^2}{2\sigma^2}}$	$\sigma e^{-\frac{\sigma^2\omega^2}{2}}$
Laplacian	$e^{-\sigma\lvert x\rvert}$	$\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$
B_{2n+1}-spline	$*^{2n+2}\chi_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$
Dirichlet	$\frac{\sin\left(\frac{(2n+1)x}{2}\right)}{\sin\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\delta(\omega-j)$
Fejér	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\left(\frac{x}{2}\right)}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$

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

MMD in terms of kernel evaluations

$$\begin{aligned}\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2 \\ &= \left\| \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2\end{aligned}$$

MMD in terms of kernel evaluations

$$\begin{aligned}
\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2 \\
&= \left\| \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2 \\
&= \langle a - b, a - b \rangle_{\mathcal{H}_k}
\end{aligned}$$

MMD in terms of kernel evaluations

$$\begin{aligned}
\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2 \\
&= \left\| \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2 \\
&= \langle a - b, a - b\rangle_{\mathcal{H}_k} \\
&= \langle \mu_k(\mathbb{P}), \mu_k(\mathbb{P})\rangle_{\mathcal{H}_k} + \langle \mu_k(\mathbb{Q}), \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k} - 2\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}
\end{aligned}$$

MMD in terms of kernel evaluations

$$
\begin{aligned}
\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2 \\
&= \left\| \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y) \right\|_{\mathcal{H}_k}^2 \\
&= \langle a - b, a - b\rangle_{\mathcal{H}_k} \\
&= \langle \mu_k(\mathbb{P}), \mu_k(\mathbb{P})\rangle_{\mathcal{H}_k} + \langle \mu_k(\mathbb{Q}), \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k} - 2\langle \mu_k(\mathbb{P}), \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k} \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{P}(x') + \int_{\mathcal{X}}\int_{\mathcal{X}} k(y,y')\mathrm{d}\mathbb{Q}(y)\mathrm{d}\mathbb{Q}(y') \\
&\quad - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{Q}(y)
\end{aligned}
$$

MMD in terms of kernel evaluations

$$\begin{aligned}
\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \|\mu_k(\mathbb{P})-\mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2 \\
&= \left\|\int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot,y)\mathrm{d}\mathbb{Q}(y)\right\|_{\mathcal{H}_k}^2 \\
&= \langle a-b, a-b\rangle_{\mathcal{H}_k} \\
&= \langle \mu_k(\mathbb{P}),\mu_k(\mathbb{P})\rangle_{\mathcal{H}_k} + \langle \mu_k(\mathbb{Q}),\mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k} - 2\langle \mu_k(\mathbb{P}),\mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k} \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{P}(x') + \int_{\mathcal{X}}\int_{\mathcal{X}} k(y,y')\mathrm{d}\mathbb{Q}(y)\mathrm{d}\mathbb{Q}(y') \\
&\quad - 2\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{Q}(y) \\
&=: \boxed{\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)}.
\end{aligned}$$

MMD in terms of characteristic functions

Using Bochner's theorem ($\mathcal{X} = \mathbb{R}^d$):

$$\text{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(\mathbf{x}, \mathbf{y}) \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{x}) \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{y})$$

MMD in terms of characteristic functions

Using Bochner's theorem ($\mathcal{X} = \mathbb{R}^d$):

$$\begin{aligned}\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(\mathbf{x},\mathbf{y})\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle}\,\mathrm{d}\Lambda(\boldsymbol{\omega})\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\end{aligned}$$

Using Bochner's theorem ($\mathcal{X} = \mathbb{R}^d$):

$$\begin{aligned}
\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(\mathbf{x}, \mathbf{y}) \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{x}) \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{y}) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y}, \omega \rangle} \mathrm{d}\Lambda(\omega) \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{x}) \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{y}) \\
&= \int_{\mathbb{R}^d} \underbrace{\left[\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}, \omega \rangle} \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{x}) \right]}_{\overline{c_{\mathbb{P}}(\omega) - c_{\mathbb{Q}}(\omega)}} \underbrace{\left[\int_{\mathbb{R}^d} e^{i\langle \mathbf{y}, \omega \rangle} \mathrm{d}(\mathbb{P} - \mathbb{Q})(\mathbf{y}) \right]}_{c_{\mathbb{P}}(\omega) - c_{\mathbb{Q}}(\omega)} \mathrm{d}\Lambda(\omega)
\end{aligned}$$

MMD in terms of characteristic functions

Using Bochner's theorem ($\mathcal{X} = \mathbb{R}^d$):

$$\begin{aligned}
\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(\mathbf{x},\mathbf{y})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle}\mathrm{d}\Lambda(\boldsymbol{\omega})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\\
&= \int_{\mathbb{R}^d} \underbrace{\left[\int_{\mathbb{R}^d} e^{-i\langle\mathbf{x},\boldsymbol{\omega}\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\right]}_{\overline{c_{\mathbb{P}}(\boldsymbol{\omega})-c_{\mathbb{Q}}(\boldsymbol{\omega})}} \underbrace{\left[\int_{\mathbb{R}^d} e^{i\langle\mathbf{y},\boldsymbol{\omega}\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\right]}_{c_{\mathbb{P}}(\boldsymbol{\omega})-c_{\mathbb{Q}}(\boldsymbol{\omega})}\mathrm{d}\Lambda(\boldsymbol{\omega})\\
&= \int_{\mathbb{R}^d}\left|c_{\mathbb{P}}(\boldsymbol{\omega})-c_{\mathbb{Q}}(\boldsymbol{\omega})\right|^2\mathrm{d}\Lambda(\boldsymbol{\omega})
\end{aligned}$$

MMD in terms of characteristic functions

Using Bochner's theorem ($\mathcal{X} = \mathbb{R}^d$):

$$
\begin{aligned}
\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(\mathbf{x},\mathbf{y})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\omega\rangle}\mathrm{d}\Lambda(\omega)\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\\
&= \int_{\mathbb{R}^d} \underbrace{\left[\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x},\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{x})\right]}_{\overline{c_{\mathbb{P}}(\omega)-c_{\mathbb{Q}}(\omega)}} \underbrace{\left[\int_{\mathbb{R}^d} e^{i\langle \mathbf{y},\omega\rangle}\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mathbf{y})\right]}_{c_{\mathbb{P}}(\omega)-c_{\mathbb{Q}}(\omega)} \mathrm{d}\Lambda(\omega)\\
&= \int_{\mathbb{R}^d} \left|c_{\mathbb{P}}(\omega)-c_{\mathbb{Q}}(\omega)\right|^2 \mathrm{d}\Lambda(\omega) = \left\|c_{\mathbb{P}}-c_{\mathbb{Q}}\right\|_{L^2(\Lambda)}^2.
\end{aligned}
$$

Simple description for shift-invariant kernels on $\mathbb{R}^d$

Theorem ([Sriperumbudur et al., 2010a])

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

Theorem ([Sriperumbudur et al., 2010a])

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

Example on $\mathbb{R}$:

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	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\chi_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$

or the Matérn kernel (next slide).

Matérn kernel

$$k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right),$$

where K_{ν}: modified Bessel function of the second kind of order ν

Matérn kernel

$$k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y}) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right)^{v} K_v\left(\frac{\sqrt{2v}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right),$$

$$\widehat{k_0}(\boldsymbol{\omega}) = \frac{2^{d+v}\pi^{\frac{d}{2}}\Gamma(v+d/2)v^{v}}{\Gamma(v)\sigma^{2v}}\left(\frac{2v}{\sigma^2}+4\pi^2\|\boldsymbol{\omega}\|_2^2\right)^{-(v+d/2)} > 0 \quad \forall \boldsymbol{\omega}\in\mathbb{R}^d,$$

where K_v: modified Bessel function of the second kind of order v, Γ: Gamma function.

Matérn kernel

$$k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right),$$

$$\widehat{k_0}(\boldsymbol{\omega}) = \frac{2^{d+\nu}\pi^{\frac{d}{2}}\Gamma(\nu+d/2)\nu^{\nu}}{\Gamma(\nu)\sigma^{2\nu}}\left(\frac{2\nu}{\sigma^2}+4\pi^2\|\boldsymbol{\omega}\|_2^2\right)^{-(\nu+d/2)} > 0 \quad \forall\boldsymbol{\omega}\in\mathbb{R}^d,$$

where K_ν: modified Bessel function of the second kind of order ν, Γ: Gamma function.

- For $\nu = \frac{1}{2}$: one gets $k(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}}$.

Matérn kernel

$$k(\mathbf{x},\mathbf{y}) = k_0(\mathbf{x}-\mathbf{y}) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right)^{v} K_v\left(\frac{\sqrt{2v}\,\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}\right),$$

$$\widehat{k_0}(\boldsymbol{\omega}) = \frac{2^{d+v}\pi^{\frac{d}{2}}\Gamma(v+d/2)v^v}{\Gamma(v)\sigma^{2v}}\left(\frac{2v}{\sigma^2}+4\pi^2\|\boldsymbol{\omega}\|_2^2\right)^{-(v+d/2)} > 0 \quad \forall\boldsymbol{\omega}\in\mathbb{R}^d,$$

where K_v: modified Bessel function of the second kind of order v, Γ: Gamma function.

- For $v=\frac{1}{2}$: one gets $k(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_2}{\sigma}}$.
- Gaussian kernel: $v\to\infty$.

Characteristic kernel: 2 notes

1. B-spline kernel type kernels on $\mathbb{R}^d$:
   - k: still continuous, bounded, shift-invariant.
   - B-spline kernel: $supp(k_0)$ is compact $\leftarrow$ practically relevant.

Characteristic kernel: 2 notes

1. B-spline kernel type kernels on $\mathbb{R}^d$:
   - k: still continuous, bounded, shift-invariant.
   - B-spline kernel: $supp(k_0)$ is compact $\leftarrow$ practically relevant.

Theorem ([Sriperumbudur et al., 2010a])

$supp(k_0)$: compact $\Rightarrow$ k is characteristic.

