Reproducing Kernel Hilbert Spaces(RKHS)

Roberto Alcover Couso

29/9/2018

1 Introduction

In this document we will study a special type of Hilbert spaces, RKHS, with a kernel which meets the reproducing property. Understanding this is key for our study due to it's relevance in statistical models and the ideas behind algorithims such as RDC,HSIC... Furthermore we will define a homogeneity test based on embeddings of probability distributions on RKHSs, where the distance between distributions corresponds to the distance between their embeddings. We will see that the unit ball of an RKHS is a rich enough space so that the expression for the discrepancy vanishes only if the two probability distributions are equal. At the same time it is restrictive enough for the empirical estimate at the discrepancy to converge quickly to its population counterpart as the sample size increases.

1.1 Preliminar knowledge

1. Feature map: ϕ is known as a feature map if is a function which maps the data to a Hilbert space $\mathcal{H}$ (feature space).

$$\begin{aligned} \phi : \mathcal{X} &\to \mathcal{H} \\ x &\mapsto \varphi \end{aligned}$$

2. Kernel function: k is called a kernel function if it is the dot product defined on a feature space.

   Then, we can rewrite the dot product of the space in terms of this mapping:

$$k : \mathcal{X} \times \mathcal{X} \to \mathcal{H}(x, x') \mapsto k(x, x') =< \phi(x), \phi(x') >$$

3. Reproducing kernel: a function k is a reproducing kernel of the Hilbert space $\mathcal{H}$ is and only if it satisfies:

   (a) k(x,.)$\in \mathcal{H}, \forall x \in \mathcal{X}$

   (b) *Reproducing property*: $< f, k(x, \cdot) >= f(x) \forall f \in \mathcal{H}, \forall x \in \mathcal{X}$

Proposition 1.1. *If k is a reproducting kernel then:* $k(x, x') =< k(x, \cdot), k(x', \cdot) >$

2 Maximum mean discrepancy

In this section it'll be shown how RKHSs can be used to define a homogeneity test in terms of the embeddings of the probability measures. This test consist

in maximizing the measure of discrepancy between functions that belong to a certain family $\mathcal{F}$ which must be rich enough to detect all the possible differences between the two probability measures.

2.1 Mean embedding

Lemma 2.1. *Given two Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$ are equal if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y)\ \forall f \in \mathcal{C}(\mathcal{X})$*

$$X \sim \mathbb{P} \text{ and } Y \sim \mathbb{Q}$$

This condition is pretty dificult to prove therefore we will keep our study in order to simplify this evaluation.

Definition 2.1. MMD

Let $\mathcal{F}$ be a class of functions f: $X \to \mathbb{R}$ the MMD based on $\mathcal{F}$ is

$$\gamma(\mathbb{P}, \mathbb{Q}) = MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \{\mathbb{E}f(X) - \mathbb{E}f(Y)\}$$

This $\mathcal{F}$ must be rich enough for it to ensure that $\mathbb{P} = \mathbb{Q} \leftrightarrow \gamma(\mathbb{P}, \mathbb{Q}) = 0$. And restrictive enough for the empirical estimate to converge quickly as the sample size increases. This will be done through RKHS with a characteristic kernel K

Definition 2.2. Riesz representation

If T is a bounded linear operator on a Hilbert space $\mathcal{H}$, then there exist some $g \in \mathcal{H}$ such that $\forall f \in \mathcal{H}$:

$$T(f) =< f, g >_{\mathcal{H}}$$

Lemma 2.2. *Given a K(s,) semi positive definite, measurable and $\mathbb{E}\sqrt{k(X,X)} < \infty$, where X$\sim$ $\mathbb{P}$ then $\mu_p \in \mathcal{H}$ exist and fulfulls the next condition $\mathbb{E}f(X) =< f, \mu_p >$ for all $f \in \mathcal{H}$*

Proof. Lets define the linear operator $T_{\mathbb{P}} f \equiv \mathbb{E}(\sqrt{k(X,X)}) < \infty \forall f \in \mathcal{H}$

$$|T_{\mathbb{P}} f| = |\mathbb{E}(f(X))| \leq \mathbb{E}(|f(X)|) = {}^{1}\mathbb{E}| < f, k(\cdot, X) >_{\mathcal{H}}| \leq {}^{2}\|f\|_{\mathcal{H}} \cdot \mathbb{E}(\sqrt{K(X,X)})^{1/2} < {}^{3}\infty$$

