# INDEPENDENCE MEASURES

Beatriz Bueno Larraz

Máster en Investigación e Innovación en Tecnologías de la Información y las Comunicaciones.
Escuela Politécnica Superior.

Máster en Matemáticas y Aplicaciones. Facultad de Ciencias.

UNIVERSIDAD AUTÓNOMA DE MADRID

---




09/03/2015

Advisors:

Alberto Suárez González
Jośe Ramón Berrendero Díaz

# Acknowledgements

This work would not have been possible without the knowledge acquired during both Master degrees. Their subjects have provided me essential notions to carry out this study.

The fulfilment of this Master's thesis is the result of the guidances, suggestions and encouragement of professors D. José Ramón Berrendero Díaz and D. Alberto Suárez Gonzalez. They have guided me throughout these months with an open and generous spirit. They have showed an excellent willingness facing the doubts that arose me, and have provided valuable observations for this research. I would like to thank them very much for the opportunity of collaborate with them in this project and for initiating me into research.

I would also like to express my gratitude to the postgraduate studies committees of both faculties, specially to the Master's degrees' coordinators. All of them have concerned about my situation due to the change of the normative, and have answered all the questions that I have had.

I shall not want to forget, of course, of my family and friends, who have supported and encouraged me during all this time.

# Contents

# List of Figures

# Chapter 1

# Introduction

The concept of distance between probability measures has been largely studied because of its numerous applications in Probability Theory, Information Theory, Bioinformatics and Statistics. 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, also named homogeneity testing. The goal in hypothesis testing is to accept or reject the null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$, versus the alternative hypothesis $H_1 : \mathbb{P} \neq \mathbb{Q}$, for a class of probability distributions $\mathbb{P}$ and $\mathbb{Q}$. For this purpose we will define a metric $\gamma$ 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 hypothesis testing, using $\mathbb{P} = \mathbb{P}_{X,Y}$ and $\mathbb{Q} = \mathbb{P}_X \cdot \mathbb{P}_Y$. Detecting dependences, especially non-linear ones, is a difficult problem in practice.

In this work we will introduce in detail three different types of independence tests. Two of them are tests in the literature, which can be derived from their corresponding homogeneity tests. One of the aims of this work is to order and summarize the published research on these tests. The third one is a novel independence test that has been developed during this Master Thesis.

In the first part of the work, which is composed of Chapters 2 to 5, we introduce a homogeneity test that consist in making embeddings of the original variables through non-linear transformations, into Hilbert spaces with reproducing kernel (RKHS). We first define RKHS's and analyse their properties. We then describe how probability distributions can be embedded into these spaces and explain how these embeddings can be used to characterize equality of the distributions. The homogeneity test that uses these embeddings is called Maximum Mean Discrepancy (MMD). We introduce also some estimators of the test and their asymptotic behaviour. Finally the original homogeneity test is adapted to define a test of independence. The independence test is a homogeneity test in which one compares the joint distribution of the variables with the product of the marginals. Finally we introduce other previous independence test which also uses kernel embeddings, but from another point of view. We show that both test are actually the same, although their formulation and estimators are different.

The second part of the work comprises Chapters 6 and 7. In the Chapter 6 we introduce a homogeneity test based on the energy distance, which is defined as a weighted $\mathcal{L}_2$ distance between probability distributions. The original definition of the test uses the Euclidean distance. This definition can be extended to general metric spaces. However, one needs to introduce some restrictions to the metric to ensure that the test characterizes equality of the distributions. In Chapter 7 we describe an independence test whose formulation is similar than the homogeneity one. However it is not the homogeneity test applied to the joint and the product distributions. Subsequently we introduce some estimators of the statistic of this independence test and their properties.

In the third part of this thesis, composed by Chapters 8 and 9, we derive relations between these methods. For instance, energy distance can be interpreted as MMD for a special choice of the kernel. To establish this equivalence between both homogeneity tests, we need to generalize the quantities used in the original tests. The equivalence is established through these generalizations. At the end of the first chapter we present a set of novel observations and results that do not appear before in the literature, which are interesting specially from a practical point of view. The second chapter is similar to the first one, but it establish the connection between the generalized independence methods.

In the final part of this report we introduce the new independence test developed in this work. This test is based on estimating the non-linear dependence between two variables through the non-Gaussianity of their one-dimensional projections. The theoretical justification of the test is given for the case of Gaussian marginals. Whether an extension to general distributions is possible remains unproven. We then introduce a novel way to characterize independence of the random variables through random projections.

Finally, we carry out some experiments, to compare the power of the proposed tests with other state-of-art independence tests.

# Chapter 2

# Reproducing Kernel Hilbert Spaces (RKHS)

RKHS's are a special type of Hilbert spaces, having a kernel that meets the reproducing property. These type of kernels are called reproducing kernels, and give the name to the spaces. These Hilbert spaces have some relevant applications in statistics, particularly in the field of statistical learning theory. The reason is that every function in an RKHS can be written as the limit of a linear combination of the kernel function with one free argument.

In addition to this, in the next chapter we will define a homogeneity test based on embeddings of probability distributions on RKHS's. The distance between distributions corresponds then to the distance between their corresponding embeddings. We will see that the unit ball of an RKHS is a rich enough space so that the expression for the discrepancy vanishes only if the two probability distributions are equal. At the same time it is restrictive enough for the empirical estimate at the discrepancy to converge quickly to its population counterpart as the sample size increases.

RKHS's are also practically useful in machine learning, for example when trying to make predictions by optimizing over a function $f$ in a Hilbert space $\mathcal{H}$. RKHS's have an advantage over common Hilbert spaces, because if $\|f_n - f\| \to 0$, where $\|\cdot\|$ is the distance derived from the inner product, then $f_n(x) \to f(x)$ for all $x$.

## 2.1 Definitions and principal properties

The classical theory of statistics is well developed for the linear case. However real world problems often require nonlinear methods. Detecting nonlinear dependencies is often important to make successful predictions. A possible way to take advantage of our knowledge of linear procedures is to transform the data into a different space so that nonlinear dependencies are transformed into linear ones. We will refer to this new space, typically high dimensional, as the feature space. Hilbert spaces $\mathcal{H}$ are often used as feature space, because they provide powerful

mathematical tools and intuitive geometric concepts. That is, we can transform the original data into a Hilbert space to equip it with a geometrical structure. A typical example to illustrate this is when the data are a set of books, which are not easy to compare. We can measure some characteristic of the books, as the number of words, the number of chapters, etc., and to translate them to a Hilbert space where we have a distance defined.

The first step is to map the data from $\mathcal{X}$, the original space, to the feature space.

**Definition 1.** *The function $\phi$ that maps the data to the feature space, a Hilbert space $\mathcal{H}$,*

$$\begin{array}{rcl} \phi : \mathcal{X} & \longrightarrow & \mathcal{H} \\ x & \mapsto & \phi(x) \end{array}$$

*is known as a* ***feature map****.*

Since we are working with Hilbert spaces, we can also define kernel functions in terms of the inner product in $\mathcal{H}$:

**Definition 2.** *A function that represents a dot product defined on a feature space is called a* ***kernel*** *function.*

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

$$\begin{array}{rcl} k : \mathcal{X} \times \mathcal{X} & \longrightarrow & \mathbb{R} \\ (x, x') & \mapsto & k(x, x') = \langle \phi(x), \phi(x') \rangle \end{array}$$

Theoretically, the function $k$ can also go to the complex numbers $\mathbb{C}$, but for most applications $\mathbb{R}$ is enough.

As we have said, in the feature space our estimation methods are linear. Thus if we are able to formulate them in terms of kernel evaluations, we never have to work explicitly in the high dimensional feature space. This idea underlies the well known classification algorithm called Support Vector Machine (SVM).

Now we will impose some restrictions to the kernel functions, to define a new class of kernels with more interesting properties.

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

1. $k(x, .) \in \mathcal{H}$, $\forall x \in \mathcal{X}$.
2. *Reproducing property:* $\langle f, k(x, .) \rangle = f(x)$, $\forall f \in \mathcal{H}$ *and* $\forall x \in \mathcal{X}$.

The name of "reproducing property" comes from the fact that the value of the function $f$ at the point $x$ is reproduced by the inner product of $f$ and $k$. From this definition it is clear that:

**Proposition 1.** *If $k$ is a reproducing kernel, then $k(x, x') = \langle k(x, .), k(x', .) \rangle$, $\forall (x, x') \in \mathcal{X} \times \mathcal{X}$.*

Using the canonical feature map $\phi(x) = k(x, \cdot)$, we see that a reproducing kernel is a kernel in the sense of Definition 2. Then we can also define a special type of Hilbert spaces based on these kernels:

**Definition 4.** *A Hilbert space of real-valued functions which possesses a reproducing kernel is called a* ***reproducing kernel Hilbert space (RKHS)****. Besides, the canonical feature map $\phi$ of an RKHS is defined in terms of the reproducing kernel $k$,*

$$\phi(x) = k(x, \cdot).$$

This means that the kernel has one argument fixed as $x$ and the second is free, so $x$ is associated with a function in $\mathcal{H}$. Some examples of RKHS's and its corresponding kernels are:

- $\mathcal{H} = \{f | f(0) = 0,\ f \text{ absolutely continuous and } f' \in \mathcal{L}^2(0,1)\}$, with $f'$ the derivative of $f$ almost everywhere, and $\mathcal{L}^2(0,1)$ the set of square integrable complex valued functions with support in $(0,1)$, is a Hilbert space with inner product:

  $$\langle f, g\rangle_{\mathcal{H}} = \int_0^1 f'\bar{g}'\mathrm{d}\lambda,$$

  where $\lambda$ is the Lebesgue measure and $\bar{g}$ denotes the complex conjugate of the function $g$. In fact this is an example of Sobolev space. For a general domain $\mathcal{X} \subset \mathbb{R}^n$, these spaces are defined as:

  $$W^{k,p}(\mathcal{X}) = \{f \in \mathcal{L}^p(\mathcal{X}) \mid D^\alpha f \in \mathcal{L}^p(\mathcal{X}), \forall \alpha \in \mathbb{N}^n \text{ such that } |\alpha| \leq k\},$$

  where $D^\alpha f$ is the multi-index partial derivative of $f$, that is, for $\alpha = (\alpha_1, \ldots, \alpha_n)$:

  $$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}},$$

  where $|\alpha| = \alpha_1 + \ldots + \alpha_n$ and $k$ is called the order of the space. In this example we are using $\mathcal{X} = [0,1]$ and $\mathcal{H} = W^{1,2}([0,1])$. This Hilbert space has reproducing kernel $k(x, x') = \min(x, x')$. The weak derivative *(generalization of the concept of the derivative for functions not assumed differentiable, $f(x) = \int_0^x f'(u)\mathrm{d}u$)* of $\min(., x)$ is $\mathbf{1}_{(0,x)}$ so it satisfies the reproducing property:

  $$\langle f, k(x, .)\rangle_{\mathcal{H}} = \int_0^x f'(x)\mathrm{d}\lambda(x) = f(x).$$

- $\mathcal{H} = H^1(\mathbb{R}) = \{f |\ f \text{ absolutely continuous and } f, f' \in \mathcal{L}^2(\mathbb{R})\}$ , where $f'$ is the derivative of $f$ almost everywhere, is a Hilbert space with inner product:

  $$\langle f, g\rangle_{\mathcal{H}} = \int_{\mathbb{R}} (f\overline{g} + f'\overline{g'})\mathrm{d}\lambda.$$

  A simple integration by parts shows that $\mathcal{H}$ has reproducing kernel $k(x, x') = \frac{1}{2}e^{-|x-x'|}$.

For more examples and the proof of the reproducing property of the second example the reader can see [2], Section 1.2.

## 2.2 Characterizing reproducing kernels

Now that the basic definitions and properties of an RKHS have been introduced, the natural question is when a complex-valued function $k$ defined on $\mathcal{X} \times \mathcal{X}$ is a reproducing kernel. We will present an important characterization of this kind of functions, which can be used also to connect RKHS's with other spaces, such as the Hilbert space generated by a random process. The main part of this section has been obtained from [2], Sections 1.2 and 1.3, and [3], Section 4.2. First we need the following definition:

**Definition 5.** *A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{C}$ is called a **positive type** function (or positive definite function) if $\forall n \geq 1$, $\forall (\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$ and $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$:*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \bar{\alpha}_j k(x_i, x_j) \geq 0. \tag{2.1}$$

If $k$ is a real positive definite function according to this definition, it is automatically symmetric, because the values $\alpha_i$ are complex. Therefore we will not indicate the symmetry condition in the following statements. Besides, we will see that being a positive definite function is equivalent to being a reproducing kernel.

**Theorem 1.** *A kernel in a Hilbert space $\mathcal{H}$ is positive definite.*

*Proof.* The result easily follows from the definition of kernel and the properties of the dot product:

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \bar{\alpha}_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \bar{\alpha}_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \alpha_i \phi(x_i), \alpha_j \phi(x_j) \rangle_{\mathcal{H}} \\
&= \langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \sum_{j=1}^{n} \alpha_j \phi(x_j) \rangle_{\mathcal{H}} \\
&= \Big\| \sum_{i=1}^{n} \alpha_i \phi(x_i) \Big\|_{\mathcal{H}}^2 \geq 0
\end{aligned}$$

□

In principle, being positive definite is a necessary but not sufficient condition for a function to be a kernel. Since a reproducing kernel is also a kernel, we have proved that:

$$\boxed{\text{Reproducing kernel} \longrightarrow \text{Kernel} \longrightarrow \text{Positive definite}}$$

Remarkably, the reverse direction also holds, that means, every positive definite function is a reproducing kernel. But first we will enunciate a simple lemma that is needed for the proof:

**Lemma 1.** *Any bilinear function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x,x) \geq 0$, satisfies the Cauchy-Schwartz inequality, i.e., $|f(x,y)|^2 \leq f(x,x)f(y,y)$.*

*Proof.* We will calculate the value $f(x-\alpha y, x-\alpha y) \geq 0$, where $\alpha \in \mathbb{R}$:

$$f(x-\alpha y, x-\alpha y) = f(x,x) + \alpha^2 f(y,y) - 2\alpha f(x,y).$$

Consider now two separate cases:

- $f(y,y) \neq 0$: We take the value $\alpha = \frac{f(x,y)}{|f(x,y)|}t$ where $t \in \mathbb{R}$. Now the product is:

$$\begin{aligned} f(x-\alpha y, x-\alpha y) &= f(x,x) + \left(\frac{f(x,y)}{|f(x,y)|}t\right)^2 f(y,y) - 2\frac{f(x,y)}{|f(x,y)|}tf(x,y) \\ &= f(x,x) + \frac{(f(x,y))^2}{|f(x,y)|^2}t^2 f(y,y) - 2\frac{(f(x,y))^2}{|f(x,y)|}t \\ &= f(x,x) + t^2 f(y,y) - 2t|f(x,y)| \geq 0. \end{aligned}$$

  This is a quadratic equation in $t$, so its solution is:

$$t = \frac{2|f(x,y)| \pm \sqrt{4|f(x,y)|^2 - 4f(x,x)f(y,y)}}{2f(y,y)}.$$

  But the parabola is always above zero, so it does not have two different real roots. This means that the square root of the solution is less or equal zero, i.e.:

$$4|f(x,y)|^2 - 4f(x,x)f(y,y) \leq 0 \Longrightarrow |f(x,y)|^2 \leq f(x,x)f(y,y).$$

- $f(y,y) = 0$: In this case the product becomes as

$$f(x-\alpha y, x-\alpha y) = f(x,x) - 2\alpha f(x,y) \geq 0 \Longrightarrow f(x,x) \geq 2\alpha f(x,y).$$

  Since the inequality holds for any value of $\alpha$, $f(x,y)$ must be zero.

□

**Theorem 2.** ***(Moore-Aronszajn)*** *Every positive definite function $k$ is the kernel of a unique RKHS $\mathcal{H}$.*

*Proof.* We define the space generated by the function $k$:

$$\mathcal{H}_k \equiv \Big\{ \sum_{i=1}^{N} a_i k(\cdot, x_i) \text{ for } N \in \mathbb{N}, a_j \in \mathbb{R}, x_j \in \mathcal{X}, j = 1, \ldots, N \Big\}. \tag{2.2}$$

Now we take two functions $f$ and $g$ from $\mathcal{H}_k$ of the form:

$$f \equiv \textstyle\sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \quad g \equiv \sum_{j=1}^{m} \beta_j k(\cdot, \tilde{x}_j).$$

We can define the function:

$$\langle f, g \rangle \equiv \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(\tilde{x}_j, x_i). \tag{2.3}$$

This definition is independent of the representation that we have chosen for $f$, since it can be written as:

$$\langle f, g \rangle = \sum_{j=1}^{m} \beta_j \sum_{i=1}^{n} \alpha_i k(\tilde{x}_j, x_i) = \sum_{j=1}^{m} \beta_j f(\tilde{x}_j).$$

$k$ is symmetric because, from (2.1), it is a real positive definite function with complex parameters $\alpha_i$, as we mentioned before. Thus the definition (2.3) is also independent of the representation of $g$:

$$\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \beta_j k(\tilde{x}_j, x_i) = \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \beta_j k(x_i, \tilde{x}_j) = \sum_{i=1}^{n} \alpha_i g(x_i). \tag{2.4}$$

Then $\langle f, g \rangle$ depends on $f$ and $g$ only through their values. This function is bilinear and symmetric, and since $k$ is positive definite, $\langle f, f \rangle \geq 0$. Moreover if $f = 0$ then $\langle f, f \rangle = 0$. To show that (2.3) is an inner product space we only need to prove that if $\langle f, f \rangle = 0$ then $f = 0$. By the previous lemma, we know that the Cauchy-Schwartz inequality holds under these conditions. So if we get a function $f$ such that $\langle f, f \rangle = 0$, then taking $g(x) = k(\cdot, x)$ in Equation (2.4) we get:

$$|f(x)|^2 = \left| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \right|^2 = |\langle f, k(\cdot, x) \rangle|^2 \leq \langle f, f \rangle \, \langle k(\cdot, x), k(\cdot, x) \rangle = 0, \quad \forall x \in \mathcal{X}.$$

Then we have that $\langle \cdot, \cdot \rangle$ is a dot product in $\mathcal{H}_k$, so that it is a pre-Hilbert space. It is also a metric space with the metric defined by the norm:

$$\|f\|_{\mathcal{H}_k} = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j).$$

It is well known that any metric space can be completed uniquely. So we denote as $\mathcal{H}$ the completion of $\mathcal{H}_k$ and as $\varphi : \mathcal{H}_k \to \mathcal{H}$ the corresponding isometric embedding. Then $\mathcal{H}$ is a Hilbert space and:

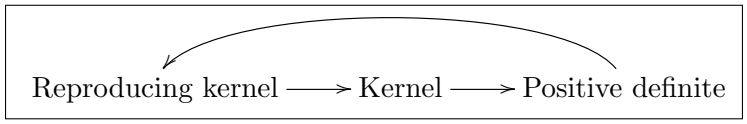
$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = \langle \varphi(k(\cdot, x)), \varphi(k(\cdot, y)) \rangle_{\mathcal{H}}.$$

Therefore if we define the feature map as $\phi(x) = \varphi(k(\cdot, x))$ we get that $k$ is a kernel in $\mathcal{H}$. Now we will prove that $k$ is actually a reproducing kernel. We use Equation (2.4) with $g = k(\cdot, x)$:

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = \langle f, g \rangle_{\mathcal{H}_k} = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x),$$

since $k$ is symmetric. This means that $k$ meets the reproducing property. □

The main part of this proof is taken from [4] (Theorem 4.16 page 118). This theorem ensures that the kernel is unique, although the feature map is not. It also shows that being a reproducing kernel is equivalent to the property of being a positive definite function. So we have that any positive definite function is a kernel in some Hilbert space and we need not to specify explicitly the spaces.




Moreover, this theorem states that any kernel function is in fact a reproducing kernel of some RKHS. This theorem also gives us a way to build RKHS spaces, using positive definite functions as reproducing kernels.

There is an alternative definition of RKHS that allows us to prove some interesting properties, with many computational applications.

**Definition 6.** *$\mathcal{H}$ is an RKHS if the evaluation operator $\delta_x$, defined as $\delta_x(f) = f(x) \in \mathbb{R}$ for $f \in \mathcal{H}$, is bounded $\forall x \in \mathcal{X}$, i.e., there exists $\lambda_x \geq 0$ such that:*

$$|f(x)| = |\delta_x(f)| \leq \lambda_x \|f\|_{\mathcal{H}} \ \forall f \in \mathcal{H}.$$

It is clear from this definition that the property of RKHS's mentioned in the introduction of the chapter holds, that is, convergence in the RKHS norm implies convergence at every point.

$$|f_n(x) - f(x)| = |\delta_x(f_n - f)| \leq \lambda_x \|f_n - f\|_{\mathcal{H}}, \ \forall f_n, f \in \mathcal{H}.$$

It is easy to see that this property does not hold for all Hilbert spaces. For example it fails to hold on $\mathcal{L}^2(\mathcal{X})$. This property is really useful in machine learning, when trying to make predictions by optimizing over $f \in \mathcal{H}$.

Now we have to see that both definitions of RKHS's are actually equivalent. But first let enunciate a well-known theorem that is needed for the proof.

**Theorem 3.** ***(Riesz Representation Theorem)*** *If $T$ is a bounded linear operator on a Hilbert space $\mathcal{H}$, then there exists some $g \in \mathcal{H}$ such that $\forall f \in \mathcal{H}$:*

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

**Theorem 4.** *$\mathcal{H}$ has bounded linear evaluation operators, $\delta_x$, if and only if $\mathcal{H}$ has a reproducing kernel.*

*Proof.* ($\Longrightarrow$) By the Riesz's Representation Theorem, since $\delta_x$ is a bounded linear operator, there exists $\varphi_x \in \mathcal{H}$ such that $\delta_x(f) = \langle f, \varphi_x \rangle$. Using the definition of $\delta_x$ we have that:

$$f(x) = \langle f, \varphi_x \rangle,$$

which is the reproducing property. So $\varphi_y(x) = k(x, y)$ is a reproducing kernel and $\mathcal{H}$ is an RKHS.

($\Longleftarrow$) The kernel $k$ of our space meets the reproducing property, $\langle f, k(x, .)\rangle = f(x)$. Hence, using the Cauchy-Schwarz inequality:

$$\begin{aligned} |\delta_x(f)| &= |f(x)| \\ &= |\langle f, k(x, .)\rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}}\|k(x, .)\|_{\mathcal{H}} \\ &= \|f\|_{\mathcal{H}}\langle k(x, .), k(x, .)\rangle_{\mathcal{H}}^{\frac{1}{2}} \\ &= \|f\|_{\mathcal{H}}k(x, x)^{\frac{1}{2}} \end{aligned}$$

Then it is sufficient to take $\lambda_x = k(x, x)^{\frac{1}{2}}$ to ensure that the evaluation operator is bounded. Moreover, in the Cauchy-Schwartz inequality, the bound is reached when $f(\cdot) = k(\cdot, x)$, and then $\|\delta_x\| = k(x, x)^{\frac{1}{2}}$. □

Another important practical question is to determine whether a given function belongs to a given RKHS. We know that any function of the RKHS is a linear combination of kernels, or a limit of such combinations (limits of Cauchy sequences). So, roughly speaking, a necessary condition for a function to belong to an RKHS is to be at least as smooth as the kernel.

In this chapter we have introduced the basic definitions and properties about RKHS's and their kernel functions. Now we will use them to define a homogeneity test based on embeddings of probability measures in these spaces. Later we will define also its corresponding independence test.

# Chapter 3

# Maximum Mean Discrepancy (MMD)

RKHS's can be used to define a homogeneity test, that can be interpreted in terms of the embeddings of the probability measures on one such space. When the homogeneity test is applied to the joint distribution and the product of the marginals yields an independence test, which is the main subject of this work.

The test consists in maximizing a measure of discrepancy between functions that belong to a certain family $\mathcal{F}$. As said earlier, such family should be rich enough to "detect" all the possible differences between the two probability measures. However, $\mathcal{F}$ should be small, so that it is possible to consistently estimate the test statistic with a reasonable sample size. RKHS's will help us to find such functional family $\mathcal{F}$.

## 3.1 Definition of MMD

We will start defining the measure of discrepancy between distribution functions in terms of embeddings in an RKHS. Let $X$ and $Y$ be random variables defined in some metric space $\mathcal{X}$ with probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively. Given two samples of these variables, obtained from $\mathbb{P}$ and $\mathbb{Q}$ independently and identically distributed (i.i.d.), we want to determine whether $\mathbb{P} \neq \mathbb{Q}$. The goal is to define a non negative discrepancy $\gamma$ that takes the value zero if and only if $\mathbb{P} = \mathbb{Q}$.

We will base our development in the following lemma. We denote the space of bounded continuous functions on $\mathcal{X}$ by $\mathcal{C}(\mathcal{X})$:

**Lemma 2.** *The Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$ are equal if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y)$, $\forall f \in \mathcal{C}(\mathcal{X})$.*

We set a general definition for the statistic, using a yet unspecified function class $\mathcal{F}$.

**Definition 7.** *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$, the* ***Maximum Mean Discrepancy***

***(MMD)*** *based on* $\mathcal{F}$ *is*

$$\gamma(\mathbb{P}, \mathbb{Q}) = MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) \equiv \sup_{f \in \mathcal{F}} \{\mathbb{E}f(X) - \mathbb{E}f(Y)\}. \tag{3.1}$$

Such $\mathcal{F}$ should be rich enough so that the expression for the MMD vanishes only if both probability distributions are equal (as we will see in Example 1 of Section 3.3). At the same time it should be restrictive enough for the empirical estimate to converge quickly as the sample size increases. As shown by Lemma 2, the class $\mathcal{C}(\mathcal{X})$ is rich enough, but too large for the finite sample setting. RKHS's will help us to find a suitable class of functions.

## 3.2 Hilbert Space Embeddings

As we have already said, we need to specify a class of functions $\mathcal{F}$ in Equation (3.1), by resorting to RKHS's. Henceforth we will use as $\mathcal{F}$ the unit ball in a reproducing kernel Hilbert space $\mathcal{H}$, as proposed in [1], which is rich enough but not too large. Then we will rewrite MMD in terms of kernel embeddings of probability measures. We will start extending the notion of feature map $\phi$ to the embedding of a probability distribution.

**Lemma 3.** *If the kernel* $k$ *is measurable and* $\mathbb{E}\sqrt{k(X,X)} < \infty$*, where* $X$ *is a random variable with distribution* $\mathbb{P}$*, then there exists* $\mu_{\mathbb{P}} \in \mathcal{H}$ *such that*

$$\mathbb{E}f(X) = \langle f, \mu_{\mathbb{P}} \rangle \text{ for all } f \in \mathcal{H}.$$

*Proof.* We define the linear operator $T_{\mathbb{P}} f \equiv \mathbb{E}f(X)$ for all $f \in \mathcal{F}$. This operator $T_{\mathbb{P}}$ is bounded under the assumptions of the lemma. It can be proved using the Jensen and Cauchy-Schwartz inequalities and the reproducing property:

$$\begin{aligned} |T_{\mathbb{P}} f| &= |\mathbb{E}f(X)| \leq \mathbb{E}|f(X)| = \mathbb{E}|\langle f, k(\cdot, X)\rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}} \mathbb{E}\langle k(\cdot, X), k(\cdot, X)\rangle_{\mathcal{H}}^{1/2} = \|f\|_{\mathcal{H}} \mathbb{E}\sqrt{k(X,X)} < \infty. \end{aligned}$$

Then, using the Riesz representation theorem applied to $T_{\mathbb{P}}$, there exists a $\mu_{\mathbb{P}} \in \mathcal{H}$ such that $T_{\mathbb{P}} f = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$. □

We can use this lemma to define the embedding:

**Definition 8.** *For a probability distribution* $\mathbb{P}$ *we define the* ***mean embedding*** *of* $\mathbb{P}$ *as an element* $\mu_{\mathbb{P}} \in \mathcal{H}$ *such that*

$$\mathbb{E}f(X) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}, \quad \text{for all } f \in \mathcal{H}.$$

This type of embeddings of probability measures in RKHS's have been widely studied. We can obtain more information about the embedding $\mu_{\mathbb{P}}$ using the reproducing property of $k$. If we set $f = \phi(t) = k(t, \cdot)$, we obtain that $\mu_{\mathbb{P}}(t) = \langle \mu_{\mathbb{P}}, k(t, \cdot)\rangle_{\mathcal{H}} = \mathbb{E}\ k(t, X)$. In other words:

**Remark 1.** *The mean embedding of* $\mathbb{P}$ *is the expectation under* $\mathbb{P}$ *of the canonical feature map.*

Now we express the MMD based on the unit ball $\mathcal{F} = \{ f \in \mathcal{H} : \ \|f\|_{\mathcal{H}} \leq 1 \}$ as a function of the mean embeddings in the RKHS $\mathcal{H}$.

**Lemma 4.** *Assume that the mean embeddings* $\mu_{\mathbb{P}}$ *and* $\mu_{\mathbb{Q}}$ *exist (e.g. under the assumptions of Lemma 3), then:*

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

*Proof.* We have only to apply the definition of this operator MMD and the previous property which says that $\mathbb{E}k(\cdot, X) = \mathbb{E}\phi(X) = \mu_{\mathbb{P}}$:

$$\begin{aligned} MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \{\mathbb{E}f(X) - \mathbb{E}f(Y)\} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \{\langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} - \langle f, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}\} \\ &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \end{aligned}$$

The last equality can be proved from the following inequalities.

($\leq$) This way can be deduced from the Cauchy-Schwarz inequality:

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \leq \sup_{\|f\|_{\mathcal{H}} \leq 1} |\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| \leq \sup_{\|f\|_{\mathcal{H}} \leq 1} \|f\|_{\mathcal{H}} \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \leq \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

($\geq$) Taking $f = \frac{\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}}{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|} \in \mathcal{H}$, with $\|f\|_{\mathcal{H}} = 1$:

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \geq \left\langle \frac{\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}}{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \right\rangle_{\mathcal{H}} = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

□

In practice it is common to use the square of the MMD because it has an unbiased estimator, as mentioned in [5], and also because it can be expressed in terms of expectations of the kernel. This link will help us to write estimators of this quantity and to connect this method with others that will be introduced later.

**Proposition 2.** *If the random variables* $X$ *and* $Y$ *are independent with distributions* $\mathbb{P}$ *and* $\mathbb{Q}$ *respectively, and* $X', Y'$ *are independent copies of* $X$ *and* $Y$*, the reproducing property of* $k$ *leads to:*

$$MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \mathbb{E}k(X, X') + \mathbb{E}k(Y, Y') - 2\mathbb{E}k(X, Y). \tag{3.2}$$

*Proof.* We start with the properties of the inner product, where $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$:

$$MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2$$

$$
\begin{aligned}
&= \left\| \mathbb{E}k(\cdot, X) - \mathbb{E}k(\cdot, Y) \right\|^2_{\mathcal{H}} \\
&= \left\langle \mathbb{E}k(\cdot, X) - \mathbb{E}k(\cdot, Y), \mathbb{E}k(\cdot, X') - \mathbb{E}k(\cdot, Y') \right\rangle_{\mathcal{H}} \\
&= \left\langle \mathbb{E}k(\cdot, X), \mathbb{E}k(\cdot, X') \right\rangle_{\mathcal{H}} + \left\langle \mathbb{E}k(\cdot, Y), \mathbb{E}k(\cdot, Y') \right\rangle_{\mathcal{H}} \\
&\quad -2\left\langle \mathbb{E}k(\cdot, X), \mathbb{E}k(\cdot, Y) \right\rangle_{\mathcal{H}}
\end{aligned}
$$

For the operator $T_{\mathbb{P}} f = \mathbb{E}f(X)$ defined in the proof of the lemma 3, we have proved that:

$$T_{\mathbb{P}} f = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} = \langle f, \mathbb{E}k(\cdot, X) \rangle_{\mathcal{H}}.$$

If we apply this property using $f = \mathbb{E}k(\cdot, Y)$ we have:

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

Using it in the previous development of MMD we obtain (3.2). □

The square of the MMD can also be expressed in integral form:

**Remark 2.** *The previous Equation* (3.2) *leads to:*

$$MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \int \int_{\mathcal{X}} k(x, y) \mathrm{d}(\mathbb{P} - \mathbb{Q})(x) \mathrm{d}(\mathbb{P} - \mathbb{Q})(y). \tag{3.3}$$

*Proof.* Writing the expectations in Proposition 2 as integrals:

$$
\begin{aligned}
MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) &= \mathbb{E}k(X, X') + \mathbb{E}k(Y, Y') - 2\mathbb{E}k(X, Y) \\
&= \int \int_{\mathcal{X}} k(x, x') \mathrm{d}\mathbb{P}(x) \mathrm{d}\mathbb{P}(x') + \int \int_{\mathcal{X}} k(y, y') \mathrm{d}\mathbb{Q}(y) \mathrm{d}\mathbb{Q}(y') \\
&\quad -2 \int \int_{\mathcal{X}} k(x, y) \mathrm{d}\mathbb{P}(x) \mathrm{d}\mathbb{Q}(y) \\
&= \int \int_{\mathcal{X}} k(x, y) \mathrm{d}(\mathbb{P} - \mathbb{Q})(x)(\mathbb{P} - \mathbb{Q})(y).
\end{aligned}
$$

□

These last two expressions of MMD are useful in later proofs. MMD has been defined in terms of reproducing kernel embeddings. However we still do not know whether it can be used to characterize equality of distributions or not. That is, whether the chosen family of functions $\mathcal{F}$ is sufficiently rich.

## 3.3 Characteristic kernels and RKHS's

Using the definitions and properties of MMD derived in the previous section, the next step is to determine whether the function $\gamma_k(\mathbb{P},\mathbb{Q}) = MMD(\mathcal{F},\mathbb{P},\mathbb{Q})$ is a metric. That is, whether $MMD(\mathcal{F},\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. By Lemma 4, this is equivalent to proving that the mean embedding is injective. This property depends on the particular RKHS considered. There are some spaces where it does not hold, as the next example, given in [6], shows.