Characteristic kernel: 2 notes

1. B-spline kernel type kernels on $\mathbb{R}^d$:
   - k: still continuous, bounded, shift-invariant.
   - B-spline kernel: $supp(k_0)$ is compact $\leftarrow$ practically relevant.

Theorem ([Sriperumbudur et al., 2010a])

$supp(k_0)$: compact $\Rightarrow$ k is characteristic.

2. Radial, bounded, continuous kernels on $\mathbb{R}^d$:

$$k(\mathbf{x}, \mathbf{y}) = k_0(\|\mathbf{x} - \mathbf{y}\|_2), \qquad k_0(z) = \int_{[0,\infty)} e^{-tz^2} \mathrm{d}\nu(t).$$

Characteristic kernel: 2 notes

1. B-spline kernel type kernels on $\mathbb{R}^d$:
   - k: still continuous, bounded, shift-invariant.
   - B-spline kernel: $supp(k_0)$ is compact $\leftarrow$ practically relevant.

Theorem ([Sriperumbudur et al., 2010a])

$supp(k_0)$: compact $\Rightarrow$ k is characteristic.

2. Radial, bounded, continuous kernels on $\mathbb{R}^d$:

$$k(\mathbf{x}, \mathbf{y}) = k_0(\|\mathbf{x} - \mathbf{y}\|_2), \qquad k_0(z) = \int_{[0,\infty)} e^{-tz^2} \mathrm{d}\nu(t).$$

Theorem ([Sriperumbudur et al., 2010a])

k is characteristic $\Leftrightarrow$ $supp(\nu) \neq \{0\}$.

Characteristic kernel: 2 notes

1. B-spline kernel type kernels on $\mathbb{R}^d$:
   - k: still continuous, bounded, shift-invariant.
   - B-spline kernel: $supp(k_0)$ is compact $\leftarrow$ practically relevant.

Theorem ([Sriperumbudur et al., 2010a])

$supp(k_0)$: compact $\Rightarrow$ k is characteristic.

2. Radial, bounded, continuous kernels on $\mathbb{R}^d$:

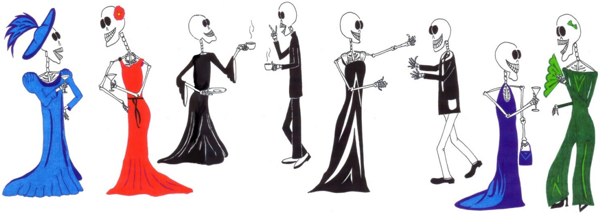
$$k(\mathbf{x},\mathbf{y}) = k_0(\|\mathbf{x}-\mathbf{y}\|_2), \qquad k_0(z) = \int_{[0,\infty)} e^{-tz^2} \mathrm{d}\nu(t).$$

Theorem ([Sriperumbudur et al., 2010a])

k is characteristic $\Leftrightarrow$ $supp(\nu) \neq \{0\}$.

We are switching to HSIC, demo first.

Cocktail party: HSIC demo



ISA reminder

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right],$$

where $\mathbf{s}^m$-s are non-Gaussian & independent.

- Goal: $\{\mathbf{x}_t\}_{t=1}^T \to \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T$,

ISA reminder

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad \mathbf{s} = \left[\mathbf{s}^1; \ldots; \mathbf{s}^M\right],$$

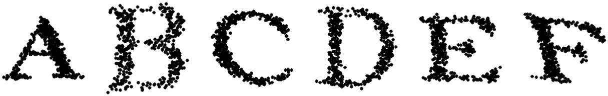
where $\mathbf{s}^m$-s are non-Gaussian & independent.

- Goal: $\{\mathbf{x}_t\}_{t=1}^T \to \mathbf{W} = \mathbf{A}^{-1}, \{\mathbf{s}_t\}_{t=1}^T$,
- Objective function:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x},$$
$$J(\mathbf{W}) = I\left(\hat{\mathbf{s}}^1, \ldots, \hat{\mathbf{s}}^M\right) \to \min_{\mathbf{W}}.$$

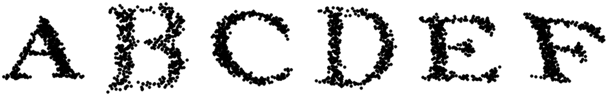
ISA: source, observation

- Hidden sources ($\mathbf{s}$):

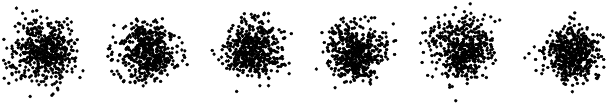


ISA: source, observation

- Hidden sources ($\mathbf{s}$):



- Observation ($\mathbf{x}$):



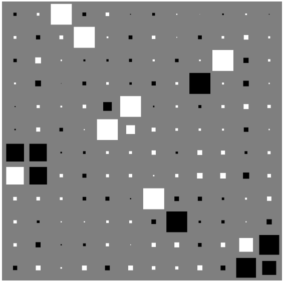
ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

ISA: estimated sources using HSIC, ambiguity

- Estimated sources ($\hat{\mathbf{s}}$):

- Performance ($\hat{\mathbf{W}}\mathbf{A}$), ambiguity:

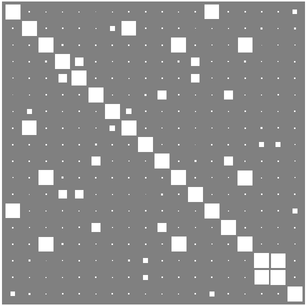


Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation.

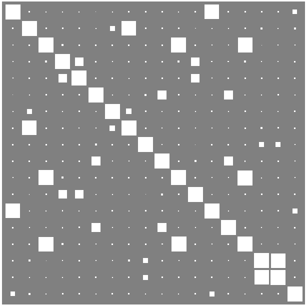
Conjecture: ISA separation theorem [Cardoso, 1998]

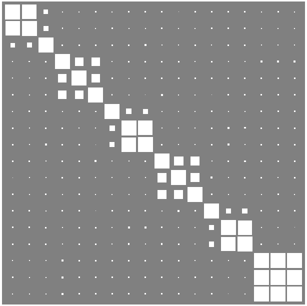
- ISA = ICA + permutation. $\widehat{\text{HSIC}}(\hat{s}_i, \hat{s}_j)$,



Conjecture: ISA separation theorem [Cardoso, 1998]

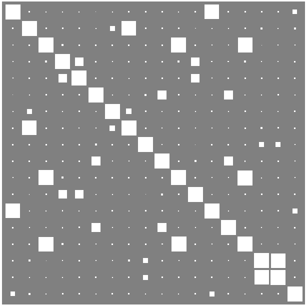
- ISA = ICA + permutation. $\widehat{\text{HSIC}}(\hat{s}_i, \hat{s}_j)$,

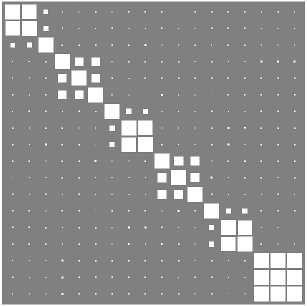




Conjecture: ISA separation theorem [Cardoso, 1998]

- ISA = ICA + permutation. $\widehat{\mathrm{HSIC}}(\hat{s}_i, \hat{s}_j)$,

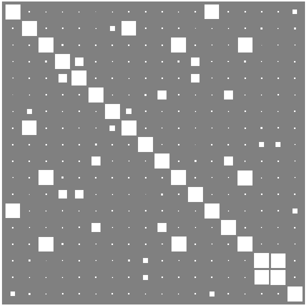


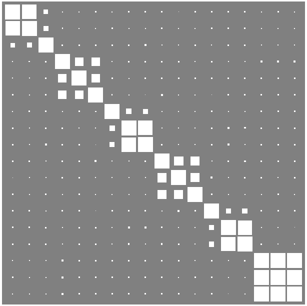


- Basis of the state-of-the-art ISA solvers.

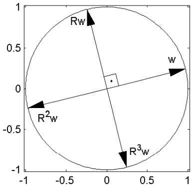
Conjecture: ISA separation theorem [Cardoso, 1998]

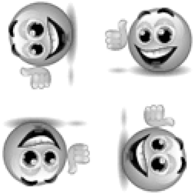
- ISA = ICA + permutation. $\widehat{\mathrm{HSIC}}(\hat{s}_i, \hat{s}_j)$,





- Basis of the state-of-the-art ISA solvers.
- Sufficient conditions [Szabó et al., 2012]:
  - $\mathbf{s}^m$: spherical [Fang et al., 1990].

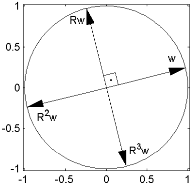


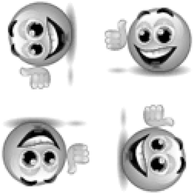


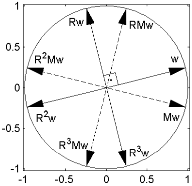
Invariance to

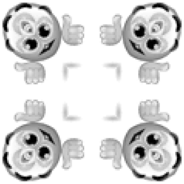
- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.

ISA separation theorem $\rightarrow$ for $dim(\mathbf{s}^m) = 2$ less is enough.





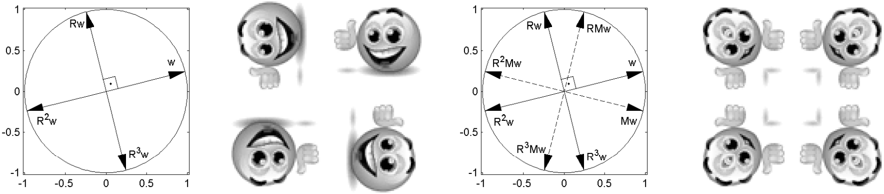




Invariance to

- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.
- permutation and sign: $f(\pm u_1, \pm u_2) = f(\pm u_2, \pm u_1)$.

ISA separation theorem → for $dim(\mathbf{s}^m) = 2$ less is enough.