Then using the Riesz representation theorem applied to T_p, there exist a $\mu_p \in \mathcal{H}$ such that $T_p f =< f, \mu_p >_{\mathcal{H}}$ □

[1] Reproducing property of the kernel
[2] Chauchy Schwarz inequality
[3] The expectation under $\mathbb{P}$ of the kernel is bounded

Definition 2.3. Mean embedding

Given a probability distribution $\mathbb{P}$ we will define the mean embedding of $\mathbb{P}$ as an element $\mu_p \in \mathcal{H}$ such that

$$\mathbb{E}(f(X)) =< f, \mu_{\mathbb{P}} >_{\mathcal{H}}, \forall f \in \mathcal{H}$$

If $f \in \mathcal{H}$ and $\mu_{\mathbb{P}} \in \mathbb{R}$ $\mathbb{E}(f(X)) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n)$

Applying the Riesz representation theorem to represent $f(x_n)$

$\forall x_n$ then:

$$f(x_n) =< f, K(\cdot, x_n) >_{\mathcal{H}}$$

then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} < f, K(\cdot, x_n >_{\mathcal{H}} =< f, \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} K(\cdot, x_n) >_{\mathcal{H}}$$

which leads to the final conclussion:

$\mu_{\mathbb{P}} \equiv \mathbb{E}_{X\sim\mathbb{P}}(K(t, X))$ $t \in [0, T]$

Second interpretation of the mean embedding

$$\mu_{\mathbb{P}} = \mathbb{E}(K(\cdot, X))$$

2.2 Introduction to MMD

Lemma 2.3. *Given the conditions of Lemma 2.2 ($\mu_{\mathbb{P}}$and$\mu_{\mathbb{Q}}$ exist) then:*

$X \sim \mathbb{P}\mu_{\mathbb{P}} \equiv \mathbb{E}_{X\sim\mathbb{P}}(K(\cdot, X))$ $Y \sim \mathbb{Q}\mu_{\mathbb{Q}} \equiv \mathbb{E}_{Y\sim\mathbb{Q}}(K(\cdot, Y))$

and:

$$MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

Proof.

$$\begin{aligned} MMD &\equiv \sup_{f\in\mathcal{H}\|f\|\leq 1} \{\mathbb{E}(f(x)) - \mathbb{E}(f(y))\} \\ &= \sup_{f\in\mathcal{H}\|f\|\leq 1} \{< f, \mu_{\mathbb{P}} > - < f, \mu_{\mathbb{Q}} >\} \\ &= \sup_{f\in\mathcal{H}\|f\|\leq 1} < f, (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}) > \qquad (1) \\ &\leq^{4} \sup_{f\in\mathcal{H}\|f\|\leq 1} \{\|f\|_{\mathcal{H}}, \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}\} \\ &\leq \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}. \end{aligned}$$

But on the other side, if we choose f as:

$$f = \frac{1}{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|}(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}})$$

then we have:

$$\sup_{f \in \mathcal{H} \ \|f\| \leq 1} \{\|f\|_{\mathcal{H}}, \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}\} \geq \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

therefore

$$MMD = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

□

Proposition 2.4. *Given: $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$ and X and Y are independent then:*

$$MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \mathbb{E}(K(X, X')) + \mathbb{E}(K(Y, Y')) - 2\mathbb{E}K(X, Y).$$

Proof.

$$\begin{aligned}
MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathcal{H}} \\
&=< \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} >_{\mathcal{H}} \\
&=< \mathbb{E}(K(\cdot, X)) - K(\cdot, Y)), \mathbb{E}(K(\cdot, X')) - K(\cdot, Y')) > \\
&= \mathbb{E}(< K(\cdot, X), K(\cdot, X') > + < K(\cdot, Y), K(\cdot, Y') > -2 < K(\cdot, X)K(\cdot, Y) >) \\
&= {}^5\mathbb{E}(K(X, X') + K(Y, Y') - 2K(X, Y)) \\
&= \mathbb{E}(K(X, X')) + \mathbb{E}(K(Y, Y')) - 2\mathbb{E}(K(X, Y)) \\
&= \int \int K(s, t) \underbrace{d(\mathbb{P} - \mathbb{Q})(s)}_{SignedMeasure} d(\mathbb{P} - \mathbb{Q})(t)
\end{aligned} \tag{2}$$

□

Prooving that MMD defines an homogeneity test

Definition 2.4. Characteristic kernel

A reproducing kernel k is a characterisctic kernel if the induced γ_k is a metric.

