

Escuela Politécnica Superior

18
19

Trabajo fin de grado

Independence tests based on embeddings in functional spaces



Roberto Alcover Couso

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C/ Francisco Tomás y Valiente nº 11

**UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR**



Grado en

TRABAJO FIN DE GRADO

**Independence tests based on embeddings in
functional spaces**

Autor: Roberto Alcover Couso

Tutor: Alberto Suárez González

abril 2019

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de la propiedad intelectual.

La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. del Código Penal*).

DERECHOS RESERVADOS

© 3 de Noviembre de 2017 por UNIVERSIDAD AUTÓNOMA DE MADRID
Francisco Tomás y Valiente, n^o 1
Madrid, 28049
Spain

Roberto Alcover Couso

Independence tests based on embeddings in functional spaces

Roberto Alcover Couso

C\ Francisco Tomás y Valiente N^o 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

AGRADECIMIENTOS

dsuijui

RESUMEN

Medir la dependencia estadística entre variables aleatorias es un problema fundamental en el área de la estadística. Los test clásicos de dependencia como el ρ de Pearson o el τ de Kendall son comúnmente aplicados debido a que son computacionalmente eficientes y están bien entendidos y estudiados, pero estos tests solamente consideran un conjunto limitado de patrones de asociación, como lineal o funciones monótonas crecientes. El desarrollo de medidas de dependencia no lineales es complejo debido a la cantidad de posibles patrones de asociación que se pueden presentar.

En este trabajo se van a presentar tres planteamientos para medir las dependencias no lineales: mediante el uso de medidas de independencia basados en kernels (HSIC), correlación canónica entre proyecciones aleatorias no lineales (RDC) y un test basado en las funciones características (DCOV)

La estructura que seguirá el proyecto es la siguiente:

Al principio de este trabajo se presenta un test de homogeneidad, MMD, basado en empotramiento de media de las variables originales mediante transformaciones no lineales a espacios de Hilbert con un núcleo reproductivo, RKHS, estos nuevos conocimientos se usarán para llegar al primer test de independencia, HSIC.

En segundo lugar estudiaremos el concepto de distancia de energía e introduciremos un segundo test de independencia, DCOV. Posteriormente se pasará a estudiar la equivalencia entre este test y MMD.

Finalmente presentaremos el último test de independencia, RDC, concluyendo en la comparación de estos tres test entre ellos y otros tests existentes.

PALABRAS CLAVE

MMD, HSIC, RDC, DCOV, DCOR, dependencia estadística, variables aleatorias, dependencia no lineal

ABSTRACT

Measuring statistical dependence between random variables is a fundamental problem in statistics. Classical tests of dependence such as Pearson's ρ or Kendall's τ are widely applied due to being computationally efficient and theoretically well understood, however they consider only a limited class of association patterns, like linear or monotonically increasing functions. The development of non-linear dependence measures is challenging because of the radically larger amount of possible association patterns.

In this work three main approaches of non-linear dependence measures will be presented: by using kernel independence measures (HSIC), canonical correlation between random non-linear projections (RDC) and a characteristic function based test (DCOV).

The structure of the work will go as it follows:

In the beginning of the work, which is composed of Chapters , an homogeneity test ,MMD, based on mean embeddings of the original variables through non-linear transformations into Hilbert spaces with reproducing kernel ,RKHS, will be introduced, this new intuitions will lead us to our first independence test ,HSIC.

Secondly we will study the concept of energy distance and introduce the second independence test ,DCOV. Followed by an study of the equivalence of this tests with MMD.

Finally RDC will be presented, concluding with a comparison of this three tests between them and with other tests.