Invariance to

- 90° rotation: $f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1)$.
- permutation and sign: $f(\pm u_1, \pm u_2) = f(\pm u_2, \pm u_1)$.
- L^p-spherical: $f(u_1, u_2) = h\left(\sum_i |u_i|^p\right) \quad (p > 0)$.

Intuition of HSIC estimator follows.

HSIC: intuition. $\mathcal{X}$: images, $\mathcal{Y}$: descriptions.

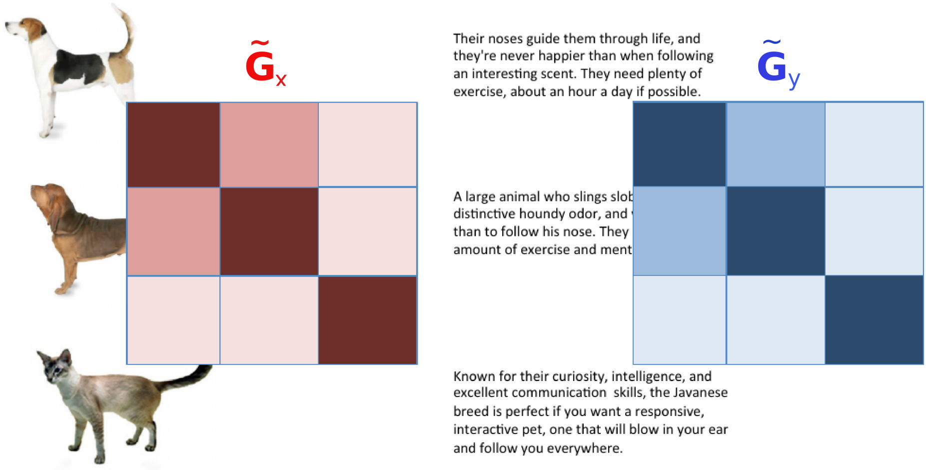
Their noses guide them through life, and they're never happier than when following an interesting scent. They need plenty of exercise, about an hour a day if possible.

A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose. They need a significant amount of exercise and mental stimulation.

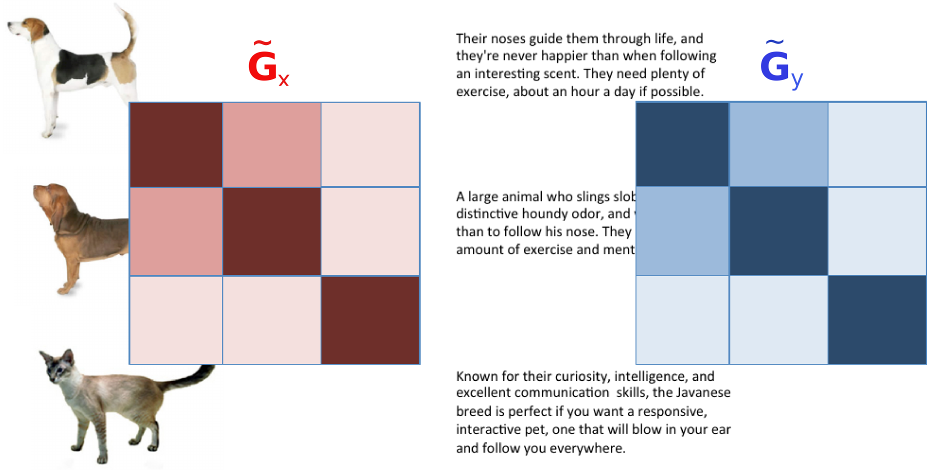
Known for their curiosity, intelligence, and excellent communication skills, the Javanese breed is perfect if you want a responsive, interactive pet, one that will blow in your ear and follow you everywhere.

Text from dogtime.com and petfinder.com

HSIC intuition: Gram matrices



HSIC intuition: Gram matrices



Empirical estimate†: easy, KCCA alternative,

$$\widehat{\mathrm{HSIC}^2}(X,Y) = \frac{1}{N^2}\left\langle \tilde{\mathbf{G}}_X, \tilde{\mathbf{G}}_Y \right\rangle_F.$$

†: Illustration credit (Arthur Gretton).

HSIC in terms of kernel evaluations

$$\mathrm{HSIC}^2(X, Y) = \|C^c_{XY}\|^2_{HS} = \|C^u_{XY} - \mu_X \otimes \mu_Y\|^2_{HS}$$

HSIC in terms of kernel evaluations

$$\begin{aligned}\mathrm{HSIC}^2(X,Y) &= \left\|C_{XY}^c\right\|_{HS}^2 = \left\|C_{XY}^u - \mu_X \otimes \mu_Y\right\|_{HS}^2 \\ &= \left\|C_{XY}^u\right\|_{HS}^2 + \left\|\mu_X \otimes \mu_Y\right\|_{HS}^2 - 2\left\langle C_{XY}^u, \mu_X \otimes \mu_Y\right\rangle_{HS}.\end{aligned}$$

HSIC in terms of kernel evaluations

$$\mathrm{HSIC}^2(X,Y) = \|C^c_{XY}\|^2_{HS} = \|C^u_{XY} - \mu_X \otimes \mu_Y\|^2_{HS}$$
$$= \|C^u_{XY}\|^2_{HS} + \|\mu_X \otimes \mu_Y\|^2_{HS} - 2\left\langle C^u_{XY}, \mu_X \otimes \mu_Y\right\rangle_{HS}.$$

First term:

$$\|C^u_{XY}\|^2_{HS} = \left\langle \mathbb{E}_{XY}\left[\varphi(X)\otimes\psi(Y)\right], \mathbb{E}_{X'Y'}\left[\varphi(X')\otimes\psi(Y')\right]\right\rangle_{HS}$$

HSIC in terms of kernel evaluations

$$\mathrm{HSIC}^2(X,Y) = \|C^c_{XY}\|^2_{HS} = \|C^u_{XY} - \mu_X \otimes \mu_Y\|^2_{HS}$$
$$= \|C^u_{XY}\|^2_{HS} + \|\mu_X \otimes \mu_Y\|^2_{HS} - 2\left\langle C^u_{XY}, \mu_X \otimes \mu_Y \right\rangle_{HS}.$$

First term:

$$\|C^u_{XY}\|^2_{HS} = \left\langle \mathbb{E}_{XY}\left[\varphi(X)\otimes\psi(Y)\right], \mathbb{E}_{X'Y'}\left[\varphi(X')\otimes\psi(Y')\right]\right\rangle_{HS}$$
$$= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\underbrace{\left\langle \varphi(X)\otimes\psi(Y), \varphi(X')\otimes\psi(Y')\right\rangle_{HS}}_{\langle\varphi(X),\varphi(X')\rangle_{\mathcal{H}_k}\langle\psi(Y),\psi(Y')\rangle_{\mathcal{H}_\ell}}$$

$$\left\langle e_1\otimes f_1, e_2\otimes f_2\right\rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)} = \left\langle e_1, e_2\right\rangle_{\mathcal{H}_1}\left\langle f_1, f_2\right\rangle_{\mathcal{H}_2}.$$

HSIC in terms of kernel evaluations

$$\begin{aligned}\mathrm{HSIC}^2(X,Y) &= \left\|C_{XY}^c\right\|_{HS}^2 = \left\|C_{XY}^u - \mu_X \otimes \mu_Y\right\|_{HS}^2 \\ &= \left\|C_{XY}^u\right\|_{HS}^2 + \left\|\mu_X \otimes \mu_Y\right\|_{HS}^2 - 2\left\langle C_{XY}^u, \mu_X \otimes \mu_Y\right\rangle_{HS}.\end{aligned}$$

First term:

$$\begin{aligned}\left\|C_{XY}^u\right\|_{HS}^2 &= \left\langle \mathbb{E}_{XY}\left[\varphi(X)\otimes\psi(Y)\right], \mathbb{E}_{X'Y'}\left[\varphi(X')\otimes\psi(Y')\right]\right\rangle_{HS} \\ &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\underbrace{\left\langle \varphi(X)\otimes\psi(Y), \varphi(X')\otimes\psi(Y')\right\rangle_{HS}}_{\langle\varphi(X),\varphi(X')\rangle_{\mathcal{H}_k}\langle\psi(Y),\psi(Y')\rangle_{\mathcal{H}_\ell}} \\ &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k(X,X')\ell(Y,Y').\end{aligned}$$

$$\left\langle e_1\otimes f_1, e_2\otimes f_2\right\rangle_{HS(\mathcal{H}_2,\mathcal{H}_1)} = \left\langle e_1, e_2\right\rangle_{\mathcal{H}_1}\left\langle f_1, f_2\right\rangle_{\mathcal{H}_2}.$$

HSIC in term of kernel evaluations – continued

$$\begin{aligned}
\mathrm{HSIC}^2(X,Y) &= \left\|C^u_{XY}\right\|^2_{HS} + \left\|\mu_X \otimes \mu_Y\right\|^2_{HS} - 2\left\langle C^u_{XY}, \mu_X \otimes \mu_Y\right\rangle_{HS},\\
\left\|C^u_{XY}\right\|^2_{HS} &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k(X,Y')\ell(Y,Y'),\\
\left\|\mu_X \otimes \mu_Y\right\|^2_{HS} &= \mathbb{E}_{XX'}k(X,X')\mathbb{E}_{YY'}\ell(Y,Y'),\\
\left\langle C^u_{XY}, \mu_X \otimes \mu_Y\right\rangle_{HS} &= \mathbb{E}_{XY}\left[\mathbb{E}_{X'}k(X,X')\mathbb{E}_{Y'}\ell(Y,Y')\right].
\end{aligned}$$

HSIC in term of kernel evaluations – continued

$$\begin{aligned}
\mathrm{HSIC}^2(X,Y) &= \|C_{XY}^u\|_{HS}^2 + \|\mu_X \otimes \mu_Y\|_{HS}^2 - 2\langle C_{XY}^u, \mu_X \otimes \mu_Y\rangle_{HS},\\
\|C_{XY}^u\|_{HS}^2 &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k(X,Y')\ell(Y,Y'),\\
\|\mu_X \otimes \mu_Y\|_{HS}^2 &= \mathbb{E}_{XX'}k(X,X')\mathbb{E}_{YY'}\ell(Y,Y'),\\
\langle C_{XY}^u, \mu_X \otimes \mu_Y\rangle_{HS} &= \mathbb{E}_{XY}\left[\mathbb{E}_{X'}k(X,X')\mathbb{E}_{Y'}\ell(Y,Y')\right].
\end{aligned}$$

Idea: given $\{(x_n,y_n)\}_{n\in[N]} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$,

- Let us estimate C_{XY}^u, μ_X, μ_Y empirically & a bit of linalg.