Theorem 2.5. *If X is a compact metric space, k is continuous and $\mathcal{H}$ is dense in $\mathcal{C}(X)$ with respect to the supremum norm, then $\mathcal{H}$ is characteristic.*

Proof. Being characteristic means that $MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = 0 \leftrightarrow \mathbb{P} = \mathbb{Q}$

$\rightarrow$

By lemma 1 we know that $\mathbb{P}$ and $\mathbb{Q}$ are equal if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y)$ $\forall f \in \mathcal{C}(\mathcal{X})$

Given that $\mathcal{H}$ is dense in $\mathcal{C}$(X) then:

$$\forall \epsilon > 0, f \in \mathcal{C}(X), \exists g \in \mathcal{H} : \|f - g\|_\infty < \epsilon$$

$$\begin{aligned}
|\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| &= |\mathbb{E}(f(X)) - \mathbb{E}(g(X)) + \mathbb{E}(g(X)) - \mathbb{E}(g(Y)) + \mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\
&\leq |\mathbb{E}(f(X)) - \mathbb{E}(g(X))| + |\mathbb{E}(g(X)) - \mathbb{E}(g(Y))| + |\mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\
&= |\mathbb{E}(f(X)) - \mathbb{E}(g(X))| + |< g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} >_{\mathcal{H}}| + |\mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\
&\leq \mathbb{E}|f(X) - g(X)| + |< g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} >_{\mathcal{H}}| + \mathbb{E}|g(Y) - f(Y)| \\
&\leq^1 \|f - g\|_\infty + |< g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} >_{\mathcal{H}}| + \|f - g\|_\infty \\
&\leq |< g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} >_{\mathcal{H}}| + 2\epsilon
\end{aligned} \tag{3}$$

By lemma 3 we know that if MMD $= 0$ then $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$. Hence:

$$|\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| \leq 2\epsilon$$

Then by lemma 1 $\mathbb{P}$ and $\mathbb{Q}$ are equal.

$\leftarrow$

By definition of MMD. $\square$

2.3 Application to independence test

From the MMD criterion we will develop an independence criterion which will be conduced by the following idea: Given $\mathcal{X} \sim \mathbb{P}$ and $\mathcal{Y} \sim \mathbb{Q}$ whose joint distribution is $\mathbb{P}_{\mathcal{XY}}$ then the test of independence between these variables will be determining if $\mathbb{P}_{\mathcal{XY}}$ is equal to the product of the marginals $\mathbb{PQ}$. Therefore:

$\mathcal{MMD}(\mathcal{F}, \mathbb{P}_{\mathcal{XY}}, \mathbb{PQ}) = 0$ if and only if $\mathcal{X}$ and $\mathcal{Y}$ are independent. To characterize this independence test we need to introduce a new RKHS, which is a tensor product of the RKHSs in which the marginal distributions of the random variables are embedded. Let $\mathcal{X}$ and $\mathcal{Y}$ be two topological spaces and let k and l be kernels on these spaces, with respective RKHS $\mathcal{H}$ and $\mathcal{G}$. Let us denote as $\upsilon((x, y), (x', y'))$ a kernel on the product space $\mathcal{X} \times \mathcal{Y}$ with RKHS $\mathcal{H}_\upsilon$. This space is known as the tensor product space $\mathcal{H} \times \mathcal{G}$. Tensor product spaces are defined as follows:

Definition 2.5. Tensor product The tensor product of Hilbert spaces$\mathcal{H}_1$ and $\mathcal{H}_2$ with inner products $< , >_1$ and $< , >_2$ is defined as the completion of the space $\mathcal{H}_1 \times \mathcal{H}_2$ with inner product $< , >_1 < , >_2$extended by linearity. The resulting space is also a Hilbert space.

