UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

18|19

Degree work

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

Campus Internacional excelencia UAM CSIC+

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Degree as

DEGREE WORK

Independence tests based on embeddings in functional spaces

Author: Roberto Alcover Couso
Advisor: Alberto Suárez González

junio 2019

Roberto Alcover Couso
Independence tests based on embeddings in functional spaces

Roberto Alcover Couso
C\ Francisco Tomás y Valiente Nº 11

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 clasicos 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 una 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 independecia, RDC, concluyendo en la comparación de estos tres test entre ellos y otros tests existentes.

Palabras clave

MMD,HSIC,RDC,DCOV,DCOR,dependencia estadistica, 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 approches 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 begining 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

Table of Contents

Lists

List of algorithms

List of codes

List of equations

List of figures

List of tables

1

INTRODUCTION

Since the 8th century, when Al-Khali(717-786) wrote the *Book of Cryptographic Messages* which contains the first use of permutations and combinations [13] humans have shown interest and studied the likelihood of events. In the eighteenth century with Jacob Bernoulli's *Ars Conjectandi* (posthumous, 1713) [14] a version of the fundamental law of large numbers was proven, which states that in a large number of trials, the average of the outcomes is likely to be very close to the expected value, probability became one of the main mathematical fields, introducing probability measures. Probability measures are widely used in hypothesis testing, density estimation, Markov chain and Monte carlo to give some examples. In this work our main focus will be hypothesis testing, mainly 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 $\mathbb{Q} = \mathbb{P}_{\mathcal{X}} \cdot \mathbb{P}_{\mathcal{Y}}$.

Measuring the existance of dependence between variables is a classical yet fundamental problem in statistics. Starting with Auguste Bravais and Francis Galton's correlation coefficient defined as a product-moment, and it's relation with linear regresion [Stigler, 1989], many techniques have been proposed, developed and studied. Nowadays this subject is of fundamental importance in scientific fields such as physics, chemistry, biology, and economics. A practical application 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 approches 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 begining 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.

1.0.1. Reproducing Kernel Hilbert Spaces (RKHS)

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 dificult 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: $X \to \mathbb{R}$ 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 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 $\forall f \in \mathcal{H}$:

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

Lemma 1.1.1. *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 $T_{\mathbb{P}} f \equiv \mathbb{E}(\sqrt{k(X,X)}) < \infty \forall f \in \mathcal{H}$

$$\begin{aligned} |T_{\mathbb{P}} f| &= |\mathbb{E}(f(X))| \\ &\leq \mathbb{E}(|f(X)|) \\ &\text{Reproducing property of the kernel} \\ &= \mathbb{E}|< f, k(\cdot, X) >_{\mathcal{H}}| \\ &\text{Chauchy Schwarz inequality} \\ &\leq \|f\|_{\mathcal{H}} \cdot \mathbb{E}(\sqrt{K(X,X)})^{1/2} \\ &\text{The expectation under } \mathbb{P} \text{ of the kernel is bounded} \\ &< \infty \end{aligned} \tag{1.1}$$

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

Definition 1.1.3. ***Mean embedding***

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

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

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:*

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

and:

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

proof

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

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

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

1.1.3. Prooving that MMD defines an homogeneity test

Definition 1.1.4. ***Characteristic kernel***

A reproducing kernel k is a characterisctic 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 $MMD(\mathcal{F},\mathbb{P},\mathbb{Q}) = 0 \leftrightarrow \mathbb{P} = \mathbb{Q}$

$\rightarrow$

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

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

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.

$\leftarrow$

By definition of MMD.

1.1.4. 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 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 $\upsilon((x,y),(x',y'))$ 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 1.1.5. ***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 1.1.5. *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:

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

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 $\mathbb{P}_{\mathcal{XY}}$ and $\mathbb{PQ}$. Most information for this chapter is taken from [1] [2] and [3]

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 $h \otimes g : \mathcal{G} \to \mathcal{H}$ is defined as:

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

Definition 1.2.2. ***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 1.2.3. ***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:

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

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

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

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:

$$\begin{aligned} \|C_{XY}\|^2_{\mathcal{HS}} &= < \bar{C}_{XY} - M_{XY}, \bar{C}_{X'Y'} - M_{X'Y'} >_{\mathcal{HS}} \\ &= < \bar{C}_{XY}, \bar{C}_{X'Y'} >_{\mathcal{HS}} + < M_{XY}, M_{X'Y'} > - 2 < \bar{C}_{XY}, M_{X'Y'} >_{\mathcal{HS}} \end{aligned} \tag{1.5}$$

Now calculating each of this products individually:

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

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

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

□

1.2.2. 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 1.2.5. ***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 1.2.2. *let* $\mathbb{E}_Z$ *denote the expectation taken over m independent copies*$(x_i, y_i$ *drawn from* $P_{\mathcal{XY}}$*. Then:*

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

.

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.

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

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

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

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

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

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

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

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

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

Normalisation by $\frac{1}{(m-1)^2}$ 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*

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

is degenerate, meaning $\mathbb{E}h = 0$. In this case, HSIC(Z) converges in distribution according to [3], 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 $(1-\alpha)$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 $m \to \infty$) 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.

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. Most of the content of this section is taken from [4] and [5]

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:*

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

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:*

$$\varepsilon(X, Y) = 2\mathbb{E}\|X - Y\|_d - \mathbb{E}\|X - X'\|_d - \mathbb{E}\|Y - Y'\|_d \tag{1.11}$$

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 $\mathbb{E}\|X\|_d + \mathbb{E}\|Y\|_d < \infty$ he energy distance between X and Y can be written as:*

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

where

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

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

To prove this proposition we need the following lemma.

Lemma 1.3.3. *$\forall x \in \mathbb{R}^d$ 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: $s = z\|x\|_d$

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{1-cos(tx)}{\|t\|_d^{d+1}}dt &= \int_{\mathbb{R}^d} \frac{1-cos(z\|x\|_d)}{\|z\|_d^{d+1}}dt \\ &= \int_{\mathbb{R}^d} \frac{1-cos(s)}{\frac{\|s\|_d}{\|x\|_d}^{d+1}\|x\|_d^d}dt \\ &= \|x\|_d \int_{\mathbb{R}^d} \frac{1-cos(s)}{\|s\|_d^{d+1}}ds \\ &= \|x\|_d \frac{\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})}\end{aligned} \tag{1.12}$$

□

Demostración. [**?**] Let $\overline{\phi_{\mathbb{P}}(t)}$ denote the complex conjugate of the characteristic function.

$$\begin{aligned}|\phi_{\mathbb{P}}(t)-\phi_{\mathbb{Q}}(t)|^2 &= = (\phi_{\mathbb{P}}(t)-\phi_{\mathbb{Q}}(t))\overline{(\phi_{\mathbb{P}}(t)-\phi_{\mathbb{Q}}(t))} \\ &= (\phi_{\mathbb{P}}(t)-\phi_{\mathbb{Q}}(t))(\overline{\phi_{\mathbb{P}}(t)}-\overline{\phi_{\mathbb{Q}}(t)}) \\ &= \phi_{\mathbb{P}}(t)\overline{\phi_{\mathbb{P}}(t)}-\phi_{\mathbb{P}}(t)\overline{\phi_{\mathbb{Q}}(t)}-\phi_{\mathbb{Q}}(t)\overline{\phi_{\mathbb{P}}(t)}+\phi_{\mathbb{Q}}(t)\overline{\phi_{\mathbb{Q}}(t)} \\ &= \mathbb{E}[e^{itX}e^{-itX'}]-\mathbb{E}[e^{itX}e^{-itY}]-\mathbb{E}[e^{itY}e^{-itX}]+\mathbb{E}[e^{itY}e^{-itY'}] \\ &= \mathbb{E}[e^{it(X-X')}-e^{it(Y-X)}-e^{it(X-Y)}+e^{it(Y-Y')} \\ &= \mathbb{E}[cos(t(X-X'))+isin(t(X-X'))-cos(t(Y-X))-isin(t(Y-X))-cos(t(X-Y \\ &- isin(t(X-Y))+cos(t(Y-Y'))+isin(t(Y-Y')) \\ &\text{sin(X) = -sin(-X), cos(X) = cos(-X), sin (x- y)=sin(x)cos(y)- cos(x)sin(y)} \\ &= \mathbb{E}[cos(t(X-X'))-2cos(t(Y-X))+cos(t(Y-Y')) \\ &+ isin(t(X-X'))+isin(t(Y-Y')) \\ &= \mathbb{E}[2(1-cos(t(Y-X)))-(1-cos(t(X-X')))-(1-cos(t(y-Y')))]\end{aligned} \tag{1.13}$$

Applying Fubini and the previous lemma:

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{|\phi_{\mathbb{P}}(t)-\phi_{\mathbb{Q}}(t)|^2}{\|t\|_d^{d+1}} &= \int_{\mathbb{R}^d} \frac{\mathbb{E}[2(1-cos(t(Y-X)))-(1-cos(t(X-X')))]}{\|t\|_d^{d+1}}dt \\ &= 2\mathbb{E}[\int_{\mathbb{R}^d} \frac{1-cos(t(Y-X))}{\|t\|_d^{d+1}}]-\mathbb{E}[\int_{\mathbb{R}^d} \frac{1-cos(t(X-X'))}{\|t\|_d^{d+1}}--\mathbb{E}[\int_{\mathbb{R}^d} \frac{1-cos(t(Y-Y')}{\|t\|_d^{d+1}} \\ &= 2\mathbb{E}[c_d\|Y-X\|]-\mathbb{E}[c_d\|X-X'\|]-\mathbb{E}[c_d\|Y-Y'\|] \\ &= c_d(2\mathbb{E}[\|Y-X\|]-\mathbb{E}[\|X-X'\|]-\mathbb{E}[\|Y-Y'\|]) \\ &= c_d\varepsilon(X,Y)\end{aligned} \tag{1.14}$$

□

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 aquired above to develop a new independence test. This new test is called distance covariance (DCOV), it's 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 $mathcalL_2$-distance, imposing rotation invariance and scale equivariance, the energy distance is:

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

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* $\nu^2(X,Y)$ *defined by:*

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

Definition 1.3.3. *The distance correlation, DCOR, between random vectors X and Y, ith* $\mathbb{E}\|X\|_{d_x} + \mathbb{E}\|Y\|_{d_y} < \infty$*, 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. *Let (X,Y),(X',Y'),(X",Y") ∼* $\mathbb{P}_{XY}$ *be iid copies of (X,Y), it holds that:*

$$\begin{aligned} \nu^2(X,Y) = {} & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|X-X'\|_{d_x}\|Y-Y'\|_{d_y} + \mathbb{E}_X\mathbb{E}_{X'}\|X-X'\|_{d_x}\mathbb{E}_Y\mathbb{E}_{Y'}\|Y-Y'\|_{d_y} \\ & - 2\mathbb{E}_{XY}[\mathbb{E}_{X'}\|X-X'\|_{d_x}\mathbb{E}_{Y'}\|Y-Y'\|_{d_y}] \end{aligned} \tag{1.15}$$

This proof is similar to the one of [?]. therefore we will leave it for the interested readers.

1.3.3. Statistics

Now we will give some estimators for both energy distance and distance covariance. Since we are interested in testing independence we will focus on the DCOV estimator. We will start with an estimator of energy distance, which as it's explained above it's a homogeneity test. Given the definition of energy distance [**?**] now we will define it's statistic as: Given two independent random samples $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_m)$, the two sample energy statistic corresponding to $\varepsilon(X, Y)$ is:

$$\varepsilon_{n,m}(x, y) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_i - y_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|y_i - y_j\|$$

Finally, an estimator of the distance covariance can be obtained directly from [**?**].For a random sample $(x, y) = ((x_1, y_1), ..., (x_n, y_n))$ of iid random vectors generated from the joint distribution of $X \in \mathbb{R}^{dx}$ and $Y \in \mathbb{R}^{dy}$, we obtain:

$$\begin{aligned}\nu^2(X, Y) = \frac{1}{n^2}\sum_{i,j=1}^{n}\|x_i - x_j\|_{d_x}\|y_i - y_j\|_{d_y} + \frac{1}{n^2}\sum_{i,j=1}^{n}\|x_i - x_j\|_{d_x}\frac{1}{n^2}\sum_{i,j=1}^{n}\|y_i - y_j\|_{d_y} \\ - \frac{2}{n^3}\sum_{i=1}^{n}[\sum_{j=1}^{n}\|x_i - x_j\|_{d_x}\sum_{j=1}^{n}\|y_i - y_j\|_{d_y}]\end{aligned} \tag{1.16}$$

As we can see this estimate cost is $O(n^2)$, that's the reason we won't calculate the distance covariance this way, our new approach will go as follows: First we compute the Euclidean distance matrix of each sample, computing all the pairwise distances between sample observations:

$$(a_{ij}) = (\|x_i - x_j\|_{dx}), (b_{ij}) = (\|y_i - y_j\|_{dy}).$$

an easy way to compute this matrix is:

$$A_{ij} = a_{ij} + \overline{a_{i\cdot}} - \overline{a_{\cdot j}} + \overline{a}, for i, j = 1, ..., n$$

where:

$$\overline{a_{i\cdot}} = \frac{1}{n}\sum_{k=1}^{n}a_{ik}, \overline{a_{\cdot j}} = \frac{1}{n}\sum_{k=1}^{n}a_{kj}, \overline{a} = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}a_{lk}$$

equivalently for B. In terms of these matrix, the distance covariance $\nu^2(x, y)$ is:

$$\nu^2(x, y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}B_{ij}$$

Finally the distance correlation is:

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

where:

$$\nu_n^2(x) = \nu_n^2(x,x) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}$$
$$\nu_n^2(y) = \nu_n^2(y,y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}$$

Now we will proove that this statistics converge almost surely when the random vectors have finite first moments.

Theorem 1.3.5. *if* $\mathbb{E}\|X\| + \mathbb{E}\|Y\| < \infty$ *then*

$$\lim_{n\to\infty} \nu_n^2(x,y) \xrightarrow{a.s} \nu^2(X,Y)$$

In order to proove this theorem we will give an alternative definition of the empirical DCOV statistic in order to make an elegant demonstration.

