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

mayo 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

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 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 $\mathbb{PQ}$

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 [2], section 5.5.2

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

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

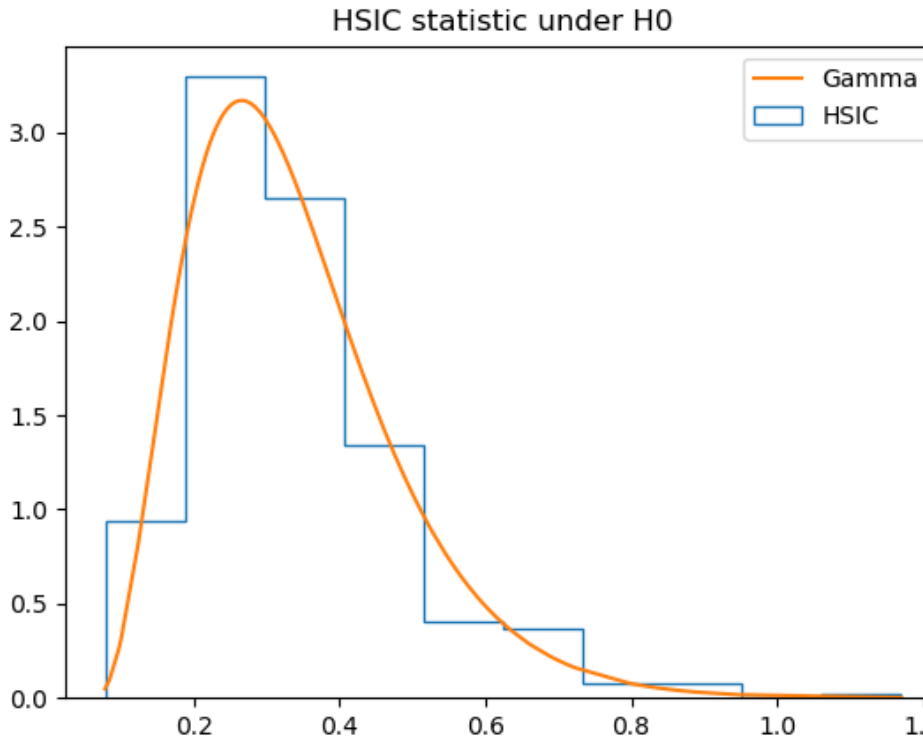
$$\lambda_l\psi_l(z_j) = \int h_{ijqr}\psi_l(z_i)dF_{iqr}$$

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

Approximating the $1-\alpha$ quantile of the null distribution

A hypothesis test using HSIC(Z) could be derived from Theorem 3.3 above by computing the $(1-\alpha)$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 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:*

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

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. 1.3.1 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 $mathcal L_2$-distance, imposing rotation invariance and scale equivariance, the energy distance is:

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

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 1.3.1. 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 1.3.1 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 1.3.3 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 $\|f_{XY}^n(t,s)\|^2$, $\|f_X^n(t)f_Y^n(s)\|^2$ and $\overline{f_{XY}^n(t,s)}f_X^n(t)f_Y^n(s)$. 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 1.3.3 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

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

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

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

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

$$\lim_{\theta\to 0} \sup \lim_{n\to\infty} \sup \|\nu^2_{n,\theta} - \nu^2_n\| = 0$$

For each $\theta > 0$

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

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 **[?] [?]***

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

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

2

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.

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

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

1.– defining the

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 [3], 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 2.1 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 2.2 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.