Lemma 2.6. *A kernel υ in the tensor product space $\mathcal{H} \times \mathcal{G}$ can be defined as:*

$$\upsilon((x, y), (x', y')) = k(x, x')l(y, y')$$

Useful definitions for the following content

$$\mathbb{E}_{\mathcal{X}} f(\mathcal{X}) = \int f(x) d\mathbb{P}(x)$$

$$\mathbb{E}_{\mathcal{Y}} f(\mathcal{Y}) = \int f(y) d\mathbb{Q}(y)$$

$$\mathbb{E}_{\mathcal{XY}} f(\mathcal{XY}) = \int f(x, y) d\mathbb{P}_{\mathcal{XY}}(x, y)$$

Using this notation, the mean embedding of $\mathbb{P}_{\mathcal{XY}}$ and $\mathbb{PQ}$ are:

$$\mu_{\mathbb{P}_{\mathcal{XY}}} = \mathbb{E}_{\mathcal{XY}} \upsilon((\mathcal{X}, \mathcal{Y}),)$$

$$\mu_{\mathbb{PQ}} = \mathbb{E}_{\mathcal{XY}} \upsilon((\mathcal{X}, \mathcal{Y}),)$$

In terms of these embeddings:

$$\mathcal{MMD}(\mathcal{F}, \mathbb{P}_{\mathcal{XY}}, \mathbb{PQ}) = \|\mathbb{P}_{\mathcal{XY}} - \mu_{\mathbb{PQ}}\|_{\mathbb{H}_\upsilon}$$

3 HSIC

In this section we will give a short overview of the cross-covariance operators between RKHSs and their Hilbert-Schmidt norms which later will be used to define the Hilbert Schmidt Independence Criterion (HSIC). After we will determine whether the dependence returned via HSIC is statistically significant by studying an hypothesis test with HSIC as its statistic and testing it empirically. Finally we will prove the ewuivalence of the HSIC test in terms of the Hilbert-Schmidt norm of the cross covariance operator in terms of the MMD between $\mathbb{P}_{\mathcal{XY}}$ and $\mathbb{PQ}$

3.1 Cross Covariance operator

Definition 3.1. Tensor product operator

Let $h \in \mathcal{H}, g \in \mathcal{G}$. The tensor product operator $h \otimes g : \mathcal{G} \to \mathcal{H}$ is defined as:

$$(h \otimes g)(f) =< g, f >_{\mathcal{G}} h, \forall f \in \mathcal{G}$$

Definition 3.2. Hilbert-Schmidt norm of a linear operator

Let $C : \mathcal{G} \to \mathcal{H}$ be a linear operator between RKHS $\mathbb{G}$ and $\mathcal{H}$ the Hilbert-Schmidt norm of C is defined as:

$$\|C\| = \sqrt{\sum < Cv_j, u_i >^2_{\mathcal{H}}}$$

Definition 3.3. Cross-Covariance operator

The cross-covariance operator associated with $\mathbb{P}_{XY}$ is the linear operator $C_{XY} : \mathcal{G} \to \mathcal{H}$ defined as:

$$C_{XY} = \mathbb{E}_{XY}[(\phi(X) - \mu_{\mathbb{P}}) \otimes (\psi(Y) - \mu_{\mathbb{Q}})] = {}^{6}\mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}$$

Which is a generalisation of the cross-covariance matrix between random vectors.

Definition 3.4. HSIC We define the Hilbert-Schmidt Independence Criterion for $\mathbb{P}_{\mathcal{XY}}$ as the squared HS norm of the associated cross-covariance operator:

$$HSIC(\mathbb{P}_{\mathcal{XY}}, \mathcal{H}, \mathcal{G}) = \|C_{XY}\|^2_{\mathcal{HS}}$$

Lemma 3.1. *If we denote $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$ then:*

$$HSIC(\mathbb{P}_{\mathcal{XY}}, \mathcal{H}, \mathcal{G}) = \mathbb{E}_{xx'yy'}[k(x,x')l(y,y')] + \mathbb{E}_{xx'}[k(x,x')]\mathbb{E}_{yy'}[l(y,y')] - 2\mathbb{E}_{xy}[\mathbb{E}_{x'}[k(x,x')]\mathbb{E}_{y'}[l(y,y')]]$$

Proof. First we will simplify the notation of C_{XY}

$$C_{XY} = \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] - \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} = \bar{C_{XY}} - M_{XY}$$

Using this notation:

$$\begin{aligned} \|C_{XY}\|^2_{\mathcal{HS}} &=< \bar{C}_{XY} - M_{XY}, \bar{C}_{X'Y'} - M_{X'Y'} >_{\mathcal{HS}} \\ &=< \bar{C}_{XY}, \bar{C}_{X'Y'} >_{\mathcal{HS}} + < M_{XY}, M_{X'Y'} > -2 < \bar{C}_{XY}, M_{X'Y'} >_{\mathcal{HS}} \end{aligned} \tag{4}$$