HSIC in term of kernel evaluations – continued

$$\begin{aligned}
\mathrm{HSIC}^2(X,Y) &= \|C_{XY}^u\|_{HS}^2 + \|\mu_X \otimes \mu_Y\|_{HS}^2 - 2\langle C_{XY}^u, \mu_X \otimes \mu_Y\rangle_{HS},\\
\|C_{XY}^u\|_{HS}^2 &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'} k(X,Y')\ell(Y,Y'),\\
\|\mu_X \otimes \mu_Y\|_{HS}^2 &= \mathbb{E}_{XX'} k(X,X')\mathbb{E}_{YY'}\ell(Y,Y'),\\
\langle C_{XY}^u, \mu_X \otimes \mu_Y\rangle_{HS} &= \mathbb{E}_{XY}\left[\mathbb{E}_{X'} k(X,X')\mathbb{E}_{Y'}\ell(Y,Y')\right].
\end{aligned}$$

Idea: given $\{(x_n,y_n)\}_{n\in[N]} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$,

- Let us estimate C_{XY}^u, μ_X, μ_Y empirically & a bit of linalg.

Result

$\widehat{\mathrm{HSIC}^2}(X,Y) = \boxed{\frac{1}{N^2}\left\langle \tilde{\mathbf{G}}_X, \tilde{\mathbf{G}}_Y \right\rangle_F}$: see the intuition.

HSIC in term of kernel evaluations – continued

$$\begin{aligned}
\mathrm{HSIC}^2(X,Y) &= \|C_{XY}^u\|_{HS}^2 + \|\mu_X \otimes \mu_Y\|_{HS}^2 - 2\langle C_{XY}^u, \mu_X \otimes \mu_Y\rangle_{HS},\\
\|C_{XY}^u\|_{HS}^2 &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'} k(X,Y')\ell(Y,Y'),\\
\|\mu_X \otimes \mu_Y\|_{HS}^2 &= \mathbb{E}_{XX'} k(X,X')\mathbb{E}_{YY'}\ell(Y,Y'),\\
\langle C_{XY}^u, \mu_X \otimes \mu_Y\rangle_{HS} &= \mathbb{E}_{XY}\left[\mathbb{E}_{X'} k(X,X')\mathbb{E}_{Y'}\ell(Y,Y')\right].
\end{aligned}$$

Idea: given $\{(x_n,y_n)\}_{n\in[N]} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$,

- Let us estimate C_{XY}^u, μ_X, μ_Y empirically & a bit of linalg.

Result

$\widehat{\mathrm{HSIC}^2}(X,Y) = \boxed{\frac{1}{N^2}\left\langle \tilde{\mathbf{G}}_X, \tilde{\mathbf{G}}_Y \right\rangle_F}$: see the intuition.

Now

Distance covariance vs. HSIC.

Distance covariance vs. HSIC

Using metric $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$:

$$\begin{aligned}
\mathrm{dCov}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X,X'\right)\rho_{\mathcal{Y}}\left(Y,Y'\right)\\
&+ \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\\
&- 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\right].
\end{aligned}$$

Distance covariance vs. HSIC

Using metric $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$:

$$\begin{aligned}\mathrm{dCov}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X,X'\right)\rho_{\mathcal{Y}}\left(Y,Y'\right)\\ &+ \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\\ &- 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\right].\end{aligned}$$

Using kernel $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$:

$$\begin{aligned}\mathrm{HSIC}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k_{\mathcal{X}}(X,X')k_{\mathcal{Y}}(Y,Y')\\ &+ \mathbb{E}_{XX'}k_{\mathcal{X}}(X,X')\mathbb{E}_{YY'}k_{\mathcal{Y}}(Y,Y')\\ &- 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}k_{\mathcal{X}}(X,X')\mathbb{E}_{Y'}k_{\mathcal{Y}}(Y,Y')\right].\end{aligned}$$

Distance covariance vs. HSIC

Using metric $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$:

$$\begin{aligned}\mathrm{dCov}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X,X'\right)\rho_{\mathcal{Y}}\left(Y,Y'\right)\\ &+ \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\\ &- 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\right].\end{aligned}$$

Using kernel $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$:

$$\begin{aligned}\mathrm{HSIC}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k_{\mathcal{X}}(X,X')k_{\mathcal{Y}}(Y,Y')\\ &+ \mathbb{E}_{XX'}k_{\mathcal{X}}(X,X')\mathbb{E}_{YY'}k_{\mathcal{Y}}(Y,Y')\\ &- 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}k_{\mathcal{X}}(X,X')\mathbb{E}_{Y'}k_{\mathcal{Y}}(Y,Y')\right].\end{aligned}$$

This suggests some relation: $k_{\mathcal{X}} \leftrightarrow \rho_{\mathcal{X}}$, $k_{\mathcal{Y}} \leftrightarrow \rho_{\mathcal{Y}}$?

HSIC vs dCov [Sejdinovic et al., 2013b]

High-level preview

- Distance and kernel techniques can be related:

$$\text{(set of) kernel(s)} \Leftrightarrow \text{semi-metric of negative type} \ni \|\cdot\|_2,$$
$$\text{characteristic kernel} \Leftrightarrow \text{semi-metric of strong negative type.}$$

- Consequence:

$$\text{energy distance} \Leftrightarrow \text{MMD}, \quad \boxed{\text{HSIC } (M=2) \Leftrightarrow \text{dCov}}.$$

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a metric on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a **metric** on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.
- Triangle inequality: $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ for $\forall x, y, z \in \mathcal{X}$.

Metric space

Definition

$\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a **metric** on $\mathcal{X}$ if

- $\rho(x, y) = 0 \Leftrightarrow x = y$.
- Symmetry: $\rho(x, y) = \rho(y, x)$ for $\forall x, y \in \mathcal{X}$.
- Triangle inequality: $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ for $\forall x, y, z \in \mathcal{X}$.

- **semi-metric**: triange inequality is dropped.
- **semi-metric of negative type**: if in addition

$$\sum_{i,j \in [N]} a_i a_j \rho(x_i, x_j) \leq 0$$

for $\forall N \geq 2$, $\forall (x_n)_{n \in [N]} \subset \mathcal{X}$ and $\forall (a_n)_{n \in [N]} \subset \mathbb{R}$ with $\sum_{n \in [N]} a_n = 0$.

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0, 1)$.

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Thus,

- 2nd part $\Rightarrow \left(\mathbb{R}^d, \|\cdot\|_2^2\right) \checkmark$

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Thus,

- 2nd part $\Rightarrow \left(\mathbb{R}^d, \|\cdot\|_2^2\right) \checkmark$
- $+$1st part $\Rightarrow \rho\left(\mathbf{x},\mathbf{y}\right) = \|\mathbf{x} - \mathbf{y}\|_2^q \checkmark$ with $q \in (0,2]$.

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Thus,

- 2nd part $\Rightarrow \left(\mathbb{R}^d, \|\cdot\|_2^2\right) \checkmark$
- +1st part $\Rightarrow \rho(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^q \checkmark$ with $q \in (0,2]$.
- Specifically: $\rho(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ is OK.

Semi-metric space of negative type

[Berg et al., 1984]:

- $\rho : \checkmark \Rightarrow \rho^a : \checkmark$ for $\forall a \in (0,1)$.
- $\Leftrightarrow$ description: $\exists m : \mathcal{X} \to \mathcal{H}$ (ilbert) injective mapping such that

$$\rho(x,y) = \|m(x) - m(y)\|_{\mathcal{H}}^2 .$$

Thus,

- 2nd part $\Rightarrow \left(\mathbb{R}^d, \|\cdot\|_2^2\right) \checkmark$
- $+$1st part $\Rightarrow \rho(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^q \checkmark$ with $q \in (0,2]$.
- Specifically: $\rho(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ is OK.
- Other example (metric space of negative type): $L^p[0,1]$ ($p \in [1,2]$), hyperbolic space [Meckes, 2013].

Semi-metric of negative type vs kernel

[Berg et al., 1984]

$(\mathcal{Z}, \rho)$: semi-metric space, $z_0 \in \mathcal{Z}$. Let

$$k(z, z') := \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').$$

The function k is kernel iff. ρ is of negative type.

Semi-metric of negative type vs kernel

[Berg et al., 1984]

$(\mathcal{Z}, \rho)$: semi-metric space, $z_0 \in \mathcal{Z}$. Let

$$k(z, z') := \rho(z, z_0) + \rho(z', z_0) - \rho(z, z').$$

The function k is kernel iff. ρ is of negative type.

Distance kernels induced by such ρ-s, scaled by 2, with z_0 varying:

$$\mathcal{K}_\rho := \left\{ k \,:\, k(z, z') = \frac{1}{2}\left[\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')\right],\, z_0 \in \mathcal{Z} \right\}.$$

Properties

1. For $k \in K_\rho$,
    1. $\exists z_0 \in \mathcal{Z}$ s.t. $k(z_0, z_0) = 0$. Note: $\not\ni k(z, z') = e^{-\gamma \|z - z'\|_2^2}$.