KEYWORDS

MMD,HSIC,RDC,DCOV,DCOR,statistical dependence, ramdom variables, non-linear dependence

ÍNDICE

1	Introduction	1
1.1	MMD	2
1.1.1	Mean embedding	2
1.1.2	Introduction to MMD	4
1.1.3	Prooving that MMD defines an homogeneity test	5
1.1.4	Application to independence test	6
1.2	HSIC	7
1.2.1	Cross Covariance operator	7
1.2.2	Statistics	9
1.3	Energy	11
1.3.1	Definitions	11
1.3.2	Application to an independence test	14

LISTAS

Lista de algoritmos

Lista de códigos

Lista de cuadros

Lista de ecuaciones

Lista de figuras

Lista de tablas

Lista de cuadros

INTRODUCTION

How to measure dependence of variables is a classical yet fundamental problem in statistics. Starting with the Galton's work of Pearson's correlation coefficient [Stigler, 1989] for measuring linear dependence, many techniques have been proposed, which are of fundamental importance in scientific fields such as physics, chemistry, biology, and economics. In Statistics, probability measures are used in a variety of applications, such as hypothesis testing, density estimation or Markov chain monte carlo. We will focus on hypothesis testing, mainly in homogeneity testing. The goal in homogeneity testing is to accept or reject the null hypothesis $\mathcal{H}_0: \mathbb{P} = \mathbb{Q}$, versus the alternative hypothesis $\mathcal{H}_1: \mathbb{P} \neq \mathbb{Q}$, for a class of probability distributions \mathbb{P} and \mathbb{Q} . For this purpose we will define a metric γ such that testing the null hypothesis is equivalent to testing for $\gamma(\mathbb{P}\mathbb{Q}) = 0$. We are specially interested in testing for independence between random vectors, which is a particular case of homogeneity testing, using $\mathbb{P} = \mathbb{P}_{\mathcal{X}\mathcal{Y}}$ and $\mathbb{Q} = \mathbb{P}_{\mathcal{X}} \cdot \mathbb{P}_{\mathcal{Y}}$. An example of a practical application of this tests is Principal Component Analysis (PCA), which is a statistical procedure that converts a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principal components.

In this work three main approaches of non-linear dependence measures will be presented: by using kernel independence measures (HSIC), canonical correlation between random non-linear projections (RDC) and a characteristic function based test (DCOV).

The structure of the work will go as it follows:

In the beginning of the work, which is composed of Chapters , an homogeneity test ,MMD, based on mean embeddings of the original variables through non-linear transformations into Hilbert spaces with reproducing kernel ,RKHS, will be introduced this new intuitions will lead us to our first independence test ,HSIC.

Secondly we will study the concept of energy distance and introduce the second independence test ,DCOV. Followed by an study of the equivalence of this tests with MMD.

Finally RDC will be presented, concluding with a comparison of this three tests between them and with other tests.

1.1. MMD

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.

1.1.1. Mean embedding

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 difficult to prove therefore we will keep our study in order to simplify this evaluation.

Definition 1.1.1. MMD

Let \mathcal{F} be a class of functions $f: X \rightarrow \mathbb{R}$ 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 1.1.2. 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 $\forall f \in \mathcal{H}$:

$$T(f) = \langle f, g \rangle_{\mathcal{H}}$$

Lemma 1.1.1. Given a $K(s, \cdot)$ 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 fulfills the next condition $\mathbb{E}f(X) = \langle f, \mu_p \rangle$ for all $f \in \mathcal{H}$

proof

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

$$|T_{\mathbb{P}}f| = |\mathbb{E}(f(X))| \\ \leq \mathbb{E}(|f(X)|)$$

Reproducing property of the kernel

$$= \mathbb{E}|\langle f, k(\cdot, X) \rangle_{\mathcal{H}}|$$

(1.1)

Chauchy Schwarz inequality

$$\leq \|f\|_{\mathcal{H}} \cdot \mathbb{E}(\sqrt{K(X, X)})^{1/2}$$

The expectation under \mathbb{P} of the kernel is bounded

$$< \infty$$

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

Definition 1.1.3. Mean embedding

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

$$\mathbb{E}(f(X)) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$$

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

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

$\forall x_n$ then:

$$f(x_n) = \langle f, K(\cdot, x_n) \rangle_{\mathcal{H}}$$

then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N f(x_n) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \langle f, K(\cdot, x_n) \rangle_{\mathcal{H}} = \langle f, \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N K(\cdot, x_n) \rangle_{\mathcal{H}}$$

which leads to the final conclusion:

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

SECOND INTERPRETATION OF THE MEAN EMBEDDING

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

1.1.2. Introduction to MMD

Lemma 1.1.2. *Given the conditions of Lemma 2.2 ($\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ exist) then:*

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

and:

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

proof

$$\begin{aligned} MMD &\equiv \sup_{f \in \mathcal{H} \|f\| \leq 1} \{\mathbb{E}(f(x)) - \mathbb{E}(f(y))\} \\ &= \sup_{f \in \mathcal{H} \|f\| \leq 1} \{ \langle f, \mu_{\mathbb{P}} \rangle - \langle f, \mu_{\mathbb{Q}} \rangle \} \\ &= \sup_{f \in \mathcal{H} \|f\| \leq 1} \langle f, (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}) \rangle \\ &\leq \sup_{f \in \mathcal{H} \|f\| \leq 1} \{ \|f\|_{\mathcal{H}}, \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \} \\ &\leq \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}. \end{aligned} \tag{1.2}$$

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 1.1.3. *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

$$\begin{aligned}
MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2 \\
&= \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\
&= \langle \mathbb{E}(K(\cdot, X)) - K(\cdot, Y), \mathbb{E}(K(\cdot, X')) - K(\cdot, Y') \rangle \\
&= \mathbb{E}(\langle K(\cdot, X), K(\cdot, X') \rangle + \langle K(\cdot, Y), K(\cdot, Y') \rangle - 2 \langle K(\cdot, X), K(\cdot, Y) \rangle) \\
&= 2\mathbb{E}(K(X, X') + K(Y, Y') - 2K(X, Y)) \\
&= \mathbb{E}(K(X, X')) + \mathbb{E}(K(Y, Y')) - 2\mathbb{E}(K(X, Y)) \\
&= \int \int K(s, t) \underbrace{d(\mathbb{P} - \mathbb{Q})(s) d(\mathbb{P} - \mathbb{Q})(t)}_{\text{Signed Measure}} \\
\end{aligned} \tag{1.3}$$

1.1.3. Proving that MMD defines an homogeneity test

Definition 1.1.4. Characteristic kernel

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

Theorem 1.1.4. 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 $MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$

→

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

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

$$\begin{aligned}
|\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| &= |\mathbb{E}(f(X)) - \mathbb{E}(g(X)) + \mathbb{E}(g(X)) - \mathbb{E}(g(Y)) + \mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\
&\leq |\mathbb{E}(f(X)) - \mathbb{E}(g(X))| + |\mathbb{E}(g(X)) - \mathbb{E}(g(Y))| + |\mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\
&= |\mathbb{E}(f(X)) - \mathbb{E}(g(X))| + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + |\mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\
&\leq \mathbb{E}|f(X) - g(X)| + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + \mathbb{E}|g(Y) - f(Y)| \\
&\leq^1 \|f - g\|_{\infty} + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + \|f - g\|_{\infty} \\
&\leq |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + 2\epsilon
\end{aligned} \tag{1.4}$$

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.

←

By definition of MMD.

1.1.4. Application to independence test

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

$\mathcal{MMD}(\mathcal{F}, \mathbb{P}_{\mathcal{X}\mathcal{Y}}, \mathbb{P}\mathbb{Q}) = 0$ 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 RKHS's 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 $v((x, y), (x', y'))$ a kernel on the product space $\mathcal{X} \times \mathcal{Y}$ with RKHS \mathcal{H}_v . This space is known as the tensor product space $\mathcal{H} \times \mathcal{G}$. Tensor product spaces are defined as follows:

Definition 1.1.5. Tensor product The tensor product of Hilbert spaces \mathcal{H}_1 and \mathcal{H}_2 with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ is defined as the completion of the space $\mathcal{H}_1 \times \mathcal{H}_2$ with inner product $\langle \cdot, \cdot \rangle_1 \otimes \langle \cdot, \cdot \rangle_2$ extended by linearity. The resulting space is also a Hilbert space.