Now calculating each of this products individually:

[6] distributive property of the tensor product

$$
\begin{aligned}
< \bar{C}_{XY}, \bar{C}_{X'Y'} >_{\mathcal{HS}} &=< \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mathbb{E}_{X'Y'}[\phi(X) \otimes \psi(Y)] > \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|\phi(X) \otimes \psi(Y)\|^2 \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|\phi(X)\|^2\|\psi(Y)\|^2 \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'Y'} < \phi(X), \phi(X') >< \psi(Y), \psi(Y) > \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k(X, X')l(Y, Y')
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
< M_{XY}, M_{X'Y'} >_{\mathcal{HS}} &=< \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} >_{\mathcal{HS}} \\
&= \|\mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}\|^2_{\mathcal{HS}} \\
&= \|\mu_{\mathbb{P}}\|^2_{\mathcal{H}}\|\mu_{\mathbb{Q}}\|^2_{\mathcal{G}} \\
&=< \mu_{\mathbb{P}}, \mu_{\mathbb{P}} >_{\mathcal{H}} < \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} >_{\mathcal{G}} \\
&=< \mathbb{E}_X k(X, \cdot), \mathbb{E}_{X'}k(X', \cdot) >_{\mathcal{H}} < \mathbb{E}_Y l(Y, \cdot), \mathbb{E}_{Y'}k(Y', \cdot) >_{\mathcal{G}} \\
&= \mathbb{E}_X\mathbb{E}_{X'}\mathbb{E}_Y\mathbb{E}_{Y'} < k(X, \cdot), k(X', \cdot) >_{\mathcal{H}} < l(Y, \cdot), l(Y', \cdot) >_{\mathcal{G}} \\
&= \mathbb{E}_X\mathbb{E}_{X'}\mathbb{E}_Y\mathbb{E}_{Y'}k(X, X')l(Y, Y')
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
< \bar{C}_{XY}, M_{XY} >_{\mathcal{HS}} &=< \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} >_{\mathcal{HS}} \\
&=< \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mathbb{E}_{X'}\phi(X') \otimes \mathbb{E}_{Y'}\psi(Y') >_{\mathcal{HS}} \\
&=< \mathbb{E}_{XY} < \mathbb{E}_{X'} < \mathbb{E}_{Y'} < \phi(X) \otimes \psi(Y), \phi(X') \otimes \psi(Y') >_{\mathcal{HS}} \\
&=< \mathbb{E}_{XY} < \mathbb{E}_{X'} < \mathbb{E}_{Y'} < \phi(X), \phi(X') >_{\mathcal{H}} < \psi(Y), \psi(Y') >_{\mathcal{G}} \\
&=< \mathbb{E}_{XY} < \mathbb{E}_{X'} < \mathbb{E}_{Y'}k(X, X')l(Y, Y').
\end{aligned} \tag{7}
$$

□

3.2 Statistics

In the previous subsection we defined the HSIC statistic.

$$HSIC(\mathbb{P}_{\mathcal{XY}}, \mathcal{H}, \mathcal{G}) = \mathbb{E}_{xx'yy'}[k(x, x')l(y, y')] + \mathbb{E}_{xx'}[k(x, x')]\mathbb{E}_{yy'}[l(y, y')] - 2\mathbb{E}_{xy}[\mathbb{E}_{x'}[k(x, x')]\mathbb{E}_{y'}[l(y, y')]]$$

In this section we will define the Empirical HSIC.

Definition 3.5. Empirical HSIC

$$HSIC(\mathbb{P}_{\mathcal{XY}}, \mathcal{H}, \mathcal{G}) = (m-1)^{-2}\mathbf{tr}KHLH$$

where: $H, K, L \in \mathbb{R}^{m \times m}$, $K_{i,j} = k(x_i, y_j)$, $L_{i,j} = l(x_i, y_j)$ and $H_{i,j} =$

$\delta_{i,j} - m^{-1}$

Theorem 3.2. *let $\mathbb{E}_Z$ denote the expectation taken over m independent copies$(x_i, y_i$ drawn from $P_{\mathcal{XY}}$. Then:*

$$HSIC(\mathbb{P}_{\mathcal{XY}}, \mathcal{H}, \mathcal{G}) = \mathbb{E}_Z[HSIC(Z, \mathcal{H}, \mathcal{G})] + O(m^{-1})$$

.