Definition 1.3.4. *Given all the introduction of this section it'd have been natural, but less elementary, to define* $\nu_n(x,y)$ *as* $\|f_{XY}^n(t,s) - f_X^n(t)f_Y^n(s)\|$ *where:*

$$f_{XY}^n(t,s) = \frac{1}{n}\sum_{k=1}^{n} exp[i < t, x_k > + i < s, y_k >]$$

is the empirical characteristic function of the sample $((x_1,y_1),...,(x_n,y_n))$ *and*

$$f_X^n(t) = \frac{1}{n}\sum_{k=1}^{n} exp[i < t, x_k >]$$

$$f_Y^n(s) = \frac{1}{n}\sum_{k=1}^{n} exp[i < s, y_k >]$$

are the marginal empirical characteristic functions of the X sample and Y sample, respectively.

The next theorem shows that the two definitions are equivalent.

Theorem 1.3.6. *If (X,Y) is a sample from the joint distribution of (X,Y), then*

$$\nu_n^2(X,Y) = \|f_{XY}^n(t,s) - f_X^n(t)f_Y^n(s)\|^2$$

Demostración. Lemma [**?**] implies that there exist constants c_p and c_q such that for all $X \in \mathbb{R}^p, y \in \mathbb{R}^q$.

$$\int_{\mathbb{R}^p} \frac{1 - exp[i < t, X >]}{\|t\|_p^{1+p}} dt = c_p \|X\|_p$$

$$\int_{\mathbb{R}^q} \frac{1-exp[i<s,Y>]}{\|s\|_p^{1+p}} dt = c_q \|Y\|_q$$
$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^q} \frac{1-exp[i<t,X>+i<s,Y>]}{\|t\|_p^{1+p}\|s\|_p^{1+p}} dt = c_q c_p \|X\|_p \|Y\|_q$$

where the integrals are understood in the principal value sense. For simplicity, consider the case p=q=1. The distance between the empirical characteristic functions in the weighted norm involves $\|f^n_{XY}(t,s)\|^2$, $\|f^n_X(t)f^n_Y(s)\|^2$ and $\overline{f^n_{XY}(t,s)}f^n_X(t)f^n_Y(s)$. Now we will give the result of evaluating this, due to the similarity to previous demostrations.

$\|f^n_{XY}(t,s)\|^2 = \frac{1}{n^2}\sum_{k,l=1}^n cos(X_k - X_l)tcos(Y_k - Y_l)s + V_1$

where V_1 represents terms that vanish when the integral $\|f^n_{XY}(t,s) - f^n_X(t)f^n_Y(s)\|^2$ is evaluated.

$\|f^n_X(t)f^n_Y(s)\|^2 = \frac{1}{n^2}\sum_{k,l=1}^n cos(X_k - X_l)t + \frac{1}{n^2}\sum_{k,l=1}^n cos(Y_k - Y_l)s + V_2$

$\overline{f^n_{XY}(t,s)}f^n_X(t)f^n_Y(s) = \frac{1}{n^3}\sum_{k,l,m=1}^n cos(X_k - X_l)tcos(Y_k - Y_l)s + V_3$

where V_2 and V_3 represent terms that vanish when the integral is evaluated. To evaluate the integral $\|f^n_{XY}(t,s) - f^n_X(t)f^n_Y(s)\|^2$, apply [**?**] and use:

$$cos(u)cos(v) = 1 - (1 - cos(u)) - (1 - cos(v)) + (1 - cos(u))(1 - cos(v))$$

After cancellation in the numerator of the integrand it remains to evaluate integrals of the type:

$$\begin{aligned}\int_{\mathbb{R}^2}(1-cos(X_k - X_l)t)(1-cos(Y_k - Y_l)s)\frac{dt}{t^2}\frac{ds}{s^2} &= \int_{\mathbb{R}}(1-cos(X_k - X_l)t)\frac{dt}{t^2}\int_{\mathbb{R}}(1-cos(Y_k - Y_l)s)\frac{ds}{s^2} \\ &= c_1^2\|X_i - X_j\|\|Y_i - Y_j\|\end{aligned} \tag{1.17}$$

where the first equality comes from applying Fubini.

For random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the same steps are applied.Thus

$$\|f^n_{XY}(t,s) - f^n_X(t)f^n_Y(s)\|^2 = S_1 + S_2 - 2S_3$$

Where:

$$S_1 = \frac{1}{n^2}\sum_{i,j=1}^n \|x_i - x_j\|_p\|y_i - y_j\|_q$$
$$S_2 = \frac{1}{n^2}\sum_{i,j=1}^n \|x_i - x_j\|_p \frac{1}{n^2}\sum_{i,j=1}^n \|x_i - x_j\|_p\|y_i - y_j\|_q$$
$$S_1 = \frac{1}{n^3}\sum_{i=1}^n\sum_{j,k=1}^n \|x_i - x_j\|_p\|y_i - y_k\|_q$$

□

Now that we have proven the equality we will proove the theorem 1.3.5

Demostración. Define

$$\zeta_n(t,s) = \frac{1}{n}\sum_{k=1}^{n} e^{i<t,X_k>+i<s,Y_k>} - \frac{1}{n}\sum_{k=1}^{n} e^{i<t,X_k>}\frac{1}{n}\sum_{k=1}^{n} e^{i<s,Y_k>}$$

so that $\nu_n^2 = \|\zeta_n(t,s)\|^2$. Then after elementary transformations: $u_k = exp(i < t, X_k >) - f_X(t)$ and $v_k = exp(i < s, Y_k >) - f_Y(s)$.

For each $\theta > 0$ define the region:

$$D(\theta) = \{(t,s) : \theta \leq \|t\|_p \leq \frac{1}{\theta}, \theta \leq \|s\|_q \leq \frac{1}{\theta}\}$$

and random variables

$$\nu_{n,\theta}^2 = \int_{D(\theta)} \|\zeta_n(t,s)\|^2 dw$$

For any fixed $\theta > 0$, the weight function w(t,s) is bounded on $D(\theta)$. Hence $\nu_{n,\theta}^2$ is a combination of V-statistics of bounded random variables, therefore by the strong law of large numbers it follows almost surely.

$$\lim_{n\to\infty} \nu_{n,\theta}^2 = \nu_{\cdot,\theta}^2 = \|f_{XY}(t,s) - f_X(t)f_Y(s)\|^2 dw$$

Clearly $\nu_{\cdot,\theta}^2$ converges to ν^2 as θ tends to zero. Now it remains to prove that almost surely

$$\lim_{\theta\to 0} \sup \lim_{n\to\infty} \sup \|\nu_{n,\theta}^2 - \nu_n^2\| = 0$$

For each $\theta > 0$

$$\begin{aligned}\|\nu_{n,\theta}^2 - \nu_n^2\| \leq & \int_{\|t\|_p \leq \theta} \|\zeta(t,s)\|^2 dw + \int_{\|t\|_p > \frac{1}{\theta}} \|\zeta(t,s)\|^2 dw \\ & + \int_{\|s\|_q \leq \theta} \|\zeta(t,s)\|^2 dw + \int_{\|s\|_q > \frac{1}{\theta}} \|\zeta(t,s)\|^2 dw\end{aligned} \tag{1.18}$$

For z = $(z_1, .., z_p)$ in $\mathbb{R}^p$ define the function

$$G(y) = \int_{\|z\|<y} \frac{1 - cos(z_1)}{\|z\|^{1+p}}$$

Clearly G(y) is bounded by c_p and $\lim_{y\to 0} G(y) = 0$. Applying the inequality $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and the following inequality.

Proposition 1.3.7. *The Cauchy–Schwarz inequality states that for all vectors u and v of an inner*

product space it is true that

$$|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle$$

where $\langle \cdot, \cdot \rangle$ is the inner product. By taking the square root of both sides, and referring to the norms of the vectors, the inequality is written as [7] [8]

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|$$

If $u_1, \ldots, u_n \in \mathbb{C}$ and $v_1, \ldots, v_n \in \mathbb{C}$, and the inner product is the standard complex inner product, then the inequality may be restated more explicitly as follows

$$|u_1 \bar{v}_1 + \cdots + u_n \bar{v}_n|^2 \leq (|u_1|^2 + \cdots + |u_n|^2)(|v_1|^2 + \cdots + |v_n|^2)$$

one can obtain that:

$$\begin{aligned} \|\zeta_n(t,s)\|^2 &\leq 2\|\frac{1}{n}\sum_{k=1}^{n} u_k v_k\|^2 + 2\|\frac{1}{n}\sum_{k=1}^{n} u_k \frac{1}{n}\sum_{k=1}^{n} v_k\|^2 \\ &\leq \frac{4}{n}\sum_{k=1}^{n}\|u_k\|^2 \frac{1}{n}\sum_{k=1}^{n}\|v_k\|^2 \end{aligned} \tag{1.19}$$

Therefore the first summand in [**?**] satisfies

$$\int_{\|t\|_p \leq \theta} \|\zeta(t,s)\|^2 dw \leq \frac{4}{n}\sum_{k=1}^{n} \int_{\|t\|_p \leq \theta} \frac{\|u_k\|^2 dt}{c_p \|t\|_p^{1+p}} \frac{1}{n}\sum_{k=1}^{n} \int_{\mathbb{R}^q} \frac{\|v_k\|^2 ds}{c_q \|s\|_q^{1+q}}$$

Here $\|v_k\|^2 = 1 + \|f_Y(s)\|^2 - exp(i < s, Y_k >)\overline{f_Y(s)} - exp(-i < s, Y_k >) f_Y(s)$, thus

$$\int_{\mathbb{R}^q} \frac{\|v_k\|^2 ds}{c_q \|s\|_q^{1+q}} = (2E_Y \|Y_k - Y\| - E\|Y - Y'\|) \leq 2(\|Y_k\| + E\|Y\|)$$

where the expectation E_Y is taken with respect to Y, and Y'$=^D Y$ is independent of Y_k. Further, after a suitable change of variables

$$\int_{\|t\|_p \leq \theta} \frac{\|u_k\|^2 dt}{c_p \|t\|_p^{1+p}} = \tag{1.20}$$

□

2

Design

In the previous chapter we've explained the tests that we will be developening. In this section we will explain the software design process, starting with analysis of the tests in order to provide a general overview of the problem for a better understanding of the choices taken.

2.0.1. Analysis

Our goal is to create an scalable and easy to modify software for different types of datasets, experiments and types of datasets.

On the grounds that we need to measure time differences between methods all software must be developed in the same language assuring that the differences in performance come from the actual algorithim and not the language difference.

Given the volume of data that we will be working with efficiency is of key importance, therefore parallelization will be heavily used to ensure a fast process of data.

As the experiments may be reused in the future and new independence tests may be added, the software has to be scalable and with a modular approach in order to handle the posible growth of the project.

2.0.2. Design

All our software is built around two classes: IndependenceTest and IndependenceTestTester. Both being abstract classes which held the code for the independence tests and the experiments respectively 2.1 shows the class diagram of our project.

IndependenceTest

This abstract class helds the main core which all independence tests will inherit.

All tests will control their own data and the progress of the experiments within themselfs, allowing

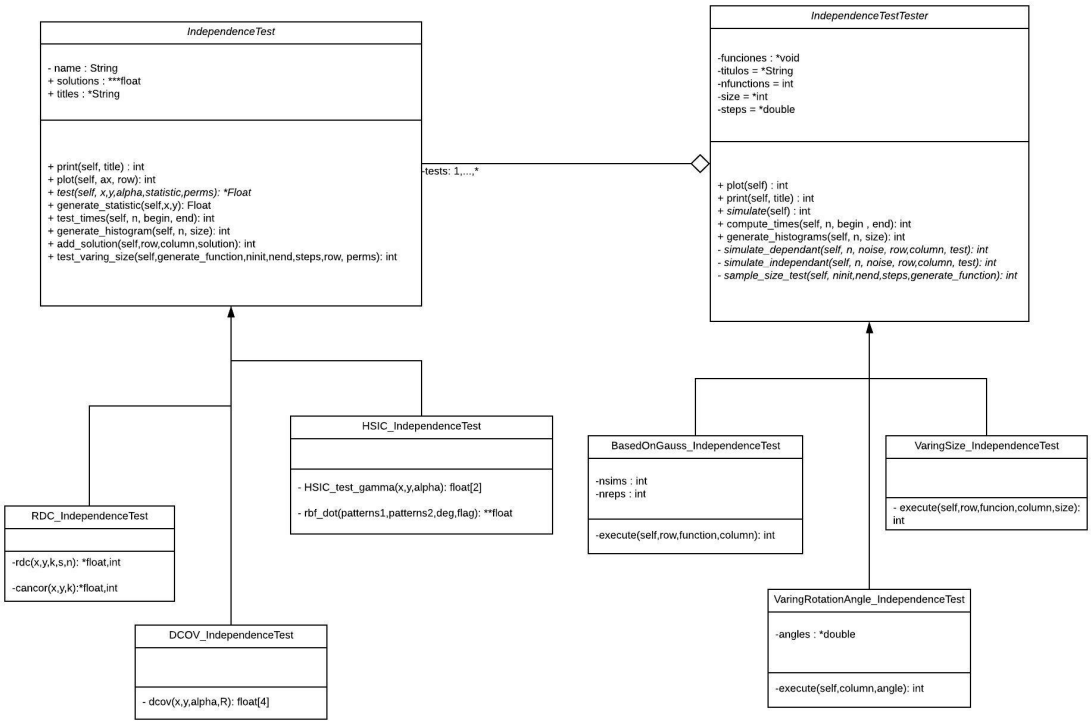

Figure 2.1: Class diagram

an easy paralelization. All tests will include a name, the titles for each subplot that will be made in the main plot, and the development of the experiment in a matrix called solutions. Furthermore this abstract class contains the functionality of plotting the results of the experiment for a given test, computing the time cost of an experiment and generating an empirical histogram of the statistic.

The specific implementation of the test will be held in an abstract function called test which will be implemented in each child object.