Properties

1. For $k \in K_\rho$,
   1. $\exists z_0 \in \mathcal{Z}$ s.t. $k(z_0, z_0) = 0$. Note: $\not\ni k(z, z') = e^{-\gamma\|z-z'\|_2^2}$.
   2. $z \mapsto \varphi(z) := k(\cdot, z)$ is injective (=:non-degenerate kernel), and
   3. k generates ρ:

$$\rho(z, z') = \|\varphi(z) - \varphi(z')\|_{\mathcal{H}_k}^2 = k(z, z) + k(z', z') - 2k(z, z').$$

Properties

1. For $k \in K_\rho$,
   1. $\exists z_0 \in \mathcal{Z}$ s.t. $k(z_0, z_0) = 0$. Note: $\nexists$ $k(z, z') = e^{-\gamma \|z - z'\|_2^2}$.
   2. $z \mapsto \varphi(z) := k(\cdot, z)$ is injective (=:non-degenerate kernel), and
   3. k generates ρ:

$$\rho(z, z') = \|\varphi(z) - \varphi(z')\|_{\mathcal{H}_k}^2 = k(z, z) + k(z', z') - 2k(z, z').$$

2. Flipping the roles: if a kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is non-generate, then

$$\boxed{\rho(z, z') = k(z, z) + k(z', z') - 2k(z, z')}$$

   generates a semi-metric of negative type. In addition, $k \in \mathcal{K}_\rho$ iff. $k(z_0, z_0) = 0$ for some $z_0 \in \mathcal{Z}$.

Let $\rho_q(\mathbf{z}, \mathbf{z}') = \|\mathbf{z} - \mathbf{z}'\|_2^q$, $q \in (0, 2]$. Then the generated distance kernel

$$k_q(\mathbf{z}, \mathbf{z}') = \|\mathbf{z}\|_2^q + \|\mathbf{z}'\|_2^q - \|\mathbf{z} - \mathbf{z}'\|_2^q$$

is the fractional Brownian motion kernel.

Examples – continued

For the Gaussian kernel $k(\mathbf{z},\mathbf{z}') = e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2}$, the induced semimetric is

$$\rho(\mathbf{z},\mathbf{z}') = 2\left[1 - e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2}\right].$$

But the ρ-induced kernel (centered at zero)

$$\tilde{k}(\mathbf{z},\mathbf{z}') = e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2} + 1 - e^{-\sigma\|\mathbf{z}\|_2^2} - e^{-\sigma\|\mathbf{z}'\|_2^2}.$$

also generates ρ.

Examples – continued

For the Gaussian kernel $k(\mathbf{z}, \mathbf{z}') = e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2}$, the induced semimetric is

$$\rho(\mathbf{z}, \mathbf{z}') = 2\left[1 - e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2}\right].$$

But the ρ-induced kernel (centered at zero)

$$\tilde{k}(\mathbf{z}, \mathbf{z}') = e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2} + \underbrace{\frac{1}{2} - e^{-\sigma\|\mathbf{z}\|_2^2}}_{f(\mathbf{z})} + \underbrace{\frac{1}{2} - e^{-\sigma\|\mathbf{z}'\|_2^2}}_{f(\mathbf{z}')}.$$

also generates ρ.

For the Gaussian kernel $k(\mathbf{z}, \mathbf{z}') = e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2}$, the induced semimetric is

$$\rho(\mathbf{z}, \mathbf{z}') = 2\left[1 - e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2}\right].$$

But the ρ-induced kernel (centered at zero)

$$\tilde{k}(\mathbf{z}, \mathbf{z}') = e^{-\sigma\|\mathbf{z}-\mathbf{z}'\|_2^2} + \underbrace{\frac{1}{2} - e^{-\sigma\|\mathbf{z}\|_2^2}}_{f(\mathbf{z})} + \underbrace{\frac{1}{2} - e^{-\sigma\|\mathbf{z}'\|_2^2}}_{f(\mathbf{z}')}.$$

also generates ρ.

There can be a lot of k-s generating ρ, but

k and $\tilde{k}$ generates ρ iff. $\tilde{k}(z, z') = k(z, z') + f(z) + f(z')$ for some shift function.

- Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semi-metric spaces of negative type, $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. $\mathbb{P}_X \in \mathcal{M}^2_{\rho_{\mathcal{X}}}$, $\mathbb{P}_Y \in \mathcal{M}^2_{\rho_{\mathcal{Y}}}$ (2nd moments $< \infty$).
- Let the corresponding distance covariance

$$\begin{aligned}\mathrm{dCov}_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X, Y) := {} & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X, X'\right)\rho_{\mathcal{Y}}\left(Y, Y'\right) \\ & + \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y, Y'\right) \\ & - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y, Y'\right)\right].\end{aligned}$$

Equivalence

- Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semi-metric spaces of negative type, $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. $\mathbb{P}_X \in \mathcal{M}^2_{\rho_{\mathcal{X}}}$, $\mathbb{P}_Y \in \mathcal{M}^2_{\rho_{\mathcal{Y}}}$ (2nd moments $< \infty$).
- Let the corresponding distance covariance

$$\begin{aligned} \mathrm{dCov}_{\rho_{\mathcal{X}}, \rho_{\mathcal{Y}}}(X, Y) := & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X, X'\right)\rho_{\mathcal{Y}}\left(Y, Y'\right) \\ & + \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y, Y'\right) \\ & - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y, Y'\right)\right]. \end{aligned}$$

- Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be any 2 kernels generating $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$.
- Let $k := k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$.

Equivalence

- Let $(\mathcal{X}, \rho_{\mathcal{X}})$ and $(\mathcal{Y}, \rho_{\mathcal{Y}})$ be semi-metric spaces of negative type, $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. $\mathbb{P}_X \in \mathcal{M}^2_{\rho_{\mathcal{X}}}$, $\mathbb{P}_Y \in \mathcal{M}^2_{\rho_{\mathcal{Y}}}$ (2nd moments $< \infty$).
- Let the corresponding distance covariance

$$\begin{aligned}\mathrm{dCov}_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X, Y) := {} & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X, X'\right)\rho_{\mathcal{Y}}\left(Y, Y'\right) \\ & + \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y, Y'\right) \\ & - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y, Y'\right)\right].\end{aligned}$$

- Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be any 2 kernels generating $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$.
- Let $k := k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$.

Then $\mathrm{dCov}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X, Y) = 4\mathrm{HSIC}^2_k(X, Y)$.

Indeed

Let $\nu = \mathbb{P}_{XY} - \mathbb{P}_Y \otimes \mathbb{P}_Y$. Notice: $\nu(\mathcal{X} \times \mathcal{Y}) = 1 - 1 = 0$.

$$\begin{aligned}\mathrm{dCov}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X,X'\right)\rho_{\mathcal{Y}}\left(Y,Y'\right)\\ &+ \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y,Y'\right) - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\right]\end{aligned}$$

Indeed

Let $\nu = \mathbb{P}_{XY} - \mathbb{P}_Y \otimes \mathbb{P}_Y$. Notice: $\nu(\mathcal{X} \times \mathcal{Y}) = 1 - 1 = 0$.

$$\text{dCov}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) = \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_\mathcal{X}\left(X, X'\right)\rho_\mathcal{Y}\left(Y, Y'\right)$$

$$+\,\mathbb{E}_{XX'}\rho_\mathcal{X}\left(X, X'\right)\mathbb{E}_{YY'}\rho_\mathcal{Y}\left(Y, Y'\right) - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_\mathcal{X}\left(X, X'\right)\mathbb{E}_{Y'}\rho_\mathcal{Y}\left(Y, Y'\right)\right]$$

$$= \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \rho_\mathcal{X}(x, x')\rho_\mathcal{Y}(y, y')\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y')$$

Indeed

Let $\nu = \mathbb{P}_{XY} - \mathbb{P}_Y \otimes \mathbb{P}_Y$. Notice: $\nu(\mathcal{X} \times \mathcal{Y}) = 1 - 1 = 0$.

$$
\begin{aligned}
\mathrm{dCov}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X, Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X, X'\right)\rho_{\mathcal{Y}}\left(Y, Y'\right) \\
&\quad + \mathbb{E}_{XX'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{YY'}\rho_{\mathcal{Y}}\left(Y, Y'\right) - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X, X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y, Y'\right)\right] \\
&= \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \rho_{\mathcal{X}}(x, x')\rho_{\mathcal{Y}}(y, y')\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y') \\
&= \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \left[k_{\mathcal{X}}(x, x) + k_{\mathcal{X}}(x', x') - \boxed{2k_{\mathcal{X}}(x, x')}\right] \\
&\qquad \times \left[k_{\mathcal{Y}}(y, y) + k_{\mathcal{Y}}(y', y') - \boxed{2k_{\mathcal{Y}}(y, y')}\right]\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y') \\
&\overset{(*)}{=}
\end{aligned}
$$

$(*)$: $\int\int g(x, y, x', y')\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y') = 0$ when g does not depend on ≥ 1 of its args, since ν also has zero marginal measures.

Indeed

Let $\nu = \mathbb{P}_{XY} - \mathbb{P}_Y \otimes \mathbb{P}_Y$. Notice: $\nu(\mathcal{X} \times \mathcal{Y}) = 1 - 1 = 0$.

$$\mathrm{dCov}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X, Y) = \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_\mathcal{X}\left(X, X'\right)\rho_\mathcal{Y}\left(Y, Y'\right)$$
$$+ \mathbb{E}_{XX'}\rho_\mathcal{X}\left(X, X'\right)\mathbb{E}_{YY'}\rho_\mathcal{Y}\left(Y, Y'\right) - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_\mathcal{X}\left(X, X'\right)\mathbb{E}_{Y'}\rho_\mathcal{Y}\left(Y, Y'\right)\right]$$
$$= \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \rho_\mathcal{X}(x, x')\rho_\mathcal{Y}(y, y')\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y')$$
$$= \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \left[k_\mathcal{X}(x, x) + k_\mathcal{X}(x', x') - \boxed{2k_\mathcal{X}(x, x')}\right]$$
$$\times \left[k_\mathcal{Y}(y, y) + k_\mathcal{Y}(y', y') - \boxed{2k_\mathcal{Y}(y, y')}\right]\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y')$$
$$\overset{(*)}{=} 4\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} k_\mathcal{X}(x, x')k_\mathcal{Y}(y, y')\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y')$$

$(*)$: $\int\int g(x, y, x', y')\mathrm{d}\nu(x, y)\mathrm{d}\nu(x', y') = 0$ when g does not depend on ≥ 1 of its args, since ν also has zero marginal measures.