Proof. By definition of H we can write:

$$\mathbf{tr}KHLH = \mathbf{tr}KL - 2m^{-1}\mathbf{1}^T KL\mathbf{1} + m^{-2}\mathbf{tr}K\mathbf{tr}L$$

where $\mathbf{1}$ is the vector of all ones.

Now we will expand each of the terms separately and take expectations with respect to Z.

1. $\mathbb{E}_Z[\mathbf{tr}KL]$:

   $$\mathbb{E}_Z[\sum_i K_{ii}L_{ii} + \sum_{(i,j)\in i_2^m} K_{ij}Lji] = O(m) + (m)_2\mathbb{E}_{XYX'Y'}[k(X,X')l(Y,Y')]$$

   Normalising terms by $\frac{1}{(m1)^2}$ yields the first term, since $\frac{m(m1)}{(m1)^2} = 1+O(m^1)$.

2. $\mathbb{E}_Z[\mathbf{1}^T KL\mathbf{1}]$:

   $$\mathbb{E}_Z[\sum_i K_{ii}L_{ii} + \sum_{(i,j)\in i_2^m}(K_{ii}Lij + K_{ij}Ljj)] + \mathbb{E}_Z[\sum_{(i,j,r)\in i_3^m} K_{ij}Ljr]$$

   $$= O(m^2) + (m)_3\mathbb{E}_{XY}[\mathbb{E}_{X'}[k(x,x')]\mathbb{E}_{Y'}[l(Y,Y')]]$$

   Again, normalising terms by $\frac{2}{(m1)^2}$ yields the second term. As before we used that $\frac{m(m1)}{(m1)^2} = 1 + O(m1)$.

3. $\mathbb{E}_Z[\mathbf{tr}K\mathbf{tr}L]$:

   $$O(m^3)+\mathbb{E}_Z[\sum_{(i,j,q,r)\in i_4^m} K_{ij}Lqr] = O(m^3)+(m)_4\mathbb{E}_{XX'}[k(x,x')]\mathbb{E}_{YY'}[l(Y,Y')]$$

   Normalisation by $\frac{1}{(m1)^2}$ takes care of the last term, which completes the proof.

□

Theorem 3.3. *Under the $\mathcal{H}_0$ the U-statistic HSIC cirresponding to the V-statistic*

$$HSIC(Z) = \frac{1}{m^4} \sum_{i,j,q,r \in_4^m} h_{ijqr}$$

is degenerate, meaning $\mathbb{E}h = 0$. In this case, HSIC(Z) converges in distribution according to [2], section 5.5.2

$$mHSIC(Z) \to \sum_{l=1} \lambda_l z_l^2$$

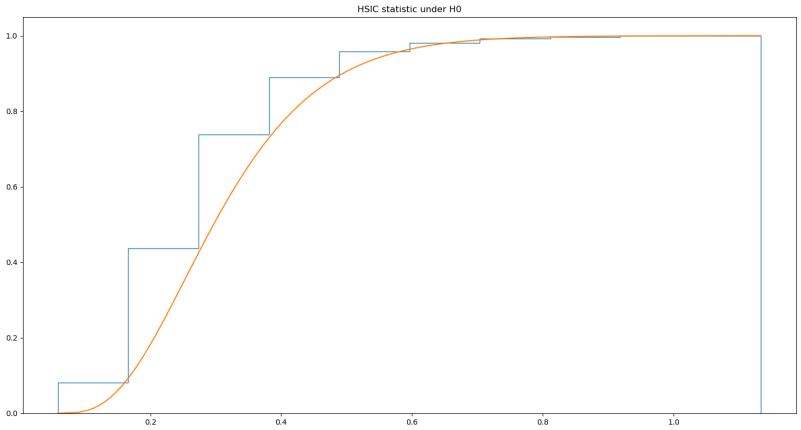
where $z_l \sim \mathcal{N}(0,1)$ i.i.d and λ_l are the solutions to the eigenvalue problem

$$\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{iqr}$$

where the integral is over the distribution of variables z_i, z_q and z_r[1]

Approximating the $1-\alpha$ quantile of the null distribution A hypothesis test using HSIC(Z) could be derived from Theorem 3.3 above by computing the (1α)th quantile of the distribution $\sum_{l=1} \lambda_l z_l^2$, where consistency of the test (that is, the convergence to zero of the Type II error for $m \to \infty$) is guaranteed by the decay as m^{-1} of the variance of HSIC(Z) under H_1 . The distribution under H_0 is complex, however: the question then becomes how to accurately approximate its quantiles.