IndependenceTestTester

This abstract class implements the general functionalities of all the performed experiments, in order to allow for any amount of tests and the future addition of new tests, this class recives a list of IndependenceTests and will perform the desired experiments by calling the functions defined by IndependenceTest.

The parameters which will be general are: the functions used in order to generate the datasets and the subtitles for the main plot.

Finally in order to ensure the minimun amount of repeated code the functionality of measuring the power of a test given X and Y is implemented in a function called simulate, letting the task of modifying the datasets as needed to each child object. Figure 2.2 shows a sequence diagram of an experiment.

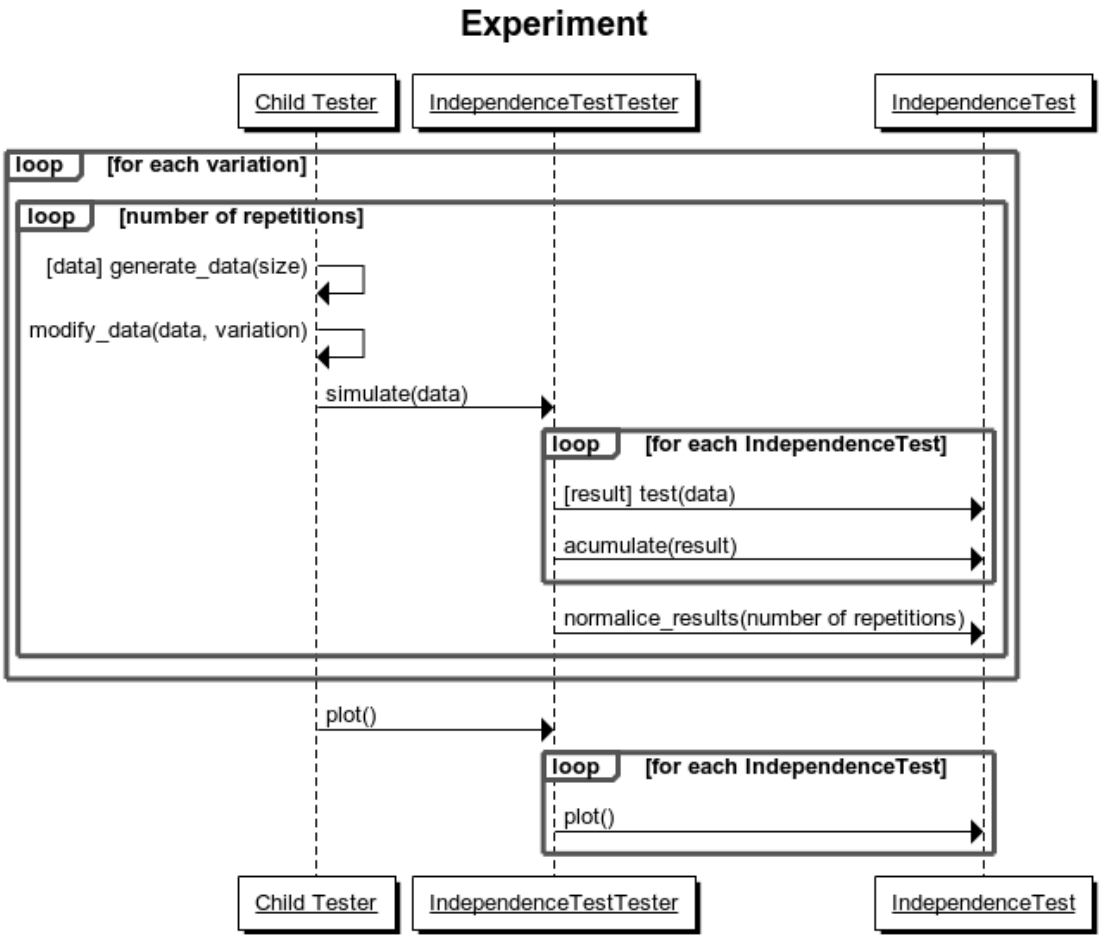

Figure 2.2: Sequence diagram

In the following section we will dive into the process of how this software was implemented, the main ploblems we found along the way and how we solved them.

3 Development

In the previous section we presented the design aspects of our project, in this section we will introduce our development experience and the reasons behind each decision. First of all we will start by how we achieved the main goals of our software. We decided to implement all the functionality in python, although all plots shown comparing distributions were plotted in R, this is because of the simplicity which R provides to perform plots comparing distributions. This will be explained in detail in 3.2.2

3.1. General aspects of implementation

3.1.1. Efficiency

As our project was implemented in python, we had access to numpy and scipy, which are libraries implemented in low levels languages, like C, Fortran and Cython, making them really efficient. Therefore, we used them whenever it was possible.

In addition to the code, parallelization was key in our project in order to obtain results in a reasonable time span. All parallelization was created by threads, given the amount of parallel lines of code we were managing, and how little work each had to made, creating process for each line wouldn't have been optimal due to the time it takes to create a new process. Most parallel process we generated a thread pool with the library concurrent.futures which provides a high level interface for asynchronously executing callables, which makes most of our simulations something as simple as shown in Code 3.1.1, which presents how to create an histogram of the statistic of all independence tests included in our independence test tester. The parameter of max workers is fixed to the amount of independence tests we have, this is because we already know how many threads we will need.

3.1.2. Modularity and Scalabillity

In order to ensure that our system was as modular as possible, we followed the design presented in Figure 2.1. In order to create abstract classes in python we used the modules abstractmethod and

Code 3.1: Code sample of how to create a thread pool with concurrent.futures

```
1 with concurrent.futures.ThreadPoolExecutor(max_workers= len(self.tests)) as executor:
2             futures = {executor.submit(test.generate_histogram,sample_size) for test in self.tests}
3             concurrent.futures.wait(futures)
```

ABCMeta from the abc package. Allowing for all main code of the test to be storaged in the child class while making all the experiments being transparent to the implementation beneath. In Code 3.1.1 is shown how for any independence test the calling mantains the same.

3.2. Specific details about each independence test implementation

In order to compute efficiently all statistics we will make use of matricial calculus which will help reducing the amount of operations needed, which will help with the overall performance. Code 3.23.2 shows how we calculate the hyperparameter for the Gaussian kernel in HSIC as an example of what we mean in the previous sentence.

Code 3.2: Sample of how to calculate the median of the distances for a sample $x \in \mathbb{R}^n$

```
1 size = len(x)
2 G = np.sum(x*x,1) #Here we calculate the square of each sample
3 Q = np.repeat(G,size).reshape(size,size) #row i contain each the square of sample i n times
4 R = Q.T #colum i contain each the square of sample i n times
5 dists = Q + R -2*np.dot(xmed,xmed.T) #we calculate (x -y)^2 = x^2+y^2-2x*y
6 dists = dists -np.tril(dists) #we remove repeated distances (x-y)^2 = (y-x)^2
7 dists = dists.reshape(size*size,1)
8 hyperparameter = np.sqrt(0.5*np.median(dists)) #Calculate the hyperparameter of our kernel
```

3.2.1. RDC

As explained in section **??** the parameter k will improve the performance of the test the largest it is, but due to numerical issues, if k is too large,then rank($\Phi(X)$) <k or rank($\Phi(Y)$) <k, so we need to find the largest k such that the eigenvalues, solutions of the canonical correlation analysis, are real-valued. As this is a problem dependant of the data, we will preform a binary search for the largest k which meets the condition. As the complexity of RDC is $O(k^2 n)$ adding a binary search to the algorithm wont affect its overall complexity as the complexity of the binary search is $O(log(k))$, which is irrelevant.

In order to compute RDC, we needed to calculate the canonical correlation analysis, which is not

implemented in python. Code 3.33.2.1 shows how we calculate the canonical correlation analysis in python.

Code 3.3: Canonical correlation analysis in python

```
def cancor(x,y,k):
    canonical_correlation_matrix = np.cov(np.hstack([x, y]).T)

    k0 = k
    lower_bound = 1 #minimum k
    upper_bound = k #maximum k
    while True:
        #Canonical correlation
        k = int(k)

        C_XX = canonical_correlation_matrix[:k,:k]
        C_YY = canonical_correlation_matrix[k0:k0+k, k0:k0+k]
        C_XY = canonical_correlation_matrix[:k, k0:k0+k]
        C_YX = canonical_correlation_matrix[k0:k0+k, :k]

        eigs = np.linalg.eigvals(np.dot(np.dot(np.linalg.inv(C_XX), C_XY),
                                        np.dot(np.linalg.inv(C_YY), C_YX)))

        #Search if K is too large
        if not (np.all(np.isreal(eigs)) and
                0 <= np.min(eigs) and
                np.max(eigs) <= 1): #Condition of being too large
            upper_bound -= 1 #reduce the maximum in 1
            k = (upper_bound + lower_bound) / 2 #search in the middle
            continue

        if lower_bound == upper_bound: break #if lower_bound == upper_bound means we found the
            optimal value for k

        lower_bound = k #as k meets the condition we set the lower bound to k

        #Set k as the middle point
        if upper_bound == lower_bound + 1:
            k = upper_bound

        else:
            k = (upper_bound + lower_bound) / 2

    return np.sqrt(eigs),k
```

3.2.2. HSIC

We've decided to implement HSIC following the mathlab implementation which makes usage of a Gaussian kernel :

$$K(x, y) = exp(-\frac{\|x-y\|^2}{\mu^2})$$

where μ is the median of the euclidian distances between samples. This kernel will be used because of the following:

As we have seen a positive definite kernel $k(x, y)$ defines an inner product $k(x, y) =< \phi(x), \phi(y) >_{\mathcal{H}}$ for feature vector ϕ constructed from the input x, and $\mathcal{H}$ is a Hilbert space. The notation $< \phi(x), \phi(y) >$ means the inner product between $\phi(x)$ and $\phi(y)$. For a better understanding, you can imagine $\mathcal{H}$ to be the usual Euclidean space, but with an infinite number of dimensions. Then take a vector which is infinitely long, like $\phi(x) = (\phi_1(x), \phi_2(x), ...)$. In kernel methods, $\mathcal{H}$ is a RKHS (explained in the introduction 1). Since we only care about the iner product of the features, we will directly evaluate the kernel k. To explain smoothness of the functions given by the Gaussian Kernel, let us consider Fourier features. As it's easy to prove, $k(x, y) = k(x - y)$, the kernel only depends on the difference of the two arguments. Let $\hat{k}$ denote the Fourier transform of k.

In this Fourier viewpoint, the features of f are given by $f = (..., \frac{\hat{f}_l}{\sqrt{\hat{k}_l}}, ...)$, this is saying that the feature representation of your function f is given by its Fourier transform divided by the Fourier transform of the kernel k. The feature representation of x, which is $\phi(x)$ is:$(..., \sqrt{\hat{k}_l} \exp(-ilx), ...)$ where $i = \sqrt{-1}$. One can show that the reproducing property holds.

Now thanks to Plancherel theorem: [12]

It states that the integral of a function's squared modulus is equal to the integral of the squared modulus of its frequency spectrum. That is, if $f(x)$ is a function on the real line, and $\widehat{f}(\xi)$ is its frequency spectrum, then ;

$$\int_{-\infty}^{\infty} |f(x)|^2 \, dx = \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2 \, d\xi$$

Hence:

$$\|f\|_H^2 = \sum_{l=-\infty}^{\infty} frac\hat{f}_l^{\,2} \hat{k}_l$$

Which as $f \in \mathcal{L}^2$ the norm is finite, the sum converges. Now as the Fourier transform of a Gaussian kernel $K(x, y) = exp(-\frac{\|x-y\|^2}{\mu^2})$ is another Gaussian where $\hat{k}_l$ decreases exponentially fast with l. So if f is to be in this space, its Fourier transform must drop even faster than that of k. This means the function will have only a few low frequency components with high weights.(A function with only low frequency components is smooth).

3.2.3. Plots

As in this work we have been working intensively with probability distributions, in order to ease the task of testing hypothesis and showcasing the asymptotic behaviour of our statistics, we decided to

make use of R in our project. All data was collected from the experiments performed in python, where we stored the results of each statistic in each transformation and experiment. In R we performed K-S test to each statistic with different sample sizes and variations which will be explained in detail in chapter 4 and saw how good our null hypothesis was. In the appendix we showcase some results obtained with R, for example Figure 4.19 shows how good of a fit is a Gamma distribution to the HSIC distribution.

3.3. Version control, repositories and continuous integration

In order to maintain control of each change throughout the project we needed to use tools in order to manage and control the advance of the project.

As a version control system we used git, which provides simplicity and comes with the advantage that hosting services for version control using git like GitHub exist. We chose GitHub because it is free and comes with all functionality for public repositories, one key functionality is that provides a version history of your code, so that previous versions are not lost with every new merge, easing removing mistakes or going back to a previous version if necessary.

In addition of version control we used https://travis-ci.org/ in order to automatically test all changes made. https://github.com/travis-ci/travis-ci/blob/master/README.md is a hosted continuous integration service used to build and test software projects hosted at GitHub, furthermore this tool is free for open projects.

Travis-CI is configured by adding a file named *.travis.yml* to the root of the repository, specifying programming language used, the desired building and testing environment, the script to run the tests, when and what to do whenever a pettition is made to the repository. Code 3.33.4 shows an example of a *.travis.yml*.

To sum up this section Figure 3.1 presents graphically the development process of this project showcasing the different tools used and their relations.