**Example 1.** *A polynomial kernel of degree two cannot distinguish between all distributions. For example, using* $k(x,y) = (1+x^Ty)^2$, $x, y \in \mathbb{R}^d$, *we have to integrate* $k(x,y) = 1+x^Tyy^Tx+2x^Ty$ *in Equation (3.3). Let* $\gamma_k^2(\mathbb{P},\mathbb{Q}) = MMD^2(\mathcal{F},\mathbb{P},\mathbb{Q})$, *then:*

$$\begin{aligned}
\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}(1+x^Tyy^Tx+2x^Ty)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}(x^Tyy^Tx)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&\quad +2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}(x^Ty)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&= 0 + I_1 + 2I_2
\end{aligned}$$

*Let* $m_{\mathbb{P}}$ *be the mean of the distribution* $\mathbb{P}$ *and* $m_{\mathbb{Q}}$ *the mean of* $\mathbb{Q}$. *The covariance matrix for* $\mathbb{P}$ *is* $\Sigma_{\mathbb{P}} = \mathbb{E}[XX^T] - m_{\mathbb{P}}m_{\mathbb{P}}^T$ *which implies* $\mathbb{E}[XX^T] = \Sigma_{\mathbb{P}} + m_{\mathbb{P}}m_{\mathbb{P}}^T$. *The integrals are:*

$$\begin{aligned}
I_1 &= \int_{\mathbb{R}^d} x^T\left(\int_{\mathbb{R}^d}(yy^T)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y)\right)x\mathrm{d}(\mathbb{P}-\mathbb{Q})(x) \\
&= \int_{\mathbb{R}^d} x^T\left(\Sigma_{\mathbb{P}} + m_{\mathbb{P}}m_{\mathbb{P}}^T - \Sigma_{\mathbb{Q}} - m_{\mathbb{Q}}m_{\mathbb{Q}}^T\right)x\mathrm{d}(\mathbb{P}-\mathbb{Q})(x) \\
&= \left\|\Sigma_{\mathbb{P}} + m_{\mathbb{P}}m_{\mathbb{P}}^T - \Sigma_{\mathbb{Q}} - m_{\mathbb{Q}}m_{\mathbb{Q}}^T\right\|_F^2,
\end{aligned}$$

*where* $\|\cdot\|_F$ *is the Frobenius norm of a matrix:*

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|a_{i,j}|^2},$$

*being* $A = (a_{i,j})$ *a* $n \times m$ *matrix. For the second integral:*

$$\begin{aligned}
I_2 &= \int_{\mathbb{R}^d}(x^T)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)\int_{\mathbb{R}^d}(y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&= (m_{\mathbb{P}}-m_{\mathbb{Q}})^T\cdot(m_{\mathbb{P}}-m_{\mathbb{Q}}) \\
&= \left\|m_{\mathbb{P}}-m_{\mathbb{Q}}\right\|_2^2
\end{aligned}$$

*Combining these expressions we get:*

$$\gamma_k(\mathbb{P},\mathbb{Q}) = \left(\left\|\Sigma_{\mathbb{P}} - \Sigma_{\mathbb{Q}} + m_{\mathbb{P}}m_{\mathbb{P}}^T - m_{\mathbb{Q}}m_{\mathbb{Q}}^T\right\|_F^2 + 2\left\|m_{\mathbb{P}}-m_{\mathbb{Q}}\right\|_2^2\right)^{1/2}.$$

*In consequence $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if and only if $m_{\mathbb{P}} = m_{\mathbb{Q}}$ and $\Sigma_{\mathbb{P}} = \Sigma_{\mathbb{Q}}$. Therefore a test based on a embedding on this RKHS cannot distinguish between two distributions with the same mean and variance but different moments of order higher than two.*

The kernels that have the property of $\gamma_k$ being a metric received a special name.

**Definition 9.** *A reproducing kernel $k$ is a **characteristic kernel** if the induced $\gamma_k$ is a metric.*

**Definition 10.** *$\mathcal{H}$ is a **characteristic RKHS** if its reproducing kernel $k$ is characteristic.*

In [1] it is given a sufficient condition for an RKHS to be characteristic. This is not the most up-to-date condition, but it is quite simple to prove.

**Theorem 5.** *If $\mathcal{X}$ is a compact metric space, $k$ is continuous and $\mathcal{H}$ is dense in $\mathcal{C}(\mathcal{X})$ with respect to the supremum norm, then $\mathcal{H}$ is characteristic.*

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

($\Longrightarrow$) By Lemma 2, $\mathbb{P} = \mathbb{Q}$ if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y)$ for all $f \in \mathcal{C}(\mathcal{X})$, where $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$. So this will be the target of the proof. As $\mathcal{H}$ is dense in $\mathcal{C}(\mathcal{X})$ with respect to the supremum norm:

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

Now we will develop the difference between these expectations, using Jensen's inequality and the expression of the supremum norm

$$\begin{aligned}
|\mathbb{E}f(X) - \mathbb{E}f(Y)| &= |(\mathbb{E}f(X) - \mathbb{E}g(X)) + (\mathbb{E}g(X) - \mathbb{E}g(Y)) + (\mathbb{E}g(Y) - \mathbb{E}f(Y))| \\
&\leq |\mathbb{E}f(X) - \mathbb{E}g(X)| + |\mathbb{E}g(X) - \mathbb{E}g(Y)| + |\mathbb{E}g(Y) - \mathbb{E}f(Y)| \\
&= |\mathbb{E}[f(X) - g(X)]| + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + |\mathbb{E}[g(Y) - f(Y)]| \\
&\leq \mathbb{E}|f(X) - g(X)| + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + \mathbb{E}|g(Y) - f(Y)| \\
&\leq \|f - g\|_\infty + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + \|f - g\|_\infty \\
&\leq |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + 2\varepsilon.
\end{aligned}$$

By Lemma 4 we know that if $MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = 0$ then $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$. Hence:

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

for all $\varepsilon > 0$ and $f \in \mathcal{C}(\mathcal{X})$, which implies that the expectations are equal.

($\Longleftarrow$) It is clear from the definition of $MMD$. □

In practice it is difficult to verify the denseness condition, and the restriction of $\mathcal{X}$ being compact implies that $\gamma_k$ induces a metric only between probabilities with compact support. We will give a less restrictive condition (obtained from [6]), for which we need the following property of functions.

**Definition 11.** *If $\mathcal{X}$ is a topological space, a measurable and bounded kernel $k$ is said to be* ***integrally strictly positive definite*** *if*

$$\int \int_{\mathcal{X}} k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) > 0$$

*for all finite non-zero signed Borel measures $\mu$ defined on $\mathcal{X}$.*

Clearly, if a kernel $k$ is integrally strictly positive definite, then it is strictly positive. However the converse is not true. A sufficient condition for characteristic kernels is:

**Theorem 6.** *If $k$ is an integrally strictly positive definite kernel on $\mathcal{X}$, then $k$ is characteristic.*

To simplify the proof we will use the following lemma, whose proof is basically the one of the theorem. It establishes a necessary and sufficient condition for a kernel to ensure that its corresponding $\gamma_k$ is characteristic. Besides, this condition is closely related to the definition of integrally strictly positive function.

**Lemma 5.** *Let $k$ be measurable and bounded on $\mathcal{X}$, $\gamma_k$ is not a metric if and only if there exists a finite non-zero signed Borel measure $\mu$ that satisfies:*

1. $\mu(\mathcal{X}) = 0$.
2. $\int \int_{\mathcal{X}} k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) = 0$,

*Proof.* ($\Longrightarrow$) We have some $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$. We define $\mu = \mathbb{P} - \mathbb{Q}$, which is a finite non-zero signed Borel measure that satisfies $\mu(\mathcal{X}) = 0$ (Condition 1). It also satisfies the condition 2 since using Equation (3.3):

$$\begin{aligned} 0 = \gamma_k^2(\mathbb{P},\mathbb{Q}) &= \int \int_{\mathcal{X}} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)(\mathbb{P}-\mathbb{Q})(y) \\ &= \int \int_{\mathcal{X}} k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) \end{aligned}$$

($\Longleftarrow$) We have a Borel measure $\mu$ that satisfies the two conditions of the lemma. By the Jordan decomposition theorem for signed measures, there exist unique positive measures $\mu^+$ and $\mu^-$, mutually singular, such that $\mu = \mu^+ - \mu^-$. By Condition 1 we have that $\mu^+(\mathcal{X}) = \mu^-(\mathcal{X}) \equiv \alpha$. Now we will define two different probability distributions $\mathbb{P}$ and $\mathbb{Q}$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$. Let:

$$\mathbb{P} = \tfrac{\mu^+}{\alpha} \quad \text{and} \quad \mathbb{Q} = \tfrac{\mu^-}{\alpha}.$$

Clearly $\mathbb{P} \neq \mathbb{Q}$, as they are mutually singular, and $\mu = \alpha(\mathbb{P} - \mathbb{Q})$. By (3.3) and the Condition 2:

$$\begin{aligned} \gamma_k^2(\mathbb{P},\mathbb{Q}) &= \int \int_{\mathcal{X}} k(x,y)\mathrm{d}(\mathbb{P}-\mathbb{Q})(x)(\mathbb{P}-\mathbb{Q})(y) \\ &= \frac{1}{\alpha^2} \int \int_{\mathcal{X}} k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) \\ &= \frac{0}{\alpha^2} = 0 \end{aligned}$$

□

Now the proof of the theorem is almost direct.

*Proof.* (Theorem 6) By the definition of integrally strictly positive definite kernel we have that

$$\int \int_{\mathcal{X}} k(x,y)\mathrm{d}\nu(x)\mathrm{d}\nu(y) > 0,$$

for any finite non-zero signed Borel measure $\nu$. This means that there does not exist a finite non-zero signed Borel measure that satisfies the condition 2 in Lemma 5. Therefore, by the same Lemma 5, $\gamma_k$ is a metric and then $k$ is characteristic. □

This condition of Theorem 6 is clearly easier to check than the previous ones of Theorem 5, which involve denseness conditions and impose excessive restrictions. The definition of integrally strictly positive definite arises intuitively from the condition to ensure that a set of vectors of a finite dimensional vector space are linearly independent. It is possible to determine whether the vectors $\{x_1, \ldots, x_n\}$ are linearly independent by analyzing their Gram matrix. The Gram matrix is a $n \times n$ matrix, given by the dot product of the vectors, $G_{i,j} = \langle x_i, x_j \rangle$. This kind of matrix is always positive definite. If the vectors are independent, then it is strictly positive definite. We do not want to determine whether a set of vectors are linearly independent, but to see if two probability distributions are the same, and both checks involve analysing scalar products. Then the integrally positive type definition could be seen as a generalization of the finite definition of strictly positive definite matrix.

We already have the homogeneity test well defined, but we will analyse it a bit more in the next section, before going into the practical applications as two-sample test.

## 3.4 Another interpretation of MMD

The MMD can be also expressed in terms of the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$. This can be useful for independence tests, where characteristic functions are often used.

**Definition 12.** *Given a probability distribution $\mathbb{P}$ and a random variable $X \sim \mathbb{P}$, the **characteristic function** of $\mathbb{P}$ is:*

$$\Phi_{\mathbb{P}}(t) = \mathbb{E}[e^{itX}].$$

Before using these functions to rewrite the MMD discrepancies, we need to define the positive definite condition for a general function defined on $\mathbb{R}^d$:

**Definition 13.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is **positive definite** if $\forall n \geq 1$ and $\forall (x_1, \ldots, x_n) \in \mathbb{R}^d$ the matrix $A = (a_{ij})_{i,j=1}^n$, where $a_{ij} = f(x_i - x_j)$, is positive semi-definite, i.e.:*

$$x^\top A x \geq 0 \ , \ \forall x \in \mathbb{R}^d.$$

Proposition 3 shows how the MMD can be rewritten in terms of characteristic functions if the kernel is translation invariant.

**Proposition 3.** *Let $\mathcal{X} = \mathbb{R}^d$ and $k(x, y) = \psi(x-y)$, where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a bounded, continuous positive definite function. Then, for any $\mathbb{P}, \mathbb{Q}$,*

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \sqrt{\int_{\mathbb{R}^d} |\Phi_{\mathbb{P}}(w) - \Phi_{\mathbb{Q}}(w)|^2 \mathrm{d}\mu(w)} \equiv \|\Phi_{\mathbb{P}} - \Phi_{\mathbb{Q}}\|_{\mathcal{L}^2(\mathbb{R}^d, \mu)},$$

*where $\Phi_{\mathbb{P}}$ and $\Phi_{\mathbb{Q}}$ are the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$ respectively, and $\mu$ is a finite non-negative Borel measure on $\mathbb{R}^d$.*

We need Bochner's theorem to prove this result:

**Theorem 7.** ***(Bochner)*** *A continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure $\mu$ on $\mathbb{R}^d$:*

$$f(x) = \int_{\mathbb{R}^d} e^{-ix^T w} d\mu(w), x \in \mathbb{R}^d.$$

Proofs of this theorem can be found, for example, in [7] or [8].

*Proof.* (Proposition 3) We will calculate again $\gamma_k^2$ using the equation (3.3). First we will use Bochner's theorem for $\psi(\cdot)$. We will also use Fubini's theorem to exchange the order of the integrals.

$$\begin{aligned}
\gamma_k^2(\mathbb{P}, \mathbb{Q}) &= \int\int_{\mathbb{R}^d} \psi(x-y) \mathrm{d}(\mathbb{P}-\mathbb{Q})(x) \mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&= \int\int \left( \int_{\mathbb{R}^d} e^{-i(x-y)^T w} \mathrm{d}\mu(w) \right) \mathrm{d}(\mathbb{P}-\mathbb{Q})(x) \mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \\
&= \int_{\mathbb{R}^d} \left( \int e^{-ix^T w} \mathrm{d}(\mathbb{P}-\mathbb{Q})(x) \right) \left( \int e^{iy^T w} \mathrm{d}(\mathbb{P}-\mathbb{Q})(y) \right) \mathrm{d}\mu(w) \\
&= \int_{\mathbb{R}^d} (\Phi_{\mathbb{P}}(w) - \Phi_{\mathbb{Q}}(w)) \left( \overline{\Phi_{\mathbb{P}}(w)} - \overline{\Phi_{\mathbb{Q}}(w)} \right) \mathrm{d}\mu(w) \\
&= \int_{\mathbb{R}^d} |\Phi_{\mathbb{P}}(w) - \Phi_{\mathbb{Q}}(w)|^2 \mathrm{d}\mu(w)
\end{aligned}$$

□

This property shows that $\gamma_k$ is a weighted $\mathcal{L}^2$-distance between the characteristic functions $\Phi_{\mathbb{P}}$ and $\Phi_{\mathbb{Q}}$. The weights are given by the Fourier transform of $\psi$. This property will help us to relate MMD, and its corresponding independence test, with other tests based on weighted distances between characteristic functions.

# Chapter 4

# Application to two-sample test

We have found discrepancies to determine whether two probability distributions are equal. To apply them to real problems we must be able to estimate them with finite samples. Then we need sample estimators of the MMD statistic. To determine whether the differences observed in the MMD are statistically significant, we also need to analyze the asymptotic behaviour of these estimators under the null hypothesis. This chapter uses material from [1], [6] and [9].

## 4.1 Some MMD estimators

A direct estimator for the general expression of the MMD (3.1) in a general function class $\mathcal{F}$ can be obtained by replacing the population expectations with the sample means. Given observations $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_m\}$ i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively:

$$MMD_b(\mathcal{F}, x, y) = \sup_{f \in \mathcal{F}} \left( \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \frac{1}{m} \sum_{i=1}^{m} f(y_i) \right).$$

However if $\mathcal{F}$ belongs to a specific RKHS, the squared MMD can be computed using Equation (3.2) of Proposition 2:

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

where $X$ and $X'$ are independent random variables with distribution $\mathbb{P}$, and $Y$ and $Y'$ independent random variables with distribution $\mathbb{Q}$. An unbiased statistic can be obtained from this expression using two U-statistics and a sample mean for the last term:

$$\begin{aligned} MMD_u^2(\mathcal{F}, x, y) \quad = \quad & \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(y_i, y_j) \\ & - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j). \end{aligned} \tag{4.1}$$

If the two samples have the same size, $m = n$, they can be put together in one sample

$z = \{z_1, \ldots, z_n\}$ from the random variable $Z = (X, Y) \sim \mathbb{P} \times \mathbb{Q}$. Then if we define:

$$h(z_i, z_j) \equiv k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i).$$

We can build an unbiased estimate of the squared MMD, simpler than (4.1).

$$MMD_u^2(\mathcal{F}, x, y) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} h(z_i, z_j).$$

This estimator is actually a U-statistic. Its value can be negative, since we have removed the terms $h(z_i, z_i)$ to avoid artificial correlations between observations. By [10] (Section 5.1.4), we know that Equation (4.1) corresponds to the minimum variance estimator for samples of the same size, $n = m$. However it is easy to see that the previous estimator is almost identical to the minimum variance one. The only difference is that the cross-terms $k(x_i, y_i)$ are present in the minimum variance estimator:

$$\begin{aligned} MMD_u^2(\mathcal{F}, x, y) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)) \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j) \\ &\quad - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, y_j). \end{aligned}$$

This estimator $MMD_u^2$ is the most commonly used in practice. We will determine its asymptotic behaviour in the following section.

## 4.2 Asymptotic distribution of MMD

Finally we will analyse asymptotic behaviour of the unbiased statistic $MMD_u^2(\mathcal{F}, x, y)$. We will use a new kernel $\widetilde{k}$ between feature space mappings from which the mean embeddings of $\mathbb{P}$ and $\mathbb{Q}$ has been subtracted:

**Definition 14.** *The **centered kernel** is defined, for $x, y \in \mathcal{X}$, as:*

$$\tilde{k}(x, y) \equiv \langle \phi(x) - \mu_{\mathbb{P}},\ \phi(y) - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}.$$

The centered kernel can be written in terms of the original kernel $k$, using the reproducing property of $k$ and recalling that the feature map, $\phi : \mathcal{X} \to \mathcal{H}$, can be written has $\phi(x) = k(x, \cdot)$. Given two random variables $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$:

$$\begin{aligned} \tilde{k}(x, y) &= \langle \phi(x) - \mu_{\mathbb{P}},\ \phi(y) - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} - \langle \mu_{\mathbb{P}}, \phi(y) \rangle_{\mathcal{H}} - \langle \phi(x), \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \end{aligned}$$

$$
\begin{aligned}
&= k(x,y) - \langle \mathbb{E}\phi(X), \phi(y)\rangle_{\mathcal{H}} - \langle \phi(x), \mathbb{E}\phi(Y)\rangle_{\mathcal{H}} + \langle \mathbb{E}\phi(X), \mathbb{E}\phi(Y)\rangle_{\mathcal{H}} \\
&= k(x,y) - \mathbb{E}k(X,y) - \mathbb{E}k(x,Y) + \mathbb{E}k(X,Y). \qquad (4.2)
\end{aligned}
$$

It is easy to see from this equation that if $k$ is bounded, then the centered kernel is square integrable, i.e. $\tilde{k} \in \mathcal{L}_2(\mathcal{X}\times\mathcal{X}, \mathbb{P}\times\mathbb{Q})$. Now we can obtain the asymptotic distribution of the statistic under the null hypothesis $H_0$ which assumes that $\mathbb{P} = \mathbb{Q}$. We will assume two conditions on the convergence of the samples sizes:

$$
\lim_{n,m\to\infty} \frac{n}{n+m} = \rho_x \qquad \lim_{n,m\to\infty} \frac{m}{n+m} = \rho_y = 1 - \rho_x,
$$

for some $\rho_x \in (0,1)$.

**Theorem 8.** *Under $H_0$ the statistics $MMD_u^2$ converges in distribution according to:*

$$
(m+n)MMD_u^2(\mathcal{F},x,y) \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i \left[ \left( \rho_x^{-1/2} Z_i - \rho_y^{-1/2} \widetilde{Z}_i \right)^2 - \frac{1}{\rho_x \rho_y} \right],
$$

*where $Z_i, \widetilde{Z}_i \sim N(0,1)$ are all independent and $\lambda_i$ are the eigenvalues of the centered kernel integral operator defined by $\widetilde{k}$, that is,*

$$
\int_{\mathcal{X}} \widetilde{k}(x,y)\psi_i(x)\mathrm{d}\mathbb{P}(x) = \lambda_i \psi_i(y), \qquad \text{for } i = 1,2,\ldots,
$$

*being $\psi_i$ the corresponding eigenfunctions.*

*Proof.* The null hypothesis means that $X, Y \sim \mathbb{P}$. First we write the expression of the unbiased estimator $MMD_u^2$ in terms of the centered kernel, using Equation (4.2), where now $Z, Z' \sim \mathbb{P}$:

$$
\begin{aligned}
MMD_u^2(\mathcal{F},x,y) &= \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i}^{n} k(x_i,x_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m}\sum_{j\neq i}^{m} k(y_i,y_j) \\
&\quad - \frac{2}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i,y_j) \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i}^{n} \left( \widetilde{k}(x_i,x_j) + \mathbb{E}_Z k(x,Z) + \mathbb{E}_Z k(Z,y) - \mathbb{E}_{Z,Z'} k(Z,Z') \right) + \\
&\quad \frac{1}{m(m-1)} \sum_{i=1}^{m}\sum_{j\neq i}^{m} \left( \widetilde{k}(y_i,y_j) + \mathbb{E}_Z k(x,Z) + \mathbb{E}_Z k(Z,y) - \mathbb{E}_{Z,Z'} k(Z,Z') \right) - \\
&\quad \frac{2}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} \left( \widetilde{k}(x_i,y_j) + \mathbb{E}_Z k(x,Z) + \mathbb{E}_Z k(Z,y) - \mathbb{E}_{Z,Z'} k(Z,Z') \right) \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i}^{n} \widetilde{k}(x_i,x_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m}\sum_{j\neq i}^{m} \widetilde{k}(y_i,y_j) \\
&\quad - \frac{2}{nm} \sum_{i=1}^{n}\sum_{j=1}^{m} \widetilde{k}(x_i,y_j). \qquad (4.3)
\end{aligned}
$$

We define now the operator $D_{\widetilde{k}} : \mathcal{L}_2(\mathbb{P}) \to \mathcal{F}$ satisfying:

$$D_{\widetilde{k}} g(x) \equiv \int_{\mathcal{X}} \widetilde{k}(x, x') g(x') \mathrm{d}\mathbb{P}(x').$$

According to [11] (Theorem VI.23), this operator is compact if and only if $\widetilde{k}$ is square integrable under $\mathbb{P}$. And it is well known that the eigenfunctions of a compact operator form an orthonormal basis of the space. Thus we can write the centered kernel as a function of the eigenfunctions $\psi_i$:

$$\widetilde{k}(x, y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y),$$

where the convergence of the sum is in $\mathcal{L}_2(\mathcal{X}\times\mathcal{X}, \mathbb{P}\times\mathbb{P})$. It is easy to prove that the eigenfunctions have zero mean and are uncorrelated, using that the U-statistics of Equation (4.3) in $\widetilde{k}(x_i, x_j)$ are degenerate $\Big($i.e. $\mathbb{E}_X \widetilde{k}(X, y) = 0\Big)$. Now we can compute the asymptotic distribution of each sum in equation (4.3):

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \widetilde{k}(x_i, x_j) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k=1}^{\infty} \lambda_k \psi_k(x_i)\psi_k(x_j) \\
&= \frac{1}{n}\sum_{k=1}^{\infty} \lambda_k \left[\left(\sum_{i=1}^{n} \psi_k(x_i)\right)^2 - \sum_{i=1}^{n} \psi_k^2(x_i)\right] \\
&\xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1),
\end{aligned}$$

where $Z_k \sim N(0, 1)$ are i.i.d., and the convergence is in distribution, which can be proved using that $\Psi_i \perp \Psi_j$ and $\mathbb{E}_X \Psi_i(X) = 0$. In addition to this, by [11] (Theorem VI.22) we know that $\sum \lambda_i^2 < \infty$, and then it can be shown that the sum $\sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1)$ converges almost surely (For example, via Kolmogorov's inequality: $\mathbb{P}\left(\max S_k \geq \mu\right) \leq \frac{1}{\mu^2} Var(S_n)$, for $1 \leq k \leq n$). It can be done also for the other two terms of the equation:

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j\neq i}^{m} \widetilde{k}(y_i, y_j) \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (\widetilde{Z}_k^2 - 1),$$

$$\frac{1}{\sqrt{nm}}\sum_{i=1}^{n}\sum_{j=1}^{m} \widetilde{k}(x_i, y_j) \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k Z_k \widetilde{Z}_k,$$

where $\widetilde{Z}_k \sim N(0, 1)$ independent of the $Z_k$. It only remains to multiply the statistic by $(n + m)$ and to substitute the expressions for $\rho_x$ and $\rho_y$. Full details of this proof can be seen in [1] (Appendix B). □

From the theorem it is clear that, if $n = m$, then the statistics converges according to:

$$nMMD_u^2(\mathcal{F}, x, y) \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i \left[\left(Z_i - \widetilde{Z}_i\right)^2 - 2\right] \overset{d}{=} 2\sum_{i=1}^{\infty} \lambda_i (X_i - 1),$$

where $X_1, X_2, \ldots \equiv \chi^2_1$, since $Z_i - \widetilde{Z}_i \sim N(0,2)$.

A natural question is why it is needed to center the kernel. Following section 5.5 of [10], the variances of any of the three statistics that appear in Equation (4.1) are:

$$\zeta_{1_x} = Var_X(\mathbb{E}_{X'}k(X,X')) = Var_X(\mu_{\mathbb{P}}(X))$$

$$\zeta_{1_y} = Var_Y(\mathbb{E}_{Y'}k(Y,Y')) = Var_Y(\mu_{\mathbb{Q}}(Y))$$

$$\zeta_{11} = Var_X(\mathbb{E}_Y k(X,Y)) = Var_X(\mu_{\mathbb{Q}}(X)) \qquad \zeta_{12} = Var_Y(\mathbb{E}_X k(X,Y)) = Var_Y(\mu_{\mathbb{P}}(Y)),$$

since the last term is a U-statistic of two samples, and the other two are only of one sample. All these variances are strictly positive, which means that any of the three statistics converges to a Normal distribution. But the third distribution depends on the previous two, so we can not ensure that the resulting distribution is also normal. Therefore we should use a variation of the original kernel $k$ such that:

$$Var_X(\mathbb{E}_{X'}\widetilde{k}(X,X')) = 0.$$

Thus we can apply the degenerate case of [10], to avoid the sum of normal distributions. One possibility is to force that $\mathbb{E}_{X'}\widetilde{k}(\cdot,X') = 0$, using for example:

$$\mathbb{E}_{X'}\widetilde{k}(\cdot,X') = \mathbb{E}_{X'}k(\cdot,X') - \mu_{\mathbb{P}}(\cdot) = 0.$$

The expression of the centered kernel can be obtained from this one using some simple tricks.

In addition, we can see the MMD density under both the null and alternative hypotheses by approximating it empirically. The left-hand side of Figure 4.1 shows the empirical distribution under $H_0$, with $\mathbb{P}$ and $\mathbb{Q}$ Gaussians with unit standard deviation, obtained by using 100 samples from each. We see that in fact it is a mixture of $\chi^2$ distributions, as we have proved before. The right-hand side shows the empirical distribution under the alternative hypothesis $H_1$. There, $\mathbb{P}$ is a Gaussian distribution with unit standard deviation, and $\mathbb{Q}$ is another Gaussian distribution with standard deviation 5, using 100 samples from each. In both cases, the histograms have been obtained using 2000 independent instances of the MMD to compute them.

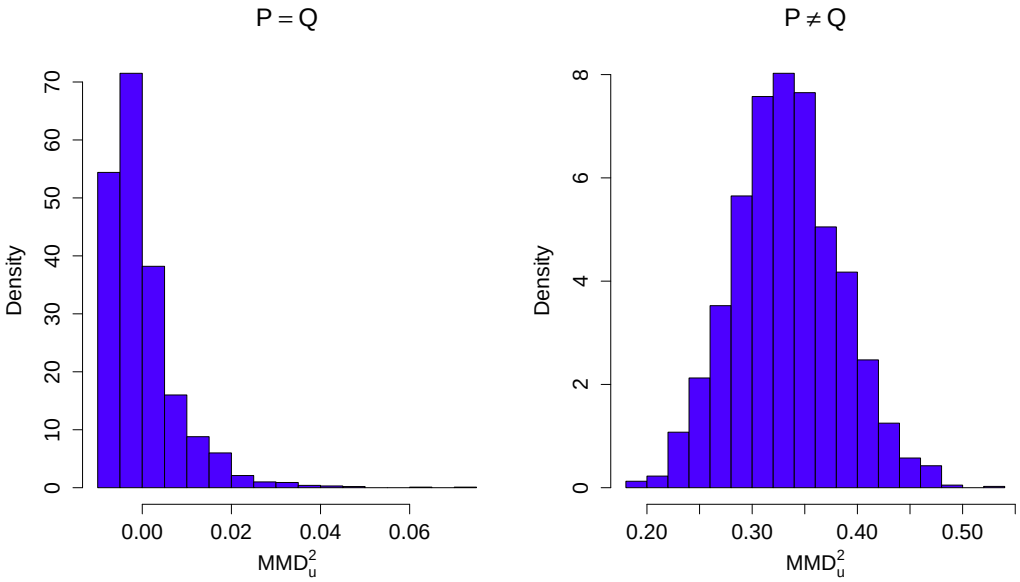



Figure 4.1: Empirical density of $MMD_u^2$ under $H_0$ and $H_1$.

# Chapter 5

# Application to independence test

Consider the random variables $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$, whose joint distribution is $\mathbb{P}_{XY}$. One way to test the independence between these variables is to determine whether the joint probability measure $\mathbb{P}_{XY}$ is equal to the product of the marginals $\mathbb{PQ}$. In the previous chapters we have seen that the MMD between two distributions is equal to zero only when the two distributions are equal. Therefore:

$$MMD(\mathcal{F}, \mathbb{P}_{XY}, \mathbb{PQ}) = 0 \text{ if and only if } X \text{ and } Y \text{ are independent.}$$

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 $\nu((x,y),(x',y'))$ a kernel on the product space $\mathcal{X} \times \mathcal{Y}$ with RKHS $\mathcal{H}_\nu$. This space is known as the tensor product space $\mathcal{H} \times \mathcal{G}$. Tensor product spaces are defined as follows:

**Definition 15.** *The **tensor product** of Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ is defined as the completion of the space $\mathcal{H}_1 \times \mathcal{H}_2$ with inner product $\langle \cdot, \cdot \rangle_1 \langle \cdot, \cdot \rangle_2$, extended by linearity. The resulting space is also a Hilbert space.*

In the following lemma we give a particular expression for the kernel of the product space:

**Lemma 6.** *A kernel $\nu$ in the tensor product space $\mathcal{H} \times \mathcal{G}$ can be defined as:*

$$\nu((x,y),(x',y')) = k(x,x')l(y,y').$$

*Proof.* This proof is taken from [4] (Lemma 4.6, page 114). To show that $k(x,x')l(y,y')$ is a kernel on the product space, we need to express it as the dot product between feature maps. We denote the feature maps to $\mathcal{H}$ and $\mathcal{G}$ by $\phi$ and $\psi$ respectively.

$$\begin{aligned} \nu((x,y),(x',y')) &= k(x,x')l(y,y') \\ &= \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \langle \psi(y), \psi(y') \rangle_{\mathcal{G}} \end{aligned}$$

$$= \langle(\phi(x), \psi(y)), (\phi(x'), \psi(y'))\rangle_{\mathcal{H}_\nu},$$

which shows that $(\phi, \psi) : \mathcal{X} \times \mathcal{Y} \to \mathcal{H} \otimes \mathcal{G}$ is in fact a feature map of $k(x, x')l(y, y')$. □

Now we need to define the mean embeddings of these two probability measures. For clarity's sake we will introduce new notation for the expectations with respect to the marginals and the joint distribution.

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

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

$$\begin{aligned}
\mu_{\mathbb{P}_{XY}} &= \mathbb{E}_{XY} \nu((X, Y), \cdot) \\
\mu_{\mathbb{PQ}} &= \mathbb{E}_X \mathbb{E}_Y \nu((X, Y), \cdot).
\end{aligned}$$

In terms of these embeddings:

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

All the estimators of the previous chapter can be applied, but now the kernel is $\nu$, the one corresponding to the product space. In the following section we will give an alternative formulation of this test in terms of generalized covariances between distances.

## 5.1 Hilbert Schmidt Independence Criterion (HSIC)

In this section, obtained mainly from [12], we will introduce an alternative formulation of the MMD independence test, based on the simplest criterion for testing linear independence, the covariance. To this end we need to introduce the cross-covariance operators for elements of general Hilbert spaces. First, let us define a new product between elements of two separable Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$.

**Definition 16.** *Let $h \in \mathcal{H}, g \in \mathcal{G}$. The **tensor product operator** $h \otimes g : \mathcal{G} \to \mathcal{H}$ is defined as:*

$$(h \otimes g)(f) = \langle g, f\rangle_{\mathcal{G}} h \text{ , for all } f \in \mathcal{G}.$$

We will see two important properties of this product, which will be useful for future proofs. First we will introduce a norm for linear operators.

**Definition 17.** *Let $C : \mathcal{G} \to \mathcal{H}$ be a linear operator between the RKHS's $\mathcal{G}$ and $\mathcal{H}$. The **Hilbert-Schmidt norm (HS)** of $C$ is defined as*

$$\|C\|_{\mathcal{HS}} = \sqrt{\sum_{i,j} \langle Cv_j, u_i \rangle^2_{\mathcal{H}}},$$

*provided that the sum converges, where $\{u_i\}$ and $\{v_j\}$ are orthonormal bases of $\mathcal{H}$ and $\mathcal{G}$ respectively.*

In terms of this norm, we define a class of operators.