Lemma 1.1.5. A kernel v in the tensor product space $\mathcal{H} \times \mathcal{G}$ can be defined as:

$$v((x, y), (x', y')) = k(x, x')l(y, y')$$

Useful definitions for the following content

$$\begin{aligned}\mathbb{E}_{\mathcal{X}}f(\mathcal{X}) &= \int f(x)d\mathbb{P}(x) \\ \mathbb{E}_{\mathcal{Y}}f(\mathcal{Y}) &= \int f(y)d\mathbb{Q}(y) \\ \mathbb{E}_{\mathcal{X}\mathcal{Y}}f(\mathcal{X}\mathcal{Y}) &= \int f(x, y)d\mathbb{P}_{\mathcal{X}\mathcal{Y}}(x, y)\end{aligned}$$

Using this notation, the mean embedding of $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ and $\mathbb{P}\mathbb{Q}$ are:

$$\begin{aligned}\mu_{\mathbb{P}_{\mathcal{X}\mathcal{Y}}} &= \mathbb{E}_{\mathcal{X}\mathcal{Y}}v((\mathcal{X}, \mathcal{Y}), \cdot) \\ \mu_{\mathbb{P}\mathbb{Q}} &= \mathbb{E}_{\mathcal{X}\mathcal{Y}}v((\mathcal{X}, \mathcal{Y}), \cdot)\end{aligned}$$

In terms of these embeddings:

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

1.2. 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 $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ and $\mathbb{P}_{\mathcal{Q}}$

1.2.1. Cross Covariance operator

Definition 1.2.1. Tensor product operator

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

$$(h \otimes g)(f) = \langle g, f \rangle_{\mathcal{G}} h, \forall f \in \mathcal{G}$$

Definition 1.2.2. Hilbert-Schmidt norm of a linear operator

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

$$\|C\| = \sqrt{\sum \langle Cv_j, u_i \rangle_{\mathcal{H}}^2}$$

Definition 1.2.3. Cross-Covariance operator

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

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

by applying the distributive property of the tensor product

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

Definition 1.2.4. HSIC We define the Hilbert-Schmidt Independence Criterion for $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ as the squared HS norm of the associated cross-covariance operator:

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

Lemma 1.2.1. *If we denote $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$ then:*

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

Demostración. 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:

$$\begin{aligned} \|C_{XY}\|_{\mathcal{HS}}^2 &= \langle \bar{C}_{XY} - M_{XY}, \bar{C}_{XY} - M_{XY} \rangle_{\mathcal{HS}} \\ &= \langle \bar{C}_{XY}, \bar{C}_{XY} \rangle_{\mathcal{HS}} + \langle M_{XY}, M_{XY} \rangle_{\mathcal{HS}} - 2 \langle \bar{C}_{XY}, M_{XY} \rangle_{\mathcal{HS}} \end{aligned} \quad (1.5)$$

Now calculating each of this products individually:

$$\begin{aligned} \langle \bar{C}_{XY}, \bar{C}_{XY} \rangle_{\mathcal{HS}} &= \langle \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] \rangle_{\mathcal{HS}} \\ &= \mathbb{E}_{XY} \mathbb{E}_{XY} [\|\phi(X) \otimes \psi(Y)\|^2] \\ &= \mathbb{E}_{XY} \mathbb{E}_{XY} [\|\phi(X)\|^2 \|\psi(Y)\|^2] \\ &= \mathbb{E}_{XY} \mathbb{E}_{XY} [\langle \phi(X), \phi(X) \rangle \langle \psi(Y), \psi(Y) \rangle] \\ &= \mathbb{E}_{XY} \mathbb{E}_{XY} [k(X, X') l(Y, Y')] \end{aligned} \quad (1.6)$$

$$\begin{aligned} \langle M_{XY}, M_{XY} \rangle_{\mathcal{HS}} &= \langle \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} \rangle_{\mathcal{HS}} \\ &= \|\mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}}\|_{\mathcal{HS}}^2 \\ &= \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 \|\mu_{\mathbb{Q}}\|_{\mathcal{G}}^2 \\ &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{G}} \\ &= \langle \mathbb{E}_X k(X, \cdot), \mathbb{E}_X k(X, \cdot) \rangle_{\mathcal{H}} \langle \mathbb{E}_Y l(Y, \cdot), \mathbb{E}_Y l(Y, \cdot) \rangle_{\mathcal{G}} \\ &= \mathbb{E}_X \mathbb{E}_X \mathbb{E}_Y \mathbb{E}_Y [\langle k(X, \cdot), k(X, \cdot) \rangle_{\mathcal{H}} \langle l(Y, \cdot), l(Y, \cdot) \rangle_{\mathcal{G}}] \\ &= \mathbb{E}_X \mathbb{E}_X \mathbb{E}_Y \mathbb{E}_Y [k(X, X') l(Y, Y')] \end{aligned} \quad (1.7)$$