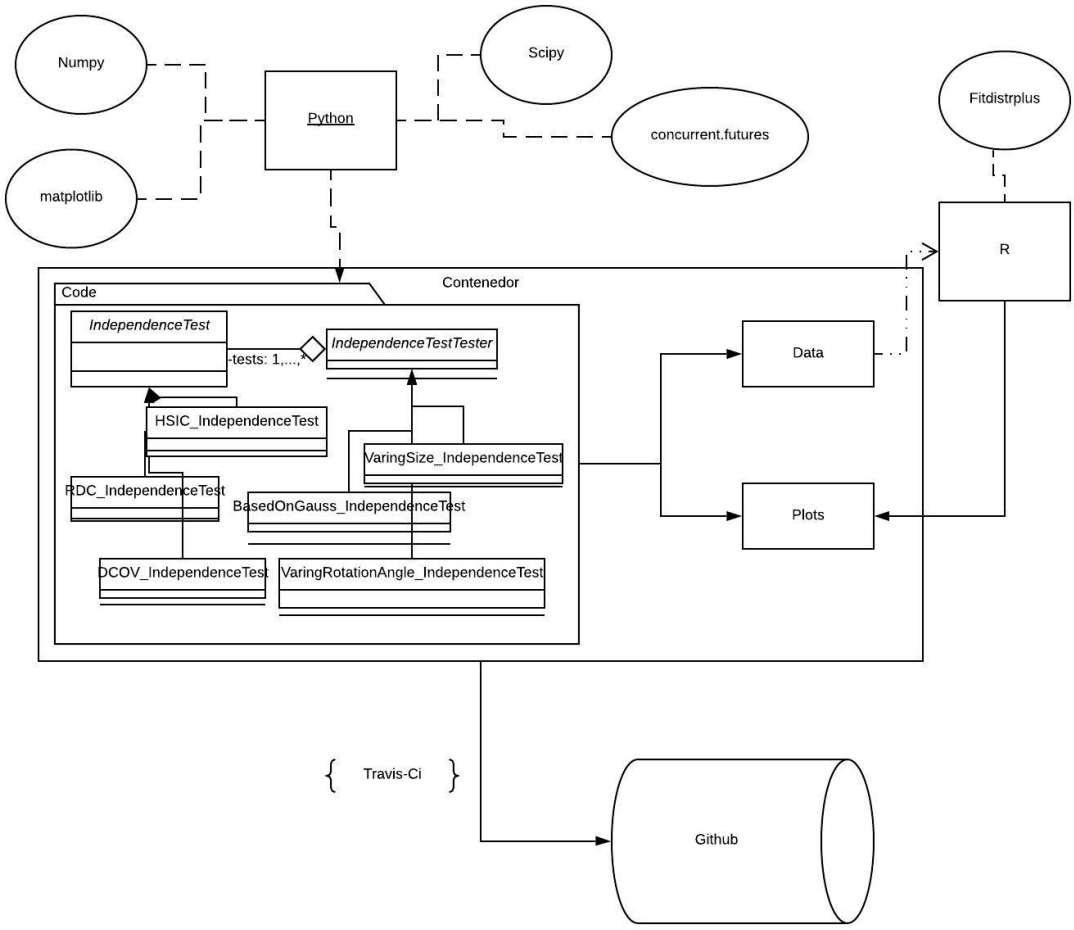

Figure 3.1: Diagram showcasing the different tools used to create this project and their relations.

Code 3.4: yml file used in order to incorporate Travis-CI in our repository.

```
language: python
python:
  -"2.7"
  -"3.4"
install:
  -pip install unittest2
script:
  python tests/HSIC_test.py
  python tests/RDC_test.py
  python tests/DCOV_test.py
  python tests/IndependenceTest_test.py
  python tests/IndependenceTestTester_test.py
notifications:
  email:
    recipients:
      -roberto.alcover@estudiante.uam.es
      -robertoalcovercouso@gmail.com
    on_success: never # default: change
    on_failure: always # default: always
```

4

Experiments

In this chapter we will present the results of various experiments in which we will compare the power of the explained tests between them and with other state-of-the-art independence tests, as well as comparing the power of these tests based on their asimptotic distribution and their empirical distribution. and [1].

In all our experiments, we set the number of random features for RDC to k = 3, and the random sampling width to s = 10^{-2}. All kernel methods make use of a Gaussian kernel with width hyper-parameter set to the median of the euclidean distances between samples of each of the input random variables.

4.1. Power

4.1.1. Real

First we will turn the issue of estimating the power of the RDC,HSIC and DCOV estimator. We define the power of a dependence measure as the percentage of times that it is able to discern between two samples with equal marginals, but one of them containing dependance.

In order to simulate the null hypothesis of our tests($\mathcal{H}_0$, the variables are independent) we will generate 500 samples under $\mathcal{H}_0$ to compute the threshold of the statistics with a signification level $\alpha = 0{,}05$. This will stand for our first group of experiments.

First we generated 500 pairs of 200 i.i.d. samples, in which the input variable was uniformly distributed on the unit interval, for each pair we generated each statistic, afterwards we calculated the 95 percentile, this will be the threshold for our test in this experiments.

To do so, we created three different experiments:

In the first one, adapted from [9], we studied 12 association patterns: linear, parabolic, quadratic, sin(4πx), sin(16πx), fourth root, circle, step, xsin(x),logarithm, Gaussian and a 2D multivariate normal distribution.Figure 4.1 shows grafically each association pattern.

Secondly for each of the 12 association patterns, we studied how Gaussian noise may affect the power of our test, with a noise increasing from 0 to 3 in 10 steps we generated 200 repetitions of 200 samples uniformly distributed on the unit interval and generated the pair with each association pattern, then we added Gaussian noise to the pair and normaliced both marginals. Figure 4.2 shows for each subplot the power obtained with each association pattern. The x axis represents how the noise increases, and the y axis the power of the tests.

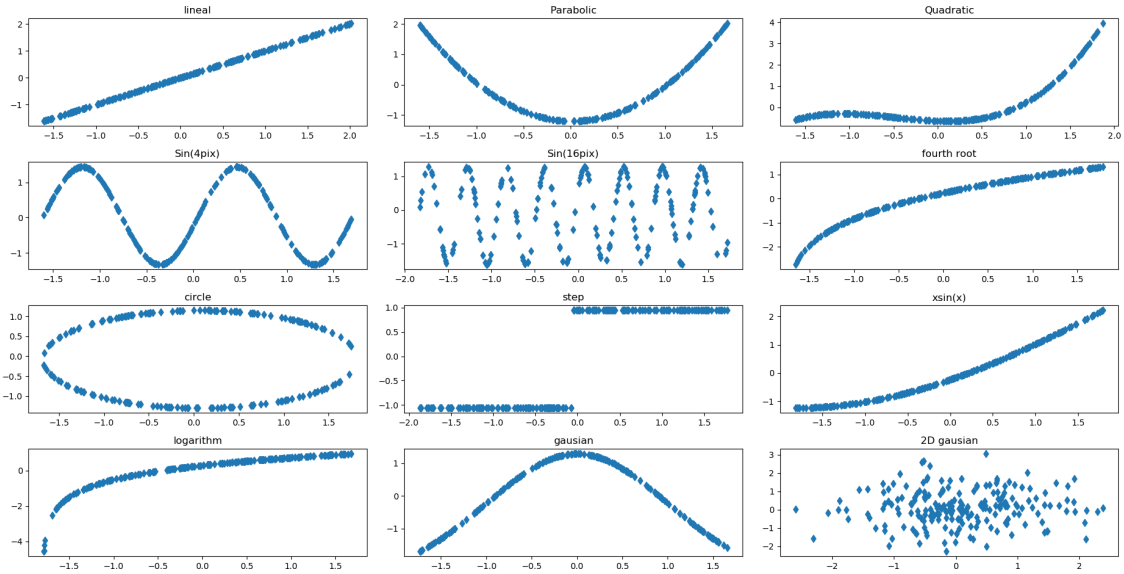

Figure 4.1: Representation of non linear dependance patterns

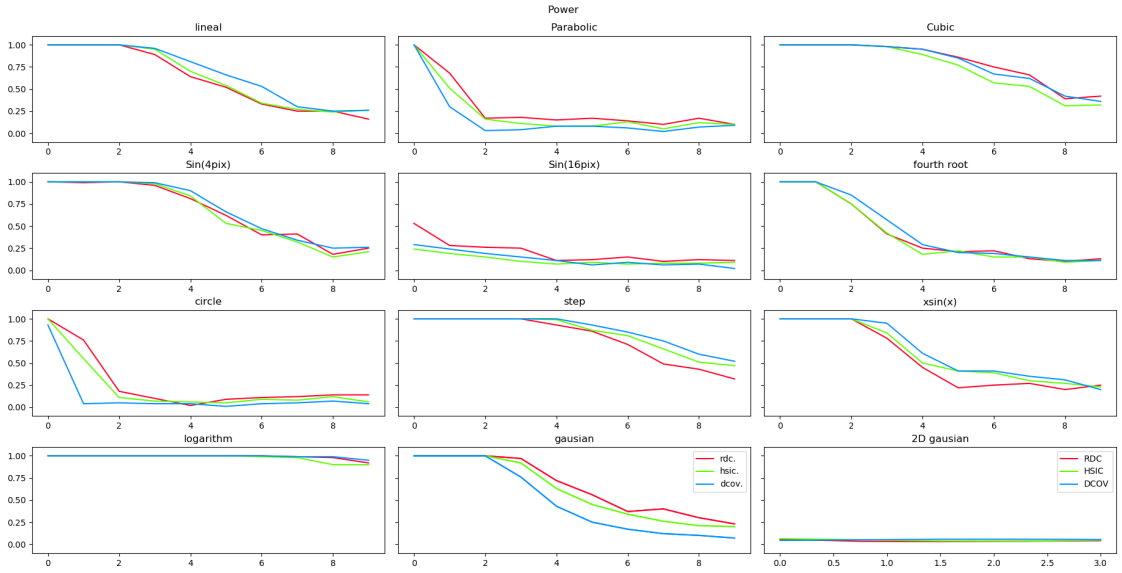

Figure 4.2: Power of tests adding Gaussian noise to marginals

In our second experiment we studied different sets of data and studied how the sample size affected the power of our tests. This test is taken from [6], the data sets are also taken from [6]. The first data set is a bivariate Gaussian with a correlation of 0.5, $(X, Y) \sim \mathcal{N}(0, \Sigma)$, where:

$$\Sigma = \begin{vmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{vmatrix}$$

For the second set we generated a uniform random variable $Z \sim U[0, 2]$. The marginals for this set will be constructed by:

$$X = ZX' \text{and} Y = ZY'$$

where $X', Y' \sim \mathcal{N}(0, 1)$, X' and Y' are independent, still X and Y are dependent due to both sharing the variable Z.

The variables X and Y in the third example are the marginals of a mixture of three bivariate Gaussians with correlations 0,0.8 and -0.8, with respective probabilities of 0.6, 0.2 and 0.2. The vector (X,Y) has density:

$0{,}6\mathcal{N}(0, \Sigma_1) + 0{,}2\mathcal{N}(0, \Sigma_2) + 0{,}2\mathcal{N}(0, \Sigma_3)$

Where

$$\Sigma_1 = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \Sigma_2 = \begin{vmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{vmatrix} \Sigma_2 = \begin{vmatrix} 1 & -0{,}8 \\ -0{,}8 & 1 \end{vmatrix}$$

The variables of the last example are generated as bivariate Gaussian random variable with correlation of 0.8 and then multiply each marginal with white Gaussian noise:

$$(X, Y) = (Z_1\epsilon_1, Z_2\epsilon_2) \text{where } Z \sim \mathcal{N}(0, \Sigma_2) and \epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \Sigma_1)$$

Below samples from this data sets are displayed in 4.3. The power is measured for sample sizes 10, 91, 173, 255, 336, 418 and 500. For this experiment and the next one, we also compared the performance of RDC,HSIC and DCOV with other state of the art independence measures, being :

1.– Energy distance to compute the non-Gaussianity of the projections, "Emean" and "Emax" denote taking the mean and the maximum of the differences respectively.

2.– MMD, where "MMDmean" and "MMDmax" denote the methods where MMD are used instead of negentropy

3.– the non-Gaussianity test when we are taking the mean of the differences of the negentropy over ρ, denoted by "gaussmean".

The results of this experiment is presented in Figure 4.4.

For this set of experiments in which we try to determine the power of the tests, we have performed a final experiment following [6] in which we studied the power of the tests and how they are affected by the rotation of the set. For this experiment we will use two independent random variables, X and Y, where X is a uniform random variable $(X \sim U[-\sqrt{3}, \sqrt{3}])$ whereas Y is a mixture of two uniform ran-

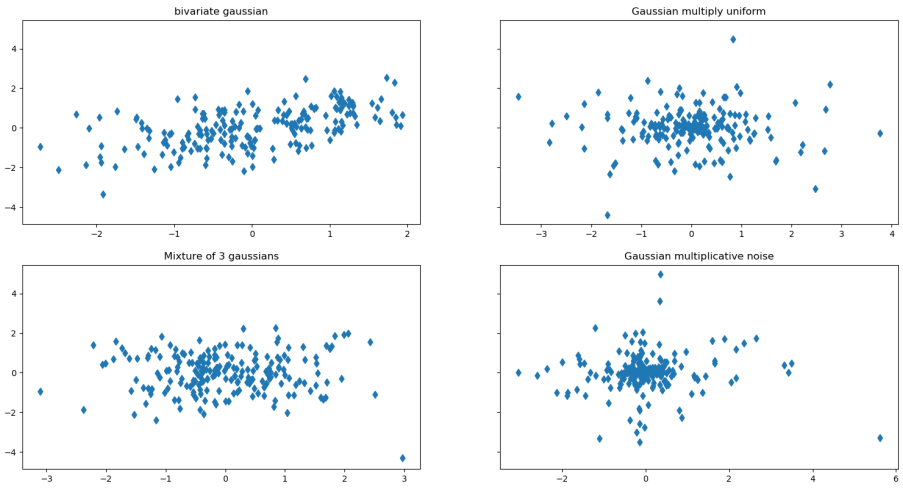

Figure 4.3: Samples from the data sets for the second experiment

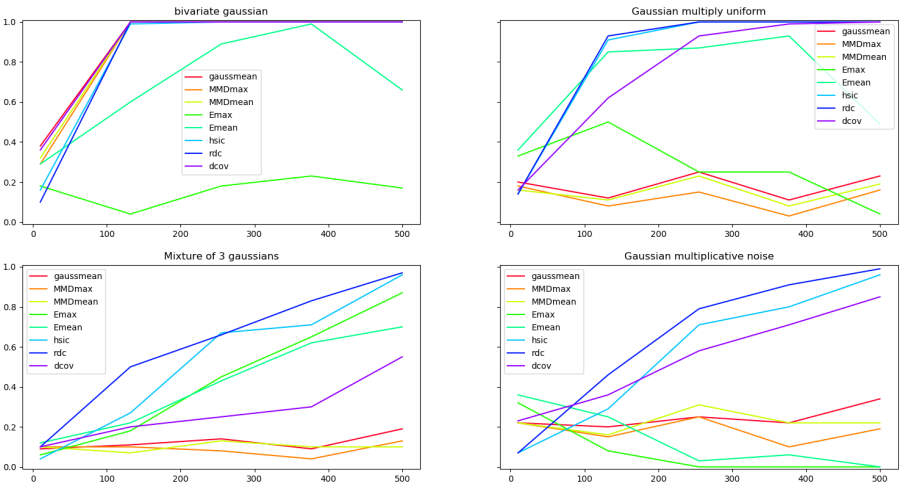

Figure 4.4: Power of tests adding Gaussian noise to marginals

dom variables, each having equal probability of ocurrence on disjoint supports. That is, Y has density: $0{,}5U[-1, 0{,}5] + 0{,}5U[0{,}5, 1]$.