**Definition 18.** *A linear operator $C : \mathcal{G} \to \mathcal{H}$ is called a **Hilbert-Schmidt operator** if its HS norm exists.*

In fact, the set of Hilbert-Schmidt operators, $\mathcal{HS}$, is a separable Hilbert space with inner product

$$\langle C, D \rangle_{\mathcal{HS}} = \sum_{i,j} \langle Cv_j, u_i \rangle_{\mathcal{H}} \langle Dv_j, u_i \rangle_{\mathcal{H}}.$$

The tensor product operator belongs to this Hilbert space, because its HS norm is finite. This can be deduced from the first point of the following proposition.

**Proposition 4.** *The tensor product between $f \in \mathcal{H}, g \in \mathcal{G}$ satisfies:*

1. *The HS norm of $f \otimes g$ is:*
$$\|f \otimes g\|_{\mathcal{HS}} = \|f\|_{\mathcal{H}} \|g\|_{\mathcal{G}}. \tag{5.1}$$

2. *The tensor product satisfies the distributive property,*

$$(f + h) \otimes (g + k) = f \otimes g + f \otimes k + h \otimes g + h \otimes k,$$

*for all $h \in \mathcal{H}$ and $k \in \mathcal{G}$.*

*Proof.* 1. We use that the representations of each of these functions in their corresponding basis are $f = \sum_i \langle f, u_i \rangle_{\mathcal{H}} u_i$ and $g = \sum_j \langle g, v_j \rangle_{\mathcal{G}} v_j$. We also use Parseval Equality, which says that $\|f\|^2_{\mathcal{H}} = \sum_i |\langle f, u_i \rangle_{\mathcal{H}}|^2$, and respectively for $g$.

$$\begin{aligned}
\|f \otimes g\|^2_{\mathcal{HS}} &= \sum_{i,j} \langle (f \otimes g) v_j, u_i \rangle^2_{\mathcal{H}} \\
&= \sum_{i,j} \langle \langle g, v_j \rangle_{\mathcal{G}} f, u_i \rangle^2_{\mathcal{H}} \\
&= \sum_{i,j} \langle f, u_i \rangle^2_{\mathcal{H}} \langle g, v_j \rangle^2_{\mathcal{G}} \\
&= \sum_i |\langle f, u_i \rangle_{\mathcal{H}}|^2 \sum_j |\langle g, v_j \rangle_{\mathcal{G}}|^2
\end{aligned}$$

$$= \|f\|_{\mathcal{H}}^2 \|g\|_{\mathcal{G}}^2.$$

Since these last two norms are finite, the HS norm of $f \otimes g$ is also finite. Hence the operator $f \otimes g$ is a Hilbert-Schmidt operator on $\mathcal{G}$.

2. The proof is direct applying the definition:

$$\begin{aligned}((f+h)\otimes(g+k))(\cdot) &= \langle g+k,\cdot\rangle_{\mathcal{G}}(f+h)\\ &= (\langle g,\cdot\rangle_{\mathcal{G}} + \langle k,\cdot\rangle_{\mathcal{G}})(f+h)\\ &= \langle g,\cdot\rangle_{\mathcal{G}} f + \langle k,\cdot\rangle_{\mathcal{G}} f + \langle g,\cdot\rangle_{\mathcal{G}} h + \langle k,\cdot\rangle_{\mathcal{G}} h\\ &= f\otimes g + f\otimes k + h\otimes g + h\otimes k.\end{aligned}$$

□

Let $\phi$ denote the feature map from $\mathcal{X}$ to the RKHS $\mathcal{H}$. Similarly let $\psi$ the feature map from $\mathcal{Y}$ to the RKHS $\mathcal{G}$.

**Definition 19.** *The **cross-covariance operator** associated with $\mathbb{P}_{XY}$ is the linear operator $C_{XY} : \mathcal{G} \to \mathcal{H}$ defined as:*

$$C_{XY} = \mathbb{E}_{XY}\left[(\phi(X) - \mu_{\mathbb{P}}) \otimes (\psi(Y) - \mu_{\mathbb{Q}})\right]. \tag{5.2}$$

Using the distributive property of the tensor product, Equation (5.2) can be also expressed as:

$$C_{XY} = \mathbb{E}_{XY}\left[(\phi(X) - \mu_{\mathbb{P}}) \otimes (\psi(Y) - \mu_{\mathbb{Q}})\right] = \mathbb{E}_{XY}\left[\phi(X)\otimes\psi(Y)\right] - \mu_{\mathbb{P}}\otimes\mu_{\mathbb{Q}}. \tag{5.3}$$

Now we define an independence criterion based on the Hilbert-Schmidt norm of this covariance operator.

**Definition 20.** *We define the **Hilbert-Schmidt Independence Criterion (HSIC)** for $\mathbb{P}_{XY}$ as the squared HS norm of the associated cross-covariance operator:*

$$HSIC(\mathbb{P}_{XY}, \mathcal{H}, \mathcal{G}) = \|C_{XY}\|_{\mathcal{HS}}^2. \tag{5.4}$$

This quantity characterizes independence when the kernels of the RKHS's are characteristics, that is, when the embeddings of probability measures in these spaces are injective.

**Proposition 5.** *Given two random variables $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$, with joint distribution $\mathbb{P}_{XY}$, and two RKHS's $\mathcal{H}$ and $\mathcal{G}$ with characteristic kernels $k$ and $l$, then $HSIC(\mathbb{P}_{XY}, \mathcal{H}, \mathcal{G}) = 0$ if and only if $\mathbb{P}_{XY} = \mathbb{P}\mathbb{Q}$, i.e. if $X$ and $Y$ are independent.*

The proof of this proposition is deferred to Section 8.1. This derivation uses a generalization of kernel embeddings for signed measures.

In addition to this, we can express HSIC in terms of kernels $k$ and $l$, from $\mathcal{H}$ and $\mathcal{G}$ respectively. This helps us to formulate estimators of this independence criterion.

**Lemma 7.** *If we denote $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$ then:*

$$\begin{aligned} HSIC(\mathbb{P}_{XY}, \mathcal{H}, \mathcal{G}) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k(X,X')l(Y,Y') + \mathbb{E}_X\mathbb{E}_Y\mathbb{E}_{X'}\mathbb{E}_{Y'}k(X,X')l(Y,Y') \\ &\quad -2\mathbb{E}_{XY}\mathbb{E}_{X'}\mathbb{E}_{Y'}k(X,X')l(Y,Y'). \end{aligned} \tag{5.5}$$

*Proof.* We will use the expression for the HS-norm of the tensor product (5.1) and the previous expression obtained for the cross-covariance operator (5.3). First, we will simplify the notation of $C_{XY}$:

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

Using this notation:

$$\begin{aligned} \|C_{XY}\|^2_{\mathcal{HS}} &= \langle \widetilde{C}_{XY} - M_{XY} \ , \ \widetilde{C}_{X'Y'} - M_{X'Y'} \rangle_{\mathcal{HS}} \\ &= \langle \widetilde{C}_{XY} \ , \ \widetilde{C}_{X'Y'} \rangle_{\mathcal{HS}} + \langle M_{XY} \ , \ M_{X'Y'} \rangle_{\mathcal{HS}} - 2\langle \widetilde{C}_{XY} \ , \ M_{X'Y'} \rangle_{\mathcal{HS}} \end{aligned}$$

We will calculate these products one by one.

$$\begin{aligned} \langle \widetilde{C}_{XY} \ , \ \widetilde{C}_{X'Y'} \rangle_{\mathcal{HS}} &= \langle \mathbb{E}_{XY}\left[\phi(X) \otimes \psi(Y)\right] \ , \ \mathbb{E}_{X'Y'}\left[\phi(X') \otimes \psi(Y')\right] \rangle_{\mathcal{HS}} \\ &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|\phi(X) \otimes \psi(Y)\|^2_{\mathcal{HS}} \\ &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|\phi(X)\|^2_{\mathcal{H}}\|\psi(Y)\|^2_{\mathcal{G}} \\ &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\langle \phi(X), \phi(X') \rangle_{\mathcal{H}} \langle \psi(Y), \psi(Y') \rangle_{\mathcal{G}} \\ &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k(X,X')l(Y,Y'). \end{aligned}$$

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

$$\begin{aligned} \langle \widetilde{C}_{XY} \ , \ M_{X'Y'} \rangle_{\mathcal{HS}} &= \langle \mathbb{E}_{XY}\left[\phi(X) \otimes \psi(Y)\right] \ , \ \mu_{\mathbb{P}} \otimes \mu_{\mathbb{Q}} \rangle_{\mathcal{HS}} \\ &= \langle \mathbb{E}_{XY}\left[\phi(X) \otimes \psi(Y)\right] \ , \ \mathbb{E}_{X'}\phi(X') \otimes \mathbb{E}_{Y'}\psi(Y') \rangle_{\mathcal{HS}} \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{XY}\mathbb{E}_{X'}\mathbb{E}_{Y'}\langle\phi(X)\otimes\psi(Y),\phi(X')\otimes\psi(Y')\rangle_{\mathcal{HS}} \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'}\mathbb{E}_{Y'}\langle\phi(X),\phi(X')\rangle_{\mathcal{H}}\langle\psi(Y),\psi(Y')\rangle_{\mathcal{G}} \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'}\mathbb{E}_{Y'}k(X,X')l(Y,Y').
\end{aligned}
$$

Expression (5.5) can be derived putting together these partial results. □

Now we include one of the biased estimators of the HSIC value, based on this lemma, which has an easy formulation. Given two samples $x$ and $y$ of the same length $n$:

$$HSIC_b = \frac{1}{n^2}\text{trace}(KHLH),$$

where $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$, being $k$ and $l$ the kernels of $\mathcal{H}$ and $\mathcal{G}$ respectively. Also $H = I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$, where $\mathbf{1}$ is an $n \times 1$ vector of ones. This estimator is obtained by substituting V-statistics in the previous HSIC formula and operating with the resulting expression. Other estimators and their asymptotic behaviour can be consulted in [13].

## 5.2 Equivalence between HSIC and MMD

In this section we prove the equivalence of the HSIC test in terms of the HS norm of the cross covariance operator and in terms of the MMD between $\mathbb{P}_{XY}$ and $\mathbb{PQ}$.

**Theorem 9.** *The squared MMD quantity between the joint distribution $\mathbb{P}_{XY}$ and the product of its marginals is equal to the Hilbert-Schmidt norm of the covariance operator between RKHSs:*

$$MMD^2(\mathcal{F}, \mathbb{P}_{XY}, \mathbb{PQ}) = HSIC(\mathbb{P}_{XY}, \mathcal{H}, \mathcal{G}).$$

*Proof.* This proof has been obtained using [9] and [14]. We will use the expression of the MMD given in Lemma 7 and properties of the tensor product kernel $\nu$ given in Lemma 6:

$$
\begin{aligned}
MMD^2(\mathcal{F}, \mathbb{P}_{XY}, \mathbb{PQ}) &= \|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{PQ}}\|^2_{\mathcal{H}_\nu} \\
&= \|\mathbb{E}_{XY}\nu((X,Y),\cdot) - \mathbb{E}_X\mathbb{E}_Y\nu((X,Y),\cdot)\|^2_{\mathcal{H}_\nu} \\
&= \|\mathbb{E}_{XY}k(X,\cdot)l(Y,\cdot) - \mathbb{E}_X\mathbb{E}_Yk(X,\cdot)l(Y,\cdot)\|^2_{\mathcal{H}_{\mathcal{H}\otimes\mathcal{G}}} \\
&= \|\mathbb{E}_{XY}k(X,\cdot)l(Y,\cdot) - \mathbb{E}_Xk(X,\cdot)\mathbb{E}_Yl(Y,\cdot)\|^2_{\mathcal{H}_{\mathcal{H}\otimes\mathcal{G}}} \\
&= \langle\mathbb{E}_{XY}k(X,\cdot)l(Y,\cdot)\ ,\ \mathbb{E}_{X'Y'}k(X',\cdot)l(Y',\cdot)\rangle_{\mathcal{H}_{\mathcal{H}\otimes\mathcal{G}}} \\
&\quad +\langle\mathbb{E}_Xk(X,\cdot)\mathbb{E}_Yl(Y,\cdot)\ ,\ \mathbb{E}_{X'}k(X',\cdot)\mathbb{E}_{Y'}l(Y',\cdot)\rangle_{\mathcal{H}_{\mathcal{H}\otimes\mathcal{G}}} \\
&\quad -2\langle\mathbb{E}_{XY}k(X,\cdot)l(Y,\cdot)\ ,\ \mathbb{E}_{X'}k(X',\cdot)\mathbb{E}_{Y'}l(Y',\cdot)\rangle_{\mathcal{H}_{\mathcal{H}\otimes\mathcal{G}}} \\
&= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}k(X,X')l(Y,Y') + \mathbb{E}_X\mathbb{E}_Y\mathbb{E}_{X'}\mathbb{E}_{Y'}k(X,X')l(Y,Y')
\end{aligned}
$$

$$-2\mathbb{E}_{XY}\mathbb{E}_{X'}\mathbb{E}_{Y'}k(X,X')l(Y,Y').$$

This is the expression for HSIC obtained in Lemma 7. □

That is, both test are equivalent, although their formulation and estimators are different. In the next two chapters we will introduce a different type of homogeneity and independence tests, using a different approach. In Chapter 8 we will show how this new type of tests are related to tests based on embeddings in RKHS's.

# Chapter 6

# Energy distance

In this chapter we describe a homogeneity test, introduced in [15]. This test will be used in the next chapter to formulate another independence test. This test is one of the most popular nowadays, because of its power and the fact that it does not depend on any parameter. In Chapter 8 we will show that this method is actually a particular case of the one introduced in the previous chapters, the MMD, for a special choice of the kernel.

In the first part of this chapter, we derive the basic homogeneity test, the energy distance. We need to introduce some theory to prove that it is a homogeneity test, i.e., that it vanishes if and only if both distributions are equal. In the second part of this chapter we generalize the energy distance to cover the cases in which the first moments are not finite.

## 6.1 Definitions and principal properties

One of the simplest distances we can define between two distributions $F$ and $G$ is the $\mathcal{L}_2$ one, although it has the drawback that the distribution of its natural estimate is not distribution-free. That is, the distribution of the estimate depends on the distribution $F$ under the null hypothesis. However, we can extend this distance easily to higher dimensions, having the property of being rotation invariant. Then energy distances can be derived as a variation of the $\mathcal{L}_2$ distance, given by the following proposition:

**Proposition 6.** *Let $F$ and $G$ be two CDFs of the independent random variables $X$ and $Y$ respectively, and $X', Y'$ two iid copies of them, then:*

$$2\int_{-\infty}^{\infty}(F(x)-G(x))^2\mathrm{d}x = 2\mathbb{E}|X-Y| - \mathbb{E}|X-X'| - \mathbb{E}|Y-Y'|. \tag{6.1}$$

*Proof.* The idea of this proof has been obtained from [16]. We will start analysing the expectations of the right hand side. We will use that for any positive random variable $Z > 0$, $\mathbb{E}Z = \int_0^\infty \mathbb{P}(Z > z)\mathrm{d}z$. We can apply this fact to $|X-Y|$, and then use Fubini's theorem:

$$\mathbb{E}|X-Y| \quad = \quad \int_0^\infty \mathbb{P}(|X-Y| > u)\mathrm{d}u$$

$$
\begin{aligned}
&= \int_0^\infty \mathbb{P}(X-Y>u)\mathrm{d}u + \int_0^\infty \mathbb{P}(X-Y<-u)\mathrm{d}u \\
&= \int_0^\infty \int_{-\infty}^\infty \mathbb{P}(X-Y>u|Y=y)\mathrm{d}G(y)\mathrm{d}u + \int_0^\infty \int_{-\infty}^\infty \mathbb{P}(X-Y<-u|X=x)\mathrm{d}F(x)\mathrm{d}u \\
&= \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(X-Y>u|Y=y)\mathrm{d}u\mathrm{d}G(y) + \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(X-Y<-u|X=x)\mathrm{d}u\mathrm{d}F(x) \\
&= \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(X>u+y)\mathrm{d}u\mathrm{d}G(y) + \int_{-\infty}^\infty \int_0^\infty \mathbb{P}(Y>u+x)\mathrm{d}u\mathrm{d}F(x).
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}|X-Y| &= \int_{-\infty}^\infty \int_y^\infty \mathbb{P}(X>z)\mathrm{d}z\mathrm{d}G(y) + \int_{-\infty}^\infty \int_x^\infty \mathbb{P}(Y>w)\mathrm{d}w\mathrm{d}F(x) \\
&= \int_{-\infty}^\infty \mathbb{P}(X>z)\int_{-\infty}^z \mathrm{d}G(y)\mathrm{d}z + \int_{-\infty}^\infty \mathbb{P}(Y>w)\int_{-\infty}^x \mathrm{d}F(x)\mathrm{d}w \\
&= \int_{-\infty}^\infty \mathbb{P}(X>z)\mathbb{P}(Y<z)\mathrm{d}z + \int_{-\infty}^\infty \mathbb{P}(Y>w)\mathbb{P}(X<w)\mathrm{d}w \\
&= \int_{-\infty}^\infty \left[(1-F(z))G(z) + (1-G(z))F(z)\right]\mathrm{d}z \\
&= -2\int_{-\infty}^\infty F(z)G(z)\mathrm{d}z + \mathbb{E}Y + \mathbb{E}X.
\end{aligned}
$$

Similarly, taking $G = F$ in the previous development:

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

Equivalently, for the last expectation:

$$
\mathbb{E}|Y-Y'| = -2\int_{-\infty}^\infty G^2(z)\mathrm{d}z + 2\mathbb{E}Y.
$$

The equality (6.1) can be obtained readily combining these partial results.

□

The expression of the right hand side of the last proposition can be directly extended to the $d$-dimensional case by replacing the absolute value with a norm. Furthermore, the resulting expression is rotation invariant and scale equivariant since it only depends on the distance between points.

**Definition 21.** *Let $X$ and $Y$ be random variables in $\mathbb{R}^d$, if $\mathbb{E}\|X\|_d + \mathbb{E}\|Y\|_d < \infty$, the **energy distance** between $X$ and $Y$ is defined as:*

$$
\mathcal{E}(X,Y) = 2\mathbb{E}\|X-Y\|_d - \mathbb{E}\|X-X'\|_d - \mathbb{E}\|Y-Y'\|_d, \tag{6.2}
$$

*where $X'$ and $Y'$ are i.i.d. copies of $X$ and $Y$ respectively.*

In dimensions greater that one, we do not have the equivalence between this distance and the differences between cumulative distribution functions.

The energy distance can be also defined in terms of the characteristic functions. In fact, it can be seen as a weighted $\mathcal{L}_2$ distance between characteristic functions, as states the following proposition. This shows that the energy distance is actually a homogeneity statistic, since the characteristic functions are equal if and only if both distributions are the same. Using the same notation as in Definition 12:

**Proposition 7.** *Given two independent $d$-dimensional random variables $X$ and $Y$, with distributions $\mathbb{P}$ and $\mathbb{Q}$ respectively and such that $\mathbb{E}\|X\|_d + \mathbb{E}\|Y\|_d < \infty$, the energy distance between them can be written as:*

$$\mathcal{E}(X,Y) = \frac{1}{c_d}\int_{\mathbb{R}^d} \frac{|\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t)|^2}{\|t\|_d^{d+1}}\mathrm{d}t, \tag{6.3}$$

*where*

$$c_d = \frac{\pi^{\frac{d+1}{2}}}{\Gamma\left(\frac{d+1}{2}\right)}, \tag{6.4}$$

*being $\Gamma(.)$ the gamma function.*

To prove this proposition for general distributions we need the following lemma. This is important and widely used for this kind of proofs because it allows to translate characteristic functions into expectations of the random variables. We have obtained the lemma and the proof from [15]. We will present here a simpler version than the one presented in the mentioned paper.

**Lemma 8.** *For all $x \in \mathbb{R}^d$, then:*

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

*where $tx$ is the inner product of $t$ and $x$ and $c_d$ is the constant (6.4).*

*Proof.* We can apply an orthogonal transformation to the variable $t$ of the integral, because orthogonal transformations preserve lengths of vectors and angles between them. That is, the value of the integral will be the same. In particular we will apply $t \to z = (z_1, z_2, \ldots, z_d) \in \mathbb{R}^d$ such that $z_1 = \frac{tx}{\|x\|_d}$. Therefore we have that $\|t\|_d = \|z\|_d$ and $tx = z_1\|x\|_d$. We will also change the variables $s = z\|x\|_d$. Then $s_1 = z_1\|x\|_d$ and the determinant of its Jacobian matrix is $\|x\|_d^{-d}$.

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{1-\cos(tx)}{\|t\|_d^{d+1}}\mathrm{d}t &= \int_{\mathbb{R}^d} \frac{1-\cos(z_1\|x\|_d)}{\|z\|_d^{d+1}}\mathrm{d}z \\ &= \int_{\mathbb{R}^d} \frac{1-\cos(s_1)}{\left(\frac{\|s\|_d}{\|x\|_d}\right)^{d+1}}\frac{1}{\|x\|_d^d}\mathrm{d}s\end{aligned}$$

$$= \|x\|_d \int_{\mathbb{R}^d} \frac{1-\cos(s_1)}{\|s\|_d^{d+1}} \mathrm{d}s$$

$$= \|x\|_d \frac{\pi^{\frac{d+1}{2}}}{\Gamma\left(\frac{d+1}{2}\right)} = \|x\|_d c_d.$$

□

The integrals of this lemma at the points $t = 0$ and $t = \infty$ are meant in the principal value sense, that is:

$$\lim_{\varepsilon \to 0} \int_{\mathbb{R}^d \setminus \{\varepsilon S^{d-1} + \varepsilon^{-1} \overline{S^{d-1}}\}} \cdot \ ,$$

where $S^{d-1}$ is the unit ball centered at 0 in $\mathbb{R}^d$ and $\overline{S^{d-1}}$ is its complement. Therefore, the integral of the energy distance is also meant in this sense. Now we can prove the previous proposition:

*Proof.* (Proposition 7) Let $\overline{\Phi_{\mathbb{P}}}(\cdot)$ denote the complex conjugate of the characteristic function. We can rewrite the numerator of the integral, using the properties of complex numbers and sines and cosines:

$$\begin{aligned}
|\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t)|^2 &= (\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t))\overline{(\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t))} \\
&= (\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t))(\overline{\Phi_{\mathbb{P}}(t)} - \overline{\Phi_{\mathbb{Q}}(t)}) \\
&= \Phi_{\mathbb{P}}(t)\overline{\Phi_{\mathbb{P}}(t)} - \Phi_{\mathbb{P}}(t)\overline{\Phi_{\mathbb{Q}}(t)} - \Phi_{\mathbb{Q}}(t)\overline{\Phi_{\mathbb{P}}(t)} + \Phi_{\mathbb{Q}}(t)\overline{\Phi_{\mathbb{Q}}(t)} \\
&= \mathbb{E}\left[e^{itX}e^{-itX'}\right] - \mathbb{E}\left[e^{itX}e^{-itY}\right] - \mathbb{E}\left[e^{itY}e^{-itX}\right] + \mathbb{E}\left[e^{itY}e^{-itY'}\right] \\
&= \mathbb{E}\left[e^{it(X-X')} - e^{it(Y-X)} - e^{it(X-Y)} + e^{it(Y-Y')}\right] \\
&= \mathbb{E}\big[\cos(t(X-X')) + i\sin(t(X-X')) - \cos(t(Y-X)) - i\sin(t(Y-X)) \\
&\quad - \cos(t(X-Y)) - i\sin(t(X-Y)) + \cos(t(Y-Y')) + i\sin(t(Y-Y'))\big] \\
&= \mathbb{E}\big[\cos(t(X-X')) - 2\cos(t(Y-X)) + \cos(t(Y-Y')) + i\sin(t(X-X')) \\
&\quad + i\sin(t(Y-Y'))\big] \\
&= \mathbb{E}\big[\cos(t(X-X')) - 2\cos(t(Y-X)) + \cos(t(Y-Y')) \\
&\quad + i(\sin(tX)\cos(tX') - \cos(tX)\sin(tX')) \\
&\quad + i(\sin(tY)\cos(tY') - \cos(tY)\sin(tY'))\big] \\
&= \mathbb{E}\left[\cos(t(X-X')) - 2\cos(t(Y-X)) + \cos(t(Y-Y'))\right] \\
&= \mathbb{E}\left[2(1-\cos(t(Y-X))) - (1-\cos(t(X-X'))) - (1-\cos(t(Y-Y')))\right].
\end{aligned}$$

Now we write the complete integral and apply Fubini and the previous lemma:

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{|\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t)|^2}{\|t\|_d^{d+1}} \mathrm{d}t &= \int_{\mathbb{R}^d} \left( \frac{\mathbb{E}\left[2(1-\cos(t(Y-X))) - (1-\cos(t(X-X')))\right]}{\|t\|_d^{d+1}} \right. \\
&\qquad \left. - \frac{\mathbb{E}\left[1-\cos(t(Y-Y'))\right]}{\|t\|_d^{d+1}} \right) \mathrm{d}t \\
&= 2\mathbb{E}\left[\int_{\mathbb{R}^d} \frac{1-\cos(t(Y-X))}{\|t\|_d^{d+1}} \mathrm{d}t\right] - \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{1-\cos(t(X-X'))}{\|t\|_d^{d+1}} \mathrm{d}t\right] \\
&\quad -\mathbb{E}\left[\int_{\mathbb{R}^d} \frac{1-\cos(t(Y-Y'))}{\|t\|_d^{d+1}} \mathrm{d}t\right] \\
&= 2\mathbb{E}\left[c_d \|Y-X\|_d\right] - \mathbb{E}\left[c_d \|X-X'\|_d\right] - \mathbb{E}\left[c_d \|Y-Y'\|_d\right] \\
&= c_d \left(2\mathbb{E}\|X-Y\|_d - \mathbb{E}\|X-X'\|_d - \mathbb{E}\|Y-Y'\|_d\right) \\
&= c_d \mathcal{E}(X,Y).
\end{aligned}$$

□

It is clear that the energy distance only vanishes when the distributions are equal, since this is equivalent to have equal characteristic functions. To design a test of independence, an extension of the energy distance needs to be given for distributions of two variables. One of the possible extensions will be carried out in Chapter 7

## 6.2 Generalized energy distance

In this section the energy distance will be generalized to any metric space, and also to distributions with infinite first moments. We will start with a remark about the origin of the weights. As we have just seen, the energy distance can be expressed as a weighted $\mathcal{L}_2$-distance between characteristic functions. It is possible to define other distance by using a different weight function $w(t)$. For simplicity we will assume that the weight function is a continuous, strictly positive function which satisfies:

$$\int |\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t)|^2 w(t) < \infty,$$

where $\Phi_{\mathbb{P}}$ and $\Phi_{\mathbb{Q}}$ are the characteristic functions of the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ respectively. We will analyse the simplest case, when the distributions are defined on $\mathbb{R}$. If we want this new distance to be scale equivariant, it has to satisfy $\forall a \in \mathbb{R}$:

$$\int |\Phi_{\mathbb{P}}(at) - \Phi_{\mathbb{Q}}(at)|^2 w(t) dt = |a| \int |\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t)|^2 w(t) dt.$$

Making the change of variables $s = at$ in this expression:

$$\int |\Phi_{\mathbb{P}}(s) - \Phi_{\mathbb{Q}}(s)|^2 \frac{w(s/a)}{|a|} ds = |a| \int |\Phi_{\mathbb{P}}(t) - \Phi_{\mathbb{Q}}(t)|^2 w(t) dt.$$

That is, the weight function has to satisfy that $\frac{w(t)}{|a|} = |a|w(at)$. If we denote $w(1) \equiv c$, we have that $w\left(\frac{1}{a}\right) = ca^2$ (using $at = 1$). That is, we obtain that the weight function should be $w(t) = \frac{c}{|t|^{d+1}}$, which is exactly the one of the energy distance. It has no sense to talk about rotational invariance in $\mathbb{R}$. However, is can be proved that the only scale and rotation invariant weighted $\mathcal{L}_2$-distance between characteristic functions is the energy distance.

Up to now we have used the euclidean distance in the definition of the energy distance (see (6.2)). In an arbitrary metric space with distance function $\delta$, the energy distance can be extended as:

$$\mathcal{E}_\delta(X, Y) = 2\mathbb{E}[\delta(X, Y)] - \mathbb{E}[\delta(X, X')] - \mathbb{E}[\delta(Y, Y')],$$

whenever these expectations exist. However, in general metric spaces $\mathcal{E}_\delta(X, Y)$ does not necessarily vanish only when the random variables have the same distribution. We will discuss a similar generalization in Section 8.2, which will be necessary to connect these ideas with the MMD methods of the previous chapters. Nevertheless, it is possible to impose restrictions on the distance function so that it characterises equality of the distributions. First we need some additional results and concepts.

**Definition 22.** *Let $\mathcal{X}$ be a nonempty set. We say that $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **negative type** function, or **negative definite function**, if $\forall n \geq 1$, $\forall (\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$ such that $\sum_{i=1}^n \alpha_i = 0$, and $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$:*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \overline{\alpha}_j h(x_i, x_j) \leq 0. \tag{6.5}$$

Moreover, $h$ is said to be **strictly negative definite** if the inequality in the definition is strict, whenever $x_1, \ldots, x_n$ are distinct and at least one of the $\alpha_1, \ldots, \alpha_n$ does not vanish. This definition has an equivalent continuous version:

**Definition 23.** *Let $\mathcal{X}$ be a metric space. We say that $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is **strictly negative definite** if it satisfies, for some probability measure $\mu$ such that $\int_{\mathcal{X}} \alpha(x) \mathrm{d}\mu(x) = 0$:*

$$\int_{\mathcal{X}} \int_{\mathcal{X}} h(x, y)\alpha(x)\alpha(y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) \leq 0,$$

*and equality holds if and only if $\alpha(x) = 0$ almost sure for $\mu$.*

We will now state an interesting result about this kind of functions:

**Proposition 8.** *The Euclidean distance $h(x, y) = \|x - y\|_d$ is strictly negative definite.*

The proof of this proposition can be seen in the Appendix of [17] (Proof of Proposition 1). From the definition of energy distance (see (6.2)) and in view of this proposition, it is straightforward to see that the energy distance can be written as a combination of expectations of a strictly negative definite function $h$,

$$\mathcal{E}(X, Y) = 2\mathbb{E}h(X, Y) - \mathbb{E}h(X, X') - \mathbb{E}h(Y, Y'). \tag{6.6}$$

We will see that this characteristic of $\delta$ (or $h$ in the last equation) of being negative definite is which gives to the energy distance the property of vanishing only when the distributions are equal.

**Proposition 9.** *Given a strictly negative definite distance function $\delta$, $\mathcal{E}_\delta(X,Y) \geq 0$ and it vanishes if and only if both random variables have the same distribution.*

*Proof.* This proof has been obtained from the discussion in [17]. Let $\mathbb{P}$ and $\mathbb{Q}$ denote the distributions of $X$ and $Y$ respectively. Now let $\mu$ be an arbitrary probability measure that dominates the previous ones: If $\mu(A) = 0$ for some measurable set $A$, then $\mathbb{P}(A) = \mathbb{Q}(A) = 0$ ($\mathbb{P}, \mathbb{Q} \ll \mu$). Define,

$$\alpha(x) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}(x) - \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mu}(x),$$

by their Radon-Nikodym derivatives. Then it is clear that $\int_{\mathcal{X}} \alpha(x)\mathrm{d}\mu(x) = 0$, because $\mathbb{P}$ and $\mathbb{Q}$ are probability measures. From the fact that the function $\delta$ is negative definite and symmetric:

$$\begin{aligned}
\mathcal{E}_\delta(X,Y) &= 2\mathbb{E}\delta(X,Y) - \mathbb{E}\delta(X,X') - \mathbb{E}\delta(Y,Y') \\
&= 2\int \delta(x,y)\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{Q}(y) - \int \delta(x,x')\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{P}(x') - \int \delta(y,y')\mathrm{d}\mathbb{Q}(y)\mathrm{d}\mathbb{Q}(y') \\
&= \int \delta(x,y)\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{Q}(y) - \int \delta(x,y)\mathrm{d}\mathbb{P}(x)\mathrm{d}\mathbb{P}(y) \\
&\quad + \int \delta(x,y)\mathrm{d}\mathbb{P}(y)\mathrm{d}\mathbb{Q}(x) - \int \delta(x,y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= -\int \delta(x,y)\mathrm{d}\mathbb{P}(x)\left[\mathrm{d}\mathbb{P}(y) - \mathrm{d}\mathbb{Q}(y)\right] + \int \delta(x,y)\mathrm{d}\mathbb{Q}(x)\left[\mathrm{d}\mathbb{P}(y) - \mathrm{d}\mathbb{Q}(y)\right] \\
&= -\int \delta(x,y)\left[\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x)\right]\left[\mathrm{d}\mathbb{P}(y) - \mathrm{d}\mathbb{Q}(y)\right] \\
&= -\int \delta(x,y)\alpha(x)\mathrm{d}\mu(x)\alpha(y)\mathrm{d}\mu(y) \geq 0.
\end{aligned}$$

And, by the definition of strictly negative, the equality holds only when $\alpha(x) = 0$ a.s., that is, when $\mathbb{P} = \mathbb{Q}$. □

Clearly this property holds for the original energy distance, when $\delta(x,y) = \|x-y\|_d$. We had already seen it by using its interpretation as the $\mathcal{L}^2$ distance between characteristic functions. This interpretation is not possible when we are using another distance function, hence the need of a different proof.