$$\begin{aligned} \langle \bar{C}_{XY}, M_{XY} \rangle_{\mathcal{HS}} &= \langle \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} \rangle_{\mathcal{HS}} \\ &= \langle \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mathbb{E}_{X'} \phi(X') \otimes \mathbb{E}_{Y'} \psi(Y') \rangle_{\mathcal{HS}} \\ &= \langle \mathbb{E}_{XY} \langle \mathbb{E}_{X'} \langle \mathbb{E}_{Y'} \langle \phi(X) \otimes \psi(Y), \phi(X') \otimes \psi(Y') \rangle_{\mathcal{HS}} \rangle_{\mathcal{H}} \rangle_{\mathcal{G}} \\ &= \langle \mathbb{E}_{XY} \langle \mathbb{E}_{X'} \langle \mathbb{E}_{Y'} \langle \phi(X), \phi(X') \rangle_{\mathcal{H}} \langle \psi(Y), \psi(Y') \rangle_{\mathcal{G}} \rangle_{\mathcal{H}} \rangle_{\mathcal{G}} \\ &= \langle \mathbb{E}_{XY} \langle \mathbb{E}_{X'} \langle \mathbb{E}_{Y'} k(X, X') l(Y, Y') \rangle_{\mathcal{G}} \rangle_{\mathcal{H}} \rangle_{\mathcal{H}} \end{aligned} \quad (1.8)$$

□

1.2.2. Statistics

In the previous subsection we defined the HSIC statistic.

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

In this section we will define the Empirical HSIC.

Definition 1.2.5. Empirical HSIC

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

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 1.2.2. let \mathbb{E}_Z denote the expectation taken over m independent copies (x_i, y_i) drawn from $P_{\mathcal{X}\mathcal{Y}}$. Then:

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

Demostración. 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 .

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

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

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

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

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

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

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

- $\mathbb{E}_Z[\text{tr}K\text{tr}L]$:

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

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

□

Theorem 1.2.3. *Under the \mathcal{H}_0 the U-statistic HSIC cirresponding to the V-statistic*

$$HSIC(Z) = \frac{1}{m^4} \sum_{i,j,q,r \in i_4^m} h_{ijqr}$$

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

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

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

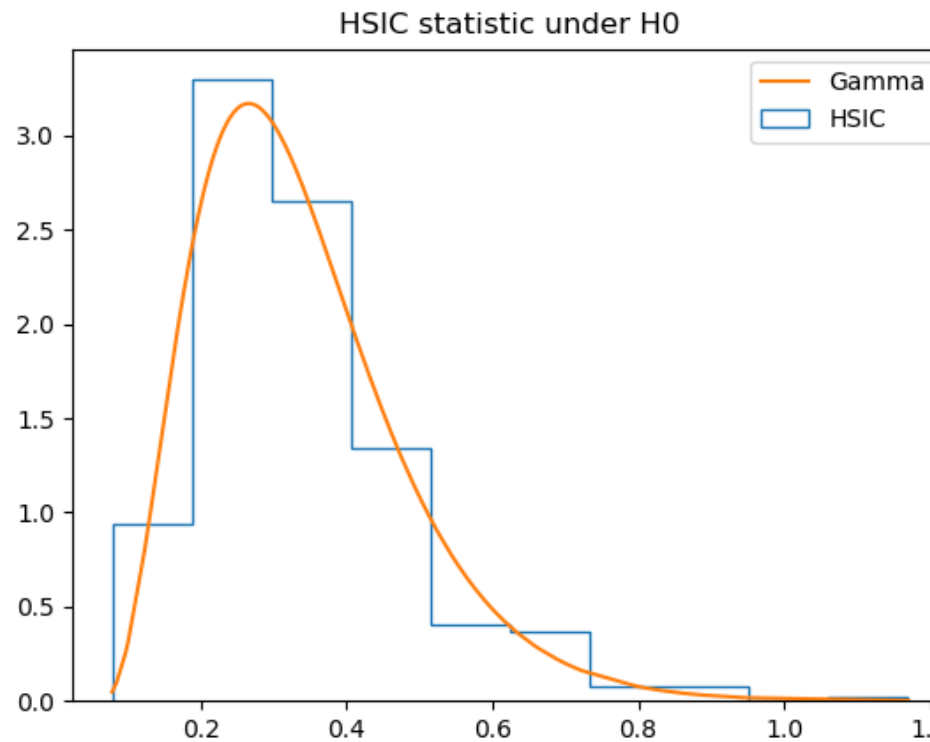
$$\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{iqr}$$

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 $(1 - \alpha)$ 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 $m \rightarrow \infty$) 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 un-



derneath is quite accurate.