We generate new pairs of random variables by rotating this random pair (X,Y). This will affect the dependence between them, this variables will be independent if and only if the angle of rotation is an integer multiple of pi, $n \cdot \pi : n \in \mathbb{Z}$. After this rotation we had scaled X,Y to have zero mean and unit variance. For this experiment we have generated 500 samples and tested the power for 100 rotation angles going from 0 to 2π, with sample size 200. In Figure 4.5 shows samples of the same data with different rotation angle. As we can see in Figure 4.6 the function power for all tests is a $\frac{\pi}{2}$ even periodic function, which confirms that the potency of our tests does not depend on the sign of the correlation.

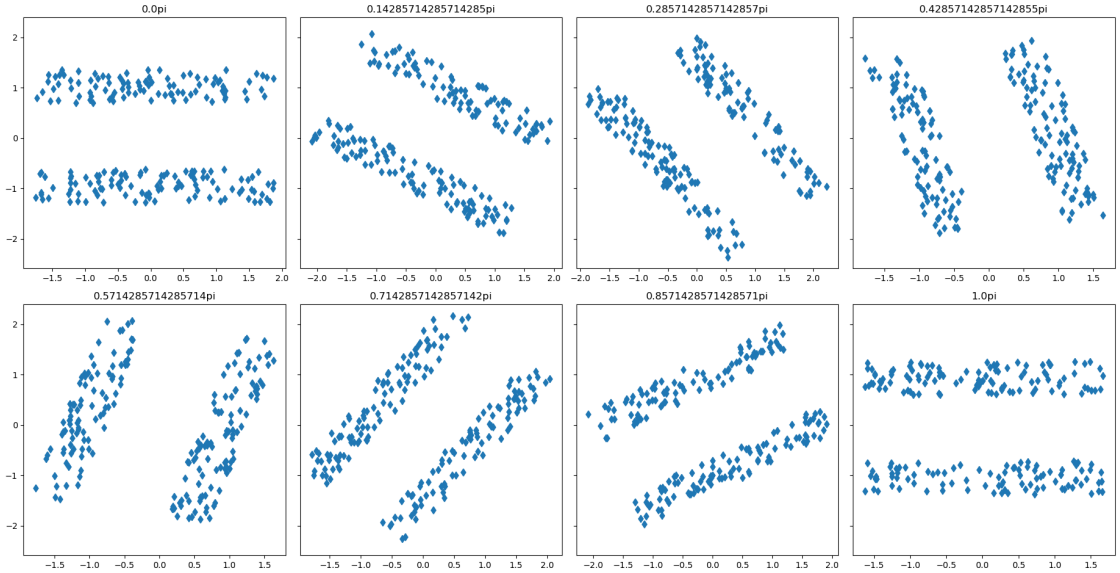

Figure 4.5: Samples from the data sets for the third experiment

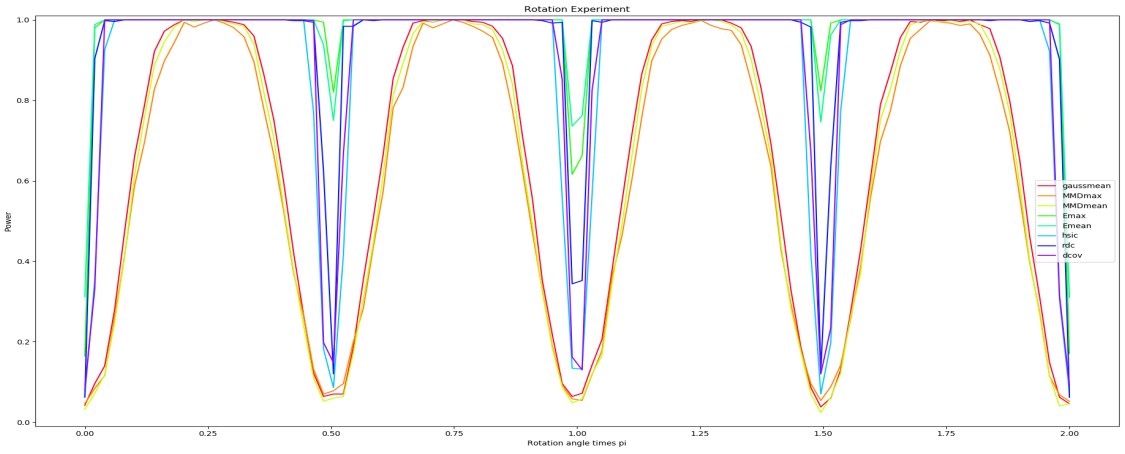

Figure 4.6: Power of the tests rotating the dataset

This concludes our first set of experiments, in the three experiments shown we can see that HSIC,DCOV and RDC are the sturdiest tests showing the best performance consistently. Among these threee tests

RDC has proven to be the most consistent test, outperforming almost everytime the other tests.

4.1.2. Asymptotic

Now for our second set of experiments we will study how the asymptotic version of the tests performs and how good the aproximations are.

For our first experiment, we will study empirically the convergence of our tests to the asymptotic distribution, or it's aproximation. For this purpose we will take bivariate gausians with correlation 0 with sample sizes 50, 100, 150 ,200, 500 and 1000, in order to decide how good or bad our aproximations are, we will perform a Kolmogorov-Smirnov homogeneity test.

First of all we will start with RDC:

For size 500, we've obtained a pvalue of 0.3564, therefore we accept the null hypothesis H_0 : $RDC \sim \mathcal{X}_9^2$ for significance levels of 0.1,0.05 and 0.01. Figure 4.25 shows the pdf, qq plot, pp plot and cdf of our statistic with the one of the $\mathcal{X}_9^2$distribution

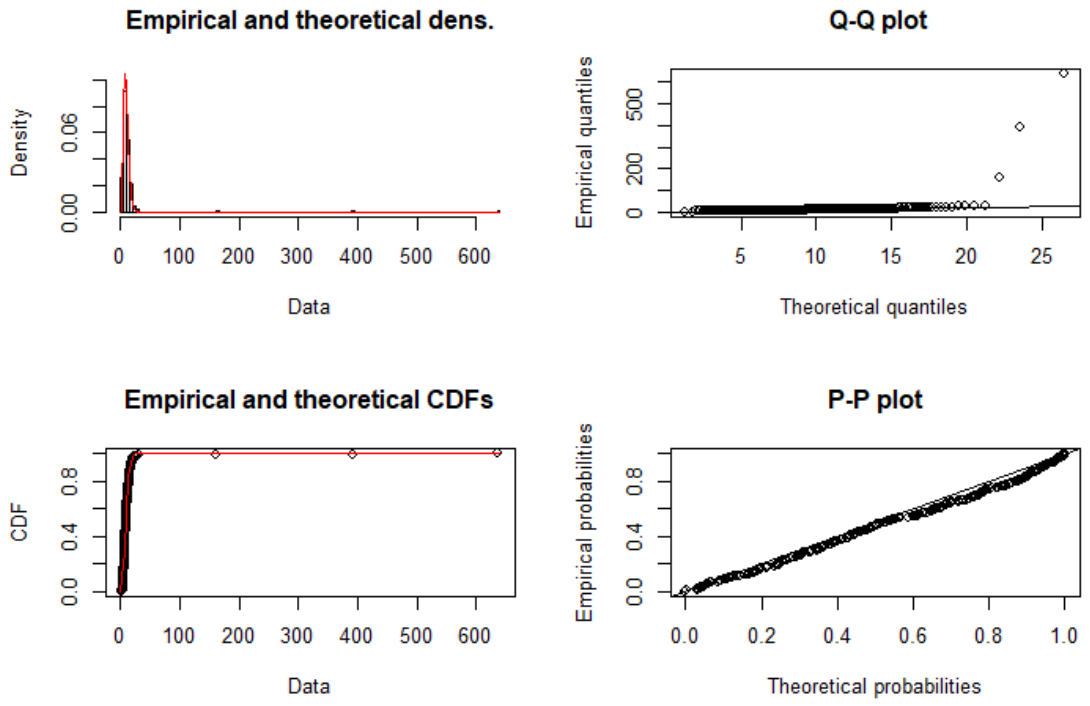

Figure 4.7: RDC statistic with a chi-squared distribution with 9 degrees of freedom

With HSIC distribution:

For size 500, we've obtained a pvalue of 0.1564, therefore we accept the null hypothesis H_0 : $RDC \sim \mathcal{X}_9^2$ for significance levels of 0.1,0.05 and 0.01. Figure 4.25 shows the pdf, qq plot, pp plot and cdf of our statistic with the one of the $\mathcal{X}_9^2$distribution

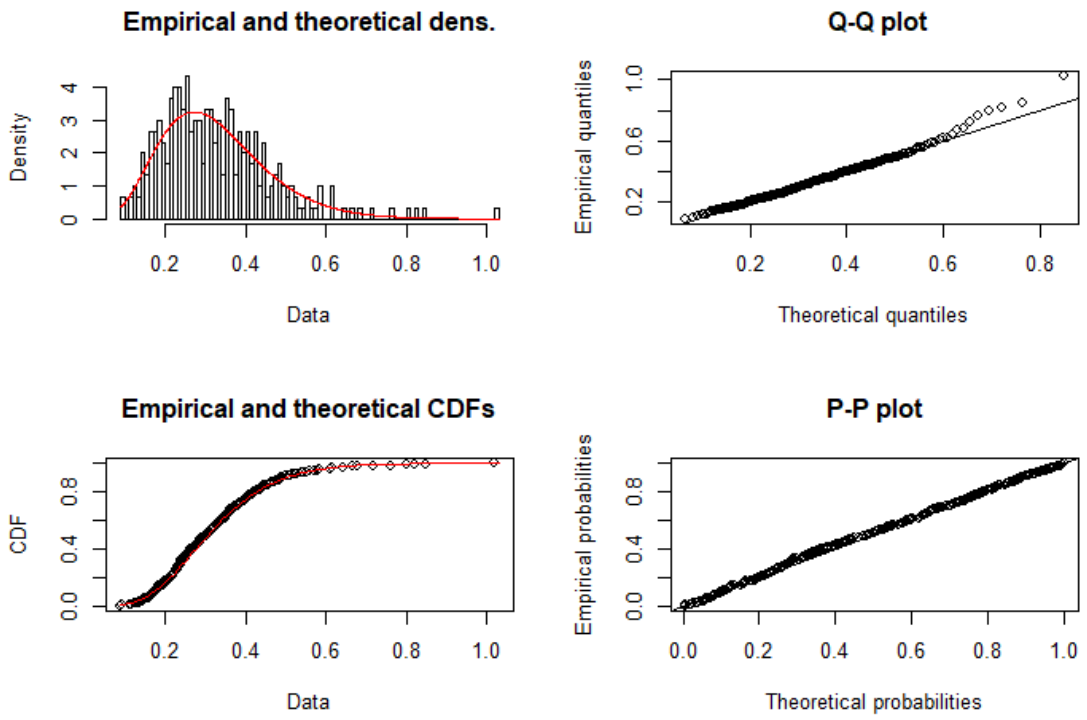

Figure 4.8: HSIC statistic with a gamma distribution

For an in depth analysis head to the Appendix **??**, we included the same experiment for different sizes and adding Gaussian noise.

Now that we have accepted our hypothesis we will analyze how good they are. We will compare the power of the asymptotic version with the *real* one on various scenarios.

In our first experiment we will analyze them with a bivariate Gaussian with sizes 50, 100, 150, 200, 500 and 1000, with different correlations 0, 0.25, 0.5, 0.75, 1.

$X, Y \sim \mathcal{N}(0, \Sigma_i)$

$$\Sigma_1 = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} \Sigma_2 = \begin{vmatrix} 1 & 0{,}25 \\ 0{,}25 & 1 \end{vmatrix} \Sigma_3 = \begin{vmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{vmatrix} \Sigma_4 = \begin{vmatrix} 1 & 0{,}75 \\ 0{,}75 & 1 \end{vmatrix}$$

$$\Sigma_4 = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix}$$

To the Y variable we will add Gaussian noise going from 0 to 3 $Y = Y + \mathcal{N}(0, noise)$.

Figures 4.9,4.10,4.11,4.12,4.13 and 4.14 show the power of DCOV, for sizes 50,100,150,200,500 and 1000 respectively.

Figures 4.15,4.16,4.17,4.18,4.19 and 4.20 show the power of HSIC, for sizes 50,100,150,200,500 and 1000 respectively.

Figures 4.21,4.22,4.23,4.24,4.25 and 4.26 show the power of RDC, for sizes 50,100,150,200,500 and 1000 respectively.