Furthermore, many important distributions do not have finite expectations, as it is imposed in the original definition of the energy distance. The following proposition gives a way to generalize the energy distance for such cases:

**Proposition 10.** *Let $X$ and $Y$ be independent $d$-dimensional random variables with characteristic functions $\Phi_\mathbb{P}$ and $\Phi_\mathbb{Q}$. If $\mathbb{E}|X|^\alpha < \infty$ and $\mathbb{E}|Y|^\alpha < \infty$, for some $0 < \alpha \leq 2$, then:*

- *For $0 < \alpha < 2$,*

$$\mathcal{E}^\alpha(X,Y) \equiv 2\mathbb{E}\|X-Y\|_d^\alpha - \mathbb{E}\|X-X'\|_d^\alpha - \mathbb{E}\|Y-Y'\|_d^\alpha = \frac{1}{C(d,\alpha)}\int_{\mathbb{R}^d} \frac{|\Phi_\mathbb{P}(t) - \Phi_\mathbb{Q}(t)|^2}{|t|^{d+\alpha}}dt,$$

*where*

$$C(d,\alpha) = 2\pi^{d/2}\frac{\Gamma(1-\frac{\alpha}{2})}{\alpha 2^{\alpha}\Gamma\left(\frac{d+\alpha}{2}\right)}.$$

- $\mathcal{E}^2(X,Y) = 2|\mathbb{E}X - \mathbb{E}Y|^2$.

These statements show that $\mathcal{E}^{\alpha}(X,Y) \geq 0$ with equality to zero if and only if the variables are equally distributed, for $0 < \alpha < 2$. But clearly the property does not hold for $\alpha = 2$. A proof of this proposition can be consulted in [15] (Proof of Proposition 2). The proof of the first point when $\alpha = 1$ is given in Proposition 7.

# Chapter 7

# Energy test of independence

In this chapter we will develop a new independence test by using similar arguments as for the energy distance. But now we have to use probability distributions depending on two vectors which can have different dimensions, because we want to measure the distance between the joint distribution and the product of the marginals. This new test is named distance covariance. It should not be confused with the covariance of the distances between the original vectors, although there is a connection between these quantities. Its name comes from the fact that it is a generalization of the classical product-moment covariance. In fact the test is almost equivalent to squared Pearson's covariance for some particular probability distributions, as we will see at the end of the first section.

In that first section we will introduce the distance covariance test, and its normalized version, the distance correlation. We will analyse some properties of the test, such as its relation with the covariance of the distances and its value for some particular distributions. In the second part of the chapter, we will introduce some estimators of energy statistics. It includes the estimators of the energy distance and of the distance covariance.

## 7.1 Distance Covariance

We will start by defining the independence test. Consider the random vectors $X$ and $Y$, with dimensions $d_x$ and $d_y$ and distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ respectively. Let $\Phi_{\mathbb{P}_X}$ and $\Phi_{\mathbb{P}_Y}$ 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, Equation (6.3), to the case in which the functions $f$ and $g$ depend on vectors of different dimension. As in the previous chapter, this expression is obtained from a weighted $\mathcal{L}_2$-distance, imposing rotation invariance and scale equivariance, along with some necessary technical conditions. The energy distance is:

$$\mathcal{E}_{d_x,d_y}(X,Y) = \frac{1}{c_{d_x}c_{d_y}} \int_{\mathbb{R}^{d_x+d_y}} \frac{|\Phi_{\mathbb{P}}(t,s) - \Phi_{\mathbb{Q}}(t,s)|^2}{\|t\|_{d_x}^{d_x+1}\|s\|_{d_y}^{d_y+1}} \mathrm{d}t\mathrm{d}s,$$

where $c_d$ is defined in (6.4). 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 24.** *The **distance covariance (dCov)** between random vectors $X$ and $Y$, with $\mathbb{E}\|X\|_{d_x} + \mathbb{E}\|Y\|_{d_y} < \infty$, is the nonnegative number $\nu(X,Y)$ defined by:*

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

As in the original definition of energy distance, the weight function $w$ is unique, imposing rotation invariance and scale equivariance. The proof of this fact can be found in Section 3 of [18]. However, it is clear that this expression is not equivalent to apply the energy distance test to the joint distribution and the product of the marginals, when we put together both variables $X$ and $Y$ in a single vector of dimension $d_x + d_y$. In fact, the relation between both methods is not direct. We will see the connection between them in Section 9.2.

By the definition of the norm it is clear that $\nu^2(X,Y) \geq 0$ and it vanishes if and only if the vectors are independent. We can define also the distance correlation using the previous definition.

**Definition 25.** *The **distance correlation (dCor)** between random vectors $X$ and $Y$, with $\mathbb{E}\|X\|_{d_x} + \mathbb{E}\|Y\|_{d_y} < \infty$, is the nonnegative number $\mathcal{R}(X,Y)$ defined by:*

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

*where $\nu^2(X,X) = \nu^2(X)$ is the distance variance.*

The distance covariance, like the energy distance, can be expressed using expectations. Using the notation introduced in the HSIC chapter:

**Lemma 9.** *Let $(X,Y),(X',Y'),(X'',Y'') \sim \mathbb{P}_{XY}$ be iid copies of $(X,Y)$, it holds that:*

$$\begin{aligned} \nu^2(X,Y) \quad = \quad & \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\|X-X'\|_{d_x}\|Y-Y'\|_{d_y} + \mathbb{E}_X\mathbb{E}_{X'}\|X-X'\|_{d_x}\mathbb{E}_Y\mathbb{E}_{Y''}\|Y-Y''\|_{d_y} \\ & -2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\|X-X'\|_{d_x}\mathbb{E}_{Y''}\|Y-Y''\|_{d_y}\right]. \end{aligned} \tag{7.1}$$

The details of the derivation of this expression can be found in the proof of Theorem 8 of [19]. This proof is similar to the one of Proposition 7, rewriting the characteristic functions in terms of cosines.

From these definitions it is clear that the distance covariance is not the covariance of distances, however both quantities are related:

**Proposition 11.** *Distance covariance can be expressed in terms of Pearson's covariance of distances:*

$$\nu^2(X,Y) = Cov(\|X-X'\|_{d_x}, \|Y-Y'\|_{d_y}) - 2Cov(\|X-X'\|_{d_x}, \|Y-Y''\|_{d_y}).$$

*Proof.* Applying Lemma 9, and exchanging $(Y-Y')$ and $(Y-Y'')$ when they are alone in an expectation:

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

□

We will now analyse a particular case, motivated by the use of the distance covariance in classification problems. Let $(X,Y)$ be a vector such that $Y \sim Bernoulli(p)$. We will see an expression of the distance covariance in this case. This result is taken from [20]:

**Theorem 10.** *The distance covariance $\nu^2(X,Y)$ can be calculated as:*

$$\nu^2(X,Y) = -2\mathbb{E}[(Y-p)(Y'-p)\|X-X'\|_d],$$

*where $X'$ and $Y'$ denote independent copies of $X$ and $Y$.*

Using simple computations we can see that, up to constants, the squared classical covariance is similar to this expression. The only change is that the value $\|X-X'\|_d$ is squared. First we will analyse the classical covariance for this case:

$$\begin{aligned}
\rho^2(X,Y) &= \mathbb{E}^2[(X-\mathbb{E}X)(Y-\mathbb{E}Y)] \\
&= \mathbb{E}[X(Y-p)]\mathbb{E}[X'(Y'-p)] \\
&= \mathbb{E}[(Y-p)(Y'-p)XX'].
\end{aligned}$$

Now we will see that the modified distance covariance is equal to this one, up to constants:

$$\begin{aligned}
-2\mathbb{E}[(Y-p)(Y'-p)\|X-X'\|_d^2] &= -2\mathbb{E}[(Y-p)(Y'-p)X^2] - 2\mathbb{E}[(Y-p)(Y'-p)X'^2] \\
&\quad +4\mathbb{E}[(Y-p)(Y'-p)XX'] \\
&= -2\mathbb{E}[Y'-p]\mathbb{E}[(Y-p)X_t^2] - 2\mathbb{E}[Y-p]\mathbb{E}[(Y'-p)X'^2] \\
&\quad +4\mathbb{E}[(Y-p)(Y'-p)XX'] \\
&= 4\mathbb{E}[(Y-p)(Y'-p)XX'] \\
&= 4\rho^2(X,Y).
\end{aligned}$$

The only difference between $\nu^2$ and $4\rho^2$ is that we use $\|X-X'\|_d$ instead of $\|X-X'\|_d^2$.

In this section we have introduced the distance covariance independence test. The statistic of this test only depends on the distances between the variables. In the next section we will introduce some empirical estimators. In contrast to HSIC, one does not need to select a particular kernel or adjust any parameters. As will be shown in the chapter on experiments, the test in general performs well, on spite of the fact that it does not adapt to the data.

## 7.2 Energy statistics

In this section we will give some estimators for both energy distance and distance covariance, to apply these test in practice. We will focus on the second type, since we are interested on testing independence. We will analyse also the asymptotic behaviour of the statistics and some interesting properties. We will start with a estimator of the energy distance, which is a test of homogeneity. Given two independent random samples $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_m)$, the two sample energy statistic corresponding to $\mathcal{E}(X,Y)$ (see (6.2)) is:

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

Obtaining the asymptotic distribution of this statistic is quite simple by applying the theory of [10]. The previous expression can be rewritten as a general V-statistic in terms of a kernel $h$:

$$\mathcal{E}_{n,m}(x,y) = \frac{1}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m}h(x_i,y_j;x_k,y_l),$$

where the kernel is given by:

$$h(x_i,y_j;x_k,y_l) = |x_i-y_j| + |x_k-y_l| - |x_i-x_k| - |y_j-y_l|.$$

Then using the same notation as in the mentioned book for a generalized statistic for two samples and under the null hypothesis (both samples have the same distribution):

$$h_{11}(z) = \mathbb{E}_X\mathbb{E}_Y\mathbb{E}_{Y'}h(z,Y;X,Y')$$

$$
\begin{aligned}
&= \mathbb{E}_Y|z-Y| + \mathbb{E}_X\mathbb{E}_Y|X-Y| - \mathbb{E}_X|z-X| - \mathbb{E}_Y\mathbb{E}_{Y'}|Y-Y'| \\
&= \mathbb{E}_X|z-X| + \mathbb{E}_{Y'}\mathbb{E}_Y|Y'-Y| - \mathbb{E}_X|z-X| - \mathbb{E}_Y\mathbb{E}_{Y'}|Y-Y'| = 0, \\
h_{12}(w) &= \mathbb{E}_X\mathbb{E}_{X'}\mathbb{E}_Y h(X,w;X',Y) \\
&= \mathbb{E}_X|X-w| + \mathbb{E}_X\mathbb{E}_Y|X-Y| - \mathbb{E}_X\mathbb{E}_{X'}|X-X'| - \mathbb{E}_Y|y-Y| \\
&= \mathbb{E}_Y|Y-w| + \mathbb{E}_X\mathbb{E}_{X'}|X-X'| - \mathbb{E}_X\mathbb{E}_{X'}|X-X'| - \mathbb{E}_Y|y-Y| = 0,
\end{aligned}
$$

since we can interchange all the random variables in the expectations. Then it is clear that:

$$\xi_{11} = Var_X(h_{11}(X)) = 0 \quad and \quad \xi_{12} = Var_Y(h_{12}(Y)) = 0.$$

Therefore the statistic for testing homogeneity is $T_{n,m} = \frac{nm}{n+m}\mathcal{E}_{n,m}$. The asymptotic distribution of this statistic is:

$$T_{n,m} = \frac{nm}{n+m}\mathcal{E}_{n,m} \quad \xrightarrow{d} \quad \sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

where $\lambda_i$ are the eigenvalues associated with the kernel $h$, and $Z_i$ are independent $\chi_1^2$ random variables. The distribution of the statistic under the null hypothesis depends on the distributions of $X$ and $Y$, through the kernel $h$. Therefore, it should be implemented using a permutation test. More details about this statistic and the consistence of the permutation test can be consulted in the Appendix of [21].

An estimator of the distance covariance can be obtained directly from Equantion (7.1). For a random sample $(x,y) = \{(x_1,y_1)\ldots,(x_n,y_n)\}$ of iid random vectors generated from the joint distribution of $X$ in $\mathbb{R}^{d_x}$ and $Y$ in $\mathbb{R}^{d_y}$, we obtain:

$$
\begin{aligned}
\nu_n^2(x,y) &= \frac{1}{n^2}\sum_{i,j=1}^{n} \|x_i-x_j\|_{d_x}\|y_i-y_j\|_{d_y} + \frac{1}{n^2}\sum_{i,j=1}^{n}\|x_i-x_j\|_{d_x}\frac{1}{n^2}\sum_{i,j=1}^{n}\|y_i-y_j\|_{d_y} \\
&\quad -\frac{2}{n^3}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\|x_i-x_j\|_{d_x}\sum_{j=1}^{n}\|y_i-y_j\|_{d_y}\right).
\end{aligned}
$$

This estimate is costly to compute. It is possible to derive a different estimator of the distance covariance. First, we get the Euclidean distance matrix of each sample, computing all the pairwise distances between sample observations:

$$(a_{ij}) = (\|x_i-x_j\|_{d_x}), \quad (b_{kl}) = (\|y_k-y_l\|_{d_y}).$$

The entries of these matrices are centered, so that their row and column means are zero. Then it can be efficiently computed.

$$A_{ij} = a_{ij} + \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, \text{ for } i,j = 1,\ldots,n,$$

where

$$\bar{a}_{i\cdot} = \tfrac{1}{n}\textstyle\sum_{k=1}^{n} a_{ik}, \quad \bar{a}_{\cdot j} = \tfrac{1}{n}\sum_{k=1}^{n} a_{kj}, \quad \bar{a}_{\cdot\cdot} = \tfrac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} a_{kl}.$$

The other centered matrix $B_{kl}$ is defined in a similar manner. In terms of these centered matrices the sample distance covariance $\nu_n^2(x,y)$ is:

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

Finally the distance correlation is:

$$\mathcal{R}_n^2(x,y) = \begin{cases} \frac{\nu_n^2(x,y)}{\sqrt{\nu_n^2(x)\nu_n^2(y)}} & if \quad \nu_n^2(x)\nu_n^2(y) > 0 \\ 0 & if \quad \nu_n^2(x)\nu_n^2(y) = 0, \end{cases}$$

where:

$$\begin{aligned} \nu_n^2(x) &= \nu_n^2(x,x) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^2, \\ \nu_n^2(y) &= \nu_n^2(y,y) = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} B_{ij}^2. \end{aligned}$$

It can be shown that these statistics converge almost surely to the population when the random vectors have finite first moments.

**Theorem 11.** *If* $\mathbb{E}\|X\| + \mathbb{E}\|Y\| < \infty$, *then*

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

*Proof.* The complete proof of this theorem can be found in Theorem 2 of [22]. Here we present an outline of the proof. First, we need other result, Theorem 1 of [22], which establishes another natural estimator of the distance covariance that uses the original definition as $\mathcal{L}_2$ distance between characteristic functions.

$$\nu_n^2(x,y) = \|\Phi^n_{\mathbb{P}_{XY}}(t,s) - \Phi^n_{\mathbb{P}_X}(t)\Phi^n_{\mathbb{P}_X}(s)\|_w^2,$$

where $\mathbb{P}_{XY}$ is the joint distribution of $X \sim \mathbb{P}_X$ and $Y \sim \mathbb{P}_Y$. The empirical characteristic functions are defined as:

$$\Phi^n_{\mathbb{P}_{XY}}(t,s) = \tfrac{1}{n}\textstyle\sum_{k=1}^n e^{itx_k+isy_k}, \qquad \Phi^n_{\mathbb{P}_X}(t) = \tfrac{1}{n}\textstyle\sum_{k=1}^n e^{itx_k}, \qquad \Phi^n_{\mathbb{P}_Y}(s) = \tfrac{1}{n}\textstyle\sum_{k=1}^n e^{isy_k}.$$

Actually $tx_k \equiv \langle t, x_k\rangle$, but we use the simplified notation. Then we can define:

$$\xi_n(t,s) = \frac{1}{n}\sum_{k=1}^{n} e^{itx_k+isy_k} - \frac{1}{n}\sum_{k=1}^{n} e^{itx_k}\frac{1}{n}\sum_{k=1}^{n} e^{isy_k},$$

and it is clear that $\nu_n^2(t,s) = \|\xi_n(t,s)\|_w^2$. After elementary computations we can check that this quantity can be rewritten as:

$$\xi_n(t,s) = \frac{1}{n}\sum_{k=1}^{n} u_k(t)v_k(s) - \frac{1}{n}\sum_{k=1}^{n} u_k(t)\frac{1}{n}\sum_{k=1}^{n} v_k(s),$$

where $u_k(t) = e^{itx_k} - \Phi_{\mathbb{P}_X}(t)$ and $v_k(s) = e^{isy_k} - \Phi_{\mathbb{P}_Y}(s)$. We now define an integration region which allows us to bound the weight function:

$$D(\delta) = \left\{ (t,s) \;\middle|\; 1 \leq |t|_{d_x}, |s|_{d_y} \leq \frac{1}{\delta} \right\},$$

for any $\delta > 0$. We then define a new random variable in this region:

$$\nu^2_{n,\delta} = \|\xi_n \chi_{D(\delta)}\|^2_w = \int_{D(\delta)} |\xi_n^2| \mathrm{d}w,$$

where $\chi_{D(\delta)}$ is the indicator function of the set $D(\delta)$. As we have mentioned, the weight function is bounded on this region for a fixed $\delta$. Then $\nu_{n,\delta}$ is combination of V-statistics of bounded random variables. Hence by the strong law of large numbers for V-statistics we have:

$$\nu^2_{n,\delta} \quad \xrightarrow[n\to\infty]{a.s} \quad \left\|(\Phi_{\mathbb{P}_{XY}} - \Phi_{\mathbb{P}_X}\Phi_{\mathbb{P}_Y})\chi_{D(\delta)}\right\|^2_w \quad \xrightarrow[\delta\to 0]{a.s} \quad \nu^2.$$

Now it only remains to prove that:

$$\limsup_{\delta\to 0} \limsup_{n\to\infty} \left|\nu^2_{n,\delta} - \nu^2\right| \overset{a.s}{=} 0.$$

This is accomplished by decomposing the expression into four integrals, one for each integration region, $|t|_{d_x} < \delta$, $|t|_{d_x} > \frac{1}{\delta}$, $|s|_{d_y} < \delta$ and $|s|_{d_y} > \frac{1}{\delta}$, and operating with the exponential functions of $u_k$ and $v_k$. □

Now that we have a consistent estimator for the distance covariance, we can analyse its distribution under the null and alternative hypotheses. On the one hand, under independence, $n\nu^2_n(x,y)$ converges in distribution to the quadratic form $\sum_{i=1}^{\infty} \lambda_i Z_i^2$, where $Z_i$ are independent standard normal variables, and $\{\lambda_i\}_{i=1}^n$ are nonnegative constants that depend on the joint distribution of the vectors. A proof can be found in Theorem 5 of [22], which uses a procedure similar to that of the previous proposition. It defines a sequence of random variables using the same integration region $D(\delta)$ and checks the convergence of the characteristic functions. On the other hand, under dependence, $n\nu^2_n(x,y) \to \infty$ as $n$ goes to infinity. Therefore a test that rejects independence for large values of the sample distance covariance is consistent.

In addition, we can see the $\nu^2_n$ density under both the null and alternative hypotheses by approximating it empirically. The left-hand side of Figure 7.1 shows the empirical distribution under $H_0$, with $X$ and $Y$ independent Gaussians variables with unit standard deviation, obtained by using 100 samples from each. The right-hand side shows the empirical distribution under the alternative hypothesis $H_1$. There, $X$ is a Gaussian variable with unit standard deviation and $Y = X^2$, using 100 samples. In both cases, the histograms have been obtained using 2000 independent instances of the $\nu^2_n$ to compute them.

We will analyse another property of this estimate. This distance covariance estimate is asymptotically unbiased. This means that $\mathbb{E}\nu^2_n(x,y) \to \nu^2(X,Y)$, in the limit when $n$ goes to infinity. Without making any assumption on the random vectors $X$ and $Y$, the expected value of the statistic is:

$$\mathbb{E}[\nu^2_n(x,y)] = \frac{(n-1)(n-2)^2}{n^3}\nu^2(X,Y) + \frac{2(n-1)^2}{n^2}\alpha - \frac{(n-1)(n-2)}{n^2}\beta\gamma,$$

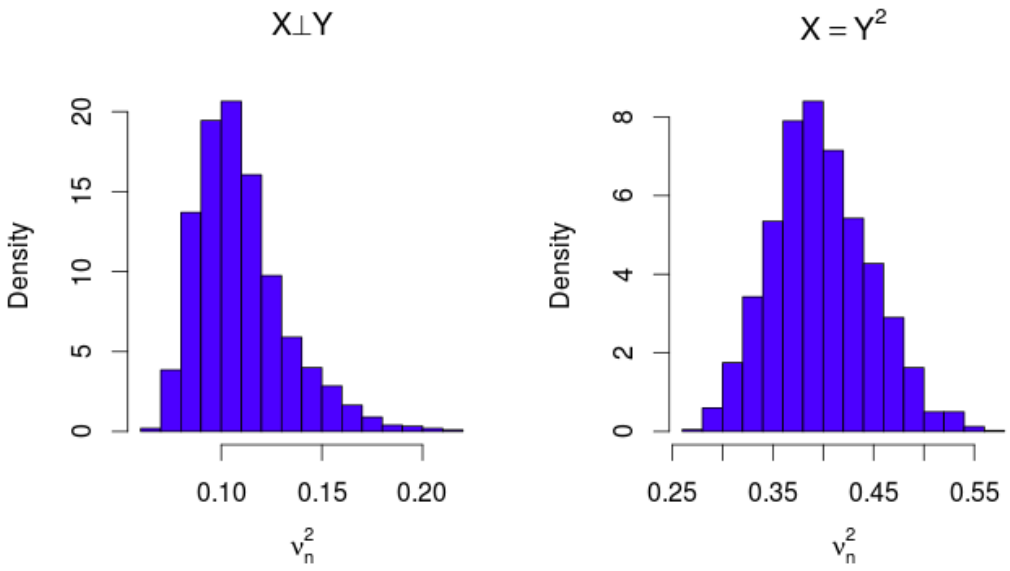



Figure 7.1: Empirical density of $\nu_n^2$ under $H_0$ and $H_1$.

where

$$\alpha = \mathbb{E}[\|X - X'\|_{d_x}\|Y - Y'\|_{d_y}], \quad \beta = \mathbb{E}\|X - X'\|_{d_x} \text{ and } \quad \gamma = \mathbb{E}\|Y - Y'\|_{d_y}.$$

In the limit $n \to \infty$, the last two terms vanish, and the constant of the distance covariance goes to one. Using this expression we can build another estimator that is unbiased for any $n$. We first replace the expressions $\alpha$ and $\beta\gamma$ for their unbiased estimators:

$$\widehat{\alpha} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij}, \quad \widehat{\beta\gamma} = \frac{1}{n(n-1)(n-2)(n-3)} \sum_{i,j=1}^{n} \sum_{k,l \notin (i,j)} a_{ij} b_{kl}.$$

Then an unbiased estimator for $\nu^2(X, Y)$ is given by:

$$U_n^2(x, y) = \frac{1}{(n-1)(n-2)^2} \left( n^3 \nu_n^2(x, y) - \frac{2}{n} \sum_{i,j=1}^{n} a_{ij} b_{ij} + \frac{1}{n(n-3)} \sum_{i,j=1}^{n} \sum_{k,l \notin (i,j)} a_{ij} b_{kl} \right).$$

But this quantity depends on $\nu_n^2(x, y)$, which should be also calculated. Furthermore, it exists a simpler and faster computing formula for an unbiased estimator, given by:

$$\mathcal{U}_n^2(x, y) = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} a_{ij} b_{ij} + \sum_{i,j=1}^{n} \frac{a_{ij}(\bar{b}_{..} - 2\bar{b}_{i.} - 2\bar{b}_{.j} + 2b_{ij})}{n(n-1)(n-2)(n-3)} - 2 \sum_{i,j=1}^{n} \frac{a_{ij}(\bar{b}_{i.} - b_{ij})}{n(n-1)(n-2)}.$$

All these results have been obtained from [15], although the technical developments are not included since they are out of the scope of this work. Being unbiased is a desirable property, but in practice it is usually used the first estimator based on the euclidean distance matrices. We have not analysed the energy distance estimators in depth since this work is mainly focused on independence tests. Nevertheless, in the same reference cited in this paragraph, the complete theory of energy distance is presented, and several statistics for different problems are formulated.

# Chapter 8

# Equivalence between MMD and energy distance

The goal of this chapter is to establish a relation between the MMD (Proposition 2):

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

and the energy distance (Definition 21):

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

This two expressions are strikingly similar. However, they have been derived in a very different ways. To connect them, they need to be generalized. In particular, we will define the general energy distance in terms of semimetrics of negative type. Then we will establish a relation between kernels and these semimetrics, which will help us to connect both generalized tests. This part of the work has been obtained mainly from [14]. Finally we will analyse the practical implications of this result, applying it to other research areas. The observations and results presented in this last part of the chapter are new in the literature.

First of all, we will introduce some notation that will help us to summarize the conditions on the measures needed to formulate the theorems presented in this chapter.

$$\begin{aligned} M(\mathcal{X}) &\equiv \{\nu \mid \nu \textit{ is a finite signed Borel measure on } \mathcal{X}\}, \\ M_+^1(\mathcal{X}) &\equiv \{\nu \mid \nu \textit{ is a Borel probability measure on } \mathcal{X}\}. \end{aligned}$$

## 8.1 Kernel embedding of signed measures

It is not possible to establish a direct relation between the MMD and energy distance homogeneity test. We have seen that both methods can be understood as $\mathcal{L}_2$ distances between characteristic functions. However, we can not find a direct correspondence using the original definitions of the tests because the weight function of energy distance is not integrable. Nevertheless, as will be shown in this chapter, generalized versions of these tests are actually equivalent.

Let us start with the MMD. In the first part of this work we have presented the notion of kernel embedding of probability measures. These embeddings can be generalized by defining them on a more general class of measures, not only on probability ones. In this section we will also describe some properties of this new types of embeddings, which will be useful in later proofs.

In previous chapters we have defined the kernel embeddings as $\mu_{\mathbb{P}} = \int k(x,\cdot)\mathrm{d}\mathbb{P}(x) = \mathbb{E}k(X,\cdot)$, where $X \sim \mathbb{P}$. We will now extend this definition to finite signed Borel measures on $\mathcal{X}$. It means, with the previous notation for the classification of the measures, that we will use $\nu \in M(\mathcal{X})$ instead of $\mathbb{P} \in M_+^1(\mathcal{X})$. Then if $k$ is a kernel on $\mathcal{X}$ we define:

**Definition 26.** *Let $\nu \in M(\mathcal{X})$. The **kernel embedding of $\nu$** into the RKHS $\mathcal{H}_k$ is $\mu_\nu \in \mathcal{H}_k$ such that, for all $f \in \mathcal{H}_k$:*

$$\int f(x)\mathrm{d}\nu(x) = \langle f, \mu_\nu \rangle_{\mathcal{H}_k}.$$

As before, the embedding can be defined also as $\mu_\nu = \int k(x,\cdot)\mathrm{d}\nu(x)$. With this definition it is possible to give the proof of Proposition 5, which was pending from Section 5.1. This proposition states that the HSIC criterion characterizes independence when the kernels of the corresponding RKHS's are characteristic.

*Proof.* (Proposition 5) We need an alternative interpretation of this criterion. If we denote the feature maps of the RKHS's as $\phi$ and $\psi$ and define the finite signed measure:

$$\theta = \mathbb{P}_{XY} - \mathbb{P}\mathbb{Q},$$

it is possible to rewrite the cross-covariance operator as:

$$C_{XY} = \int_{\mathcal{X}\times\mathcal{Y}} \psi(y) \otimes \phi(x)\mathrm{d}\theta(x,y).$$

Then we will prove that $C_{XY} = 0$ if and only if $\mathbb{P}_{XY} = \mathbb{P}\mathbb{Q}$.

($\Longrightarrow$) $\theta$ is a signed measure on $\mathcal{X}\times\mathcal{Y}$, therefore to see that $\theta = 0$, it is enough to check that $\theta(A\times B) = 0$ for all Borel sets $A \in \mathcal{B}(\mathcal{X})$ and $B \in \mathcal{B}(\mathcal{Y})$, where $\mathcal{B}(\cdot)$ denote the Borel $\sigma$-algebras of the spaces. We start defining a finite signed Borel measure for every $f \in \mathcal{H}$ and $B \in \mathcal{B}(\mathcal{Y})$:

$$\begin{aligned} \nu_f(B) &= \int_{\mathcal{X}\times\mathcal{Y}} \langle \phi(x), f\rangle_{\mathcal{H}} \chi_B(y)\mathrm{d}\theta(x,y) \\ &= \int_{\mathcal{X}\times B} \langle \phi(x), f\rangle_{\mathcal{H}} \mathrm{d}\theta(x,y), \end{aligned}$$

where $\chi_B(\cdot)$ is the indicator of the set $B$. From the last expression, we can interpret $\langle \phi(x), f\rangle_{\mathcal{H}}$ as the Radon-Nikodym derivative of $\nu_f$ with respect to $\theta$. That is:

$$\langle \phi(x), f\rangle_{\mathcal{H}} = \frac{\partial \nu_f}{\partial \theta}.$$

We can write the embedding of this measure $\nu_f$ on $\mathcal{G}$:

$$\mu_{\nu_f} = \int_{\mathcal{Y}} l(y,\cdot)\mathrm{d}\nu_f(y) = \int_{\mathcal{Y}} \psi(y)\mathrm{d}\nu_f(y)$$

$$
\begin{aligned}
&= \int_{\mathcal{X}\times\mathcal{Y}} \psi(y)\langle\phi(x), f\rangle_{\mathcal{H}} \mathrm{d}\theta(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} (\psi(y)\otimes\phi(x))(f) \mathrm{d}\theta(x,y) \\
&= \left(\int_{\mathcal{X}\times\mathcal{Y}} \psi(y)\otimes\phi(x) \mathrm{d}\theta(x,y)\right)(f) \\
&= C_{XY}(f) = 0,
\end{aligned}
$$

where we have used the definition of the tensor product operator. Since the kernel $l$ is characteristic, the embedding is injective. This implies that $\nu_f = 0$ for all $f \in \mathcal{H}$. Therefore, by the definition of the reproducing kernel $k$:

$$
\begin{aligned}
\int_{\mathcal{X}\times\mathcal{Y}} \langle\phi(x), \cdot\rangle_{\mathcal{H}} \chi_B(y) \mathrm{d}\theta(x,y) &= \int_{\mathcal{X}\times\mathcal{Y}} k(x,\cdot)\chi_B(y)\mathrm{d}\theta(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \phi(x)\chi_B(y)\mathrm{d}\theta(x,y) = 0.
\end{aligned}
$$

The last expression can be interpreted as the kernel embedding to $\mathcal{H}$ of the finite signed measure $\nu_B(A) = \theta(A\times B)$:

$$
\mu_{\nu_B} = \int_{\mathcal{X}} \phi(x)\mathrm{d}\nu_B(x) = \int_{\mathcal{X}\times\mathcal{Y}} \phi(x)\chi_B(y)\mathrm{d}\theta(x,y) = 0.
$$

Since the kernel $k$ is characteristic, this embedding is injective and then $\nu_B(A) = \theta(A\times B)$ is zero for every Borel sets $A$ and $B$. Thus $\mathbb{P}_{XY} = \mathbb{P}\mathbb{Q}$.

($\Longleftarrow$) $\mathbb{P}_{XY} = \mathbb{P}\mathbb{Q}$ is equivalent to $\theta = 0$. Then it is clear that the cross-covariance operator is also zero. □

In addition to his, if $k$ is bounded, $\mu_\nu$ exists for all $\nu \in M(\mathcal{X})$. This property is equivalent to the one given for probability measures. However, if $k$ is not bounded, there will always exist measures for which the integral $\int k(x,\cdot)\mathrm{d}\nu(x)$ diverges. Henceforth we will assume that the kernels are continuous, and hence measurable, but not necessarily bounded. This means that embeddings will not be defined for some measures. Then we need to define a new group of measures for which the embeddings are well defined.

$$
M_k^\theta(\mathcal{X}) \equiv \left\{ \nu \in M(\mathcal{X}) \;\middle|\; \int k^\theta(x,x)\mathrm{d}|\nu|(x) < \infty \right\}, \text{ with } \theta > 0.
$$

These spaces have some interesting properties, depending on the value of $\theta$, which will be useful in a little while for some proofs.

**Proposition 12.** *Let $\theta_1, \theta_2 > 0$ and $\theta_1 \leq \theta_2$, then $M_k^{\theta_2}(\mathcal{X}) \subseteq M_k^{\theta_1}(\mathcal{X})$.*

**Proposition 13.** *The kernel embedding $\mu_\nu$ is well defined for all $\nu \in M_k^{0.5}(\mathcal{X})$.*

*Proof.* This proof is similar to the one of Lemma 3. We first define the operator $T_\nu f = \int f(x)\mathrm{d}\nu(x)$, for all $f \in \mathcal{H}_k$. Using Jensen and Cauchy-Schwartz inequalities and the reproducing property of $k$ it is possible to show that this operator is bounded under the assumptions of the statement:

$$|T_\nu f| \leq \|f\|_{\mathcal{H}_k} \int \sqrt{k(x,x)}\mathrm{d}|\nu|(x) < \infty.$$

Applying the Riesz representation theorem to $T_\nu$ we know that there exists a $\mu_\nu \in \mathcal{H}_k$ such that $T_\nu f = \langle f, \mu_\nu \rangle_{\mathcal{H}_k}$. □

This generalization of the kernel embeddings allows us to define a more general MMD test and will help us to establish a correspondence with the energy distance. Actually, the relation will not be in terms of the original energy distance, but with another generalization, which will be defined in the next section.