Indeed

Let $\nu = \mathbb{P}_{XY} - \mathbb{P}_Y \otimes \mathbb{P}_Y$. Notice: $\nu(\mathcal{X} \times \mathcal{Y}) = 1 - 1 = 0$.

$$\begin{aligned}
\mathrm{dCov}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}(X,X')\rho_{\mathcal{Y}}(Y,Y') \\
&\quad + \mathbb{E}_{XX'}\rho_{\mathcal{X}}(X,X')\mathbb{E}_{YY'}\rho_{\mathcal{Y}}(Y,Y') - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}(X,X')\mathbb{E}_{Y'}\rho_{\mathcal{Y}}(Y,Y')\right] \\
&= \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \rho_{\mathcal{X}}(x,x')\rho_{\mathcal{Y}}(y,y')\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&= \int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} \left[k_{\mathcal{X}}(x,x) + k_{\mathcal{X}}(x',x') - \boxed{2k_{\mathcal{X}}(x,x')}\right] \\
&\qquad \times \left[k_{\mathcal{Y}}(y,y) + k_{\mathcal{Y}}(y',y') - \boxed{2k_{\mathcal{Y}}(y,y')}\right]\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&\overset{(*)}{=} 4\int_{\mathcal{X}\times\mathcal{Y}}\int_{\mathcal{X}\times\mathcal{Y}} k_{\mathcal{X}}(x,x')k_{\mathcal{Y}}(y,y')\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&= 4\mathrm{HSIC}^2_k(X,Y).
\end{aligned}$$

$(*)$: $\int\int g(x,y,x',y')\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') = 0$ when g does not depend on ≥ 1 of its args, since ν also has zero marginal measures.

Energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Sejdinovic et al., 2013b] or N-distance [Zinger et al., 1992, Klebanov, 2005]

- $(\mathcal{X}, \rho)$: semi-metric space of negative type.
- Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
- Finite a-moment w.r.t. ρ and k ($a > 0$):

$$\mathcal{M}_\rho^a(\mathcal{X}) = \left\{ \mathbb{P} \in \mathcal{M}_b(\mathcal{X}) \; : \; \exists x_0 \in \mathcal{X} \text{ s.t. } \int_{\mathcal{X}} \rho^a(x, x_0) \mathrm{d}\mathbb{P}(x) < \infty \right\},$$

$$\mathcal{M}_k^a(\mathcal{X}) = \left\{ \mathbb{P} \in \mathcal{M}_b(\mathcal{X}) \; : \; \int_{\mathcal{X}} k^a(x, x) \mathrm{d}\mathbb{P}(x) < \infty \right\}.$$

Energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Sejdinovic et al., 2013b] or N-distance [Zinger et al., 1992, Klebanov, 2005]

- $(\mathcal{X}, \rho)$: semi-metric space of negative type.
- Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
- Finite a-moment w.r.t. ρ and k ($a > 0$):

$$\mathcal{M}_{\rho}^{a}(\mathcal{X}) = \left\{ \mathbb{P} \in \mathcal{M}_b(\mathcal{X}) \, : \, \exists x_0 \in \mathcal{X} \text{ s.t. } \int_{\mathcal{X}} \rho^a(x, x_0) \mathrm{d}\mathbb{P}(x) < \infty \right\},$$

$$\mathcal{M}_{k}^{a}(\mathcal{X}) = \left\{ \mathbb{P} \in \mathcal{M}_b(\mathcal{X}) \, : \, \int_{\mathcal{X}} k^a(x, x) \mathrm{d}\mathbb{P}(x) < \infty \right\}.$$

- Recall: $\exists \mu_k(\mathbb{P}) \Leftrightarrow \mathbb{P} \in \mathcal{M}_k^{\frac{1}{2}}(\mathcal{X})$.

Energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005, Sejdinovic et al., 2013b] or N-distance [Zinger et al., 1992, Klebanov, 2005]

- $(\mathcal{X}, \rho)$: semi-metric space of negative type.
- Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
- Finite a-moment w.r.t. ρ and k ($a > 0$):

$$\mathcal{M}_\rho^a(\mathcal{X}) = \left\{ \mathbb{P} \in \mathcal{M}_b(\mathcal{X}) \,:\, \exists x_0 \in \mathcal{X} \text{ s.t. } \int_{\mathcal{X}} \rho^a(x, x_0) \mathrm{d}\mathbb{P}(x) < \infty \right\},$$

$$\mathcal{M}_k^a(\mathcal{X}) = \left\{ \mathbb{P} \in \mathcal{M}_b(\mathcal{X}) \,:\, \int_{\mathcal{X}} k^a(x, x) \mathrm{d}\mathbb{P}(x) < \infty \right\}.$$

- Recall: $\exists \mu_k(\mathbb{P}) \Leftrightarrow \mathbb{P} \in \mathcal{M}_k^{\frac{1}{2}}(\mathcal{X})$.

Moments comparison

Let k generate ρ and $n \in \mathbb{N}$. Then $\mathcal{M}_\rho^{\frac{n}{2}}(\mathcal{X}) = \mathcal{M}_k^{\frac{n}{2}}(\mathcal{X})$.

Energy distance – continued

Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^1_\rho(\mathcal{X})$, $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$. Energy distance of $\mathbb{P}$ and $\mathbb{Q}$ [Sejdinovic et al., 2013b]:

$$D^2_{E,\rho}(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{X,Y}\rho(X, Y) - \mathbb{E}_{X,X'}\rho(X, X') - \mathbb{E}_{Y,Y'}\rho(Y, Y').$$

- ρ: negative type $\Rightarrow D_{E,\rho}(\mathbb{P}, \mathbb{Q}) \geq 0$.
- ρ: strong negative type if for $\forall \mathbb{P}, \mathbb{Q} \in \mathcal{M}^1(\mathcal{X})$:

$$\mathbb{P} \neq \mathbb{Q} \Rightarrow D_{E,\rho}(\mathbb{P}, \mathbb{Q}) \neq 0.$$

Energy distance – continued

Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_\rho^1(\mathcal{X})$, $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$. Energy distance of $\mathbb{P}$ and $\mathbb{Q}$ [Sejdinovic et al., 2013b]:

$$D_{E,\rho}^2(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{X,Y}\rho(X, Y) - \mathbb{E}_{X,X'}\rho(X, X') - \mathbb{E}_{Y,Y'}\rho(Y, Y').$$

- ρ: negative type $\Rightarrow D_{E,\rho}(\mathbb{P}, \mathbb{Q}) \geq 0$.
- ρ: strong negative type if for $\forall \mathbb{P}, \mathbb{Q} \in \mathcal{M}^1(\mathcal{X})$:

$$\mathbb{P} \neq \mathbb{Q} \Rightarrow D_{E,\rho}(\mathbb{P}, \mathbb{Q}) \neq 0.$$

Notes:

- $\Rightarrow D_{E,\rho}$ can distinguish probability measures.

Energy distance – continued

Let $\mathbb{P},\mathbb{Q} \in \mathcal{M}^1_\rho(\mathcal{X})$, $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$. Energy distance of $\mathbb{P}$ and $\mathbb{Q}$ [Sejdinovic et al., 2013b]:

$$D^2_{E,\rho}(\mathbb{P},\mathbb{Q}) = 2\mathbb{E}_{X,Y}\rho(X,Y) - \mathbb{E}_{X,X'}\rho(X,X') - \mathbb{E}_{Y,Y'}\rho(Y,Y').$$

- ρ: negative type $\Rightarrow D_{E,\rho}(\mathbb{P},\mathbb{Q}) \ge 0$.
- ρ: strong negative type if for $\forall \mathbb{P},\mathbb{Q} \in \mathcal{M}^1(\mathcal{X})$:

$$\mathbb{P} \ne \mathbb{Q} \Rightarrow D_{E,\rho}(\mathbb{P},\mathbb{Q}) \ne 0.$$

Notes:

- $\Rightarrow D_{E,\rho}$ can distinguish probability measures.
- Example [Lyons, 2013]: every separable Hilbert space.

Energy distance – continued

Let $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^1_\rho(\mathcal{X})$, $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$. Energy distance of $\mathbb{P}$ and $\mathbb{Q}$ [Sejdinovic et al., 2013b]:

$$D^2_{E,\rho}(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{X,Y}\rho(X, Y) - \mathbb{E}_{X,X'}\rho(X, X') - \mathbb{E}_{Y,Y'}\rho(Y, Y').$$

- ρ: negative type $\Rightarrow D_{E,\rho}(\mathbb{P}, \mathbb{Q}) \geq 0$.
- ρ: strong negative type if for $\forall \mathbb{P}, \mathbb{Q} \in \mathcal{M}^1(\mathcal{X})$:

$$\mathbb{P} \neq \mathbb{Q} \Rightarrow D_{E,\rho}(\mathbb{P}, \mathbb{Q}) \neq 0.$$

Notes:

- $\Rightarrow D_{E,\rho}$ can distinguish probability measures.
- Example [Lyons, 2013]: every separable Hilbert space.
- $\text{dCov}_{\rho_\mathcal{X}, \rho_\mathcal{Y}}$ is a valid independence measure $\Leftrightarrow \rho_\mathcal{X}$ and $\rho_\mathcal{Y}$: metric of negative type [Lyons, 2013].