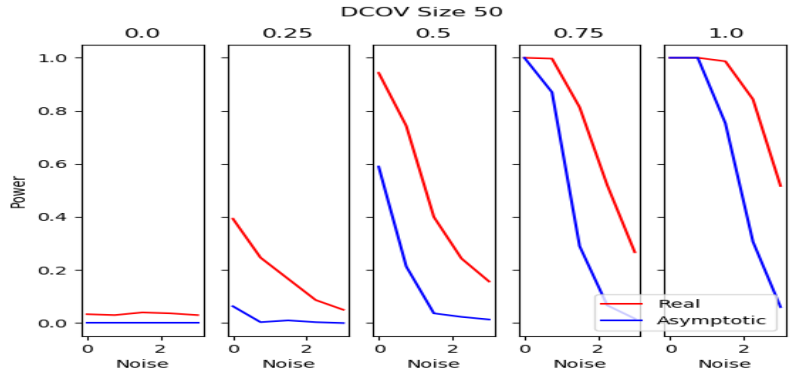

Figure 4.9: Power comparison between the asymptotic and the real version of DCOV for sample size 50

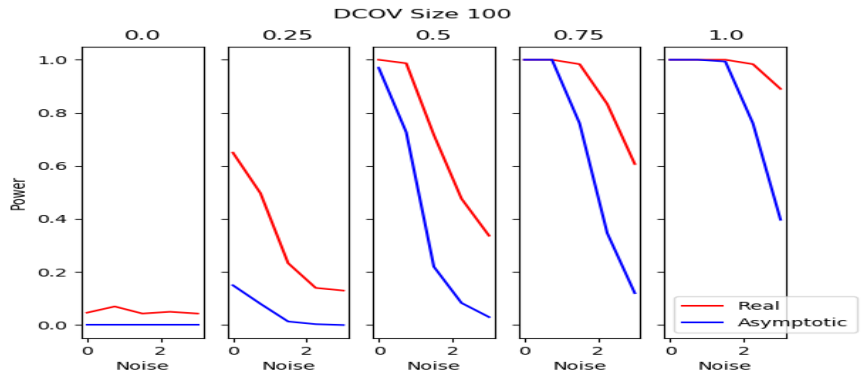

Figure 4.10: Power comparison between the asymptotic and the real version of DCOV for sample size 100

For all tests we've seen how our null hypothesis is always conservative, this is always useful for situations where computation time is critical and the asymptotic version may be better because it minimizes type 1 error.

Now we will study the differences one can see in the previous experiments if we perform the test with the asymptotic version instead of the *real* one.

Starting with the first experiment 4.1.1, we reproduced the same experiment, with a significance level of 0.05, sample size of 200 and Gaussian noise going from 0 to 3. Figures 4.27, 4.28, 4.29 show the asymptotic behaviour of DCOV, HSIC and RDC against the original one respectively for the relation patterns shown in Figure 4.1

In the second experiment where we studied how good our test was for different sizes and , in this experiment we will see how RDC will outperform the other two tests in their asymptotic behaviour, this may be explained by the fact that HSIC and DCOV asymptotics distributions used for the test were good

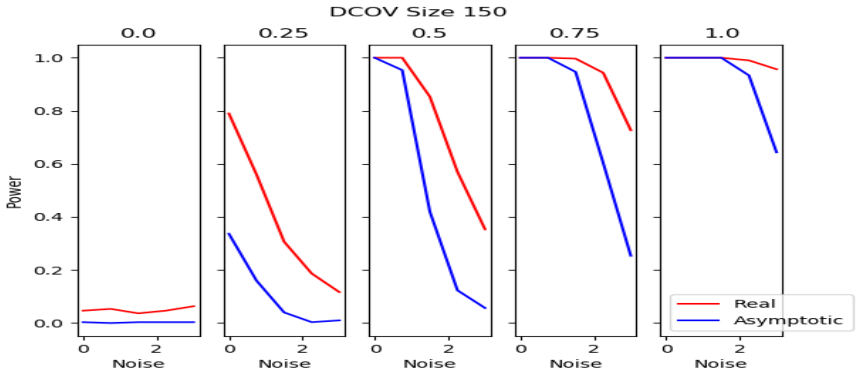

Figure 4.11: Power comparison between the asymptotic and the real version of DCOV for sample size 150

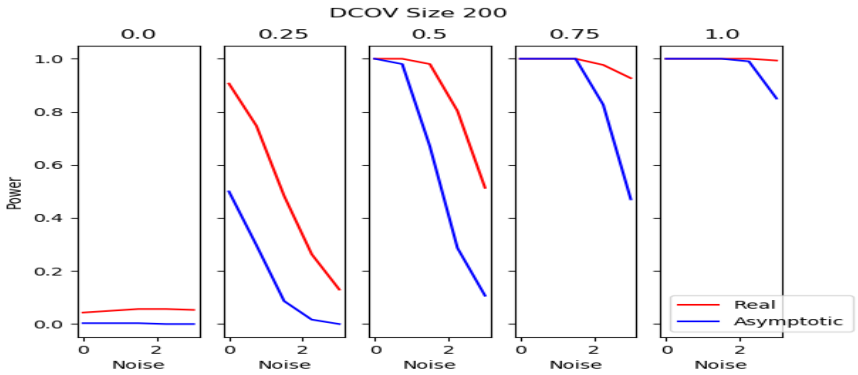

Figure 4.12: Power comparison between the asymptotic and the real version of DCOV for sample size 200

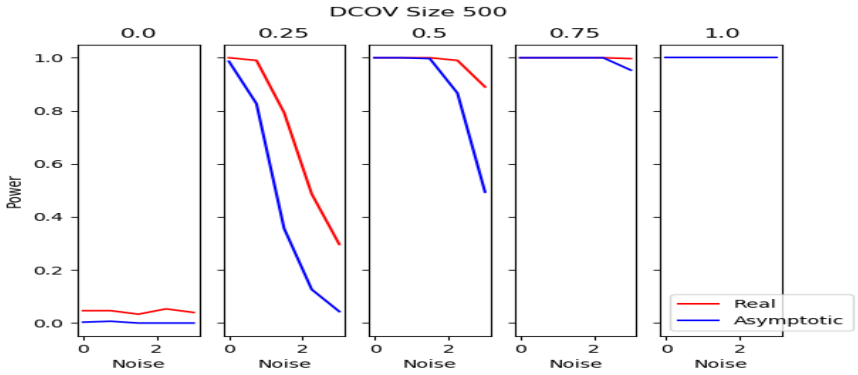

Figure 4.13: Power comparison between the asymptotic and the real version of DCOV for sample size 500

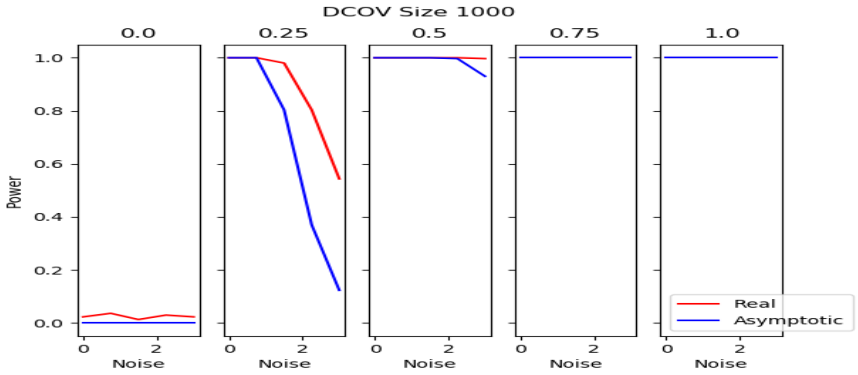

Figure 4.14: Power comparison between the asymptotic and the real version of DCOV for sample size 1000

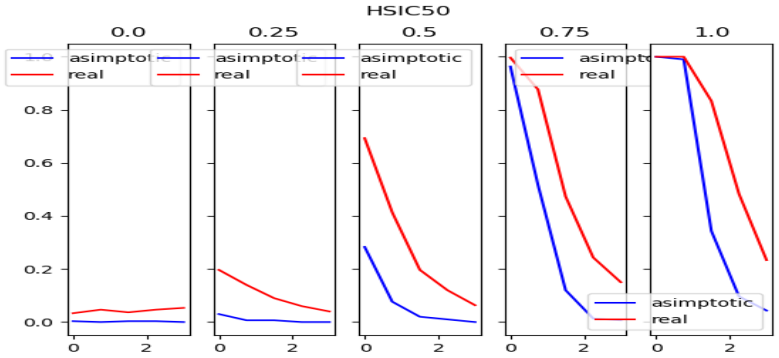

Figure 4.15: Power comparison between the asymptotic and the real version of HSIC for sample size 50

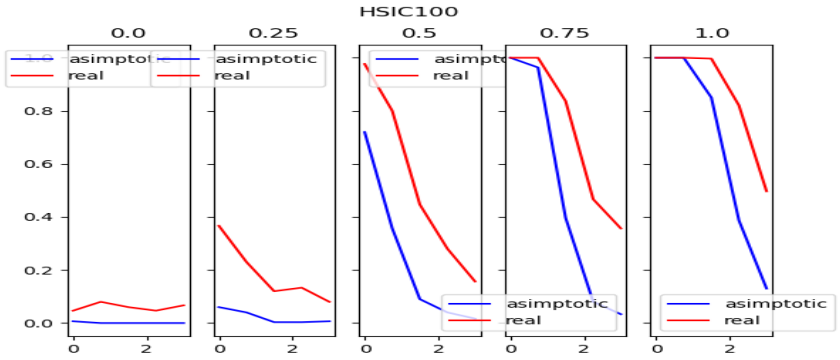

Figure 4.16: Power comparison between the asymptotic and the real version of HSIC for sample size 100

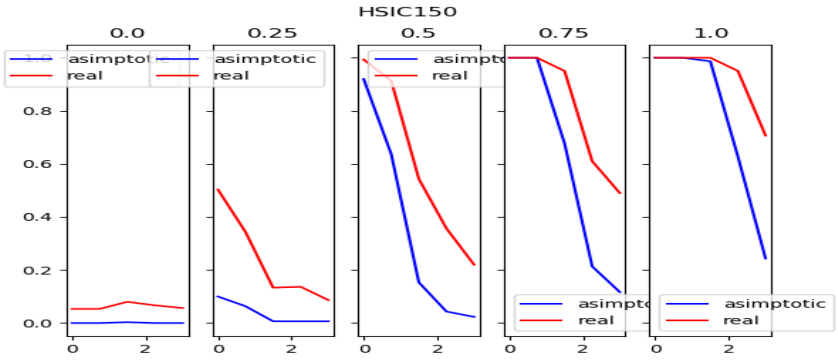

Figure 4.17: Power comparison between the asymptotic and the real version of HSIC for sample size 150

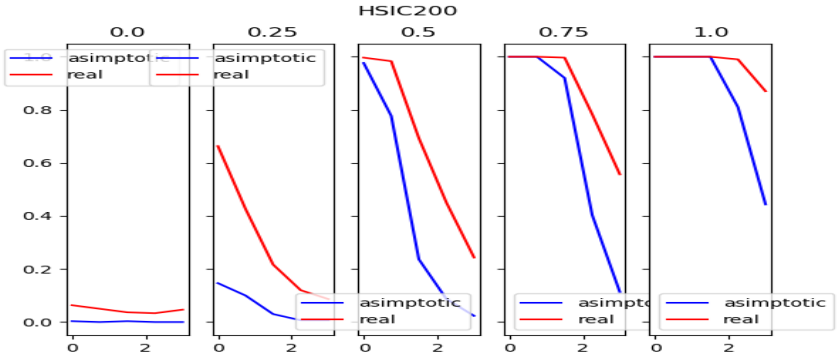

Figure 4.18: Power comparison between the asymptotic and the real version of HSIC for sample size 200

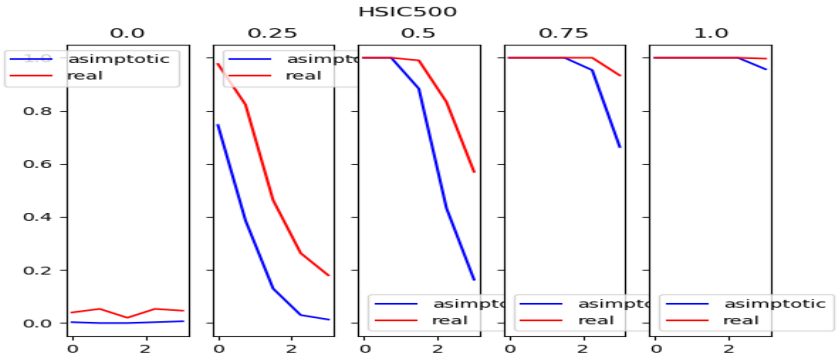

Figure 4.19: Power comparison between the asymptotic and the real version of HSIC for sample size 500

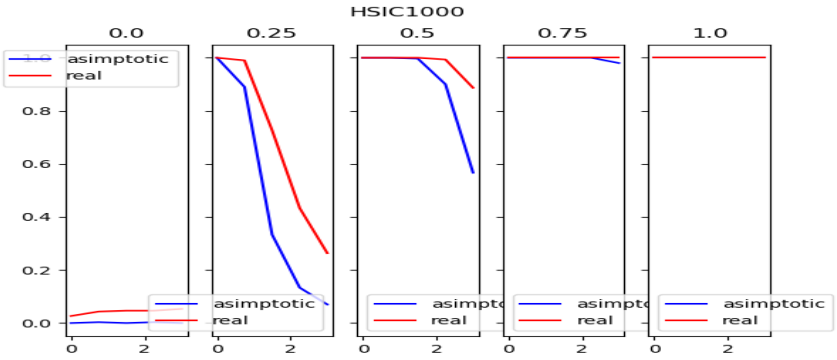

Figure 4.20: Power comparison between the asymptotic and the real version of HSIC for sample size 1000