## 8.2 Energy distance with negative type semimetrics

We will now generalize the energy distance test. Although the formulation of this test is simpler that the MMD, its generalization requires more previous knowledge. We will introduce the notion of semimetric of negative type, and use it to give a more general version of the energy distance, similar to the one described in Section 6.2. A semimetric is similar to a distance function but it is not required to satisfy the triangle inequality:

**Definition 27.** *Let $\mathcal{X}$ be a nonempty set, $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a **semimetric** on $\mathcal{X}$ if:*

1. *$\rho(x, x') = 0$ if and only if $x = x'$,*
2. *$\rho(x, x') = \rho(x', x)$.*

*Also $(\mathcal{X}, \rho)$ is said to be a **semimetric space**.*

The semimetric $\rho$ is said to be **of negative type** if it also meets the equation in Definition 22, that is, $\forall n \geq 1$, $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\forall (\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$ such that $\sum_{i=1}^n \alpha_i = 0$:

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha}_j \rho(x_i, x_j) \leq 0$$

If $\rho$ is of negative type, $\rho^a$ with $0 < a < 1$ is also of negative type. We can extend Proposition 8 to characterize these types of functions in general Hilbert spaces:

**Proposition 14.** *A function $\rho$ is a semimetric of negative type if and only if there exists a Hilbert space $\mathcal{H}$ and an injective map $\phi : \mathcal{X} \to \mathcal{H}$ such that:*

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

Using these two last properties it is clear that all Euclidean spaces are of negative type. Furthermore, $\rho^{1/2}$ is a metric if $\rho$ is a semimetric of negative type. If we want to generalize

the previous theory about energy distance by using semimetrics we will need a new moment condition with respect to (w.r.t) a semimetric:

**Definition 28.** *A finite signed Borel measure $\nu$ is said to have a* ***finite $\theta$-moment****, for $\theta > 0$, w.r.t. a semimetric $\rho$ of negative type if there exists $x_0 \in \mathcal{X}$ such that:*

$$\int \rho^\theta(x, x_0) \mathrm{d}|\nu|(x) < \infty.$$

We denote this new moment condition in a similar way as the previous ones of this chapter:

$$M_\rho^\theta(\mathcal{X}) \equiv \{\nu \in M(\mathcal{X}) \mid \nu \text{ has finite } \theta\text{-moment w.r.t } \rho\}.$$

We can use this negative type semimetrics to generalize the energy distance, by replacing the Euclidean distance with a suitable semimetric in its definition. This generalization is similar to the one made in Section 6.2, where we defined the energy distance in an arbitrary metric space with distance function $\delta$. Now we will use a semimetric $\rho$ instead of a distance function:

**Definition 29.** *Let $(\mathcal{X}, \rho)$ be a semimetric space of negative type, $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X}) \cap M_\rho^1(\mathcal{X})$ and $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$ iid. The* ***energy distance w.r.t. $\rho$*** *between $X$ and $Y$ is:*

$$\mathcal{E}_\rho(X, Y) = 2\mathbb{E}\rho(X, Y) - \mathbb{E}\rho(X, X') - \mathbb{E}\rho(Y, Y'). \tag{8.1}$$

When $\rho$ is the Euclidean metric, the condition of $\mathbb{P} \in M_\rho^1(\mathcal{X})$ is equivalent to having finite first moments. If $\rho$ is only a semimetric, and not a metric, we still do not have the tools to show that the conditions imposed are sufficient to ensure that the expectations exist. This will be shown in Section 8.4. This general version of the energy distance has an integral form:

$$\mathcal{E}_\rho(X, Y) = -\int \rho \mathrm{d}([\mathbb{P} - \mathbb{Q}] \times [\mathbb{P} - \mathbb{Q}]). \tag{8.2}$$

The minus sign means that $\mathcal{E}_\rho(X, Y)$ is nonnegative, since $\rho$ is negative definite. In Section 6.2 we proved that this generalization is well defined. From this expression, we see that the semimetric used in the original definition of the energy distance is simply the standard $d$-dimensional norm, $\rho_{\mathcal{E}}(X, X') = \|X - X'\|_d$, if $X, X'$ have dimension $d$. We will use this semimetric in Section 8.5 to establish a equivalence between the original definitions of MMD and energy distance tests.

## 8.3 Kernels and Semimetrics

To connect the generalized versions of MMD and energy distance we will establish a relation between reproducing kernels and semimetrics of negative type. Specifically, we will see that kernels can be defined in terms of semimetrics and vice versa. This means that semimetrics of negative type and symmetric positive definite kernels are closely related.

### 8.3.1 Kernels induced by semimetrics

We will first show how to generate kernels from semimetrics. Then we will see how to derive semimetrics from the kernels induced from them. Let $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a semimetric on a nonempty set $\mathcal{X}$. The following lemma gives a method to define positive definite functions taking $\rho$ as starting point:

**Lemma 10.** *Consider a semimetric $\rho$. Let $x_0 \in \mathcal{X}$. Define the function $k(x, x') = \rho(x, x_0) + \rho(x', x_0) - \rho(x, x')$. The function $k$ is positive definite if and only if $\rho$ is of negative type.*

*Proof.* This proof has been adapted from the one of Lemma 2.1 of [23] (Page 74).

($\Longrightarrow$) We simply have to apply the definition of positive definite function, $\forall n \in \mathbb{N}$, $n \geq 1$, $\forall \{\alpha_i\}_{i=1}^n \in \mathbb{C}^n$ such that $\sum_{i=1}^n \alpha_i = 0$ and $\forall \{x_i\}_{i=1}^n \in \mathcal{X}^n$:

$$\begin{aligned}
\sum_{i,j=1}^{n} \alpha_i \overline{\alpha}_j k(x_i, x_j) &= \sum_{i,j=1}^{n} \alpha_i \overline{\alpha}_j (\rho(x_i, x_0) + \rho(x_j, x_0) - \rho(x_i, x_j)) \\
&= \sum_{i=1}^{n} \alpha_i \rho(x_i, x_0) \sum_{j=1}^{n} \overline{\alpha}_j + \sum_{j=1}^{n} \overline{\alpha}_j \rho(x_j, x_0) \sum_{i=1}^{n} \alpha_i - \sum_{i,j=1}^{n} \alpha_i \overline{\alpha}_j \rho(x_i, x_j) \\
&= -\sum_{i,j=1}^{n} \alpha_i \overline{\alpha}_j \rho(x_i, x_j) \geq 0.
\end{aligned}$$

($\Longleftarrow$) Now we will add the point $x_0$ to the definition of the negative definiteness of $\rho$ and we will use $\alpha_0 = -\sum_{i=1}^n \alpha_i$, which ensures that the sum of all the parameters is zero.

$$\begin{aligned}
\sum_{i,j=0}^{n} \alpha_i \overline{\alpha}_j \rho(x_i, x_j) &= \sum_{i,j=1}^{n} \alpha_i \overline{\alpha}_j \rho(x_i, x_j) + \sum_{i=1}^{n} \alpha_i \overline{\alpha}_0 \rho(x_i, x_0) + \sum_{j=1}^{n} \alpha_0 \overline{\alpha}_j \rho(x_0, x_j) + |\alpha_0|^2 \rho(x_0, x_0) \\
&= \sum_{i,j=1}^{n} \alpha_i \overline{\alpha}_j \rho(x_i, x_j) + \sum_{i=1}^{n} \alpha_i \left( -\sum_{j=1}^{n} \overline{\alpha}_j \right) \rho(x_i, x_0) \\
&\quad + \sum_{j=1}^{n} \left( -\sum_{i=1}^{n} \alpha_i \right) \overline{\alpha}_j \rho(x_0, x_j) + |\alpha_0|^2 0 \\
&= \sum_{i,j=0}^{n} \alpha_i \overline{\alpha}_j \left( \rho(x_i, x_j) - \rho(x_i, x_0) - \rho(x_0, x_j) \right) \\
&= -\sum_{i,j=0}^{n} \alpha_i \overline{\alpha}_j (\rho(x_i, x_0) + \rho(x_0, x_j) - \rho(x_i, x_j)) \\
&= -\sum_{i,j=0}^{n} \alpha_i \overline{\alpha}_j k(x_i, x_j) \leq 0.
\end{aligned}$$

Note that $\rho(x_0, x_0) = 0$, since $\rho$ is a semimetric. □

Actually the original lemma says that it is not necessary for $\rho$ to be a semimetric. It is sufficient that it satisfies $\rho(x_0, x_0) \geq 0$. This can be deduced from the proof if $\rho(x_0, x_0)$ is not removed from the development. Then, it is clear that we can build reproducing kernels from semimetrics of negative type. For convenience we will work with such kernels scaled by $\frac{1}{2}$, since it will avoid to carry along a 2 in subsequent proofs. Let's define them explicitly:

**Definition 30.** *Let $\rho$ be a semimetric of negative type on $\mathcal{X}$ and $x_0 \in \mathcal{X}$. The kernel*

$$k(x, x') = \frac{1}{2}[\rho(x, x_0) + \rho(x', x_0) - \rho(x, x')] \tag{8.3}$$

*is the **distance-induced kernel** (or simply distance kernel), induced by $\rho$ and centered at $x_0$.*

In this definition the kernel depends on the center point $x_0$. It is therefore natural to define a family of kernels induced by the same semimetric by varying this point:

**Definition 31.** *We define the **family of distance-induced kernels** induced by $\rho$ as:*

$$\mathcal{K}_\rho = \left\{ \frac{1}{2}[\rho(x, x_0) + \rho(x', x_0) - \rho(x, x')] \right\}_{x_0 \in \mathcal{X}}.$$

It can be shown that all the kernels in this family are non-degenerate (i.e. their feature maps are injective) if $\rho$ is of negative type. There exists a simply way to see it by using the following proposition. From the equation of Definition 30, the semimetric can be defined in terms of any of the kernels that it induces:

**Proposition 15.** *Let $(\mathcal{X}, \rho)$ be a semimetric space of negative type and $k \in \mathcal{K}_\rho$, then:*

$$\rho(x, x') = \|k(\cdot, x) - k(\cdot, x')\|^2_{\mathcal{H}}.$$

*Proof.* If we set $x = x'$ in Equation (8.3) we get:

$$k(x, x) = \frac{1}{2}[\rho(x, x_0) + \rho(x, x_0) - \rho(x, x)] = \rho(x, x_0).$$

Now we clear the semimetric in (8.3) and use this result:

$$\begin{aligned}
\rho(x, x') &= \rho(x, x_0) + \rho(x', x_0) - 2k(x, x') \\
&= k(x, x) + k(x', x') - 2k(x, x') \\
&= \langle k(\cdot, x), k(\cdot, x)\rangle_{\mathcal{H}} + \langle k(\cdot, x'), k(\cdot, x')\rangle_{\mathcal{H}} - 2\langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}} \\
&= \langle k(\cdot, x) - k(\cdot, x'), k(\cdot, x) - k(\cdot, x')\rangle_{\mathcal{H}} \\
&= \|k(\cdot, x) - k(\cdot, x')\|^2_{\mathcal{H}}.
\end{aligned}$$

□

This expression is very similar to the one used to characterize negative definite functions, by expressing them in terms of injective maps to Hilbert spaces (see Proposition 14). Using Proposition 14, since $\rho$ is a semimetric of negative type, the map $k(\cdot, x) = \phi(\cdot) : \mathcal{X} \to \mathcal{H}$ is injective. Hence all the kernels in $\mathcal{K}_\rho$ are non-degenerate.

In view of this proposition, the natural question is if $\rho$ is also a semimetric when the kernel $k$ involved in the norm does not belong to the family $\mathcal{K}_\rho$. This would clarify completely the connection between kernels and semimetrics of negative type. This is the property that will be shown in the next subsection.

### 8.3.2 Semimetrics generated by kernels

We have just established a way to generate kernels from semimetrics of negative type, but in view of the previous Proposition 15, we can think about generating also semimetrics using kernels.

**Corollary 1.** *Let $k$ be a non-degenerate kernel on $\mathcal{X}$. Then we can define a semimetric $\rho$ of negative type as:*

$$\rho(x, x') = \|k(\cdot, x) - k(\cdot, x')\|^2_{\mathcal{H}}.$$

This result is actually a corollary of Proposition 14. If $k$ is non-degenerate, its characteristic map is injective, and then we can apply it in the statement of the proposition. This shows that $\rho$ is effectively a semimetric of negative type. We say then that ***k* generates *ρ***. It is not the same as proposition 15, as we have said, since in that previous proposition the kernel $k$ belongs to $\mathcal{K}_\rho$, and now it can be a general one. Moreover, we can see in the aforesaid proposition that we can write a semimetric depending on any of the kernels of the family $\mathcal{K}_\rho$, what leads us to define a new relation between kernels:

**Definition 32.** *If two kernels generate the same semimetric, we say that they are **equivalent kernels.***

It is clear that all the kernels $k \in \mathcal{K}_\rho$ generate $\rho$. Therefore, they are equivalent kernels. But, in fact, they are not the only kernels that generates the same semimetric.

**Proposition 16.** *Two kernels $k$ and $h$ on $\mathcal{X}$ are equivalent if and only if they can be written as:*

$$h(x, x') = k(x, x') + f(x) + f(x'),$$

*for some function $f : \mathcal{X} \to \mathbb{R}$.*

*Proof.* ($\Longrightarrow$) If the two kernels are equivalent, they generate the same semimetric, and then:

$$\|k(\cdot, x) - k(\cdot, x')\|^2_{\mathcal{H}} = \|h(\cdot, x) - h(\cdot, x')\|^2_{\mathcal{H}},$$

which is equivalent to

$$k(x, x) + k(x', x') - 2k(x, x') = h(x, x) + h(x', x') - 2h(x, x').$$

Manipulating this equation:

$$\begin{aligned} h(x,x') &= k(x,x') + \frac{1}{2}[h(x,x) + h(x',x') - k(x,x) - k(x,x')] \\ &= k(x,x') + \frac{1}{2}[h(x,x) - k(x,x)] + \frac{1}{2}[h(x',x') - k(x,x')] \\ &= k(x,x') + f(x) + f(x'). \end{aligned}$$

Therefore, $f(x) = \frac{1}{2}(h(x,x) - k(x,x))$.

($\Longleftarrow$) Let us denote as $\rho_k$ the semimetric generated by the kernel $k$. Replacing the kernel in the previous corollary by the expression of $h$ in the statement:

$$\begin{aligned} \rho_h(x,x') &= \|h(\cdot,x) - h(\cdot,x')\|^2_{\mathcal{H}} \\ &= h(x,x) + h(x',x') - 2h(x,x') \\ &= [k(x,x) + f(x) + f(x)] + [k(x',x') + f(x') + f(x')] - 2[k(x,x') + f(x) + f(x')] \\ &= k(x,x) + k(x',x') - 2k(x,x') \\ &= \rho_k(x,x'). \end{aligned}$$

□

Since $k$ and $h$ need to be positive definite, not all the functions $f$ are valid. We can use, for exaple, functions from $\mathcal{H}_k$. If $g \in \mathcal{H}_k$, we can use $f(x) = \frac{1}{2}\|g\|^2_{\mathcal{H}_k} - g(x)$ as the function in the previous proposition to generate equivalent kernels.

The complete relation between kernels and semimetrics can be shown graphically in Figure 8.1. From this figure it is clear that a group of nondegenerated kernels is associated to a single semimetric of negative type. This group can be seen as an equivalence class. Besides, the group of distance kernels induced by this semimetric is a proper subset of the equivalence class.

In the next section we will use this correspondence between kernels and semimetrics to connect the generalizations of the energy distance and the MMD defined in the previous sections. Later we will also show the relation between the original methods without these generalizations.

## 8.4 MMD and energy distance

The goal of this chapter is to connect the generalized energy distance and MMD using semimetrics of negative type. The first step is to harmonize the requirements for the existence of these two quantities. On the one hand, in Proposition 13 of Section 8.1 about kernel embeddings of signed measures, we have seen that a sufficient condition for $\mu_\nu$ to exist is that $\nu \in M_k^{0.5}(\mathcal{X})$. On the other hand, in Definition 29 of Section 8.2, we have seen that a condition for $\mathcal{E}_\rho(X,Y)$

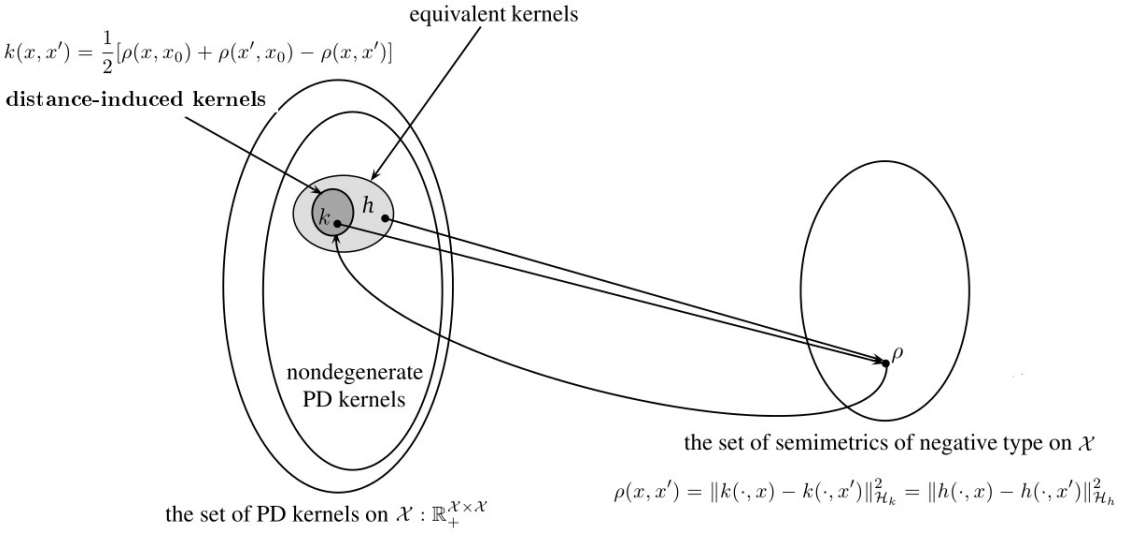



Figure 8.1: Relationship between kernels and semimetrics. From [14].

to exist is that $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X}) \cap M_\rho^1(\mathcal{X})$, where $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$. We will first find the relation between the spaces $M_k^\theta(\mathcal{X})$ and $M_\rho^\theta(\mathcal{X})$ using the relation between kernels and semimetrics.

**Proposition 17.** *Let $k$ be a kernel that generates the semimetric $\rho$, then $M_k^{n/2}(\mathcal{X}) = M_\rho^{n/2}(\mathcal{X})$, for all $n \in \mathbb{N}$.*

*Proof.* $\left[M_k^{n/2}(\mathcal{X}) \subseteq M_\rho^{n/2}(\mathcal{X})\right]$ We will prove this direction for a more general case, using $\theta \geq \frac{1}{2}$. We take $\nu \in M_k^\theta(\mathcal{X})$, which means that $\int k^\theta(x,x)\mathrm{d}|\nu|(x) < \infty$. Using the expression of a semimetric induced by a kernel.

$$\begin{aligned}\int \rho^\theta(x,x_0)\mathrm{d}|\nu|(x) &= \int \|k(\cdot,x) - k(\cdot,x_0)\|_{\mathcal{H}}^{2\theta}\mathrm{d}|\nu|(x) \\ &\leq \int (\|k(\cdot,x)\|_{\mathcal{H}} + \|k(\cdot,x_0)\|_{\mathcal{H}})^{2\theta}\mathrm{d}|\nu|(x)\end{aligned}$$

We will now use that the function $x^{2\theta}$ is convex:

$$(xt + y(1-t))^{2\theta} \leq x^{2\theta}t + y^{2\theta}(1-t),$$

for $t \in [0,1]$. Taking $t = \frac{1}{2}$:

$$\left(\frac{x}{2} + \frac{y}{2}\right)^{2\theta} \leq \frac{x^{2\theta}}{2} + \frac{y^{2\theta}}{2} \Longrightarrow (x+y)^{2\theta} \leq \frac{2^{2\theta}}{2}(x^{2\theta} + y^{2\theta}) = 2^{2\theta-1}(x^{2\theta} + y^{2\theta}).$$

Then applying this inequality in the previous integral:

$$\begin{aligned}\int \rho^\theta(x,x_0)\mathrm{d}|\nu|(x) &\leq 2^{2\theta-1}\left(\int \|k(\cdot,x)\|_{\mathcal{H}}^{2\theta}\mathrm{d}|\nu|(x) + \int \|k(\cdot,x_0)\|_{\mathcal{H}})^{2\theta}\mathrm{d}|\nu|(x)\right) \\ &= 2^{2\theta-1}\left(\int k^\theta(x,x)\mathrm{d}|\nu|(x) + k^\theta(x_0,x_0)\int \mathrm{d}|\nu|(x)\right) \\ &= 2^{2\theta-1}\left(\int k^\theta(x,x)\mathrm{d}|\nu|(x) + |\nu|(\mathcal{X})k^\theta(x_0,x_0)\right) < \infty.\end{aligned}$$

By the definition of $M_k^\theta(\mathcal{X})$, we know that $|\nu|(\mathcal{X}) < \infty$.

$\left[M_k^{n/2}(\mathcal{X}) \supseteq M_\rho^{n/2}(\mathcal{X})\right]$ The relation can be shown by induction. We will actually prove that $M_\rho^\theta(\mathcal{X}) \subseteq M_k^{n/2}(\mathcal{X})$ for $\theta \geq \frac{n}{2}$.

- Let $n = 1$: Let $\theta \geq \frac{1}{2}$ and assume that $\nu \in M_\rho^\theta(\mathcal{X})$, which means that $\int \rho^\theta(x, x_0)\mathrm{d}|\nu|(x) < \infty$. We will use first the inverse triangle inequality and later the Jensen's inequality, since $|\cdot|^{2\theta}$ for $\theta \geq \frac{1}{2}$ is a convex function:

$$\begin{aligned}
\infty > \int \rho^\theta(x, x_0)\mathrm{d}|\nu|(x) &= \int \|k(\cdot, x) - k(\cdot, x_0)\|_{\mathcal{H}}^{2\theta}\mathrm{d}|\nu|(x) \\
&\geq \int \left|\|k(\cdot, x)\|_{\mathcal{H}} - \|k(\cdot, x_0)\|_{\mathcal{H}}\right|^{2\theta}\mathrm{d}|\nu|(x) \\
&= \int \left|k^{\frac{1}{2}}(x, x) - k^{\frac{1}{2}}(x_0, x_0)\right|^{2\theta}\mathrm{d}|\nu|(x) \\
&\geq \left|\int \left(k^{\frac{1}{2}}(x, x) - k^{\frac{1}{2}}(x_0, x_0)\right)\mathrm{d}|\nu|(x)\right|^{2\theta} \\
&= \left|\int k^{\frac{1}{2}}(x, x)\mathrm{d}|\nu|(x) - k^{\frac{1}{2}}(x_0, x_0)|\nu|(\mathcal{X})\right|^{2\theta}.
\end{aligned}$$

  It implies that $\int k^{\frac{1}{2}}(x, x)\mathrm{d}|\nu|(x)$ is finite, i.e. $\nu \in M_k^{\frac{1}{2}}(\mathcal{X})$ and then $M_\rho^\theta(\mathcal{X}) \subseteq M_k^{1/2}(\mathcal{X})$.

- Assume the statement holds for $\theta \geq \frac{n-1}{2}$, that is, $M_\rho^\theta(\mathcal{X}) \subseteq M_k^{(n-1)/2}(\mathcal{X})$. We will prove it for $\theta \geq \frac{n}{2}$. Let $\nu \in M_\rho^\theta(\mathcal{X})$ for $\theta \geq \frac{n}{2}$ and use the same inequalities than before and the Newton's binomial formula:

$$\begin{aligned}
\int \rho^\theta(x, x_0)\mathrm{d}|\nu|(x) &= \int \|k(\cdot, x) - k(\cdot, x_0)\|_{\mathcal{H}}^{2\theta}\mathrm{d}|\nu|(x) \\
&= \int \left(\|k(\cdot, x) - k(\cdot, x_0)\|_{\mathcal{H}}^n\right)^{2\theta/n}\mathrm{d}|\nu|(x) \\
&\geq \left|\int \left(\|k(\cdot, x)\|_{\mathcal{H}} - \|k(\cdot, x_0)\|_{\mathcal{H}}\right)^n\mathrm{d}|\nu|(x)\right|^{2\theta/n} \\
&= \left|\int \sum_{i=0}^n (-1)^i \binom{n}{i} \|k(\cdot, x)\|_{\mathcal{H}}^{n-i}\|k(\cdot, x_0)\|_{\mathcal{H}}^i\mathrm{d}|\nu|(x)\right|^{2\theta/n} \\
&= \left|\int \|k(\cdot, x)\|_{\mathcal{H}}^n + \sum_{i=1}^n (-1)^i \binom{n}{i} \|k(\cdot, x)\|_{\mathcal{H}}^{n-i}\|k(\cdot, x_0)\|_{\mathcal{H}}^i\mathrm{d}|\nu|(x)\right|^{2\theta/n} \\
&= \Bigg|\int k^{n/2}(x, x)\mathrm{d}|\nu|(x) + \dots \\
&\quad + \sum_{i=1}^n (-1)^i \binom{n}{i} k^{i/2}(x_0, x_0) \int k^{(n-i)/2}(x, x)\mathrm{d}|\nu|(x)\Bigg|^{2\theta/n}.
\end{aligned}$$

Since $\theta \geq \frac{n}{2} \geq \frac{n-1}{2} \geq \ldots \geq \frac{1}{2}$, all the terms in the last summation are finite. Furthermore, since $\nu \in M_\rho^\theta(\mathcal{X})$, the first integral of the development is also finite. Therefore, the first term of the last equation is finite. That is, $\nu \in M_k^{n/2}(\mathcal{X})$, which implies $M_\rho^\theta(\mathcal{X}) \subseteq M_k^{n/2}(\mathcal{X})$.

□

From this result it is direct to see that if two kernels $k_1$ and $k_2$ are equivalent (they induce the same semimetric), then $M_{k_1}^{n/2}(\mathcal{X}) = M_{k_2}^{n/2}(\mathcal{X})$. The conditions on kernels embeddings can be expressed in terms of moments w.r.t. $\rho$. Let $\mu_\mathbb{P}$ be the kernel embedding of a probability measure in the RKHS with kernel $k$. If $\rho$ is the semimetric generated by $k$, $\mu_\mathbb{P}$ exists for every $\mathbb{P} \in M_\rho^{0.5}(\mathcal{X})$ (It has finite half-moment w.r.t. $\rho$). The MMD between $\mathbb{P}$ and $\mathbb{Q}$, $\gamma_k(\mathbb{P}, \mathbb{Q})$, is well defined whenever $\mathbb{P}, \mathbb{Q} \in M_\rho^{0.5}(\mathcal{X})$.

The following proposition states that the conditions imposed in the definition of the energy distance w.r.t. $\rho$ (Definition 29) are sufficient to ensure that the expectations involved exist.

**Proposition 18.** *Let $\mathbb{P}, \mathbb{Q} \in M_\rho^1(\mathcal{X})$, then $\mathcal{E}_\rho(X, Y) < \infty$.*

*Proof.* Since $\rho$ is a general semimetric, and not necessarily a metric, we can not use the triangle inequality, which would simplify the proof. We have to prove that:

$$\mathcal{E}_\rho(X, Y) = 2\mathbb{E}\rho(X, Y) - \mathbb{E}\rho(X, X') - \mathbb{E}\rho(Y, Y') < \infty.$$

Let $k$ be any kernel that generates $\rho$:

$$\begin{aligned} \mathbb{E}\rho(X, Y) &= \mathbb{E}\|k(\cdot, X) - k(\cdot, Y)\|_\mathcal{H}^2 \\ &= \mathbb{E}[k(X, X) + k(Y, Y) - 2k(X, Y)] \\ &= \mathbb{E}k(X, X) + \mathbb{E}k(Y, Y) - 2\mathbb{E}k(X, Y). \end{aligned}$$

The first two terms are finite because $\mathbb{P}, \mathbb{Q} \in M_k^1(\mathcal{X})$, since this space is equal to $M_\rho^1(\mathcal{X})$ by the previous proposition (using $n = 2$). For the last term we only have to use the Cauchy-Schwarz inequality:

$$|k(X, Y)| \leq k^{\frac{1}{2}}(X, X)k^{\frac{1}{2}}(Y, Y).$$

By Proposition 12 we know that $M_k^1(\mathcal{X}) \subset M_k^{\frac{1}{2}}(\mathcal{X})$, that is, $\mathbb{P}, \mathbb{Q} \in M_k^{\frac{1}{2}}(\mathcal{X})$. Therefore the last term is also finite. The derivation is similar for the remaining two expectations . □

We now establish the equivalence between the generalized versions of MMD and energy distance.

**Theorem 12.** *Let $(\mathcal{X}, \rho)$ be a semimetric space of negative type and let $k$ be a kernel that generates $\rho$. Then for all $\mathbb{P}, \mathbb{Q} \in M_+^1(\mathcal{X}) \cap M_\rho^1(\mathcal{X})$ such that $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$:*

$$\mathcal{E}_\rho(X, Y) = 2\gamma_k^2(\mathbb{P}, \mathbb{Q}).$$

*Proof.* In Section 8.2 we have seen that the generalized energy distance can be written in integral form (Equation (8.2)). Integrating with respect to $(\mathbb{P} - \mathbb{Q}) \times (\mathbb{P} - \mathbb{Q})$.

$$\mathcal{E}_\rho(X, Y) = -\int \rho(x, y) \mathrm{d}\nu(x) \mathrm{d}\nu(y),$$

where $\nu = \mathbb{P} - \mathbb{Q}$. We have also seen an integral form for the Maximum Mean Discrepancy, Equation (3.2), as:

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \int k(x, y) \mathrm{d}\nu(x) \mathrm{d}\nu(y).$$

Also since $k$ generates $\rho$, we can use that:

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

We can rewritte the energy distance using Equation (8.2). Notice that $\nu(\mathcal{X}) = \mathbb{P}(\mathcal{X}) - \mathbb{Q}(\mathcal{X}) = 1 - 1 = 0$.

$$\begin{aligned}
\mathcal{E}_\rho(X, Y) &= -\int \rho(x, y) \mathrm{d}\nu(x) \mathrm{d}\nu(y) \\
&= -\int (k(x, x) + k(y, y) - 2k(x, y)) \mathrm{d}\nu(x) \mathrm{d}\nu(y) \\
&= -\nu(\mathcal{X}) \left( \int k(x, x) \mathrm{d}\nu(x) + \int k(y, y) \mathrm{d}\nu(y) \right) + 2 \int k(x, y) \mathrm{d}\nu(x) \mathrm{d}\nu(y) \\
&= 2 \int k(x, y) \mathrm{d}\nu(x) \mathrm{d}\nu(y) \\
&= 2\gamma_k^2(\mathbb{P}, \mathbb{Q}).
\end{aligned}$$

□

By this theorem it is clear that equivalent kernels have the same Maximum Mean Discrepancy, MMD. We have imposed the condition $\mathbb{P}, \mathbb{Q} \in M_\rho^1(\mathcal{X})$ to ensure the existence of the energy distance. But the Maximum Mean Discrepancy also exists when $\mathbb{P}$ and $\mathbb{Q}$ have finite half-moments w.r.t. $\rho$, that is, $\mathbb{P}, \mathbb{Q} \in M_\rho^{\frac{1}{2}}(\mathcal{X})$. In that case the energy distance can be infinite and no equivalence can be established.

In Section 6.2, we have seen that the energy distance in a general metric spaces characterises independence if and only if the metric $\delta$ is strict negative definite (Definition 23). This condition is also necessary when we use semimetrics of negative type. In fact:

**Proposition 19.** *Let $k$ be a kernel that generates $\rho$. Then $\rho$ is strict negative definite if and only if $k$ is characteristic to $M_+^1(\mathcal{X}) \cap M_k^1(\mathcal{X})$.*

Thus, the problem of checking whether a semimetric is of strong negative type is equivalent to checking whether its associated kernel is characteristic. Since the MMD with the kernel $k$ is equivalent to the general energy distance with the semimetric $\rho$, the general energy distance characterises independence if and only if $\rho$ is strict negative definite.

## 8.5 Practical implications

In this section we will present some consequences of the connection established in this chapter between MMD and energy distance. We will show the equivalence between the original definitions of energy distance and MMD. In addition to this, at the end of the section, we will introduce an improvement of Theorem 2 of [24]. This is not a consequence of Theorem 12, but is related to semimetrics of negative type.