1.3. Energy

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.

1.3.1. Definitions

Proposition 1.3.1. *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:*

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

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

$$\begin{aligned}
\mathbb{E}|X - Y| &= \int_0^\infty \mathbb{P}(|X - Y| > u) du \\
&= \int_0^\infty \mathbb{P}(X - Y > u) du + \int_0^\infty \mathbb{P}(X - Y < u) du \\
&= \int_0^\infty \int_{-\infty}^\infty \mathbb{P}(X - Y > u | Y = y) d\mathcal{G}(y) du + \int_0^\infty \int_{-\infty}^\infty \mathbb{P}(X - Y < u | X = x) d\mathcal{F}(x)(y) du \\
&= 3 \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(X - Y > u | Y = y) du \mathcal{G}(y) + \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(X - Y < u | X = x) du \mathcal{F}(x) \\
&= \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(X > u + y) du \mathcal{G}(y) + \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(Y > u + x) du \mathcal{F}(x)
\end{aligned} \tag{1.9}$$

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:

$$\begin{aligned}
\mathbb{E}|X - Y| &= \int_{-\infty}^\infty \int_y^\infty \mathbb{P}(X > z) dz \mathcal{G}(y) + \int_{-\infty}^\infty \int_x^\infty \mathbb{P}(Y > w) dw \mathcal{F}(x) \\
&= \int_{-\infty}^\infty \mathbb{P}(X > z) dz \int_y^\infty \mathcal{G}(y) + \int_{-\infty}^\infty \mathbb{P}(Y > w) dw \int_x^\infty \mathcal{F}(x) \\
&= \int_{-\infty}^\infty \mathbb{P}(X > z) \mathbb{P}(Y < z) dz + \int_{-\infty}^\infty \mathbb{P}(Y > w) \mathbb{P}(X < w) dw \\
&= \int_{-\infty}^\infty [(1 - \mathcal{F}(z)) \mathcal{G}(z) + (1 - \mathcal{G}(z)) \mathcal{F}(z)] dz \\
&= -2 \int_{-\infty}^\infty \mathcal{F}(z) \mathcal{G}(z) dz + \mathbb{E}|X| + \mathbb{E}|Y|
\end{aligned} \tag{1.10}$$

Taking $\mathcal{F} = \mathcal{G}$ in the previous development:

$$\mathbb{E}|X - X'| = -2 \int_{-\infty}^\infty \mathcal{F}^2(z) dz + 2\mathbb{E}|X|$$

Equivalently for Y . Combining these partial results concludes the proof. □

Definition 1.3.1. Let X and Y be random variables in \mathbb{R}^d of $\mathbb{E}\|X\|_d + \mathbb{E}\|Y\|_d < \infty$ the energy distance between X and Y is defined as:

$$\varepsilon(X, Y) = 2\mathbb{E}\|X - Y\|_d - \mathbb{E}\|X - X'\|_d - \mathbb{E}\|Y - Y'\|_d$$

where X' and Y' are i.i.d copies of X and Y respectively. The energy distance can also be defined in terms of the characteristic functions. In fact, it can be seen as a weighted \mathcal{L}_2 distance between characteristic functions.

Proposition 1.3.2. Given two independent d -dimensional random variables X and Y , with distributions \mathbb{P} and \mathbb{Q} respectively such that $\mathbb{E}\|X\|_d + \mathbb{E}\|Y\|_d < \infty$ the energy distance between X and Y can be written as:

$$\varepsilon(X, Y) = \frac{1}{c_d} \int_{\mathbb{R}^d} \frac{|\phi_{\mathbb{P}}(t) - \phi_{\mathbb{Q}}(t)|^2}{\|t\|_d^{d+1}} dt$$

where

$$c_d = \frac{\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})}$$

being $\Gamma(\cdot)$ the gamma function

To prove this proposition we need the following lemma.