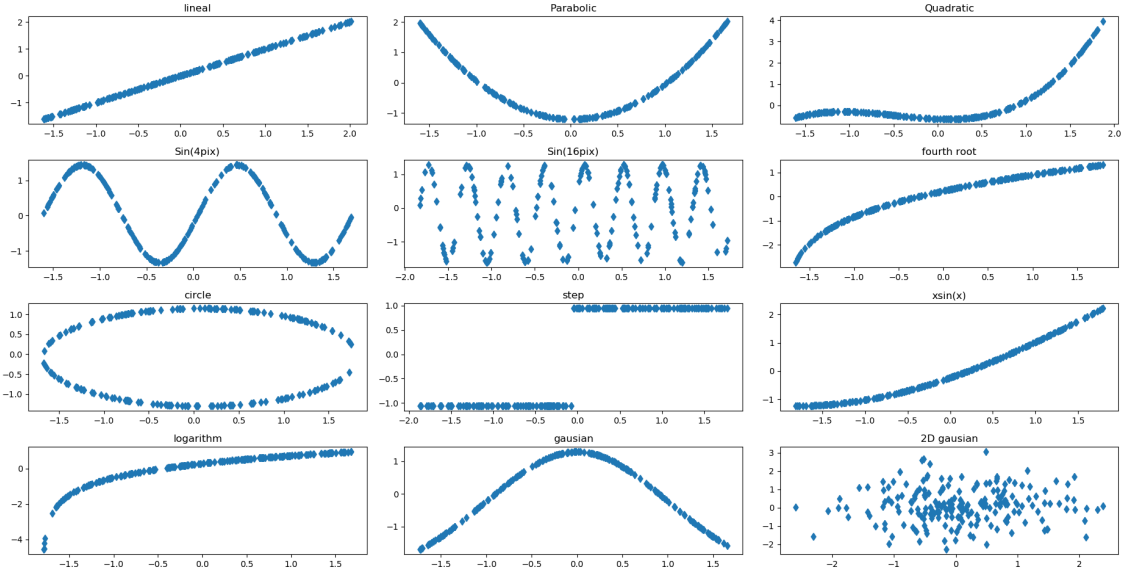


Figure 2.1: Representation of non linear dependance patterns

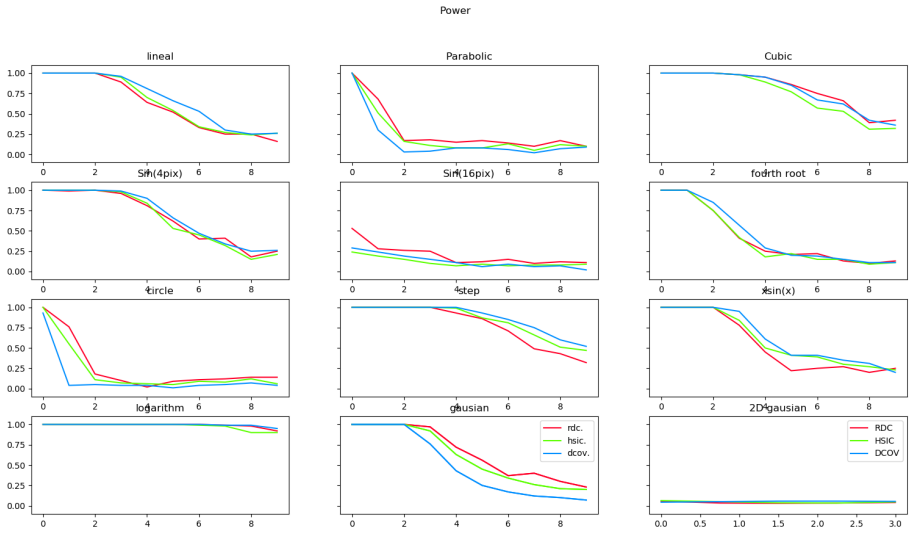


Figure 2.2: Power of tests adding gaussian noise to marginals

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 [4] 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 2.3. 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 2.4.

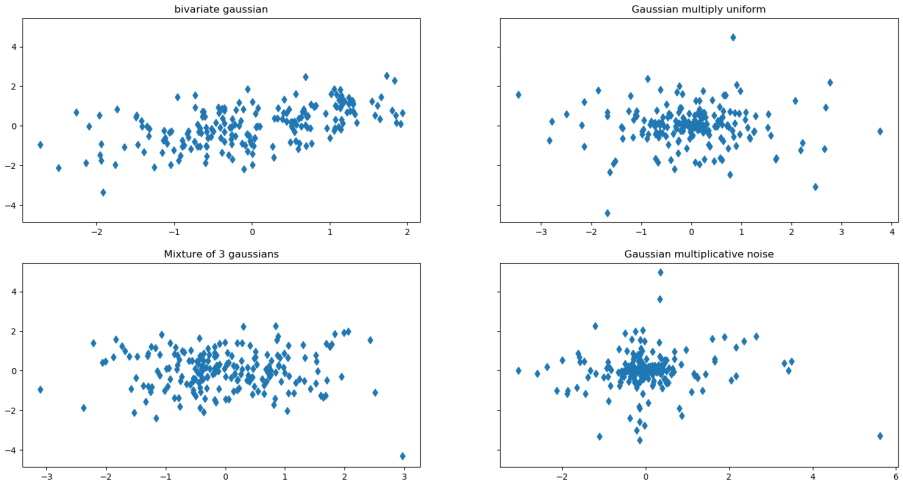


Figure 2.3: Samples from the data sets for the second experiment

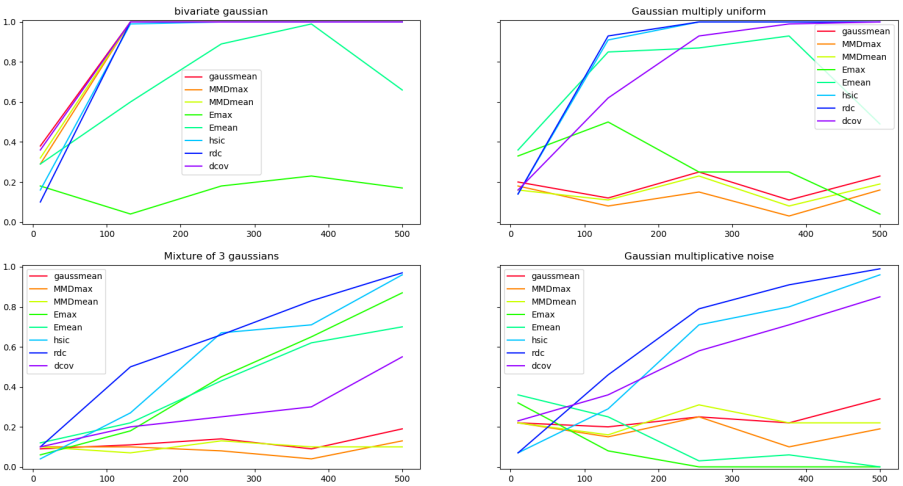


Figure 2.4: Power of tests adding gaussian noise to marginals

Bibliography

[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*

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

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

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

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


UAM
UNIVERSIDAD AUTONOMA
DE MADRID