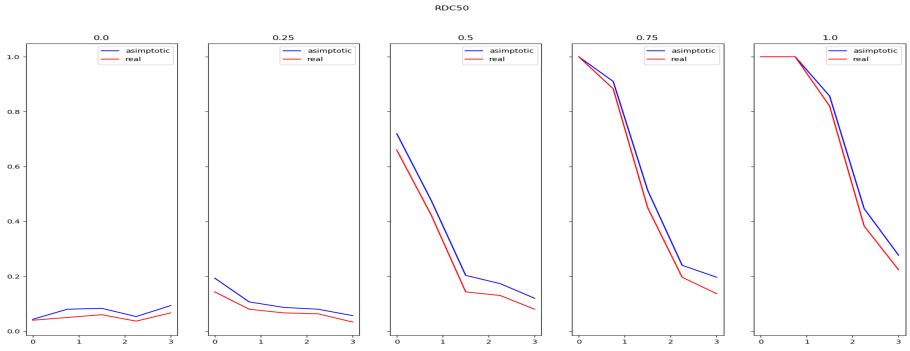

Figure 4.21: Power comparison between the asymptotic and the real version of RDC for sample size 50

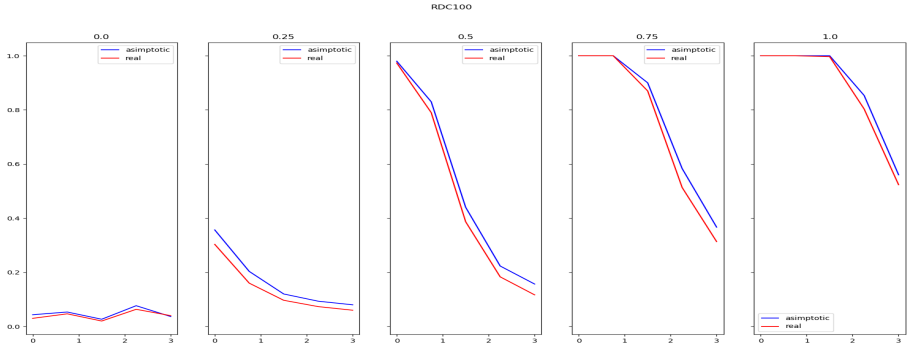

Figure 4.22: Power comparison between the asymptotic and the real version of RDC for sample size 100

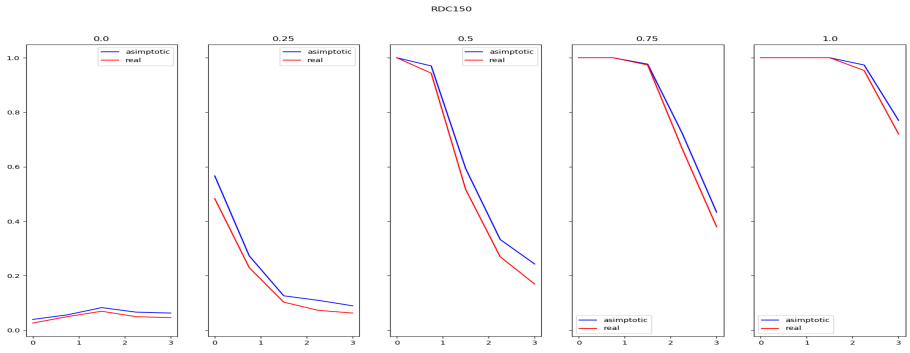

Figure 4.23: Power comparison between the asymptotic and the real version of RDC for sample size 150

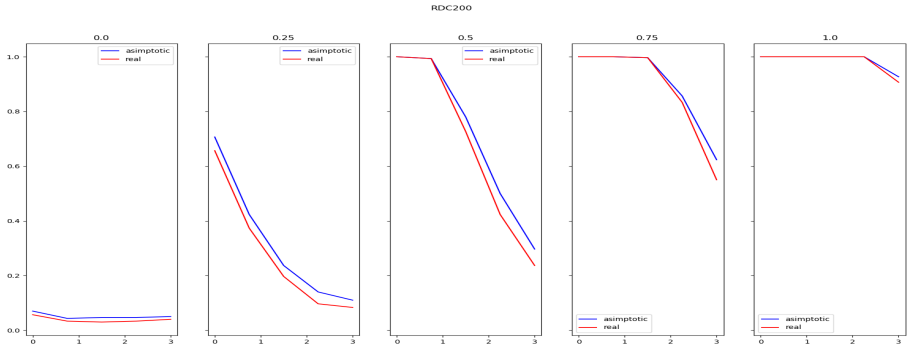

Figure 4.24: Power comparison between the asymptotic and the real version of RDC for sample size 200

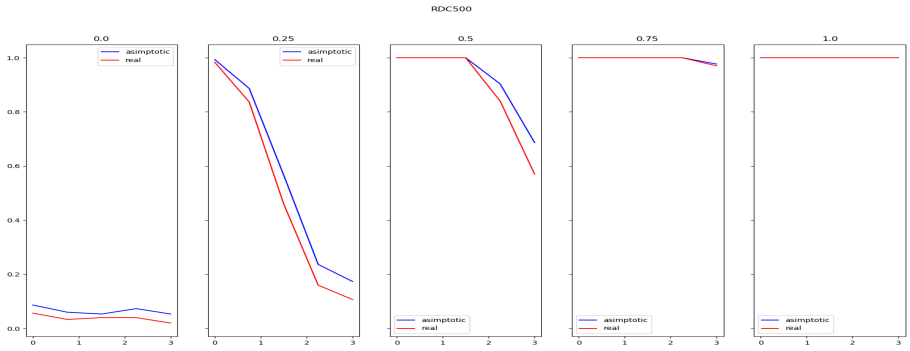

Figure 4.25: Power comparison between the asymptotic and the real version of RDC for sample size 500

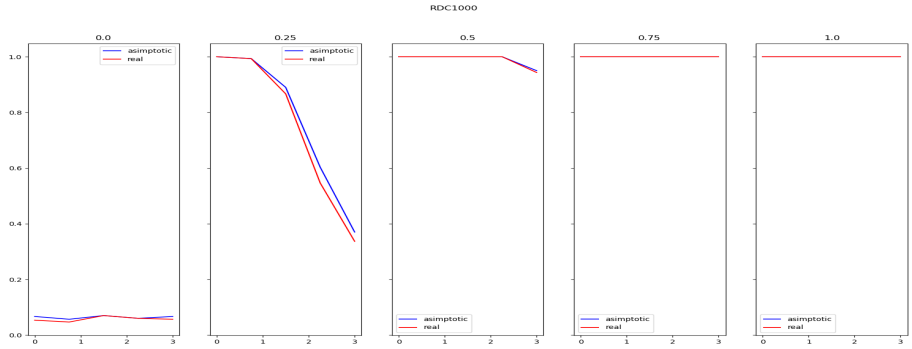

Figure 4.26: Power comparison between the asymptotic and the real version of RDC for sample size 1000

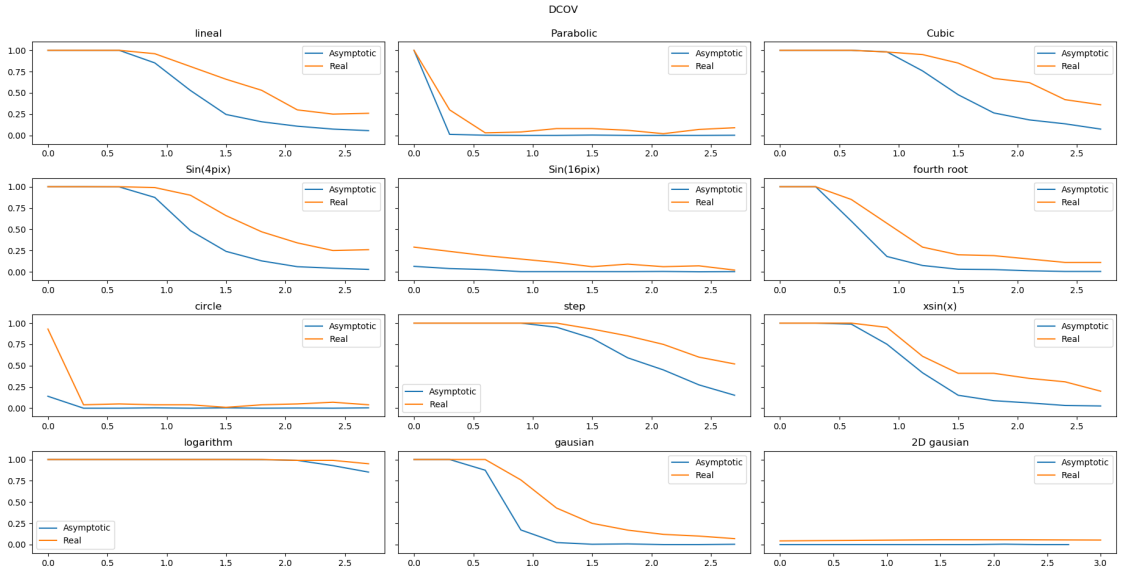

Figure 4.27: Power comparison between the asymptotic and the real version of DCOV for different relation patterns with sample sizes of 200, significance level of 0.05 and Gaussian noise from 0 to 3

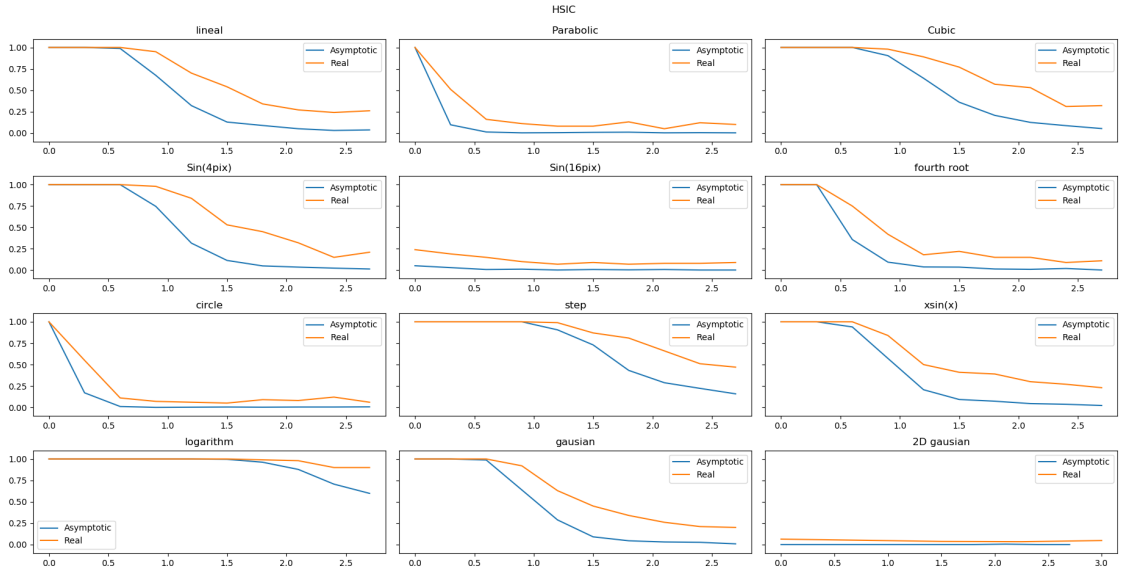

Figure 4.28: Power comparison between the asymptotic and the real version of HSIC for different relation patterns with sample sizes of 200, significance level of 0.05 and Gaussian noise from 0 to 3

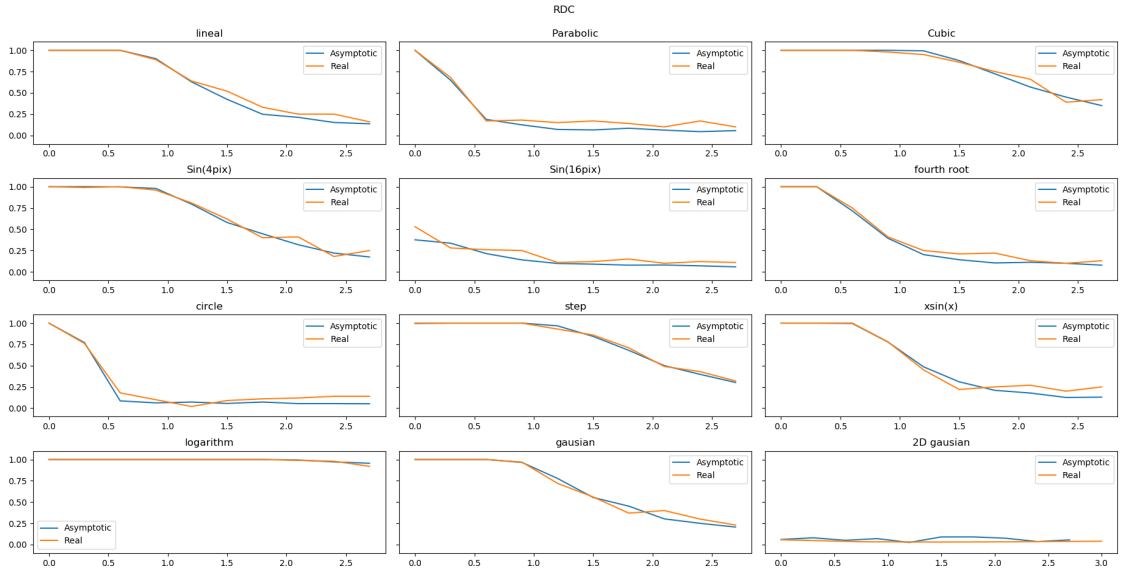

Figure 4.29: Power comparison between the asymptotic and the real version of RDC for different relation patterns with sample sizes of 200, significance level of 0.05 and Gaussian noise from 0 to 3

aproximations of the real one, while in RDC we used the actual asymptotic distribution. Figures 4.30, 4.31, 4.32 show respectively the obtained results.