Energy distance $\Leftrightarrow$ MMD

[Sejdinovic et al., 2013b]

Let $(\mathcal{X}, \rho)$ be a semi-metric space of negative type, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ any kernel that generates ρ, and $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^1_\rho(\mathcal{X})$. Then

$$D^2_{E,\rho}(\mathbb{P}, \mathbb{Q}) = 2\mathrm{MMD}^2_k(\mathbb{P}, \mathbb{Q}).$$

Energy distance $\Leftrightarrow$ MMD

[Sejdinovic et al., 2013b]

Let $(\mathcal{X}, \rho)$ be a semi-metric space of negative type, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ any kernel that generates ρ, and $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_\rho^1(\mathcal{X})$. Then

$$D_{E,\rho}^2(\mathbb{P}, \mathbb{Q}) = 2\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}).$$

Consequence

Let k any kernel that generates ρ. Then

$(\mathcal{X}, \rho)$ is of strong negative type $\Leftrightarrow$ k is characteristic on $\mathcal{M}_\rho^1(\mathcal{X})$.

Energy distance $\Leftrightarrow$ MMD

[Sejdinovic et al., 2013b]

Let $(\mathcal{X}, \rho)$ be a semi-metric space of negative type, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ any kernel that generates ρ, and $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_\rho^1(\mathcal{X})$. Then

$$D^2_{E,\rho}(\mathbb{P}, \mathbb{Q}) = 2\text{MMD}^2_k(\mathbb{P}, \mathbb{Q}).$$

Consequence

Let k any kernel that generates ρ. Then

$(\mathcal{X}, \rho)$ is of strong negative type $\Leftrightarrow$ k is characteristic on $\mathcal{M}_\rho^1(\mathcal{X})$.

Validness of HSIC and MMD follow.

Validness of HSIC and MMD (product space)

Central in applications

characteristic / $\mathcal{I}$-characteristic property!

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

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

Validness of HSIC and MMD (product space)

Central in applications

characteristic / $\mathcal{I}$-characteristic property!

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

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

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

Validness of HSIC and MMD (product space)

Central in applications

characteristic / $\mathcal{I}$-characteristic property!

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

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

- $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic.
- Wanted: Characteristic properties of $\otimes_{m=1}^{M} k_m$ in terms of k_m-s!

Validness of HSIC and MMD (product space)

Central in applications

characteristic / $\mathcal{I}$-characteristic property!

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

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

- $\otimes_{m=1}^{M} k_m$: universal $\Rightarrow$ characteristic $\Rightarrow$ $\mathcal{I}$-characteristic.
- Wanted: Characteristic properties of $\otimes_{m=1}^{M} k_m$ in terms of k_m-s!
- Known ($M = 2$, [Blanchard et al., 2011, Lyons, 2013]):

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

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

- Characteristic property:

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

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

- Characteristic property:

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

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

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

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

- Characteristic property:

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

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

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

- Witness construction:

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

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

- Characteristic property:

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

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

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

- Witness construction:

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

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

- Characteristic property:

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

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

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

- Witness construction:

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

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

Theorem (characteristic property)

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

Results [Szabó and Sriperumbudur, 2018]

Theorem (characteristic property)

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

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

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

Results – continued

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

The followings are equivalent:

(i) $(k_m)_{m=1}^M$-*s are characteristic.*

(ii) $\otimes_{m=1}^M k_m$: $\mathcal{I}$-*characteristic.*

(iii) $\otimes_{m=1}^M k_m$: *characteristic.*

Results – continued

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

The followings are equivalent:

(i) $(k_m)_{m=1}^M$-*s are characteristic.*

(ii) $\otimes_{m=1}^M k_m$: $\mathcal{I}$-*characteristic.*

(iii) $\otimes_{m=1}^M k_m$: *characteristic.*

Theorem (Universality)

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

Results – continued

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

The followings are equivalent:

(i) $(k_m)_{m=1}^M$*-s are characteristic.*

(ii) $\otimes_{m=1}^M k_m$*:* $\mathcal{I}$*-characteristic.*

(iii) $\otimes_{m=1}^M k_m$*: characteristic.*

Theorem (Universality)

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

These results settle MMD and HSIC. Now: hypothesis testing.

Asymptotic null distribution ($M \ge 2$ [Pfister et al., 2018])

- Domain: $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$.

Asymptotic null distribution ($M \geq 2$ [Pfister et al., 2018])

- Domain: $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$.
- h_2: core function of the V-statistic based HSIC estimator. Associated estimator: $\widehat{\text{HSIC}}_N$ (we saw it for $M = 2$).

- Domain: $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$.
- h_2: core function of the V-statistic based HSIC estimator. Associated estimator: $\widehat{\text{HSIC}}_N$ (we saw it for $M = 2$).
- T_{h_2}: integral operator associated to h_2

$$(T_{h_2}f)(x) := \int_{\mathcal{X}} h_2(x,y)f(y)\mathrm{d}\mathbb{P}(y), \quad f \in L^2(\mathbb{P}).$$

Asymptotic null distribution ($M \geq 2$ [Pfister et al., 2018])

- Domain: $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$.
- h_2: core function of the V-statistic based HSIC estimator. Associated estimator: $\widehat{\mathrm{HSIC}}_N$ (we saw it for $M = 2$).
- T_{h_2}: integral operator associated to h_2

$$(T_{h_2}f)(x) := \int_{\mathcal{X}} h_2(x, y)f(y)\mathrm{d}\mathbb{P}(y), \quad f \in L^2(\mathbb{P}).$$

- $(\lambda_n)_{n\in\mathbb{N}}$: eigenvalues of T_{h_2}.

- Domain: $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$.
- h_2: core function of the V-statistic based HSIC estimator. Associated estimator: $\widehat{\text{HSIC}}_N$ (we saw it for $M = 2$).
- T_{h_2}: integral operator associated to h_2

$$(T_{h_2}f)(x) := \int_{\mathcal{X}} h_2(x,y)f(y)\mathrm{d}\mathbb{P}(y), \quad f \in L^2(\mathbb{P}).$$

- $(\lambda_n)_{n\in\mathbb{N}}$: eigenvalues of T_{h_2}.
- $(Z_n)_{n\in\mathbb{N}}$: sequence of independent $N(0,1)$ variables.

Asymptotic null distribution ($M \geq 2$ [Pfister et al., 2018])

- Domain: $\mathcal{X} = \times_{m\in[M]}\mathcal{X}_m$.
- h_2: core function of the V-statistic based HSIC estimator. Associated estimator: $\widehat{\text{HSIC}}_N$ (we saw it for $M = 2$).
- T_{h_2}: integral operator associated to h_2

$$(T_{h_2} f)(x) := \int_{\mathcal{X}} h_2(x,y) f(y) \mathrm{d}\mathbb{P}(y), \quad f \in L^2(\mathbb{P}).$$

- $(\lambda_n)_{n\in\mathbb{N}}$: eigenvalues of T_{h_2}.
- $(Z_n)_{n\in\mathbb{N}}$: sequence of independent $N(0,1)$ variables.

Then

$N\widehat{\text{HSIC}}_N \xrightarrow{w} \binom{2M}{2} \sum_{n\in\mathbb{N}} \lambda_n Z_n^2$ as $N \to \infty$.

3 commonly applied null approximations

1. Permutation.
2. Bootstrap.
3. Gamma:
    - motivated by the form of the asymptotic null.
    - fast, but no guarantee.

Resampling schemes

$\ni$ permutation, bootstrap.

Accept/reject the null

- Observation: $\{\mathbf{x}_n\}_{n\in[N]} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$. Null and alternative:

$$H_0 : \mathbb{P} = \otimes_{m\in[M]}\mathbb{P}_m, \qquad H_1 : \mathbb{P} \neq \otimes_{m\in[M]}\mathbb{P}_m.$$

- Decision function: reject the null if

$$\varphi_N(\mathbf{x}_1,\ldots,\mathbf{x}_N) := \mathbb{I}_{\{N\widehat{\text{HSIC}}_N(\mathbf{x}_1,\ldots,\mathbf{x}_N) > c_N(\mathbf{x}_1,\ldots,\mathbf{x}_N)\}}.$$

- Threshold $c_N(\mathbf{x}_1,\ldots,\mathbf{x}_N)$: to be specified (later).

Desired guarantees

- Level: let $\alpha \in (0,1)$ fixed.
  - Ideally: the test has (valid) level α, i.e. for all $\mathbb{P} \in H_0$ and $N \in \mathbb{Z}^+$

$$\underbrace{\mathbb{P}(\varphi(\mathbf{X}_1, \ldots, \mathbf{X}_N) = 1)}_{\mathbb{P}(\text{reject } H_0 \mid H_0)} \leq \alpha.$$

  - 'OK': the test respects the level asymptotically (it has pointwise asymptotic level), i.e. for every $\mathbb{P} \in H_0$

$$\limsup_{N \to \infty} \mathbb{P}(\varphi(\mathbf{X}_1, \ldots, \mathbf{X}_N) = 1) \leq \alpha.$$

- Pointwise consistency: If for all $\mathbb{P} \in H_1$

$$\lim_{N \to \infty} \underbrace{\mathbb{P}(\varphi(\mathbf{X}_1, \ldots, \mathbf{X}_N) = 1)}_{\mathbb{P}(\text{reject } H_0 \mid H_1)} = 1.$$

- Goal: approximate the

$$\boxed{\text{distribution of } \widehat{\text{HSIC}}_N(\mathbf{X}_1, \ldots, \mathbf{X}_N)}$$

using the available data $\{\mathbf{x}_n\}_{n\in[N]}$.

Resampling schemes

- Goal: approximate the

$$\boxed{\text{distribution of } \widehat{\text{HSIC}}_N(\mathbf{X}_1, \ldots, \mathbf{X}_N)}$$

  using the available data $\{\mathbf{x}_n\}_{n \in [N]}$.
- Resampling trick:
  - 'shuffling' functions: $\psi_m \in B_N := \{[N] \to [N] \text{ functions}\}$, $m \in [M]$.