Lemma 1.3.3. For all $x \in \mathbb{R}^d$ then:

$$\int_{\mathbb{R}^d} \frac{1 - \cos(tx)}{\|t\|_d^{d+1}} dt = c_d \|x\|_d$$

where tx is the inner product of t and x .

Demostración. We will begin by applying the following transformation: $z_1 = \frac{tx}{\|x\|_d}$ followed by the following change of variables: $s = z\|x\|_d$

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{1 - \cos(tx)}{\|t\|_d^{d+1}} dt &= \int_{\mathbb{R}^d} \frac{1 - \cos(z\|x\|_d)}{\|z\|_d^{d+1}} dz \\ &= \int_{\mathbb{R}^d} \frac{1 - \cos(s)}{\frac{\|s\|_d}{\|x\|_d}^{d+1} \|x\|_d^d} ds \\ &= \|x\|_d \int_{\mathbb{R}^d} \frac{1 - \cos(s)}{\|s\|_d^{d+1}} ds \\ &= \|x\|_d \frac{\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})} \end{aligned} \tag{1.11}$$

□

Demostración. (Proposition 1.3.2) Let $\overline{\phi_{\mathbb{P}}(t)}$ denote the complex conjugate of the characteristic function.

$$\begin{aligned}
|\phi_{\mathbb{P}}(t) - \phi_{\mathbb{Q}}(t)|^2 &= (\phi_{\mathbb{P}}(t) - \phi_{\mathbb{Q}}(t))\overline{(\phi_{\mathbb{P}}(t) - \phi_{\mathbb{Q}}(t))} \\
&= (\phi_{\mathbb{P}}(t) - \phi_{\mathbb{Q}}(t))(\overline{\phi_{\mathbb{P}}(t)} - \overline{\phi_{\mathbb{Q}}(t)}) \\
&= \phi_{\mathbb{P}}(t)\overline{\phi_{\mathbb{P}}(t)} - \phi_{\mathbb{P}}(t)\overline{\phi_{\mathbb{Q}}(t)} - \phi_{\mathbb{Q}}(t)\overline{\phi_{\mathbb{P}}(t)} + \phi_{\mathbb{Q}}(t)\overline{\phi_{\mathbb{Q}}(t)} \\
&= \mathbb{E}[e^{itX}e^{-itX'}] - \mathbb{E}[e^{itX}e^{-itY}] - \mathbb{E}[e^{itY}e^{-itX}] + \mathbb{E}[e^{itY}e^{-itY'}] \\
&= \mathbb{E}[e^{it(X-X')} - e^{it(Y-X)} - e^{it(X-Y)} + e^{it(Y-Y')}] \\
&= \mathbb{E}[\cos(t(X-X')) + i\sin(t(X-X')) - \cos(t(Y-X)) - (i)\sin(t(Y-X)) - \cos(t(X-Y)) \\
&\quad - i\sin(t(X-Y)) + \cos(t(Y-Y')) + i\sin(t(Y-Y'))] \\
&\text{sin}(X) = -\sin(-X), \cos(X) = \cos(-X), \sin(x-y) = \sin(x)\cos(y) - \cos(x)\sin(y) \\
&= \mathbb{E}[\cos(t(X-X')) - 2\cos(t(Y-X)) + \cos(t(Y-Y')) \\
&\quad + i\sin(t(X-X')) + i\sin(t(Y-Y'))] \\
&=
\end{aligned} \tag{1.12}$$

Applying Fubini and the previous lemma:

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{|\phi_{\mathbb{P}}(t) - \phi_{\mathbb{Q}}(t)|^2}{\|t\|_d^{d+1}} &= \int_{\mathbb{R}^d} \frac{\mathbb{E}[2(1 - \cos(t(Y-X))) - (1 - \cos(t(X-X')))]}{\|t\|_d^{d+1}} dt \\
&= 2\mathbb{E}\left[\int_{\mathbb{R}^d} \frac{1 - \cos(t(Y-X))}{\|t\|_d^{d+1}}\right] - \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{1 - \cos(t(X-X'))}{\|t\|_d^{d+1}}\right] - \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{1 - \cos(t(Y-Y'))}{\|t\|_d^{d+1}}\right] \\
&= 2\mathbb{E}[c_d\|Y-X\|] - \mathbb{E}[c_d\|X-X'\|] - \mathbb{E}[c_d\|Y-Y'\|] \\
&= c_d(2\mathbb{E}[\|Y-X\|] - \mathbb{E}[\|X-X'\|] - \mathbb{E}[\|Y-Y'\|]) \\
&= c_d\varepsilon(X, Y)
\end{aligned} \tag{1.13}$$