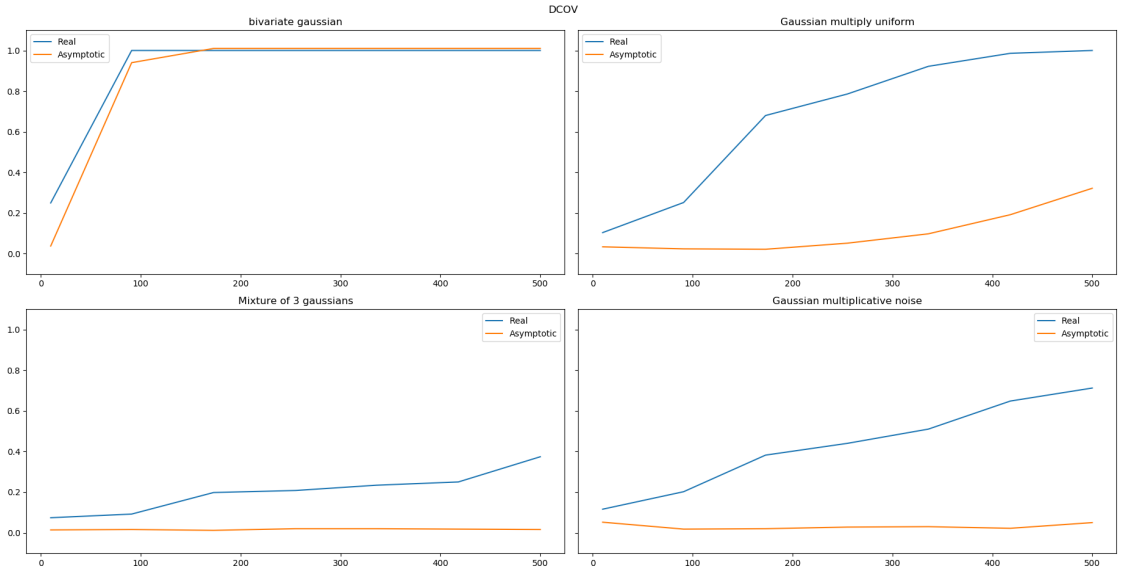

Figure 4.30: Power comparison between the asymptotic and the real version of DCOV for different relation patterns with sample sizes of varing from 10 to 500, significance level of 0.05

For our last experiment we studied how rotating variables may affect the power of our tests, Figure 4.5 shows samples of the same data with different rotation angle.

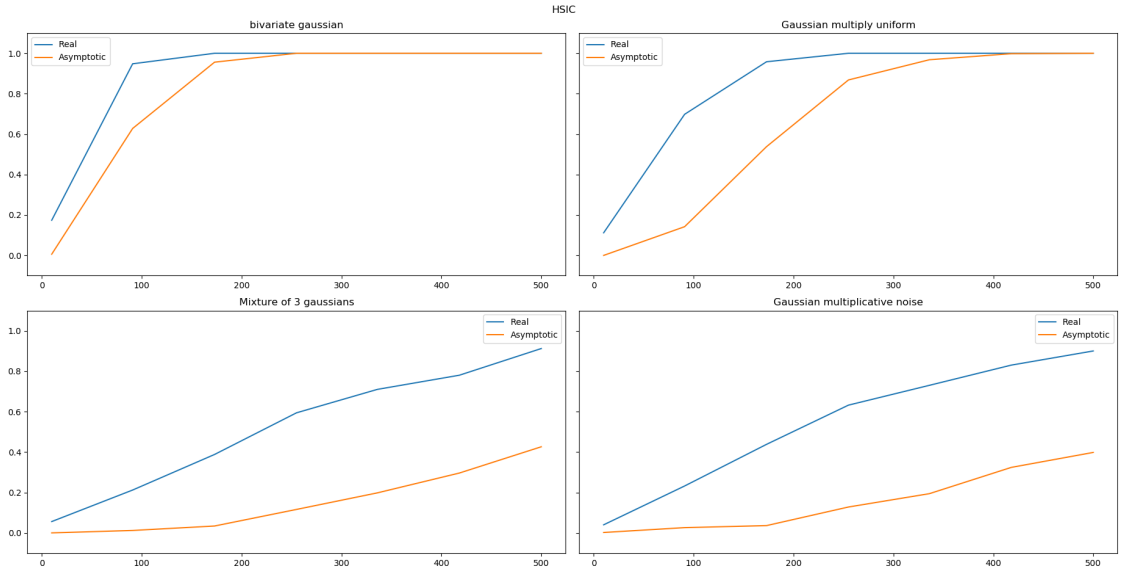

Figure 4.31: Power comparison between the asymptotic and the real version of HSIC ffor different relation patterns with sample sizes of varing from 10 to 500, significance level of 0.05

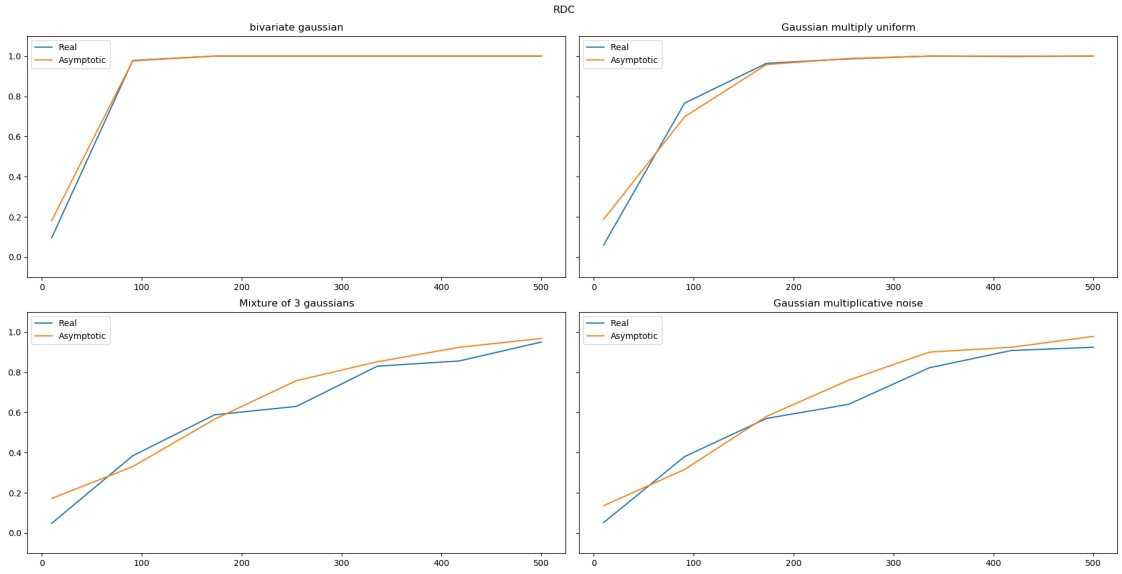

Figure 4.32: Power comparison between the asymptotic and the real version of RDC for different relation patterns with sample sizes of varing from 10 to 500, significance level of 0.05

4.2. Time

As we have seen in this experiments generally RDC outperforms the rest of tests, both in it's *real* and asymptotic version. Now to conclude this set of experiments, we will compare the complexity and the times to calculate the statistic for sample sizes going from 10 to 1000.

The following table, Table 4.1, contains different charasteristics of the studied tests, as well as Pearson'sρ to compare, taken from [9].

Coefficient	Non-Linear	N dimensional	Complexity
Pearson'sρ	✗	✗	O(n)
HSIC	✓	✓	O(n^2)
DCOV	✓	✓	O(n^2)
RDC	✓	✓	O(k^2n)

Table 4.1: Table with differences between the statistics and other relevant independence statistics.

Finally Figure 4.33 showcase how RDC is considerably faster than HSIC and DCOV. In the figure we can see how around 500 the time curve changes slope, that is because for samples larger than 500 for a bivariate Gaussian the optimal k changes from 3 to 4, therefore the slope increases on the basis of $\frac{16}{9}$. Figure 4.34 showcases a polinomic aproximation for the times, where we can see how HSIC follows a quadratic form with respect to the sample size , while RDC follows a linear form with respect to the sample size.

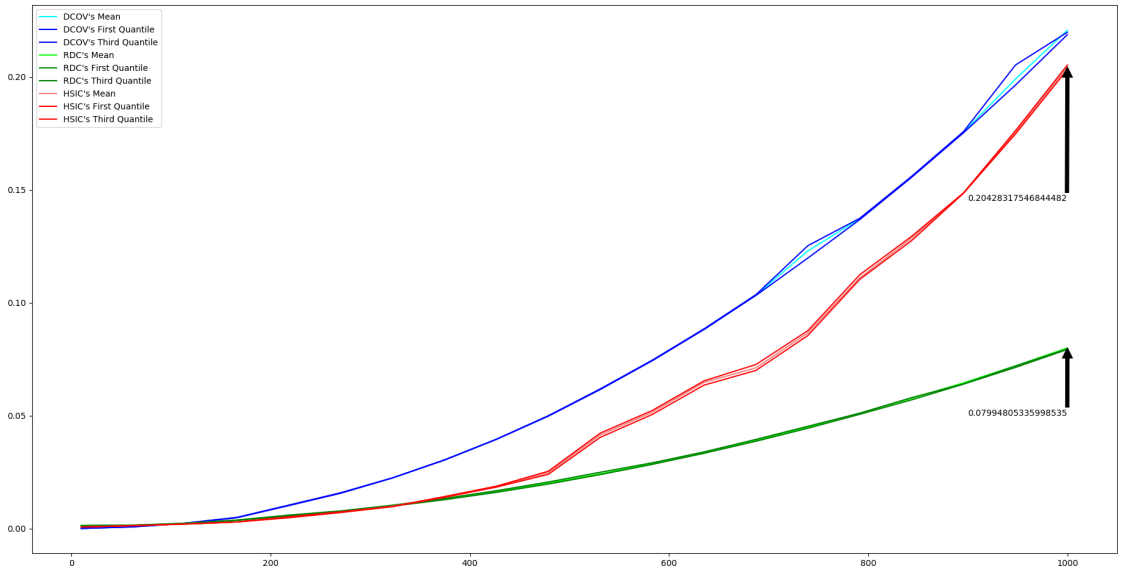

Figure 4.33: Comparison of time needed to calculate each statistic with different sample sizes, going from 10 to 1000

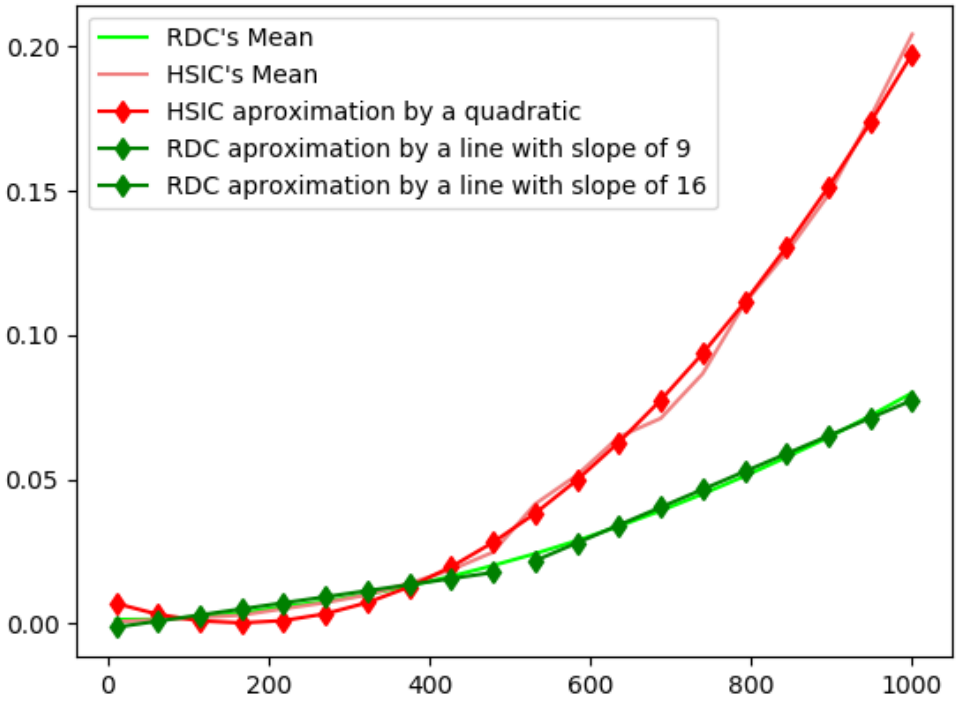

Figure 4.34: Theorical polinomical aproximation of HSIC and RDC put against the mean of the time needed to calculate the statistic

4.3. Conclusion

In general RDC performs better than the other statistics, there are few case scenarios where it may be better to use another test, the most relevant one being the relation pattern step, where DCOV outperformed RDC. This may be important to notice because this relation pattern is two variables X,Y, where Y = heaviside(X), this relation pattern is of key importance various scientific fields, such as differential equations, where it represents a signal which switches on at a specified time and stays switched on indefinitely.

Therefore for general purposes RDC will be the best answer, because it's more time efficent and generally performs better than the other tests, but if there is previous knowledge of the relation pattern that our data may follow then DCOV or HSIC may be a better solution.

5

Appendix

Bibliography

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

[2] Fukumizu K, Gretton A, Sun X, Schölkopf B. (2007) *Kernel measures of conditional dependence*

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

[4] Gabor J.Székely, Maria L. Rizzo and Nail K. Bakirov (2007) *Measuring and testing dependence by correlation of distances*

[5] Gabor J.Székely and Maria L. Rizzo (2009) *Brownian Distance Covariance*

[6] Rao M., Seth S., Xu J., Chen Y., Tagare H., and Príncipe J.C (2011) *A test of independence based on a generalized correlation function*

[7] Strang, Gilbert *Linear Algebra and its Applications* (2005)

[8] Hunter, John K.; Nachtergaele, Bruno *Applied Analysis* (2001)

[9] David Lopez-Paz and Bernhard Schölkopf. (2013) *The Randomized Dependence Coefficient*

[10] David Lopez-Paz, Philipp Hennig and Bernhard Schölkopf. (2013) *The Randomized Dependence Coefficient*

[11] H. Gebelein *Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung.* (1941)

[12] Par M.Michel Plancherel *Contribution À L'Étude de la reprÉsentation D'une fonction arbitraire par des intÉgrales dÉfinies* (Genève, 1910)

[13] Lyle D.Broemeling *An Account of Early Statistical Inference in Arab Cryptology* (2010)

[14] Jacob Bernoulli *Ars Conjectandi* (1713)


UAM
UNIVERSIDAD AUTONOMA
DE MADRID