One approach taken by [1] is by using a Gamma distribution, which as we can see in the figure underneath is quite accurate.



4 Energy distance

In this section we will define energy distance and we will use it to define a homogeneity test. This knowledge will be used in order to formulate another independence test based on energy distance, distance covariance and distance correlation. This test is one of the most popular nowadays because of its power and the fact that it does not depend on any parameter.

4.1 Definitions

One of the simplest distances we can define between two distributions F and G is the L 2 one, although it has the drawback that the distribution of its natural estimate is not distribution-free. That is, the distribution of the estimate depends on the distribution F under the null hypothesis. However, we can extend this distance easily to higher dimensions, having the property of being rotation invariant. Then energy distances can be derived as a variation of the L 2 distance, given by the following proposition:

Proposition 4.1. *Let $\mathcal{F}$ and $\mathcal{G}$ be two CDFs of the independent random variables X,Y respectively and X',Y' two iid copies of them, then:*

$$2\int_{-\infty}^{\infty}(\mathcal{F}(x)-\mathcal{G}(x))^2dx = 2\mathbb{E}|X-Y| - \mathbb{E}|X-X'| - \mathbb{E}|Y-Y'|$$

Proof. We will start analysing the expectations of the right hand side. We will use that for any positive random variable $Z > 0$, $\mathbb{E}Z = \int_0^\infty \mathbb{P}(Z > z)dz$

$$\begin{aligned}
\mathbb{E}|X-Y| &= \int_0^\infty \mathbb{P}(|X-Y|>u)du \\
&= \int_0^\infty \mathbb{P}(X-Y>u)du + \int_0^\infty \mathbb{P}(X-Y<u)du \\
&= \int_0^\infty\int_{-\infty}^\infty \mathbb{P}(X-Y>u|Y=y)d\mathcal{G}(y)du + \int_0^\infty\int_{-\infty}^\infty \mathbb{P}(X-Y<u|X=x)d\mathcal{F}(x)(y)du \\
&= {}^{7}\int_{-\infty}^\infty\int_0^\infty \mathbb{P}(X-Y>u|Y=y)du\mathcal{G}(y) + \int_{-\infty}^\infty\int_0^\infty \mathbb{P}(X-Y<u|X=x)du\mathcal{F}(x) \\
&= \int_{-\infty}^\infty\int_0^\infty \mathbb{P}(X>u+y)du\mathcal{G}(y) + \int_{-\infty}^\infty\int_0^\infty \mathbb{P}(Y>u+x)du\mathcal{F}(x)
\end{aligned} \tag{8}$$

Now we use the change of variables z = u + y for the first integral, and w = u + x for the second one. Applying Fubini again:

$$
\begin{aligned}
\mathbb{E}|X-Y| &= \int_{-\infty}^{\infty}\int_{y}^{\infty}\mathbb{P}(X>z)dz\mathcal{G}(y) + \int_{-\infty}^{\infty}\int_{x}^{\infty}\mathbb{P}(Y>w)dw\mathcal{F}(x) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X>z)dz\int_{y}^{\infty}\mathcal{G}(y) + \int_{-\infty}^{\infty}\mathbb{P}(Y>w)dw\int_{x}^{\infty}\mathcal{F}(x) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X>z)\mathbb{P}(Y<z)dz + \int_{-\infty}^{\infty}\mathbb{P}(Y>w)\mathbb{P}(X<w)dw \qquad (9) \\
&= \int_{-\infty}^{\infty}[(1-\mathcal{F}(z))\mathcal{G}(z) + (1-\mathcal{G}(z))\mathcal{F}(z)]dz \\
&= -2\int_{-\infty}^{\infty}\mathcal{F}(z)\mathcal{G}(z)dz + \mathbb{E}|X| + \mathbb{E}|Y|
\end{aligned}
$$

□

References

[1] Gretton A,Fukumizu K, Teo H.C., Song L, Schlkopf B, Smola J.A. (2007) *A Kernel Statistical Test of Independence*

[2] Serfling R. (Wiley, New York, 1980) *Approximation Theorems of Mathematical Statistics*