□

It is easy to see that the energy distance only vanishes when the distributions are equal, since it is equivalent to having equal characteristic functions.

1.3.2. Application to an independence test

In this subsection we will use the knowledge acquired above to develop a new independence test. This new test is called distance covariance (DCOV), its name comes from the fact that it is a generalization of the classical product-moment covariance.

We will start by defining the independence test. Given the random vectors $X \in \mathbb{R}^{dx}, Y \in \mathbb{R}^{dy}$, distributions \mathbb{P}_X and \mathbb{P}_Y respectively. Let $\phi_{\mathbb{P}_X}, \phi_{\mathbb{P}_Y}$ denote their characteristic functions and $\phi_{\mathbb{P}_{XY}}$ the characteristic function of the joint distribution. X and Y are independent if and only if $\phi_{\mathbb{P}_X}\phi_{\mathbb{P}_Y} = \phi_{\mathbb{P}_{XY}}$. The covariance energy test is based on measuring a distance between these functions.

First we need to generalize the energy distance expression for random vectors of different dimensions. As defined earlier this expression is obtained from a weighted L_2 -distance, imposing rotation invariance and scale equivariance, the energy distance is:

$$\varepsilon(X, Y) = \frac{1}{c_{d_x} c_{d_y}} \int_{\mathbb{R}^{d_x+d_y}} \frac{|\phi_{\mathbb{P}}(t) - \phi_{\mathbb{Q}}(t)|^2}{\|t\|_{d_x}^{d_x+1} \|s\|_{d_y}^{d_y+1}} dt ds$$

Where c_d is defined as before. The distance covariance is defined by replacing $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ in the previous formula with characteristic functions of the joint distribution and the product of the marginals respectively.

Definition 1.3.2. *The distance covatiance, DCOV, between random vectors X and Y , with $\mathbb{E}\|X\|_{d_x} + \mathbb{E}\|Y\|_{d_y} < \infty$, is the nonnegative number $\nu(X, Y)$ defined by:*

$$\nu(X, Y) = \|\phi_{\mathbb{P}_{X,Y}}(t, s) - \phi_{\mathbb{P}_X}(t)\phi_{\mathbb{P}_Y}(s)\|_w^2 = \frac{1}{c_{d_x} c_{d_y}} \int_{\mathbb{R}^{d_x+d_y}} \frac{|\phi_{\mathbb{P}_{X,Y}}(t, s) - \phi_{\mathbb{P}_X}(t)\phi_{\mathbb{P}_Y}(s)|^2}{\|t\|_{d_x}^{d_x+1} \|s\|_{d_y}^{d_y+1}} dt ds$$

Definition 1.3.3. *The distance correlation, DCOR, between random vectors X and Y , with $\mathbb{E}\|X\|_{d_x} + \mathbb{E}\|Y\|_{d_y} < \infty$, is the nonnegative number $\mathcal{R}(X, Y)$ defined by:*

$$\mathcal{R}(X, Y) = \begin{cases} \frac{\nu^2(X, Y)}{\sqrt{\nu^2(X)\nu^2(Y)}} & \text{if } \nu^2(X)\nu^2(Y) > 0 \\ 0 & \text{if } \nu^2(X)\nu^2(Y) = 0 \end{cases}$$

The distance covariance, like the energy distance, can be expressed using expectations.

Lemma 1.3.4.

BIBLIOGRAFÍA

- [1] Gretton A, Fukumizu K, Teo H.C., Song L, Schölkopf B, Smola J.A. (2007) *A Kernel Statistical Test of Independence*
- [2] Serfling R. (Wiley, New York, 1980) *Approximation Theorems of Mathematical Statistics*