Let us start by deriving the equivalence between the original MMD and energy distance, defined in chapters 3 and 6 respectively. We only have to determine the semimetric of negative type of the original energy distance. Comparing the definition of $\mathcal{E}_\rho$ with the one of $\mathcal{E}$ it is clear that the semimetric $\rho_{\mathcal{E}} : \mathbb{R}^d \to \mathbb{R}$ of the original energy distance is:

$$\rho_{\mathcal{E}}(x, y) = \|x - y\|_d.$$

Now we want to determine which kernels produces an MMD equivalent to the energy distance. In view of the assumptions of Theorem 12 it is clear that both methods will be equal for any kernel $k$ that generates $\rho_{\mathcal{E}}$. As we know the semimetric, we can build the family of kernels generated by it:

$$\begin{aligned} \mathcal{K}_{\rho_{\mathcal{E}}} &= \left\{ \frac{1}{2}[\rho_{\mathcal{E}}(x, x_0) + \rho_{\mathcal{E}}(x', x_0) - \rho_{\mathcal{E}}(x, x')] \right\}_{x_0 \in \mathcal{X}} \\ &= \left\{ \frac{1}{2}(\|x - x_0\|_d + \|x' - x_0\|_d - \|x - x'\|_d) \right\}_{x_0 \in \mathcal{X}} \end{aligned}$$

Since all these kernels produces the same value of $\text{MMD}_k$, and hence the same value as energy distance, any equivalent kernel can be chosen. The MMD value is equal to the original energy distance for any kernel in $\mathcal{K}_{\rho_{\mathcal{E}}}$. However, MMD and energy distance suggest two different estimates for the same quantity. A natural question is which of these two estimates is better. That is, which of the estimates approximates better the real value. Specifically, we will compare the estimators $MMD_u^2$ and $\nu_n^2$, obtained from the original definitions of MMD and energy distance, in Sections 4.1 and 7.2 respectively. This theoretical value is difficult to calculate in general, but we can obtain it at least when the random variables are unidimensional Gaussians. We will use the expression of the energy distance to make the calculation. If $X \sim \mathcal{N}(\mu, \sigma_X)$ and $Y \sim \mathcal{N}(\mu, \sigma_Y)$, where $\sigma_X$ and $\sigma_Y$ are the standard deviations, the energy distance between them is equal to:

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

We know that:

$$X - Y \sim \mathcal{N}\left(0, \sqrt{\sigma_X^2 + \sigma_Y^2}\right), \quad X - X' \sim \mathcal{N}\left(0, \sigma_X\sqrt{2}\right), \quad Y - Y' \sim \mathcal{N}\left(0, \sigma_Y\sqrt{2}\right).$$

To obtain the expectation of the absolute value, we will use that if $\phi_Z(z)$ is the density function of a Gaussian variable $Z \sim \mathcal{N}(\mu, \sigma)$, then $2\phi_Z(z)$ is the one of $|Z|$:

$$\mathbb{E}|Z| = \int_0^\infty \frac{2z}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z = \frac{2}{\sigma\sqrt{2\pi}} \int_0^\infty z e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z$$


Figure 8.2: Mean Square Error to the real value, sample size 50.

$$= \frac{-2\sigma}{\sqrt{2\pi}} \int_0^\infty \frac{-z}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z = \frac{-2\sigma}{\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma^2}} \Big|_0^\infty = \sigma\sqrt{\frac{2}{\pi}}.$$

Therefore if we apply this expression to the expectations of the energy distance:

$$\begin{aligned} \mathcal{E}(X,Y) &= 2\sqrt{\sigma_X^2+\sigma_Y^2}\sqrt{\frac{2}{\pi}} - \sigma_X\sqrt{2}\sqrt{\frac{2}{\pi}} - \sigma_Y\sqrt{2}\sqrt{\frac{2}{\pi}} \\ &= \frac{2}{\sqrt{\pi}}\left(\sqrt{2(\sigma_X^2+\sigma_Y^2)} - \sigma_X - \sigma_Y\right) \\ &= 2\gamma_k^2(\mathcal{N}(\mu,\sigma_X),\mathcal{N}(\mu,\sigma_Y)). \end{aligned}$$

We have carried out several simulations for this case and the observed behaviour is that as the sample size increases, both estimates tend to the real value, as expected. But when the sample size is small, MMD is closer to the real value most of the times, although the difference between both values is small. For example, in Figure 8.2 we can see the Mean Square Error of both estimates value when $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,5)$, with sample size 50 for 500 replications. We can see that the MSE of the MMD is slightly closer to zero than the one of the energy distance. However, in Figure 8.3 we can see the results for sample size 200. In this case both values of the MSE are almost identical.

It is possible that this behaviour only occurs in this case, when the marginal distributions

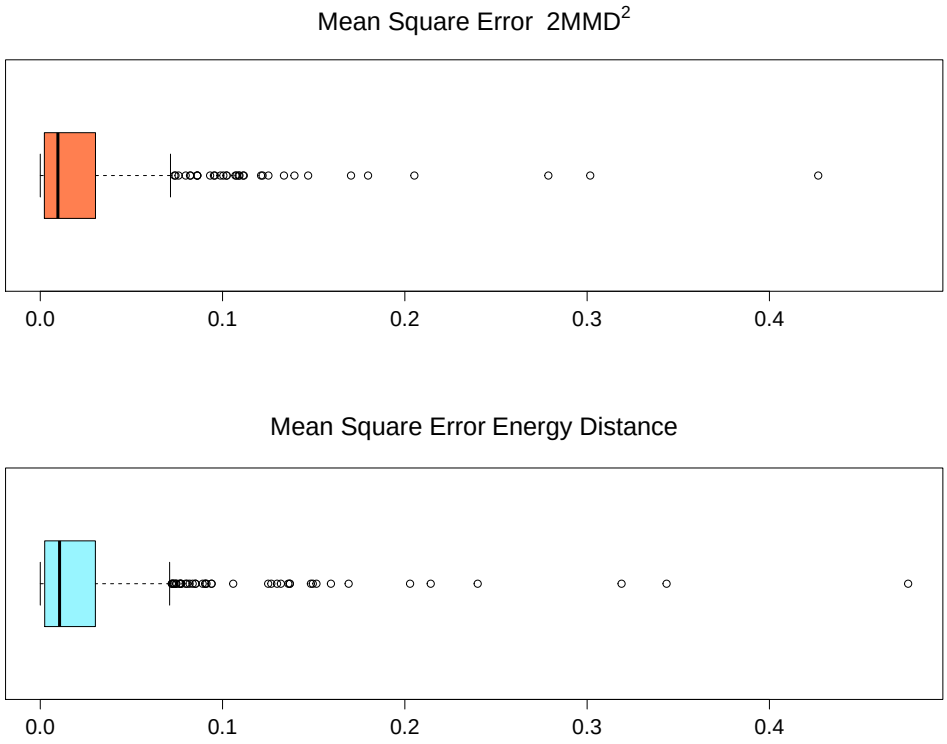



Figure 8.3: Mean Square Error to the real value, sample size 200.

of the variables are Gaussian, but it is difficult to obtain theoretical real values for other distributions.

We have shown that the kernels that generate energy distance belong to the family $\mathcal{K}_{\rho_\mathcal{E}}$, whose elements are similar to the exponent of the Laplacian kernel:

$$k_L(x,y) = e^{-\frac{\|x-y\|_d}{\sigma}}.$$

In fact, we can see that the kernels in $\mathcal{K}_{\rho_\mathcal{E}}$ are a limit of the Lalacian when their width became large. First, we introduce the family of kernels which are equivalent to the Laplacian one. To do this, we obtain the semimetric induced by the Laplacian kernel:

$$\rho_L(x,y) = e^{-\frac{\|x-x\|_d}{\sigma}} + e^{-\frac{\|y-y\|_d}{\sigma}} - 2e^{-\frac{\|x-y\|_d}{\sigma}} = 2 - 2e^{-\frac{\|x-y\|_d}{\sigma}}.$$

Therefore, the family of kernels induced by this semimetric is:

$$\begin{aligned}\mathcal{K}_{k_L} &= \left\{\frac{1}{2}\left(2 - 2e^{-\frac{\|x-x_0\|_d}{\sigma}} - 2 - 2e^{-\frac{\|y-y_0\|_d}{\sigma}} + 2 + 2e^{-\frac{\|x-y\|_d}{\sigma}}\right)\right\}_{x_0\in\mathcal{X}} \\ &= \left\{1 - e^{-\frac{\|x-x_0\|_d}{\sigma}} - e^{-\frac{\|y-y_0\|_d}{\sigma}} + e^{-\frac{\|x-y\|_d}{\sigma}}\right\}_{x_0\in\mathcal{X}}.\end{aligned}$$

We will develop one of these general Laplacian kernels, $\widetilde{k}_L(x,y) \in \mathcal{K}_{k_L}$, using Taylor series with respect to $\frac{1}{\sigma}$ around the point 0:

$$\widetilde{k}_L(x,y) = 1 - e^{-\|x-x_0\|_d\frac{1}{\sigma}} - e^{-\|y-y_0\|_d\frac{1}{\sigma}} + e^{-\|x-y\|_d\frac{1}{\sigma}}$$

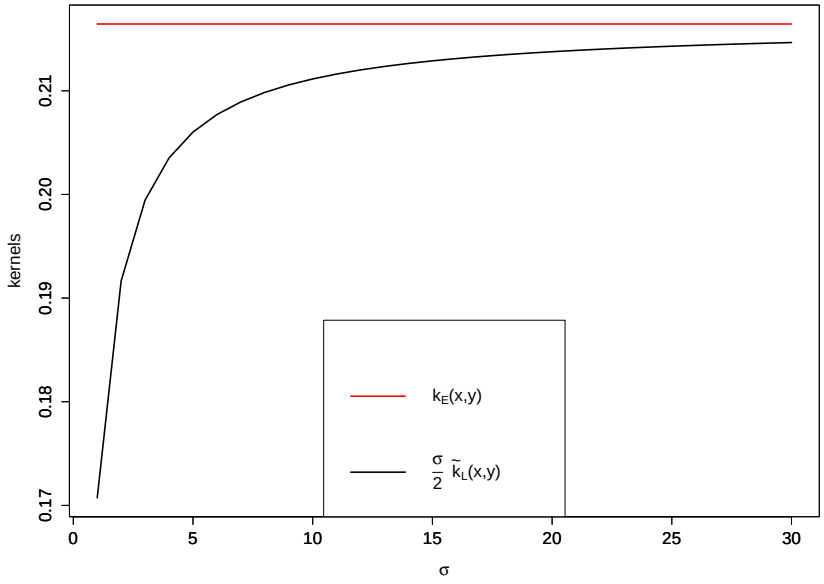



Figure 8.4: Convergence of the value of a scaled Laplacian kernel to the energy distance one as $\sigma \to \infty$.

$$
\begin{aligned}
&= 1 - \left[1 - \frac{1}{\sigma}\|x - x_0\|_d + O\left(\frac{1}{\sigma^2}\right)\right] - \left[1 - \frac{1}{\sigma}\|y - x_0\|_d + O\left(\frac{1}{\sigma^2}\right)\right] \\
&\quad + \left[1 - \frac{1}{\sigma}\|x - y\|_d + O\left(\frac{1}{\sigma^2}\right)\right] \\
&= \frac{1}{\sigma}(\|x - x_0\|_d + \|y - x_0\|_d - \|x - y\|_d) + O\left(\frac{1}{\sigma^2}\right) \\
&= \frac{2}{\sigma} k_{\mathcal{E}}(x, y) + O\left(\frac{1}{\sigma^2}\right),
\end{aligned}
$$

where $k_{\mathcal{E}} \in \mathcal{K}_{\mathcal{E}}$. That is:

$$k_{\mathcal{E}}(x, y) = \frac{\sigma}{2}\widetilde{k}_L(x, y) + O\left(\frac{1}{\sigma}\right).$$

If we take the limit $\sigma \to \infty$, the term $O\left(\frac{1}{\sigma}\right)$ goes to zero, and the remaining term of the expression has to converge since it is equal to the energy distance kernel. We have checked empirically that this convergence holds, as we can see in Figure 8.4, where we have used two random points of $\mathbb{R}^2$ to evaluate both kernels. That is, we can interpret the energy distance as a kernel embedding using a scaled Laplacian kernel with "infinite" width. In empirical results, that we will present later, we have observed that the energy distance is smother than MMD. This can be explained using this relation, because the wider the kernel is, the smother the statistics in which it is involved are.

We have mentioned that establishing a relation between the original versions of the test poses some difficulties if we do not use their generalizations in terms of semimetrics of negative type. However, we have just shown that the original methods are equivalent. The MMD and energy

distance can be written as $\mathcal{L}_2$ distances between characteristic functions:

$$\begin{aligned}\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d} |\Phi_{\mathbb{P}}(w)-\Phi_{\mathbb{Q}}(w)|^2 \mathrm{d}\mu(w),\\ \mathcal{E}(X,Y) &= \frac{1}{c_d}\int_{\mathbb{R}^d} \frac{|\Phi_{\mathbb{P}}(t)-\Phi_{\mathbb{Q}}(t)|^2}{\|t\|_d^{d+1}}\mathrm{d}t.\end{aligned}$$

At first glance, the main problem that difficulties the connection between these quantities is that the weight function of the energy distance is not integrable at zero. However, we have seen that in this limit the energy distance should be understood as Cauchy's principal value sense. For instance, if we take $d=1$:

$$\lim_{\varepsilon\to 0}\int_{\mathbb{R}\setminus(-\varepsilon,\varepsilon)} \frac{|\Phi_{\mathbb{P}}(t)-\Phi_{\mathbb{Q}}(t)|^2}{t^2}\mathrm{d}t = \lim_{\varepsilon\to 0}\int_{\mathbb{R}} \frac{|\Phi_{\mathbb{P}}(t)-\Phi_{\mathbb{Q}}(t)|^2}{t^2+\varepsilon^2}\mathrm{d}t.$$

With this notation, the weights are integrable in $\mathbb{R}$ and the connection can be made without problems. Moreover, this is what explains the meaning of the kernel of the energy distance, $k_{\mathcal{E}}(x,y)$, as a limit of kernels:

$$k_{\mathcal{E}}(x,y) = \lim_{\sigma\to\infty}\frac{\sigma}{2}\left(1-e^{-\frac{\|x-x_0\|_d}{\sigma}}-e^{-\frac{\|y-x_0\|_d}{\sigma}}+e^{-\frac{\|x-y\|_d}{\sigma}}\right).$$

In addition, we have found an application of kernels and semimetrics of negative type in a different problem: dimensionality reduction. A popular technique for reducing the dimensionality in comparing two samples from $X\sim\mathbb{P}$ and $Y\sim\mathbb{Q}$ is to analyse distributions of interpoint comparisons based on a univariate discrepancy function $h$. The theoretical foundation of this technique is given in [24]. First, they assume some restrictive conditions for the density functions of the data and the function $h$. Given $X$ and $Y$ of dimension $d$ with densities $f$ and $g$, the conditions are:

- $\int f^2(x)\mathrm{d}x, \int g^2(y)\mathrm{d}y < \infty$,
- The vector 0 is a Lebesgue point of the function $u(y)=\int_{\mathbb{R}^d} g(x+y)f(x)\mathrm{d}x$, that is:

  $$\lim_{r\to 0^+}\frac{1}{|B(0,r)|}\int_{B(0,r)}|u(y)-u(0)|\mathrm{d}y = 0,$$

  where $B(0,r)$ is a ball centred at 0 with radius $r>0$, and $|B(0,r)|$ is its Lebesgue measure,
- The function $h:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ is nonnegative and continuous,
- $h(x,y)=0$ if and only if $x=y$,
- $h(ax+b,ay+b)=|a|h(x,y)\ \forall a\in\mathbb{R},\forall b\in\mathbb{R}^d$.

Then, Theorem 2 of the mentioned article says:

**Theorem 13.** *Let $X, X'$ be iid d-dimensional random variables with density $f$ and cdf $F$ and let $Y, Y'$ be iid d-dimensional variables with density $g$ and cdf $G$. Assume that the $X$'s and $Y$'s are independent. If the densities $f$ and $g$ and the function $h$ satisfy the previous conditions, then:*

$$h(X, X') \overset{d}{=} h(Y, Y') \overset{d}{=} h(X, Y) \text{ if and only if } F = G,$$

*where $\overset{d}{=}$ means that the distributions are equal.*

The main problem with this theorem is that it imposes restrictions on the density functions. Therefore, it is not valid for all the distributions. In practice, the distribution of the data is often unknown. We can improve the result by using the expectations involved in the MMD expression, that is applicable to any distribution. The extension of the theorem in terms of kernels is:

**Theorem 14.** *Let $X, X' \sim \mathbb{P}$ iid d-dimensional random variables and let $Y, Y' \sim \mathbb{Q}$ be also iid, independent from the $X$'s. Given any characteristic kernel $k$:*

$$\mathbb{E}k(X, X') = \mathbb{E}k(Y, Y') = \mathbb{E}k(X, Y) \text{ if and only if } \mathbb{P} = \mathbb{Q}.$$

*Proof.* We only have to apply the expression of the MMD when the variables are independent:

($\Longrightarrow$) If all the expectations are equal, then:

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}k(X, X) + \mathbb{E}k(Y, Y') - 2\mathbb{E}k(X, Y) = 0.$$

And then, as MMD, based on a characteristic kernel, characterizes equality of distributions, $\mathbb{P} = \mathbb{Q}$.

($\Longleftarrow$) It is clear that if $\mathbb{P} = \mathbb{Q}$, all the expectations are equal. □

As well as working for every distributions, Theorem 14 involves checking only equality of expectations, which is much easier than checking the equality of the whole distributions. However, it imposes more restrictions over the function $k$ than the original theorem. Nevertheless, this does not pose a problem since we have many well-known characteristic kernels at our disposal.

At the beginning of the chapter we have noticed the similarity between the expressions of MMD and energy distance. So, we can use also the generalized definition of the energy distance to write an equivalent theorem:

**Theorem 15.** *Let $X, X' \sim \mathbb{P}$ iid d-dimensional random variables and let $Y, Y' \sim \mathbb{Q}$ be also iid, independent from the $X$'s. Given any strict negative type semimetric $\rho$:*

$$\mathbb{E}\rho(X, X') = \mathbb{E}\rho(Y, Y') = \mathbb{E}\rho(X, Y) \text{ if and only if } \mathbb{P} = \mathbb{Q}.$$

The proof is identical to the previous one but using the energy distance. There are many examples of this kind of semimetrics, including the Euclidean distance used in the original definition of the energy distance.

# Chapter 9

# HSIC and distance covariance

In this chapter we will carry out similar derivations than in the previous one, but in this case with the independence tests. Our goal is to establish a relation between the HSIC and the distance covariance. The first step will be to generalize distance covariance to negative type semimetrics. Then we will show the equivalence between the generalized methods. Moreover, this equivalence allows us to find a direct connection between distance covariance and energy distance.

## 9.1 Generalizations

Let us start generalizing both methods, as we did for the energy distance and the MMD. As in the previous chapter, we need to generalize the distance covariance using semimetrics of negative type. Let $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \tau)$ be semimetric spaces of negative type. The extended definition for the distance covariance is:

**Definition 33.** *Let $X \sim \mathbb{P} \in M_\rho^2(\mathcal{X})$, $Y \sim \mathbb{Q} \in M_\tau^2(\mathcal{Y})$, with joint distribution $\mathbb{P}_{XY}$. The generalized squared Distance Covariance between $X$ and $Y$ is:*

$$\begin{aligned}\nu_{\rho,\tau}^2(X,Y) &= \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho(X,X')\tau(Y,Y') + \mathbb{E}_X\mathbb{E}_{X'}\rho(X,X')\mathbb{E}_Y\mathbb{E}_{Y'}\tau(Y,Y') \\ &\quad -2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho(X,X')\mathbb{E}_{Y'}\tau(Y,Y')\right].\end{aligned}$$

As with the energy distance, the moment conditions ensure that the expectations in this expression are finite.

**Proposition 20.** *Let $X \sim \mathbb{P} \in M_\rho^2(\mathcal{X}), Y \sim \mathbb{Q} \in M_\tau^2(\mathcal{Y})$, then $\nu_{\rho,\tau}(X,Y) < \infty$.*

*Proof.* Applying the Cauchy-Schwartz inequality, we can write the square of the first expectation of the distance covariance as:

$$\begin{aligned}\left[\mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho(X,X')\tau(Y,Y')\right]^2 &\leq \left|\mathbb{E}_{XY,X'Y'}\rho(X,X')\tau(Y,Y')\right|^2 \\ &\leq \mathbb{E}_{XY,X'Y'}\rho^2(X,X')\mathbb{E}_{XY,X'Y'}\tau^2(Y,Y')\end{aligned}$$

$$= \mathbb{E}_X \mathbb{E}_{X'} \rho^2(X, X') \mathbb{E}_Y \mathbb{E}_{Y'} \tau^2(Y, Y').$$

Since $\mathbb{P} \in M_\rho^2(\mathcal{X})$, we know that $\exists x_0 \in \mathcal{X}$ such that:

$$\mathbb{E}_X \rho^2(X, x_0) < \infty,$$

and equivalently there exists an $y_0$ for $\tau$. By Proposition 14, we can apply the triangle inequality to each of the expectations because $\rho$ and $\tau$ are of negative type:

$$\begin{aligned}
\mathbb{E}_X \mathbb{E}_{X'} \rho^2(X, X') &= \mathbb{E}_X \mathbb{E}_{X'} \left( \rho^{\frac{1}{2}}(X, X') \right)^4 \\
&\leq \mathbb{E}_X \mathbb{E}_{X'} \left( \rho^{\frac{1}{2}}(X, x_0) + \rho^{\frac{1}{3}}(x_0, X') \right)^4 \\
&= \mathbb{E}_X \rho(X, x_0)^2 + 4\mathbb{E}_X \rho^{\frac{1}{2}}(X, x_0) \mathbb{E}_{X'} \rho^{\frac{3}{2}}(x_0, X') + 6\mathbb{E}_X \rho(X, x_0) \mathbb{E}_{X'} \rho(x_0, X') \\
&\quad + 4\mathbb{E}_X \rho^{\frac{3}{2}}(X, x_0) \mathbb{E}_{X'} \rho^{\frac{1}{2}}(x_0, X') + \mathbb{E}_{X'} \rho(x_0, X')^2 \\
&= 2\mathbb{E}_X \rho(X, x_0)^2 + 8\mathbb{E}_X \rho^{\frac{1}{2}}(X, x_0) \mathbb{E}_X \rho^{\frac{3}{2}}(X, x_0) + 6 \left[ \mathbb{E}_X \rho(X, x_0) \right]^2 < \infty,
\end{aligned}$$

since $X$ and $X'$ have the same distribution, and by Propositions 12 and 17 we have that $M_\rho^2(\mathcal{X}) \subseteq M_\rho^{\frac{3}{2}}(\mathcal{X}) \subseteq M_\rho^1(\mathcal{X}) \subseteq M_\rho^{\frac{1}{2}}(\mathcal{X})$. That is, all the expectations in the expression above are finite. Equivalently for the expectation of $\tau$. A similar reasoning can be made to show that the remaining terms of the distance covariance are finite. □

This generalized distance covariance can be also expressed in integral form:

$$\nu_{\rho,\tau}^2(X, Y) = \int \rho\tau \mathrm{d}([\mathbb{P}_{XY} - \mathbb{P}, \mathbb{Q}] \times [\mathbb{P}_{XY} - \mathbb{P}, \mathbb{Q}]), \tag{9.1}$$

where $\mathbb{P}_{XY}$ is the joint distribution of $X$ and $Y$ and $\rho\tau$ is viewed as a function on $(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$.

This generalized distance covariance do not characterizes independence for every $\rho$ and $\tau$. That is, we may have $\nu_{\rho,\tau}(X, Y) = 0$ for some different distributions $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$. Theorem 3.11 of [25] shows that if the semimetrics are of strong negative type, then distance covariance characterizes independence. This is the same property as for the energy distance. Using Proposition 19, we can rewrite this restriction in terms of kernel properties. We do not need to generalize HSIC, because it is already defined for two different kernels.

## 9.2 Equivalence between methods

We can state a result similar to the previous Theorem 12, now between HSIC and distance covariance. Let $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \tau)$ be semimetric spaces of negative type, and let $k_x$ and $k_y$ be two kernels on $\mathcal{X}$ and $\mathcal{Y}$, with RKHS's $\mathcal{H}$ and $\mathcal{G}$, that generate $\rho$ and $\tau$ respectively. The product of $k_x$ and $k_y$ kernel in the tensor product space $\mathcal{H} \times \mathcal{G}$ is defined as:

$$k((x, y), (x', y')) = k_x(x, x') k_y(y, y').$$

Using this definition:

**Theorem 16.** *Let $X \sim \mathbb{P} \in M_\rho^2(\mathcal{X})$, $Y \sim \mathbb{Q} \in M_\tau^2(\mathcal{Y})$, with joint distribution $\mathbb{P}_{XY}$. Let $k(x,y)$ the product kernel of $k_x(x)$ and $k_y(y)$ that generate $\rho$ and $\tau$ respectively, then:*

$$\nu_{\rho,\tau}^2(X,Y) = 4HSIC(\mathbb{P}_{XY}, \mathcal{H}, \mathcal{G}).$$

*Proof.* This proof is similar to the one of Theorem 12, which establishes the equivalence between energy distance and MMD. Define $\nu = \mathbb{P}_{XY} - \mathbb{P}\mathbb{Q}$. We will use that:

$$\nu(\mathcal{X}\times\mathcal{Y}) = \mathbb{P}_{XY}(\mathcal{X}\times\mathcal{Y}) - \mathbb{P}\mathbb{Q}(\mathcal{X}\times\mathcal{Y}) = \mathbb{P}_{XY}(\mathcal{X}\times\mathcal{Y}) - \mathbb{P}(\mathcal{X})\mathbb{Q}(\mathcal{Y}) = 1 - 1 = 0.$$

We will also use that $\nu$ has zero marginal measures, that is:

$$\begin{aligned}
\nu_X(x) &= \int (\mathbb{P}_{XY}(x,y) - \mathbb{P}(x)\mathbb{Q}(y))\mathrm{d}y \\
&= \int \mathbb{P}_{XY}(x,y)\mathrm{d}y - \mathbb{P}(x)\int \mathbb{Q}(y)\mathrm{d}y \\
&= \mathbb{P}(x) - \mathbb{P}(x)\mathbb{Q}(\mathcal{Y}) = 0,
\end{aligned}$$

$$\begin{aligned}
\nu_Y(y) &= \int (\mathbb{P}_{XY}(x,y) - \mathbb{P}(x)\mathbb{Q}(y))\mathrm{d}x \\
&= \int \mathbb{P}_{XY}(x,y)\mathrm{d}x - \mathbb{Q}(y)\int \mathbb{P}(x)\mathrm{d}x \\
&= \mathbb{Q}(y) - \mathbb{Q}(y)\mathbb{P}(\mathcal{X}) = 0.
\end{aligned}$$

This means that $\int g(x,y,x',y')\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') = 0$ if $g$ does not depend on one or more of its arguments. Let's use the integral form of the distance covariance, given by Equation (9.1). We will also use the fact that the semimetrics are generated by the kernels $k_x$ and $k_y$ and the equivalence between kernel methods, $\gamma_k^2(\mathbb{P}_{XY}, \mathbb{P}\mathbb{Q}) \equiv MMD^2(\mathcal{F}, \mathbb{P}_{XY}, \mathbb{P}\mathbb{Q}) = HSIC(\mathbb{P}_{XY}, \mathcal{H}, \mathcal{G})$:

$$\begin{aligned}
\nu_{\rho,\tau}^2(X,Y) &= \int \rho(x,x')\tau(y,y')\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&= \int (k_x(x,x) + k_x(x',x') - 2k_x(x,x'))(k_y(y,y) + k_y(y',y') - 2k_y(y,y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&= \int k((x,y),(x,y))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') + \int k((x,y'),k(x,y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&\quad -2\int k((x,y),(x,y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') + \int k((x',y),(x',y))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&\quad + \int k((x',y'),(x',y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') - 2\int k((x',y),(x',y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&\quad -2\int k((x,y),(x',y))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') - 2\int k((x,y'),(x',y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&\quad +4\int k((x,y),(x',y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y') \\
&= \nu(\mathcal{X}\times\mathcal{Y})\left(\int k((x,y),(x,y))\mathrm{d}\nu(x,y) + \int k((x',y'),(x',y'))\mathrm{d}\nu(x',y')\right) \\
&\quad +0 + 4\int k((x,y),(x',y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y')
\end{aligned}$$

$$= 4\int k((x,y),(x',y'))\mathrm{d}\nu(x,y)\mathrm{d}\nu(x',y')$$

$$= 4\gamma_k^2(\mathbb{P}_{XY},\mathbb{P}\mathbb{Q})$$

$$= 4HSIC(\mathbb{P}_{XY},\mathcal{H},\mathcal{G}).$$

The two integrals that involve the term $\nu(\mathcal{X}\times\mathcal{Y})$ are finite due to the moment conditions on the marginals, $\mathbb{P}\in M_\rho^2(\mathcal{X}) = M_{k_x}^2(\mathcal{X}) \subseteq M_{k_x}(\mathcal{X})$ and $\mathbb{Q}\in M_\tau^2(\mathcal{Y}) = M_{k_y}^2(\mathcal{Y}) \subseteq M_{k_y}(\mathcal{Y})$. □

From Theorems 12 and 16, we can establish also a relation between the energy distance and the distance covariance. From the original definitions of both methods it suggests that they are closely related, but up to now it is not clear whether the distance covariance is equal to $\mathcal{E}_{\widetilde{\rho}}$ for some semimetric $\widetilde{\rho}$ on $\mathcal{X}\times\mathcal{Y}$. We will clarify it in the next corollary of the previous theorem:

**Corollary 2.** *Let $(\mathcal{X},\rho)$ and $(\mathcal{Y},\tau)$ be semimetric spaces of negative type, and let $X\sim\mathbb{P}\in M_\rho^2(\mathcal{X})$ and $Y\sim\mathbb{Q}\in M_\tau^2(\mathcal{Y})$, with joint distribution $\mathbb{P}_{XY}$. Then:*

$$\nu_{\rho,\tau}^2(X,Y) = \mathcal{E}_{\widetilde{\rho}}(Z,W),$$

*where $Z\sim\mathbb{P}_{XY}$ and $W\sim\mathbb{P}\mathbb{Q}$ and the semimetric $\frac{\widetilde{\rho}}{2}$ is generated by the product kernel $k(x,y) = k_x(x)k_y(y)$.*

This result can be seen directly applying the previous results which estate the relation between all the methods. From these results the relation between the different methods is apparent.

Since we are working in the tensor product space $\mathcal{H}\times\mathcal{G}$, we could think about distinguishing probability distributions on this space. However, the product kernel $k((x,y),(x',y')) = k_x(x,x')k_y(y,y')$ may not be characteristic even if $k_x$ and $k_y$ are characteristic. We can formulate a simple example of this behaviour. We assume that $k_x$ and $k_y$ are bounded, so that we can consider embeddings of all probability measures. Let $k_x$ be a kernel centered at $x_0$ and induced by the semimetric $\rho$. This implies that $k((x_0,y),(x_0,y')) = k_x(x_0,x_0)k_y(y,y') = 0$. Then for every two distinct $\mathbb{Q}_1,\mathbb{Q}_2\in M_+^1(\mathcal{Y})$, we have that $\gamma_k^2(\delta_{x_0}\mathbb{Q}_1,\delta_{x_0}\mathbb{Q}_2) = 0$, where $\delta_{x_0}$ is the Dirac delta function. If we assume that $\rho$ and $\tau$ are strictly negative definite, $\nu_{\rho,\tau}$ characterises independence. By Theorem 16, $\gamma_k$ also characterises independence, but not equality of probability measures on the product space. That is, it is easier to distinguish $\mathbb{P}_{XY}$ from $\mathbb{P}\mathbb{Q}$ than two-sample testing on the product space.

In conclusion, all the methods introduced up to now in this work are closely related. Besides, in the following chapter we will introduce a novel independence test, whose power will be compared in Chapter 11 with independence tests based on discrepancies introduced in the previous chapters.

# Chapter 10

# Independence test based on non-Gaussianity

In this chapter we will introduce a class of new independence tests based on measures of non-Gaussianity. We prove that the proposed tests characterize independence when the marginals of the random variables are Gaussian. However, we have not been able to extend this derivation to the general case, when the marginal distributions are non-Gaussian. We will describe the problems that we have found during this extension. At the end of the chapter we present some modifications that could be made to the test. In the following chapter we analyse the power of all the possible considered modifications, and compare them with other independence tests proposed in the literature.

## 10.1 Basic idea and theoretical foundation

In this section we will introduce the ideas that lead to the new independence test, and some theoretical results. The first one is the observation that, in many cases, the sum of two random variables is more "Gaussian" if the variables are independent than if they are not. In Figure 10.1 we can see an illustration of this behaviour. Here $X \sim U[-1,1]$ and $Y = X^2$ with sample size 300, scaled later to have zero mean and unit variance. To obtain a sample of observations $\widetilde{Y}$ with the same distribution of $Y$ and independent of $X$, we permute the sample of $Y$ (formally, the marginals of $X$ and $Y$ are exchangeable, not independent, but in practice the difference is not relevant). In this example, we present histograms of the samples as a simple way to asses the Gaussianity. It is clear that the second sum is much closer to the Gaussian distribution than the first one.

This idea of analysing the non-Gaussianity of a sum of two variables arises from the central limit theorem, since the sum of an increasing number of independent variables,with finite variance, approaches the Gaussian distribution. However this need not be the case if the variables are dependent. In Section 10.3 (Figure 10.2) we provide an example for which the sum of dependent random variables is less Gaussian than the sum of independent random variables with

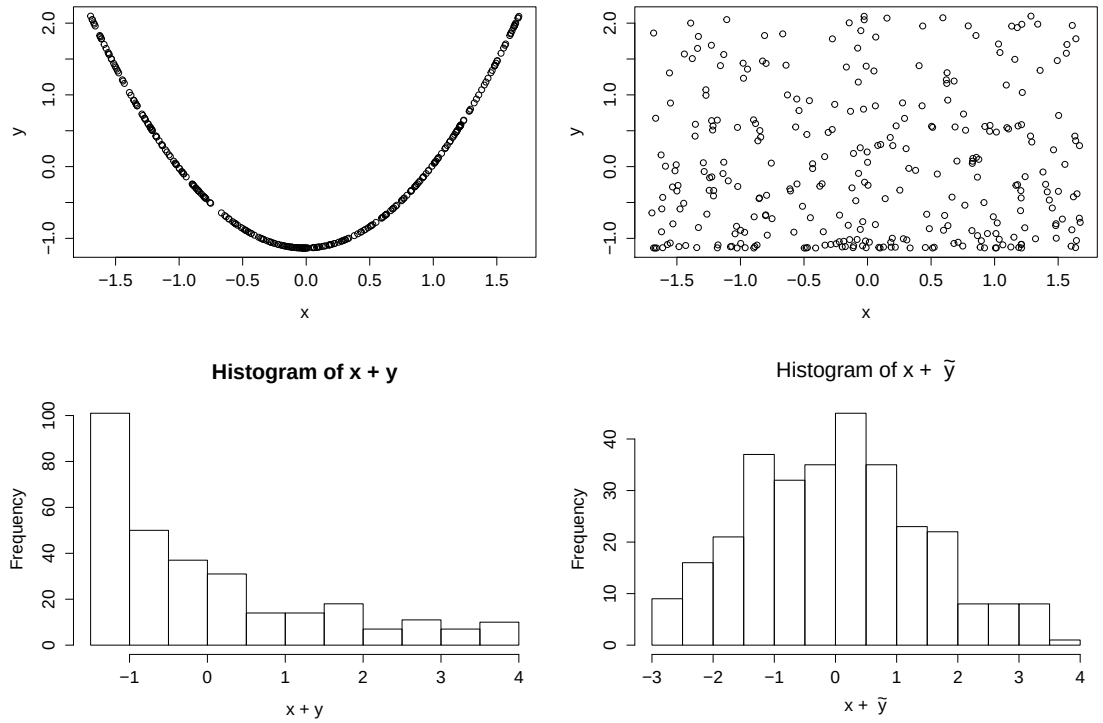



Figure 10.1: Histograms of the sum of two variables, being $X \perp \widetilde{Y}$.

the same marginals.