Resampling schemes

- Goal: approximate the

$$\boxed{\text{distribution of } \widehat{\mathrm{HSIC}}_N(\mathbf{X}_1, \ldots, \mathbf{X}_N)}$$

  using the available data $\{\mathbf{x}_n\}_{n \in [N]}$.
- Resampling trick:
  - 'shuffling' functions: $\psi_m \in B_N := \{[N] \to [N] \text{ functions}\}$, $m \in [M]$. Effect of ψ_m:

$$x_{m,1}, \ldots, x_{m,N} \xrightarrow{\psi_m} x_{m,\psi_m(1)}, \ldots, x_{m,\psi_m(N)}.$$

Resampling schemes

- Goal: approximate the

$$\boxed{\text{distribution of } \widehat{\text{HSIC}}_N(\mathbf{X}_1, \ldots, \mathbf{X}_N)}$$

  using the available data $\{\mathbf{x}_n\}_{n \in [N]}$.
- Resampling trick:
  - 'shuffling' functions: $\psi_m \in B_N := \{[N] \to [N] \text{ functions}\}$, $m \in [M]$. Effect of ψ_m:

$$x_{m,1}, \ldots, x_{m,N} \xrightarrow{\psi_m} x_{m,\psi_m(1)}, \ldots, x_{m,\psi_m(N)}.$$

  - shuffled samples: $g_{N,\psi}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $\psi := (\psi_m)_{m \in [M]}$.

- Goal: approximate the

$$\boxed{\text{distribution of } \widehat{\text{HSIC}}_N(\mathbf{X}_1, \ldots, \mathbf{X}_N)}$$

  using the available data $\{\mathbf{x}_n\}_{n\in[N]}$.
- Resampling trick:
  - 'shuffling' functions: $\psi_m \in B_N := \{[N] \to [N] \text{ functions}\}$, $m \in [M]$. Effect of ψ_m:

$$x_{m,1}, \ldots, x_{m,N} \xrightarrow{\psi_m} x_{m,\psi_m(1)}, \ldots, x_{m,\psi_m(N)}.$$

  - shuffled samples: $g_{N,\psi}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $\psi := (\psi_m)_{m\in[M]}$.
- Resampling method: $g := (g_{N,\psi})_{\psi\in A_N}$, $A_N \subseteq B_N^M$.

Permutation and bootstrap

- Permutation: $A_N = (S_N)^M$, $S_N =$ permutations of $[N]$, $|A_N| = (N!)^M$.
- Bootstrap: $A_N = B_N^M$, $|A_N| = N^{NM}$.

Permutation and bootstrap

- Permutation: $A_N = (S_N)^M$, S_N = permutations of $[N]$, $|A_N| = (N!)^M$.
- Bootstrap: $A_N = B_N^M$, $|A_N| = N^{NM}$.
- In both cases:
  - estimated cdf

$$\hat{R}_N(\mathbf{x}_1,\ldots,\mathbf{x}_N)(t) := \frac{1}{|A_N|}\sum_{\psi\in A_N}\mathbb{I}_{\{N\widehat{\mathrm{HSIC}}_N(g_{N,\psi}(\mathbf{x}_1,\ldots,\mathbf{x}_N))\le t\}}.$$

Permutation and bootstrap

- Permutation: $A_N = (S_N)^M$, S_N = permutations of $[N]$, $|A_N| = (N!)^M$.
- Bootstrap: $A_N = B_N^M$, $|A_N| = N^{NM}$.
- In both cases:
  - estimated cdf

$$\hat{R}_N(\mathbf{x}_1, \ldots, \mathbf{x}_N)(t) := \frac{1}{|A_N|} \sum_{\psi \in A_N} \mathbb{I}_{\{N\widehat{\mathrm{HSIC}}_N(g_{N,\psi}(\mathbf{x}_1,\ldots,\mathbf{x}_N)) \le t\}}.$$

  - estimated threshold: $(1-\alpha)$-quantile of $\hat{R}_N$, i.e.

$$\boxed{c_N(\mathbf{x}_1, \ldots, \mathbf{x}_N)} := \hat{R}_N(\mathbf{x}_1, \ldots, \mathbf{x}_N)^{-1}(1-\alpha).$$

Results [Pfister et al., 2018, Rindt et al., 2021]

Level & consistency:

Independence test	level	consistency
permutation	valid	pointwise

Results [Pfister et al., 2018, Rindt et al., 2021]

Level & consistency:

Independence test	level	consistency
permutation	valid	pointwise
bootstrap	pointwise asymptotic	pointwise

Results [Pfister et al., 2018, Rindt et al., 2021]

Level & consistency:

Independence test	level	consistency
permutation	valid	pointwise
bootstrap	pointwise asymptotic	pointwise
Gamma	no guarantee	no guarantee

Results [Pfister et al., 2018, Rindt et al., 2021]

Level & consistency:

Independence test	level	consistency
permutation	valid	pointwise
bootstrap	pointwise asymptotic	pointwise
Gamma	no guarantee	no guarantee

Notes (in practice): as $|A_N|$ can be large

- B shuffling are generated instead of $|A_N|$. Optimal B: open.
- permutation test: still has valid level.

Summary

- Focus: independence measures & testing.
- Applications.
- Techniques:
    - copula,
    - maximum correlation,
    - distance,
    - kernel.

Thank you for the attention!

Bach, F. and Jordan, M. (2002).
Kernel independent component analysis.
Journal of Machine Learning Research, 3:1–48.

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

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

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

Berg, C., Christensen, J. P. R., and Ressel, P. (1984).

Harmonic Analysis on Semigroups.
Springer-Verlag.

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

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

Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. (2017).
Domain generalization by marginal transfer learning.
Technical report.
(https://arxiv.org/abs/1711.07910).

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

Generalizing from several related classification tasks to a new unlabeled sample.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., and Riec, B. (2020).
Graph kernels: State-of-the-art and future challenges.
Foundations and Trends in Machine Learning, 13(5-6):531–712.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. (2006).
Integrating structured biological data by kernel maximum mean discrepancy.
Bioinformatics, 22(14):e49–e57.

Borgwardt, K. M. and Kriegel, H.-P. (2005).
Shortest-path kernels on graphs.

In *International Conference on Data Mining (ICDM)*, pages 74–81.

Cardoso, J.-F. (1998).
Multidimensional independent component analysis.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1941–1944.

Carmeli, C., Vito, E. D., Toigo, A., and Umanitá, V. (2010).
Vector valued reproducing kernel Hilbert spaces and universality.
Analysis and Applications, 8:19–61.

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.

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.

Draief, M., Kutzkov, K., Scaman, K., and Vojnovic, M. (2018).
KONG: Kernels for ordered-neighborhood graphs.
Technical report.
(https://arxiv.org/abs/1805.10014).

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.

Fang, K.-T., Kotz, S., and Ng, K. W. (1990).
Symmetric multivariate and related distributions.
Chapman and Hall.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007).

Statistical consistency of kernel canonical correlation analysis.
Journal of Machine Learning Research, 8(14):361–383.

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.

Gaißer, S., Ruppert, M., and Schmid, F. (2010).
A multivariate version of Hoeffding's phi-square.
Journal of Multivariate Analysis, 101:2571–2586.

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. (2005a).
  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.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b).
Kernel methods for measuring independence.
Journal of Machine Learning Research, 6(70):2075–2129.

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

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

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

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

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)*, 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.

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. (2018).
Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces.
Technical report.
(https://arxiv.org/abs/1712.01572).

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

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

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2004).
Profile-based string kernels for remote homology detection and motif extraction.
Journal of Bioinformatics and Computational Biology, 13(4):527–550.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016).
Persistence weighted Gaussian kernel for topological data analysis.
In *International Conference on Machine Learning (ICML)*, pages 2004–2013.

Kybic, J. (2004).
High-dimensional mutual information estimation for image registration.
In *IEEE International Conference on Image Processing (ICIP)*, pages 1779–1782.

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.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993).
Canonical correlation analysis when the data are curves.
Journal of the Royal Statistical Society, Series B (Methodological), 55(3):725–740.

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

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J., and Ghahramani, Z. (2014).
Automatic construction and natural-language description of nonparametric regression models.
In *AAAI Conference on Artificial Intelligence*, pages 1242–1250.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).
Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015).
Towards a learning theory of cause-effect inference.

International Conference on Machine Learning (ICML), 37:1452–1461.

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

Meckes, M. W. (2013).
Positive definite metric spaces.
Positivity, 17:733–757.

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.

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

Neemuchwala, H., Hero, A., Zabuawala, S., and Carson, P. (2007).
Image registration methods in high dimensional space.
International Journal of Imaging Systems and Technology, 16:130–145.

Nelsen, R. B. (2006).
An Introduction to Copulas (Springer Series in Statistics).
Springer.

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.

Peng, H., Long, F., and Ding, C. (2005).
Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8):1226–1238.

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)*.

Rényi, A. (1959).
On measures of dependence.
Acta Mathematica Academiae Scientiarum Hungaricae, 10:441–451.

Rindt, D., Sejdinovic, D., and Steinsaltz, D. (2021).
Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC.
Stat, 10(1):e364.

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.

Saitoh, S. and Sawano, Y. (2016).
Theory of Reproducing Kernels and Applications.
Springer Singapore.

Schmid, F., Schmidt, R., Blumentritt, T., Gaißer, S., and Ruppert, M. (2010).
Copula Theory and Its Applications, chapter Copula based Measures of Multivariate Association.
Lecture Notes in Statistics. Springer.

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.

Schweizer, B. and Wolff, E. F. (1981).
On nonparametric measures of dependence for random variables.
The Annals of Statistics, 9:879–885.

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.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018).
Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions.
Journal of Machine Learning Research, 44:1–29.

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

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. (2011).
Kernel belief propagation.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715.

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

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2010b).
On the relation between universality, characteristic kernels and RKHS embedding of measures.
In *International Conference on AI and Statistics (AISTATS)*, pages 781–788.

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.

Szabó, Z., Póczos, B., and Lorincz, A. (2012).
Separation theorem for independent subspace analysis and its consequences.
Pattern Recognition, 45(4):1782–1791.

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.

van der Vaart, A. (1998).
Asymptotic Statitics.
Cambridte University Press.

van der Vaart, A. and Wellner, J. (2000).
Weak Convergence and Empirical Processes: With Applications to Statistics.

Springer.

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

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

Wendland, H. (2005).
Scattered Data Approximation.
Cambridge University Press.

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.

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.