The basis of our analysis is a result given in [27] which links the negentropy, a measure of non-Gaussianity, with a well known measure of the independence, the mutual information. Let us start defining this independence criterion, which characterizes independence by measuring the Kullback-Leibler divergence between the joint distribution and the product of the marginals:

**Definition 34.** *The **Kullback-Leibler divergence** of two probability function distributions $F$ and $G$, with density functions $f$ and $g$ respectively, is defined as:*

$$D_{KL}(F\|G) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x.$$

The Kullback-Leibler divergence is not symmetric. That is, even if $D_{KL}(F\|G) = 0$ if and only if $f = g$, it does not define a distance. The mutual information of two random variables is defined as follows:

**Definition 35.** *Given two random variables $X$ and $Y$ with distribution functions $F_X$ and $F_Y$ and joint distribution $F_{XY}$, we define their **mutual information** as:*

$$I(X, Y) = D_{KL}(F_{XY}\|F_X F_Y).$$

This quantity measures the dependence between $X$ and $Y$. It actually characterizes independence:

**Proposition 21.** *The mutual information is non-negative, $I(X,Y) \geq 0$, and it is zero if and only if $X$ and $Y$ are independent.*

*Proof.* The first claim can be readily derive applying Jensen's inequality to the minus logarithm.

$$\begin{aligned} I(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\mathrm{d}x\mathrm{d}y \\ &= \mathbb{E}_{XY}\log\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} \\ &= \mathbb{E}_{XY}\left[-\log\frac{f_X(x)f_Y(y)}{f_{XY}(x,y)}\right] \\ &\geq -\log\mathbb{E}_{XY}\frac{f_X(x)f_Y(y)}{f_{XY}(x,y)} \\ &= -\log\int_{-\infty}^{\infty} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y \\ &= -\log 1 = 0. \end{aligned}$$

For the other one we have to prove both directions, although one of them is direct:

($\Longrightarrow$) Since $f_{XY}$ is a density function, it is non-negative and its integral is equal to one. Therefore it can not be zero everywhere. If $I(X,Y) = 0$, the logarithm should be zero at least in the support of $f_{XY}$.

$$f_{XY}(x,y)\chi_D(x,y) = f_{XY}(x,y) = f_X(x)f_Y(y)\chi_D(x,y),$$

where $D = supp(f_{XY}) \subseteq \mathbb{R}^2$ and $\chi_D$ denotes the indicator function of the set $D$. This means that if $f_{XY} \neq 0$, it is equal to the product of the marginals. We now prove that, if the joint distribution is zero, then the product of the marginals is also zero. For this we will integrate the expression above with respect to $x$ and $y$:

$$\begin{aligned} f_Y(y) &= \textstyle\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x &&= f_Y(y)\int_{-\infty}^{\infty} f_X(x)\chi_D(x,y)\mathrm{d}x, \\ f_X(x) &= \textstyle\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y &&= f_X(x)\int_{-\infty}^{\infty} f_Y(y)\chi_D(x,y)\mathrm{d}y. \end{aligned}$$

Then we have that:

$$\begin{aligned} 1 &= \textstyle\int_{-\infty}^{\infty} f_X(x)\chi_D(x,y)\mathrm{d}x \leq \int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1 \quad \forall y, \\ 1 &= \textstyle\int_{-\infty}^{\infty} f_Y(y)\chi_D(x,y)\mathrm{d}y \leq \int_{-\infty}^{\infty} f_Y(y)\mathrm{d}y = 1 \quad \forall x. \end{aligned}$$

This means that $supp(f_X) \subseteq supp(f_{XY})$ and $supp(f_Y) \subseteq supp(f_{XY})$, i.e., $f_{XY}(x,y) = 0$ implies $f_X(x)f_Y(y) = 0$.

($\Longleftarrow$) If the variables are independent then $f_{XY}(x,y) = f_X(x)f_Y(y)$, which implies that:

$$I(X,Y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\log(1)\mathrm{d}x\mathrm{d}y = 0.$$

□

A measure of non-Gaussianity can be given in terms of the entropy of a random variable:

**Definition 36.** *Let $X$ be a random variable with density function $f$. The **differential entropy**, or simply entropy, of $X$ is:*

$$H(X) = -\int f(x)\log f(x)\mathrm{d}x.$$

The entropy depends on the distribution of the variable. It can be shown that the Gaussian distribution maximizes it for a given covariance matrix. In particular, the entropy for a normal random variable with covariance matrix $\Sigma$ and any mean vector is equal to $\log((2\pi e)^{n/2}|\Sigma|^{1/2})$, where $n$ is the dimension of the variable. This value is the maximum for any possible distribution whose covariance matrix is $\Sigma$. The Kullback-Leibler divergence between two distribution functions $F$ and $G$, with density functions $f$ and $g$ respectively, is:

$$\begin{aligned} D_{KL}(F\|G) &= \int f(x)\log(f(x))\mathrm{d}x - \int f(x)\log(g(x))\mathrm{d}x \\ &= -H(X) - \int f(x)\log(g(x))\mathrm{d}x, \end{aligned} \tag{10.1}$$

where $X \sim F$. The mutual information can be expressed in terms of the entropy as well:

**Proposition 22.** *Let $X$ and $Y$ be two random variables. Let $Y_\perp$ be another random variable with the same distribution as $Y$ but independent of $X$. Then, we can rewrite their mutual information as:*

$$I(X,Y) = H(X,Y_\perp) - H(X,Y).$$

*Proof.* We just have to apply the definition of the mutual information and the previous Equation (10.1):

$$\begin{aligned} I(X,Y) &= -H(X,Y) - \int f_{X,Y}(x,y)\log(f_X(x)f_Y(y))\mathrm{d}x\mathrm{d}y \\ &= -H(X,Y) - \int f_{X,Y}(x,y)\log(f_X(x))\mathrm{d}x\mathrm{d}y - \int f_{X,Y}(x,y)\log(f_Y(y))\mathrm{d}x\mathrm{d}y \\ &= -H(X,Y) - \int f_X(x)\log(f_X(x))\mathrm{d}x - \int f_Y(y)\log(f_Y(y))\mathrm{d}y \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

Now if we take the other random variable $Y_\perp$, we know that $I(X,Y_\perp) = 0$, because $X$ and $Y_\perp$ are independent. Therefore, using the fact that the entropy only depends on the distribution of the variable, we get:

$$0 = H(X) + H(Y_\perp) - H(X,Y_\perp) \implies H(X) + H(Y) = H(X,Y_\perp).$$

And it only remains to substitute this in the previous expression. □

The mutual information can be expressed as the sum of two terms: one that measures non-linear dependences and other linear ones. Define the random vector $Z = (X, Y)$, with $X$ and $Y$ the variables whose dependences we want to measure. The linear dependence between $X$ and $Y$ is determined by the covariance matrix $\Sigma_{XY}$. Let $\Sigma_Z$ be the covariance matrix of $Z$:

$$\Sigma_Z = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_Y \end{pmatrix},$$

where $\Sigma_X$ and $\Sigma_Y$ are the covariance matrices of the variables, and $\Sigma_{XY}$ the matrix with entries $\mathbb{E}[(X_i - \mathbb{E}X_i)(Y_j - \mathbb{E}Y_j)]$. If the variables $X$ and $\widetilde{Y}$ are linearly independent, $\Sigma_{X\widetilde{Y}}$ is the null matrix, and then, denoting $\widetilde{Z} = (X, \widetilde{Y})$:

$$\Sigma_{\widetilde{Z}} = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}.$$

Then we can define:

**Definition 37.** *Given the random vectors $Z = (X, Y)$ and $\widetilde{Z} = (X, \widetilde{Y})$, where $Y, \widetilde{Y}$ have the same distribution and $X$ and $\widetilde{Y}$ are linearly independent. The linear dependence between $X$ and $Y$ can be measured by:*

$$C(Z) = C(X, Y) = D_{KL}(\mathcal{N}(\mu_Z, \Sigma_Z) \| \mathcal{N}(\mu_Z, \Sigma_{\widetilde{Z}})),$$

*where $\mu_Z$ denotes the vector composed of the mean vectors $\mu_X$ and $\mu_Y$ of each of the random variables.*

Along the same lines we define another function that measures the non-Gaussianity of a distribution. Using also the Kullback-Leibler divergence, it quantifies the "distance" from the distribution to a Gaussian one with the same mean and covariance matrix.

**Definition 38.** *Given a random vector $Z = (X, Y)$ with mean vector $\mu$ and covariance matrix $\Sigma$, the **negentropy** of the vector is defined as:*

$$G(Z) = G(X, Y) = D_{KL}(F_{XY} \| \mathcal{N}(\mu, \Sigma)),$$

*where $F_{XY}$ is the joint distribution.*

This quantity depends only on the distribution of the variables and has some desirable properties. For example it is always non-negative ($G(Z) \geq 0$) and invariant with respect to invertible affine transformations ($AZ + b$ with $A$ an invertible matrix). Since it also depends on the Kullback-Leibler divergence, we can express it in terms of the entropy, as we did before with the mutual information. Now we transform it by using $\mathbb{Q} \equiv \mathcal{N}(0, \Sigma)$ in the Kullback-Leibler definition, and hence $g(x) = \phi(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$, where $n$ is the dimension of the variable. Actually it is equal to the difference between the entropies of both distributions:

**Proposition 23.** *Let $Z$ be a $\mathcal{N}(\mu, \Sigma)$ variable and $X$ be some random variable with the same mean and variance. Then the negentropy of $X$ is equal to:*

$$G(X) = H(Z) - H(X).$$

*Proof.* We only have to decompose the expression of the Kullback-Leibler divergence like in Equation (10.1) and operate with the resulting integrals:

$$\begin{aligned} G(X) &= -H(X) - \int_{\mathbb{R}^n} f(x) \log\left(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}\right) \mathrm{d}x \\ &= -H(X) + \frac{1}{2}\int_{\mathbb{R}^n} f(x)(x-\mu)^\top \Sigma^{-1}(x-\mu)\mathrm{d}x + \log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right)\int_{\mathbb{R}^n} f(x)\mathrm{d}x \\ &= \log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) + \frac{n}{2} - H(X) \\ &= \log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right) + \frac{1}{2}\log(e^n) - H(X) \\ &= \log\left((2\pi e)^{n/2}|\Sigma|^{1/2}\right) - H(X) \\ &= H(Z) - H(X). \end{aligned}$$

□

This result implies that the Gaussian has maximum entropy. Since the negentropy, which is defined in terms of the Kullback-Leibler divergence, is always non-negative:

$$H(Z) - H(X) \geq 0 \quad \Longrightarrow \quad H(Z) \geq H(X).$$

Now that we have all the definitions and principal properties of all the needed quantities, we can introduce the central result on which the test is based, which has been taken from [27]. It brings to light the relation between non-Gaussianity and linear and nonlinear dependencies of these variables, through mutual information:

**Proposition 24.** *The mutual information between two random variables $X$ and $Y$ can be expressed as:*

$$I(X,Y) = G(X,Y) - G(X) - G(Y) + C(X,Y). \tag{10.2}$$

*Proof.* For a random variable $Z = (X,Y)$ with distribution $F_{XY}$, mean vector $\mu_Z$ and covariance matrix $\Sigma_Z$ we define the matrix:

$$\Sigma_{\widetilde{Z}} = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}.$$

We want to combine the mutual information, which compares a joint distribution with the product of the marginals, and the negentropy, which compares this joint distribution with a Gaussian. Then we will measure the Kullback-Leibler divergence between a joint distribution $F_{XY}$ of two random variables $X$ and $Y$ and a product of two independent Gaussians of the same dimensions as $X$ and $Y$ respectively. We denote this new product distribution as:

$$F^{F_{XY}\wedge G} \equiv \mathcal{N}\left(\mu_Z, \Sigma_{\widetilde{Z}}\right).$$

This decomposition can be done by two ways:

- We will start adjusting to $F_{XY}$ a general product of distributions $Q_X Q_Y$, and later we will transform these $Q_X, Q_Y$ into Gaussian.

$$\begin{aligned} D_{KL}\left(F_{XY} \| Q_X Q_Y\right) &= D_{KL}\left(F_{XY} \| F_X F_Y\right) + D_{KL}\left(F_X F_Y \| Q_X Q_Y\right) \\ &= I(X,Y) + D_{KL}\left(F_X \| Q_X\right) + D_{KL}\left(F_Y \| Q_Y\right), \end{aligned}$$

where $F_X$ and $F_Y$ are the distribution functions of $X$ and $Y$ respectively. Now if we take $Q_X = \mathcal{N}(\mu_X, \Sigma_X)$ and $Q_Y = \mathcal{N}(\mu_Y, \Sigma_Y)$ independent, we ensure that the product is $Q_X Q_Y = \mathcal{N}(\mu_Z, \Sigma_{\widetilde{Z}})$. Then we have:

$$\begin{aligned} D_{KL}\left(F_{XY} \| F^{F_{XY} \wedge G}\right) &= D_{KL}\left(F_{XY} \| Q_X Q_Y\right) \\ &= I(X,Y) + D_{KL}\left(F_X \| \mathcal{N}(\mu_X, \Sigma_X)\right) + D_{KL}\left(F_Y \| \mathcal{N}(\mu_Y, \Sigma_Y)\right) \\ &= I(X,Y) + G(X) + G(Y). \end{aligned}$$

- Now we will start adjusting a general Gaussian $\mathcal{N}(\mu, \Sigma)$ to $F_{XY}$ and later we will express it as an independent product:

$$\begin{aligned} D_{KK}(F_{XY} \| \mathcal{N}(\mu, \Sigma)) &= D_{KL}(F_{XY} \| \mathcal{N}(\mu_Z, \Sigma_Z)) + D_{KL}(\mathcal{N}(\mu_Z, \Sigma_Z) \| \mathcal{N}(\mu, \Sigma)) \\ &= G(X,Y) + D_{KL}(\mathcal{N}(\mu_Z, \Sigma_Z) \| \mathcal{N}(\mu, \Sigma)). \end{aligned}$$

Taking $\mu = \mu_Z$ and $\Sigma = \Sigma_{\widetilde{Z}}$:

$$\begin{aligned} D_{KL}\left(F_{XY} \| F^{F_{XY} \wedge G}\right) &= D_{KL}\left(F_{XY} \| \mathcal{N}(\mu_Z, \Sigma_{\widetilde{Z}}\right) \\ &= G(X,Y) + D_{KL}(\mathcal{N}(\mu_Z, \Sigma_Z) \| \mathcal{N}(\mu_Z, \Sigma_{\widetilde{Z}}) \\ &= G(X,Y) + C(X,Y). \end{aligned}$$

Equating both equations of $D_{KL}\left(F_{XY} \| F^{F_{XY} \wedge G}\right)$ we obtain the result. □

Since the mutual information characterizes independence and the term $C$ only detects linear dependencies, the remainder part of the expression is the measure of the non-linear ones. So we will analyse this term. If we take another random variable $Y_\perp$ with the same distribution as $Y$ but independent of $X$, their mutual information will be zero, and also their linear dependencies. Then:

$$G(X, Y_\perp) - G(X) - G(Y_\perp) = 0 \quad \longrightarrow \quad G(X, Y_\perp) = G(X) + G(Y_\perp).$$

The negentropy only depends on the distribution, so $G(Y) = G(Y_\perp)$. Then we can rewrite the mutual information as:

$$I(X,Y) = (G(X,Y) - G(X, Y_\perp)) + C(X,Y). \tag{10.3}$$

This means that we can measure the non-linear dependencies of the variables by comparing how Gaussian is the joint distribution with respect to the independent one. This is already close to the original idea given at the beginning of the section, since it involves the non-Gaussianity of a joint distribution an its corresponding with independent marginals. In fact, from this expression we can deduce that:

**Theorem 17.** *Given two uncorrelated random variables $X_{LI}$ and $Y_{LI}$, their joint distribution is farther from the Gaussian than the joint one of the corresponding independent random variables with the same marginal distributions $X_\perp$ and $Y_\perp$. That is:*

$$G(X_{LI}, Y_{LI}) \geq G(X_\perp, Y_\perp).$$

*Proof.* It is direct by using Equation (10.3). If the variables are uncorrelated, $C(X_{LI}, Y_{LI}) = 0$. Therefore, since the mutual information is always equal or greater than zero, we obtain the result. □

However, we still do not have considered the sum of the variables, which was part of the original idea. To fix it we will analyse the simplest case in the next section, when the distributions of $X$ and $Y$ are both Gaussian. We will use it to derive a test of independence based on one dimensional projections of the variables, which we will apply later to arbitrary distributions.

## 10.2 Gaussian marginals

In this section we will define a new way to compute the non-linear dependencies between two random variables, instead of using the negentropy. We will develop it only when the marginals are Gaussian. Without loss of generality, we shall henceforth assume that the random variables $X$ and $Y$ are standardized. Then their negentropy is zero and the expression of the mutual information given in Equation (10.2) is:

$$I(X, Y) = G(X, Y) + C(X, Y).$$

We will see that we can measure the non-linear dependencies of the sample by computing the nongaussianity of the projections $\rho X + \sqrt{1-\rho^2}Y$, for $\rho \in [-1, 1]$. With this change we recover the sums involved in the original idea of the test. It is also interesting given that it is always easier to work in one dimension.

**Proposition 25.** *Given two uncorrelated normal random variables, then $G(X, Y) = 0$ if and only if $G(\rho X + \sqrt{1-\rho^2}Y) = 0$, $\rho \in [-1, 1]$.*

*Proof.* $C(X, Y) = 0$ because $X$ and $Y$ are uncorrelated, i.e. they do not have linear dependencies. Since both variables are normal, their mutual information in directly their negentropy.

$$I(X, Y) = G(X, Y).$$

($\Longrightarrow$) We have already proven that the mutual information characterizes the independence. Therefore $G(X,Y) = 0$ implies that the variables are independent. It is well known that if two normal variables are independent, their linear combinations are also normal. Then $\rho X + \sqrt{1-\rho^2}Y$ are normal for all $\rho \in [-1,1]$, which implies that $G(\rho X + \sqrt{1-\rho^2}Y) = 0$.

($\Longleftarrow$) If $G(\rho X + \sqrt{1-\rho^2}Y) = 0$ for all $\rho$, all these linear combinations are normal. Then by the Cramer-Wold Theorem, which states that a Borel probability measure on $\mathbb{R}^k$ is uniquely determined by the totality of its one-dimensional projections, we obtain the result. □

In addition, we know that the non-Gaussianity of the independent pair is equal zero, and therefore $G(\rho X + \sqrt{1-\rho^2}Y_\perp) = 0$ for all $\rho \in [-1,1]$ when $X \perp Y_\perp$. This means that the main idea of the test presented at the beginning of the previous section holds when the marginals are Gaussian:

$$G(\rho X + \sqrt{1-\rho^2}Y) \geq G(\rho X + \sqrt{1-\rho^2}Y_\perp) = 0. \tag{10.4}$$

In other words, the linear combination of two independent Gaussian random variables is more Gaussian than the corresponding one when the variables are not independent (It is clear since if they are independent their sum is directly a Gaussian).

In view of this last proposition, we have a family of negentropy values, $\{G(\rho X+\sqrt{1-\rho^2}Y)\}_{\rho\in[-1,1]}$. We will use the mean as a representative of the family. It is clear from the proof of the proposition that the result holds also when using the mean over $\rho$ instead of all $\rho \in [-1,1]$. We could have used the maximum instead of the mean, but considering the practical results, we decided to use this definition. That is, we will use the following quantity to measure the non-linear dependencies:

$$NLD(X,Y) = \int_{-1}^{1} G(\rho X + \sqrt{1-\rho^2}Y)\mathrm{d}\rho.$$

We can use Proposition 23 to rewrite the negentropy of $\rho X + \sqrt{1-\rho^2}Y$ in terms of the entropy, which gives an easy way to compute it. It is easy to see that the standard deviation of this sum, denoted by $\sigma_\rho$, is equal to $\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right)^{1/2}$, where $\rho_{XY}$ is the linear correlation of the variables (it is equal to the covariance since the variables have unit variance).

$$\begin{aligned} G(\rho X + \sqrt{1-\rho^2}Y) &= \log(\sigma_\rho\sqrt{2\pi e}) - H(\rho X + \sqrt{1-\rho^2}Y) \\ &= \frac{1}{2}\log(\sigma_\rho^2 2\pi e) - H(\rho X + \sqrt{1-\rho^2}Y) \\ &= \frac{1}{2}\log\left(2\pi e(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY})\right) - H(\rho X + \sqrt{1-\rho^2}Y). \end{aligned}$$

But if we want to use $NLD(X,Y)$ to substitute the negentropy in the mutual information, we need to rewrite also the measure for the linear dependencies. That is, we need to obtain a new expression for $C(X,Y)$ in terms of the projections. This new expression should be such that it recovers the mutual information if we sum it with the measure of the non-linear dependencies, like in Equation (10.2). To do it we will use the expression of the mutual information in terms

of the entropy, given by Proposition 22. If we make the non-linear dependencies equal to zero, we will obtain the expression of the linear ones. Using the above formulation for $G$:

$$G(\rho X + \sqrt{1-\rho^2}Y) = 0 \implies H(\rho X + \sqrt{1-\rho^2}Y) = \frac{1}{2}\log\left(2\pi e(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY})\right).$$

Therefore, if we substitute this result in the Equation of Proposition 22 we get:

$$\begin{aligned} C(\rho X + \sqrt{1-\rho^2}Y) &= H(\rho X + \sqrt{1-\rho^2}Y_\perp) - H(\rho X + \sqrt{1-\rho^2}Y) \\ &= \frac{1}{2}\log(2\pi e) - \left[\frac{1}{2}\log\left(2\pi e(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY})\right)\right] \\ &= \frac{1}{2}\log(2\pi e) - \frac{1}{2}\log(2\pi e) - \frac{1}{2}\log\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right) \\ &= -\frac{1}{2}\log\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right). \end{aligned}$$

As before, we get a family of values depending on $\rho$, so we will use their maximum as a representative of the family. So our new measure for the linear dependencies is:

$$LD(X,Y) = -\frac{1}{2}\max_{\rho\in[-1,1]}\log\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right).$$

But we can calculate explicitly the value of this maximum. We have to derive this expression with respect to $\rho$ to obtain the extreme points:

$$\begin{aligned} \frac{\partial}{\partial\rho}\log\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right) &= \frac{\frac{\partial}{\partial\rho}\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right)}{1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}} \\ &= \frac{2\rho_{XY}}{1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}}\left(\sqrt{1-\rho^2} - \frac{\rho^2}{\sqrt{1-\rho^2}}\right) \\ &= \frac{2\rho_{XY}}{\sqrt{1-\rho^2} + 2\rho(1-\rho^2)\rho_{XY}}\left(1-2\rho^2\right) = 0. \end{aligned}$$

So the maximum values of $LD(X,Y)$ are obtained in:

$$\rho = \pm\frac{1}{\sqrt{2}}.$$

We obtain the final expression by substituting this values in the original function:

$$LD(X,Y) = -\frac{1}{2}\log(1 - |\rho_{XY}|).$$

Using this expression and the previous one developed for the non-linear dependencies we can define the independence measure for the Gaussian case:

$$I^*(X,Y) = NLD(X,Y) + LD(X,Y)$$

$$= \int_{-1}^{1} \left[ \frac{1}{2} \log \left( 2\pi e(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}) \right) - H(\rho X + \sqrt{1-\rho^2}Y) \right] \mathrm{d}\rho$$
$$-\frac{1}{2}\log(1 - |\rho_{XY}|). \tag{10.5}$$

We have seen along this section and the previous one that this is a valid independence test, that is, this expression characterises independence since mutual information does. But right now we only have a measure which assumes that the marginal distributions of the variables are both Gaussian, which is not the case in most applications. In the next section we will define the general independence test that does not require that the marginals to be Gaussian. The final expression is similar to this one, but without removing the non-Gaussianity of the independent pair.

## 10.3 Non Gaussian marginals

In this section we will introduce a new independence test for general random variables, without making any assumption about their marginal distributions. However in this case the theoretical foundations of the test are not completely developed.

The first attempt would be to see if any linear combination of the variables are more Gaussian if they are independent. This property holds when the marginals are Gaussian, as shows Equation (10.4). However, if the random variables can have any distribution, there could be directions for which the sum of the independent samples of the random variables is less Gaussian than the sum of the original dependent ones. Anyway, in practice it is difficult to find an example where the mentioned property does not hold.

We still have not delimited the class of distributions for which Equation (10.4) does not hold for all $\rho \in [-1, 1]$, but we have obtained some clues from the experiments. The most clear example is displayed in Figure 10.2, where $X \sim U[0, 1]$ and $Y \sim Pareto(1, 1)$ is obtained from $X$ as:

$$Y = \frac{1}{1-X}, \quad \textit{where } 1 - X \sim U[0, 1].$$

The explicit relation between the data can be seen in Figure 10.3, where it is clearly reflected that the variable $Y$ is almost zero. The Pareto distribution has heavy tails, so it keeps its shape after standardization. We divided by $1 - X$ instead of $X$ to have positive correlation between the variables, which leads to a more clear interchange between the order of the negentropies. In this example we see that for $\rho > 0$ the negentropy of the sum of the random variables is larger when they are independent that when they are dependent. That is, the independent sum is more distant from the Gaussian distribution according to the Kullback-Leibler divergence.

Although Equation (10.4) only holds for $\rho < 0$ in this example, the difference between both values of the negentropy is really small for the values of $\rho$ where it does not hold. So we believe that the mean of all the differences are greater than zero. That is, we guess that:

**Conjecture 1.** *Given three random variables $X \sim \mathbb{P}$ and $Y, Y_\perp \sim \mathbb{Q}$, such that $X$ is independent*

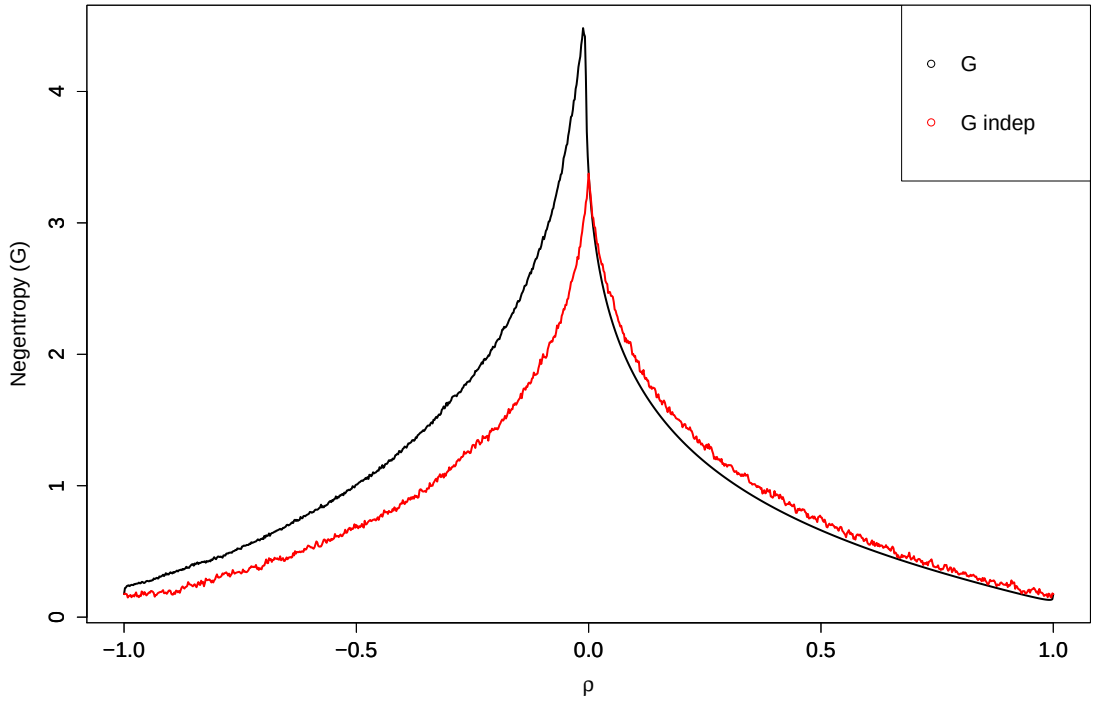



Figure 10.2: Negentropy of $\rho X + \sqrt{1-\rho^2}Y$ depending on whether the variables are independent.

*from $Y_\perp$, then for some reasonable rotation invariant function $\mu_{NG}$ that measures the non-Gaussinity of a distribution:*

$$\int_{-1}^{1} \mu_{NG}\left(\rho X + \sqrt{1-\rho^2}Y\right) \mathrm{d}\rho \geq \int_{-1}^{1} \mu_{NG}\left(\rho X + \sqrt{1-\rho^2}Y_\perp\right) \mathrm{d}\rho.$$

In other words, if we think of $\rho$ as a uniform random variable over $[-1,1]$ ($P \sim U[-1,1]$), we conjecture that:

$$\mathbb{E}_P\left[\mu_{NG}\left(PX + \sqrt{1-P^2}Y\right)\right] \geq \mathbb{E}_P\left[\mu_{NG}\left(PX + \sqrt{1-P^2}Y_\perp\right)\right].$$

Expressing it with expectations allows us to work mathematically with it better, although we have not obtained remarkable results for the mean up to now. However, we have obtained a proof for the maximum of the differences of the negentropy when the variables are uncorrelated:

**Theorem 18.** *Given two uncorrelated random variables $X$ and $Y$:*

$$\max_{\rho\in[-1,1]} \left(G(\rho X + \sqrt{1-\rho^2}Y) - G(\rho X + \sqrt{1-\rho^2}Y_\perp)\right) \geq 0,$$

*where $Y_\perp$ is a random variable with the same distribution as $Y$ but independent of $X$.*

*Proof.* We can change $\rho = \cos(\theta)$, and then we have to prove:

$$\max_{\theta\in[0,\pi]} \left(G(X\cos\theta + Y\sin\theta) - G(X\cos\theta + Y_\perp\sin\theta)\right) \geq 0.$$

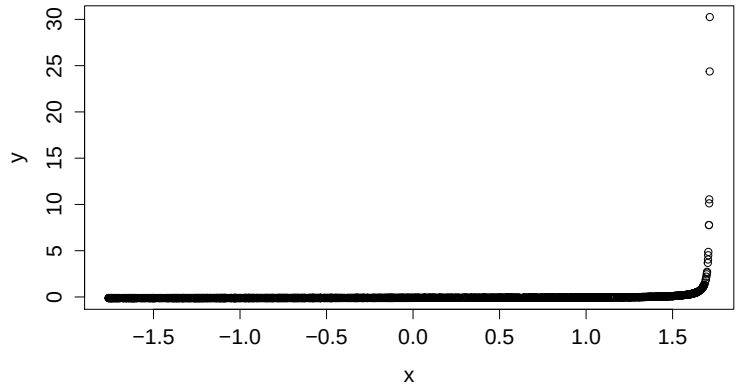



Figure 10.3: Counterexample where $X \sim U[-1,1]$ and $Y = \frac{1}{1-X} \sim Pareto(1,1)$.

To simplify the notation, we will denote the projections as:

$$\begin{aligned} Z(\theta) &= X\cos\theta + Y\sin\theta, \\ Z_\perp(\theta) &= X\cos\theta + Y_\perp\sin\theta. \end{aligned}$$

Since $X$ and $Y$ are uncorrelated, the variances of $Z(\theta)$ and $Z_\perp(\theta)$ do not depend on $\theta$, so this expression can be rewritten as:

$$\begin{aligned} \max_{\theta\in[0,\pi]} \left(G(Z(\theta)) - G(Z_\perp(\theta))\right) &= \max_{\theta\in[0,\pi]} \left([\log(2\pi e) - H(Z(\theta))] - [\log(2\pi e) - H(Z_\perp(\theta))]\right) \\ &= \max_{\theta\in[0,\pi]} \left(H(Z_\perp(\theta)) - H(Z(\theta))\right) \geq 0. \end{aligned}$$

Which is equivalent to:

$$\min_{\theta\in[0,\pi]} \left(H(Z(\theta)) - H(Z_\perp(\theta))\right) \leq 0. \tag{10.6}$$

We will analyse the term with the independent variables. We can apply the entropy power inequality, which says that, for two independent random variables $W$ and $W_\perp$:

$$e^{2H(W+W_\perp)} \geq e^{2H(W)} + e^{2H(W_\perp)}.$$

We will use also the following property of the entropy, for a $d$-dimensional random variable $W$ and a matrix $a$:

$$H(aW) = H(W) + \log\det(a).$$

Therefore we can establish a bound for the entropy:

$$\begin{aligned} H(X\cos\theta + Y_\perp\sin\theta) &= \frac{1}{2}\log\left(e^{2H(X\cos\theta + Y_\perp\sin\theta)}\right) \\ &\geq \frac{1}{2}\log\left(e^{2H(X\cos\theta)} + e^{2H(Y_\perp\sin\theta)}\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\log\left(e^{2H(X)+2\log\cos\theta} + e^{2H(Y_\perp)+2\log\sin\theta}\right)\\
&= \frac{1}{2}\log\left(e^{2H(X)}\cos^2\theta + e^{2H(Y_\perp)}\sin^2\theta\right)\\
&\geq \min_{\theta\in[0,\pi]}\frac{1}{2}\log\left(e^{2H(X)}\cos^2\theta + e^{2H(Y_\perp)}\sin^2\theta\right)\\
&= \min(H(X), H(Y)).
\end{aligned}
$$

The last minimum has been computed as usual. Since this inequality holds for all $\theta \in [0, \pi]$, and there is no $\theta$ is the final expression, we can write:

$$\min_{\theta\in[0,\pi]} H(X\cos\theta + Y_\perp \sin\theta) \geq \min(H(X), H(Y)).$$

Besides, it is clear that the inverse inequality also holds, therefore:

$$\min_{\theta\in[0,\pi]} H(X\cos\theta + Y_\perp \sin\theta) = \min(H(X), Y(Y)).$$

Therefore, for the dependent variables $(X, Y)$ we get:

$$\min_{\theta\in[0,\pi]} H(X\cos\theta + Y\sin\theta) \leq \min(H(X), Y(Y)) = \min_{\theta\in[0,\pi]} H(X\cos\theta + Y_\perp\sin\theta).$$

Since the right-hand term does not depend on $\theta$ and $H(Z_\perp(\theta)) \geq \min(H(X), Y(Y))$ for all $\theta \in [0, \pi]$, this expression is equivalent to Equation 10.6, that we want to prove. □

In fact, we could have used the interval $\left[0, \frac{\pi}{2}\right]$ in the previous proof, instead of $[0, \pi]$. The interval $\theta \in \left[0, \frac{\pi}{2}\right]$ corresponds to $\rho \in [0, 1]$. We have seen empirically (Figure 10.2) that the strict inequality does not hold in this interval. Therefore, we have decided to keep the complete interval $[0, \pi]$, since our goal is to prove the strict inequality.

Now using this conjecture we can define a new way to measure the non-linear dependencies of the sample as the mean difference between the negentropies. However, since we do not know whether the conjecture is true or not, we will use the absolute value of the differences to define the test. In most of the cases this absolute value will not change the result, because the previous conjecture is true for the negentropy in the commonly used relations between variables. However we ensure with it that the $NLD$ term for general distributions is positive.

$$
\begin{aligned}
NLD^*(X,Y) &= \int_{-1}^{1}\left|G(\rho X + \sqrt{1-\rho^2}Y) - G(\rho X + \sqrt{1-\rho^2}Y_\perp)\right|\mathrm{d}\rho\\
&= \int_{-1}^{1}\Bigg|\left(\frac{1}{2}\log\left(2\pi e(1+2\rho\sqrt{1-\rho^2}\rho_{XY})\right) - H(\rho X + \sqrt{1-\rho^2}Y)\right)\\
&\qquad - \left(\frac{1}{2}\log\left(2\pi e(1+2\rho\sqrt{1-\rho^2}\rho_{XY_\perp})\right) - H(\rho X + \sqrt{1-\rho^2}Y_\perp)\right)\Bigg|\mathrm{d}\rho
\end{aligned}
$$

$$= \int_{-1}^{1} \Big| H(\rho X + \sqrt{1-\rho^2} Y_\perp) - H(\rho X + \sqrt{1-\rho^2} Y) + \frac{1}{2} \log\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right) \Big| \mathrm{d}\rho.$$

Therefore we can define a new independence measure by replacing $NLD$ with $NLD^*$ in the definition of the test for the Gaussian marginals.

**Definition 39.** *We define a measure of independence between the random variables $X$ and $Y$ named* ***non-Gaussianity independence measure*** *as:*

$$I^*(X,Y) = NLD^*(X,Y) + LD(X,Y) \tag{10.7}$$

$$= \int_{-1}^{1} \Big| H(\rho X + \sqrt{1-\rho^2} Y_\perp) - H(\rho X + \sqrt{1-\rho^2} Y) + \frac{1}{2} \log\left(1 + 2\rho\sqrt{1-\rho^2}\rho_{XY}\right) \Big| \mathrm{d}\rho - \frac{1}{2}\log(1 - |\rho_{XY}|).$$

Although this measure lacks a complete theoretical foundation, the underlying assumptions hold in empirically. As we have mentioned before in this chapter, we have used the negentropy to measure the Gaussianity of the projections, but there are other Gaussianity measures which could be used instead. In the next section we will apply other measures to define alternative ways to measure independence,for which we could try to prove the conjecture of the present section.

## 10.4 Other non-Gaussianity measures

In this section we will develop two different versions of our non-Gaussianity independence test by using other measures for the non-linear dependencies. In the original definition of the test we measure the difference between the negentropies, but this measure is not smooth and has a lot of peaks, specially for small samples, since the differential entropy involved in its expression is difficult to estimate. Now we will change this quantity by using the two homogeneity tests previously presented in this work, which have better properties. That is, we will measure the distance to the Gaussian distribution using the MMD and the energy distance.

The first attempt would be to use directly energy distance or MMD instead of negentropy in the original formula of Equation (10.2). That is, comparing the joint distribution of the random variables $X$ and $Y$ with a multivariate Gaussian, without using unidimensional projections. But, for the set of problems that we have tested, this option performs worse than the other option based on projections.

Therefore, we will use MMD and energy distance instead of negentropy in the term $NLD^*$ of the definition of the measure. But using these measures requires taking into account some technical issues. The main problem with using different non-Gaussianity measures in $NLD^*$ is that we have to sum it with the linear dependencies, $LD$. Therefore we have to adjust the scale of the new measure. If we do not do this, it could be that one of the two quantities was

negligible. For example, the scale of MMD is ten times smaller than the scale of negentropy. Therefore if we sum the MMD with $LD$, the measure of the linear dependencies will be so large compared to the MMD that we will not detect the non-linear dependencies. Conversely for the linear dependencies if the scale of the new non-linear measure is very large in comparison. Then the new tests would be, for $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$:

$$\begin{aligned} I_{\mathcal{E}}(X,Y) &= C_{\mathcal{E}} \max_{\rho\in[-1,1]} \left|\mathcal{E}(Z,\rho X+\sqrt{1-\rho^2}Y)-\mathcal{E}(Z,\rho X+\sqrt{1-\rho^2}Y_\perp)\right| \\ &\quad -\frac{1}{2}\log(1-|\rho_{XY}|), \end{aligned} \tag{10.8}$$

$$\begin{aligned} I_{k}(X,Y) &= C_{k} \max_{\rho\in[-1,1]} \left|MMD(\mathcal{F},Z,\rho X+\sqrt{1-\rho^2}Y)-MMD(\mathcal{F},Z,\rho X+\sqrt{1-\rho^2}Y_\perp)\right| \\ &\quad -\frac{1}{2}\log(1-|\rho_{XY}|), \end{aligned} \tag{10.9}$$

where $Z \sim \mathcal{N}(0,1)$ and $\mathcal{F}$ is the unit ball of the RKHS $\mathcal{H}_k$. We have to determine the constants $C_{\mathcal{E}}$ and $C_k$ to adjust the scale. We have written here the maximum over $\rho$ instead of the mean, that we used with the negentropy, but both methods are possible. In the following chapter we will explore both possibilities, because neither of them is uniformly better than the others.

The first approach is to adjust the scales of the differences empirically, and this is the option chosen in the following practical chapter. In particular, we have stablish a value for the constant for each particular sample. As the real value of the scale is given by the negentropy, we will use it to obtain the constant, although it could introduce a bias in some cases when the negentropy is not able to detect the dependences. However in practice this adjustment performs well. In particular, we use the median of the quotients:

$$\frac{G(\rho X+\sqrt{1-\rho^2}Y)}{\mathcal{E}(Z_\rho,\rho X+\sqrt{1-\rho^2}Y)} \quad \textit{and} \quad \frac{G(\rho X+\sqrt{1-\rho^2}Y)}{MMD(\mathcal{F},Z_\rho,\rho X+\sqrt{1-\rho^2}Y)},$$

where $Z_\rho$ is a Gaussian variable with the same mean an variance of the projection $\rho X + \sqrt{1-\rho^2}Y$. We have checked empirically that this selection of the constants do not have problems under the null hypothesis. It is necessary to adjust the constants for every sample to avoid problems when the marginal distributions of the variables are Gaussian.

Along the experiments we observed that the shape of the negentropy is similar to the shapes of the MMD and the energy distance. This could point out that there exists a theoretical relationship between them, as it happens between the MMD and the energy distance. In Figure 10.4 we can see an example of this similitude with both methods, where we used a quadratic relationship with sample size 300. In particular, we take $X \sim U([0,1])$ and $Y = X^2$. We scale the variables to have zero mean and unit variance and whiten the data to remove the linear dependences. We have represented $G(Z,\rho X+\sqrt{1-\rho^2}Y)$ and $G(Z,\rho X+\sqrt{1-\rho^2}Y_\perp)$ in the first subfigure, where $Z$ is a Gaussian distribution, and the equivalent functions for the energy distance and MMD in the other two subfigures. We have used a standard Gaussian kernel for the MMD. We can also observe that, indeed, negentropy is not smooth, in contrast to MMD or energy distance. Moreover, it is clear that these last two methods are connected.

But we have not explored this option yet, so we still do not know if there exists really a theoretical relationship between the measures. In this example we have used a Gaussian kernel

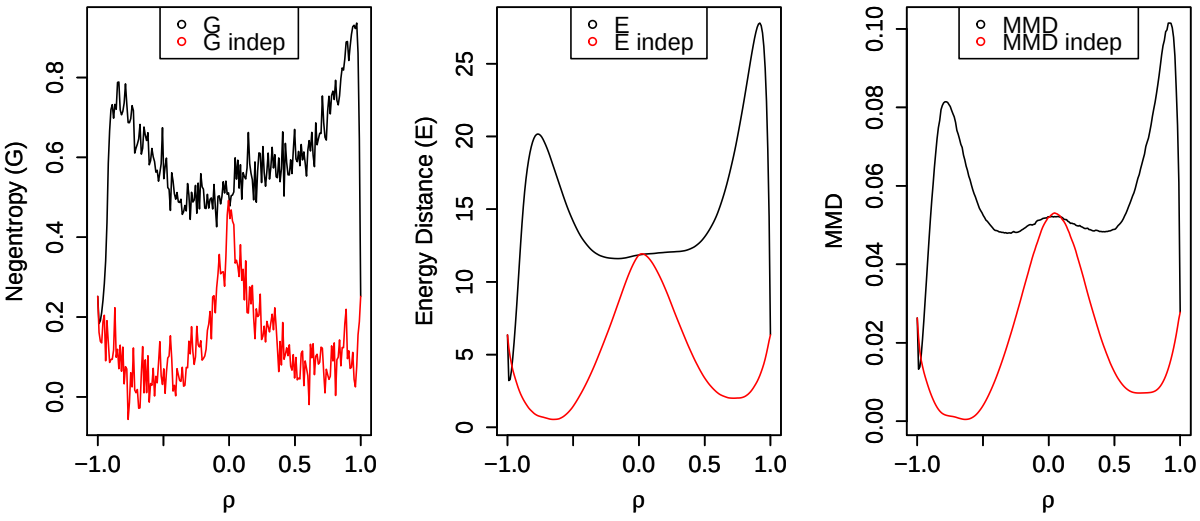



Figure 10.4: Comparison between negentropy, energy distance and MMD for the non-linear dependence obtained by whitening a quadratic one ($Y = X^2$).

with unit width, because adjusting the parameter entails analyse the samples for every value of $\rho$. Besides it is enough for seeing the similar shape of the quantities.

Nevertheless whitening the data can help us to solve our original problem. If we remove the linear dependencies from our samples, then $LD(X, Y) = 0$, so we have not to worry about adjusting the scale of the nonlinear ones. But then we do not have a measure of independence, but a nonlinear dependence one. However, it can help us to determine which test is more powerful, since detecting the linear dependencies is an easy task. One option would be to make a two-step test, one step to detect the linear dependencies and, if it does not determine the dependence of the sample, applying a second step to detect the non-linear dependencies. But if we want to keep the significance level of the global test, we have to reduce the significance level in each step, which may lead to diminish the total power.

On another front, when studying results about random projections, we found a connection with these measures based on projections of the variables. We need some previous notation:

**Definition 40.** *Given a closed subspace $L$ of $\mathcal{H}$, we denote by $\mathbb{P}_L$ the projection of the probability distribution $\mathbb{P}$ onto $L$, namely the probability measure on $L$ given by:*

$$\mathbb{P}_L(B) = \mathbb{P}(\pi_L^{-1}(B)),$$

*where $B \subset L$ is a Borel set.*

Corollary 3.3 of [28] states:

**Corollary 3.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be Borel probability measures of $\mathbb{R}^d$, where $d \geq 2$. Assume that:*

- *The absolute moments $m_n = \int |x|^n \mathrm{d}\mathbb{P}(x)$ are finite and satisfy $\sum_{n\geq 1} m_n^{-\frac{1}{n}} = \infty$;*

- $\mathbb{P}_L = \mathbb{Q}_L$ *for infinite many hyperplanes* $L$ *in* $\mathbb{R}^d$.

*Then* $\mathbb{P} = \mathbb{Q}$.

For $d = 2$, the one-dimensional projections are the same as those we are using. To show independence, we can build the MMD statistic with a characteristic kernel, or energy distance, to determine whether the distributions of $\rho X + \sqrt{1-\rho^2}Y$ and $\rho X + \sqrt{1-\rho^2}Y_\perp$ are equal for a given value of $\rho \in [-1, 1]$. For example:

$$MMD^2_\rho(\mathcal{F}, X, Y) \equiv MMD^2\left(\mathcal{F}, \rho X + \sqrt{1-\rho^2}Y, \rho X + \sqrt{1-\rho^2}Y_\perp\right).$$

Therefore, using the previous corollary, the random variables $X$ and $Y$ are independent if and only if $MMD^2_\rho(\mathcal{F}, X, Y) = 0$ for infinitely many values of $\rho$.

This is not exactly what we are doing. In fact, using the reverse triangle inequality:

$$\begin{aligned}\left|MMD^2\left(\mathcal{F}, Z, \rho X + \sqrt{1-\rho^2}Y_\perp\right) - MMD^2\left(\mathcal{F}, \rho X + \sqrt{1-\rho^2}Y_\perp, Z\right)\right| \\ \leq MMD^2\left(\mathcal{F}, \rho X + \sqrt{1-\rho^2}Y, \rho X + \sqrt{1-\rho^2}Y_\perp\right),\end{aligned}$$

where $Z$ is a Gaussian variable. However, this result is interesting by itself, since it also involves characterising independence of the variables through analysing their projections. Moreover, with this measure it is not necessary to check the property for all $\rho$, but only for a infinity number of it.

We have presented in this section several possible improvements of the original test, defined in the previous section. We can not choose one option empirically, because some of them are better for some problems than the others. The only option discarded is the maximum of the negentropy, because it performs clearly worse that the mean for the tested problems. Therefore, in the following chapter we will compare tests based on these options with other state-of-the-art independence tests.

# Chapter 11

# Comparison between methods

In this chapter we will carry out some experiments on benchmark data from [29] and [30]. In the first group of experiments we assess the robustness of the tests on independence when noise is injected in the data. The goal of the second group of experiments is to measure the power of the test as a function of sample size. We compare the independence tests introduced in this work, and other state-of-the-art test, which are introduced in the original papers of the experiments.

## 11.1 Approximate Correntropy Independence

In this section we introduce an independence criterion based on a generalization of the concept of correlations for non-linear projections of the random variables, taken from [30]. To this end we will use a real valued, continuous, symmetric, non-negative definite and translationally invariant kernel $k$. This will allow us to use Bochner's Theorem.

**Definition 41.** *Given two random variables $X$ and $Y$, their **correntropy** is defined as:*

$$V(X,Y) = \mathbb{E}k(X-Y) = \int k(x-y)\mathrm{d}\mathbb{P}_{XY}(x,y),$$

*where $\mathbb{P}_{XY}$ is the joint probability distribution.*

Correntropy is a generalization of the concept of correlation that extracts not only second order information, but also higher order moments of the joint distribution.

**Definition 42.** *Given two random variables $X$ and $Y$, their **centered correntropy** is defined as:*

$$U(X,Y) = \mathbb{E}_{XY}k(X-Y) - \mathbb{E}_X\mathbb{E}_Y k(X-Y) = \int k(x-y)(\mathrm{d}\mathbb{P}_{XY}(x,y) - \mathrm{d}\mathbb{P}_X(x)\mathbb{P}_Y(y)),$$

*where $\mathbb{P}_X$ and $\mathbb{P}_Y$ are the marginal distributions of $X$ and $Y$, respectively, and $\mathbb{P}_{XY}$ their joint distribution.*

Correntropy and centered correntropy exhibit similar properties as correlations and covariances. Zero centered correntropy does not imply independence. To characterize independence we need a more general quantities:

**Definition 43.** *Given two random variables $X$ and $Y$, the* ***parametric centered correntropy*** *is defined as:*

$$\begin{aligned} U_{a,b}(X,Y) &= \mathbb{E}_{XY}k(aX+b-Y) - \mathbb{E}_X\mathbb{E}_Y k(aX+b-Y) \\ &= \int k(ax+b-y)(\mathrm{d}\mathbb{P}_{XY}(x,y) - \mathrm{d}\mathbb{P}_X(x)\mathbb{P}_Y(y)), \end{aligned}$$

*where $a, b \in \mathbb{R}$ and $a \neq 0$.*

It can be shown that given two random variables $X$ and $Y$, the variables are independent if and only if $U_{a,b}(X,Y) = 0\ \forall a, b \in \mathbb{R}$. The original independence test given in [30] is formulated in terms of the supremum of the absolute value of the parametric centered correntropy. The search of this supremum involves evaluating this quantity on a grid in $\mathbb{R}^2$, which is computationally expensive. We will use a simplified version, also introduced in [30].

**Definition 44.** *Given two random variables $X$ and $Y$, the* ***Approximate Correntropy Independence (ACI)*** *measure is defined as:*

$$\gamma(X,Y) = \max(|U(X,Y)|, |U(-X,Y)|) = \max(|U_{1,0}(X,Y)|, |U_{-1,0}(X,Y)|).$$

This is not really a measure of independence, however this test characterizes independence under the assumption that that the joint density is a mixture of Gaussian distributions with the same mean and Gaussian marginals.

Instead of using the projections $aX + b$, we have developed a novel generalization that uses more intuitive and general ones. Consider the set of one-dimensional projections of a $d$-dimensional random variable $X$:

$$Z(w) = w^\top X,$$

where $w \in \mathbb{R}^d$ is a unit vector, whose norm is equal to one. We first define a homogeneity test:

**Proposition 26.** *Given two distributions $\mathbb{P}$ and $\mathbb{Q}$, $\mathbb{P} = \mathbb{Q}$ if and only if:*

$$U_{w,b}(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} k(w^\top x + b)(\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x)) = 0,$$

*for all $w \in \mathbb{R}^d$ of unit norm and $b \in \mathbb{R}$.*

*Proof.* This proof is similar to the one for Lemma 1 of [30].

($\Longrightarrow$) It is clear that if $\mathbb{P} = \mathbb{Q}$, then $U_{w,b}(\mathbb{P},\mathbb{Q}) = 0$ for all $w$ and $b$.

($\Longleftarrow$) We can rewrite $U_{w,b}(\mathbb{P},\mathbb{Q})$, using Bochner's and Fubini's Theorems, as:

$$\begin{aligned} U_{w,b}(\mathbb{P},\mathbb{Q}) &= \int_{\mathbb{R}^d} k(w^\top x + b)(\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x)) \\ &= \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}} e^{-i\alpha(w^\top x + b)}\, \mathrm{d}\mu(\alpha) \right) (\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x)) \end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}} e^{-i\alpha b} \left( \int_{\mathbb{R}^d} e^{-i\alpha w^\top x} \left( \mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x) \right) \right) \mathrm{d}\mu(\alpha) \\
&= \int_{\mathbb{R}} e^{-i\alpha b} \left( \int_{\mathbb{R}^d} e^{-iw_\alpha^\top x} \left( \mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x) \right) \right) \mathrm{d}\mu(\alpha),
\end{aligned}
$$

where $\mu$ is a finite positive measure and $w_\alpha = \alpha w$. If $U_{w,b}(\mathbb{P}, \mathbb{Q}) = 0$ for all $b \in \mathbb{R}$, by the properties of the Fourier transform:

$$
\int_{\mathbb{R}^d} e^{-iw_\alpha^\top x} \left( \mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x) \right) = 0, \quad \forall w_\alpha \in \mathbb{R}^d.
$$

This expression can be written in terms of the characteristic functions of the distributions:

$$
\begin{aligned}
\int_{\mathbb{R}^d} e^{-iw_\alpha^\top x} \left( \mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x) \right) &= \int_{\mathbb{R}^d} e^{-iw_\alpha^\top x} \, \mathrm{d}\mathbb{P}(x) - \int_{\mathbb{R}^d} e^{-iw_\alpha^\top x} \, \mathrm{d}\mathbb{Q}(x) \\
&= \Phi_{\mathbb{P}}(w_\alpha) - \Phi_{\mathbb{Q}}(w_\alpha) = 0, \quad \forall w_\alpha \in \mathbb{R}^d.
\end{aligned}
$$

The equality of the characteristic functions implies equality in distribution $\mathbb{P} = \mathbb{Q}$. □

For the two dimensional casi, this expression is similar to the parametric centered correntropy, $\mathbb{P}(x, y) = \mathbb{P}_{XY}(x, y)$ and $\mathbb{Q}(x, y) = \mathbb{P}_X(x)\mathbb{P}_Y(y)$. Then the projections used are of the type:

$$
\cos\theta x + \sin\theta y + b,
$$

It is possible to formulate a different test in terms of one dimensional projections only:

**Proposition 27.** *Given two distributions $\mathbb{P}$ and $\mathbb{Q}$, $\mathbb{P} = \mathbb{Q}$ if and only if:*

$$
U_w(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(w^\top (x - x'))(\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x))(\mathrm{d}\mathbb{P}(x') - \mathrm{d}\mathbb{Q}(x')) = 0,
$$

*for all unitary $w \in \mathbb{R}^d$.*

*Proof.* This proof combines ideas of [30] and the tests based on embeddings in RKHS.

($\Longrightarrow$) It is clear that if $\mathbb{P} = \mathbb{Q}$, then $U_w(\mathbb{P}, \mathbb{Q}) = 0$ for all $w$.

($\Longleftarrow$) Using Bochner's and Fubini's theorems:

$$
\begin{aligned}
U_w(\mathbb{P}, \mathbb{Q}) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(w^\top (x - x'))(\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x))(\mathrm{d}\mathbb{P}(x') - \mathrm{d}\mathbb{Q}(x')) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}} e^{-i\alpha w^\top (x - x')} \mathrm{d}\mu(\alpha) \right) (\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x))(\mathrm{d}\mathbb{P}(x') - \mathrm{d}\mathbb{Q}(x')) \\
&= \int_{\mathbb{R}} \left[ \int_{\mathbb{R}^d} e^{-i\alpha w^\top x} \left( \int_{\mathbb{R}^d} e^{-i\alpha w^\top x'} (\mathrm{d}\mathbb{P}(x') - \mathrm{d}\mathbb{Q}(x')) \right) (\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x)) \right] \mathrm{d}\mu(\alpha)
\end{aligned}
$$

$$= \int_{\mathbb{R}} \left( \int_{\mathbb{R}^d} e^{-i\alpha w^\top x} (\mathrm{d}\mathbb{P}(x) - \mathrm{d}\mathbb{Q}(x)) \right) \left( \int_{\mathbb{R}^d} e^{-i\alpha w^\top x'} (\mathrm{d}\mathbb{P}(x') - \mathrm{d}\mathbb{Q}(x')) \right) \mathrm{d}\mu(\alpha)$$

$$= \int_{\mathbb{R}} \left| \Phi_{\mathbb{P}}(w_\alpha) - \Phi_{\mathbb{Q}}(w_\alpha) \right|^2 \mathrm{d}\mu(\alpha),$$

where $\mu$ is a finite positive measure, $w_\alpha = \alpha w$ and $\Phi_{\mathbb{P}}(\cdot)$ and $\Phi_{\mathbb{Q}}(\cdot)$ are the characteristic functions of the distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively. If $U_w(\mathbb{P}, \mathbb{Q}) = 0$ for all $w \in \mathbb{R}^d$ of unit norm, then:

$$\Phi_{\mathbb{P}}(w_\alpha) = \Phi_{\mathbb{Q}}(w_\alpha), \quad \forall w_\alpha \in \mathbb{R}^d.$$

Since the characteristic functions are equal, the corresponding distributions are also equal. □

For $d = 2$, we can take $\mathbb{P}(x, y) = \mathbb{P}_{XY}(x, y)$ and $\mathbb{Q}(x, y) = \mathbb{P}_X(x)\mathbb{P}_Y(y)$ to define a independence test. Since we are using projections of the variables, we can use some of the results of random projections, like the one presented in the previous chapter, to reduce the range of $w$ for which we have to check the test.

## 11.2 Randomized Dependence Coefficient

In this section we will introduce another independence coefficient introduced in [29]. This coefficient is a scalable estimator with the same properties as the following one:

**Definition 45.** *Given two random variables $X$ and $Y$, the **Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (HGR)** is the supremum of Pearson's correlation coefficient $\rho$ over all Borel-measurable functions $f$ and $g$ of finite variance:*

$$hgr(X, Y) = \sup_{f,g} \rho(f(X), g(Y)).$$

The HGR correlation coefficient is difficult to use in practice, because it is the supremum over an infinite-dimensional space. Instead we define the RDC, which measures the dependence in terms of the largest canonical correlation between $n$ random non-linear projections of the copula transformation of the variables.

**Definition 46.** *Given a $d$-dimensional random vector $X = (X_1, \ldots, X_d)$ with continuous marginal cumulative distribution functions $F_1, \ldots, F_d$, the vector $U = (U_1, \ldots, U_d) = (F_1(X_1), \ldots, F_d(X_d))$ whose marginals are $U[0, 1]$, is known as the **copula transformation**.*

The first step of the RDC method is to compute the copula transformation using the empirical cumulative distribution function. The second one is to augment these empirical transformations with non-linear projections, so that linear methods can be used to capture non-linear dependences in the original data. The choice of the non-linear projections $\phi$ is equivalent to the choice of the spaces of features. In the original paper, sine and cosine projections are used:

$$\phi(X) = \left( \cos(w^\top X + b),\ \sin(w^\top X + b) \right),$$

where $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. In particular, in [29] these parameters are samples of the random variables $W \sim \mathcal{N}(0, sI)$ and $B \sim U[-\pi, \pi]$, where $I$ is the identity matrix. Choosing $W$ to be Gaussian is analogous to the use of a Gaussian kernel for the projections. The parameter $s$ plays the role of the kernel width. Let the set of $n$ random projections:

$$\Phi_{n,s}(X) = \left(\phi(w_i^\top X + b_i),\ \ldots,\ \phi(w_n^\top X + b_n)\right).$$

Canonical Correlation Analysis is the calculation of pairs of directions $(\alpha, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$ such that the projections $\alpha^\top X$ and $\beta^\top Y$, of the $d$-dimensional random variables $X$ and $Y$, are maximally correlated. These projections are named canonical, therefore:

**Definition 47.** *Given two $d$-dimensional random variables $X$ and $Y$, the **canonical correlations** between them are the correlations between the random projections $\alpha^\top X$ and $\beta^\top Y$, for $(\alpha, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$.*

The Randomized Dependence Coefficient is defined in terms of the canonical correlations of the random projections:

**Definition 48.** *Given two $d$-dimensional random variables $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$, and parameters $n \in \mathbb{N}$ and $s \in \mathbb{R}$ $(n, s > 0)$, the **Randomized Dependence Coefficient** between the variables is defined as:*

$$rdc(X, Y; n, s) = \sup_{\alpha,\beta}\ \rho\left(\alpha^\top \Phi_{n,s}(\mathbb{P}(X)),\ \beta^\top \Phi_{n,s}(\mathbb{Q}(Y))\right).$$

## 11.3 Experiments

In this section we present the result of experiment in which we compare the power of the tests introduced in Chapter 10 with state-of-the-art independence tests. These experiments have been adapted from [29] and [30].

Some of the methods that we will test require to adjust parameters. For the HSIC test we will use a Gaussian kernel. The width of the kernel are set to be the median of $\|(X, Y) - (X', Y')\|_d^2$, where $(X', Y')$ is an independent copy of $(X, Y)$. For the RDC we will use ten random projections. The random sampling parameters $(s_X, s_Y)$ are set to the median of $\|X - X'\|_d^2$ and $\|Y - Y'\|_d^2$, where $X'$ and $Y'$ are independent copies of $X$ and $Y$, respectively. These parameters are computed independently for each of the two random samples.

The first group of experiments consist in computing the power of the methods as a function of the level noise injected. We will use nine different dependences between the variables, the first eight are taken from [29], the last one is a mixture of two crossed bivariate Gaussian. These dependences can be seen in Figure 11.1. For the first eight data sets, the variable $X$ follow a distribution $U([0, 1])$, and the variable $Y$ is a function of $X$. In the last example, both marginal distributions are Gaussian. Before applying the tests the data are standardized so that they have zero mean and unit variance. Then we inject additive Gaussian noise to the variable $Y$ with standard deviation varying from $0.1a$ to $3a$, where a is problem dependent. We will use samples of size 320. In particular, the value $a$ for each problem is:

| Problem | a |
|---|---|
| Linear | 1 |
| Parabolic | 1 |
| Cubic | 10 |
| Sin($4\pi$ x) | 2 |
| Sin($16\pi$ x) | 1 |
| Fourth root | 1 |
| Circle | 0.25 |
| Step | 5 |
| 2D cross Gaussian | 1 |

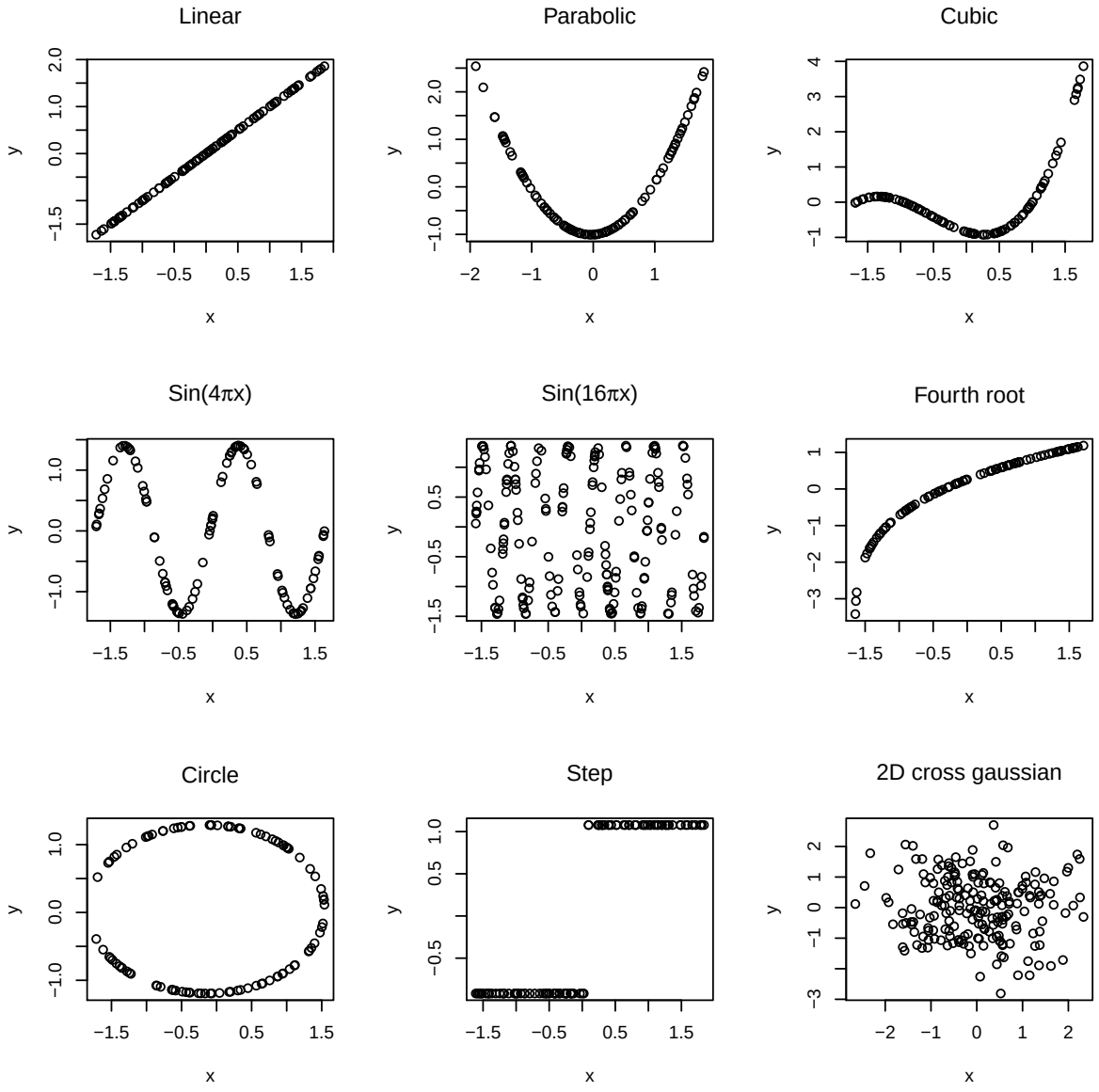



Figure 11.1: Data dependences for the first experiment.

To simulate using the null hypothesis ($H_0$, the variables are independent) we will generate

a new sample of $X$, since we know its real distribution. First we use 500 independent samples generated under $H_0$ to compute the value of the statistic beyond which the null hypothesis is rejected, at a signification level $\alpha = 0.05$. Then we simulate another 500 dependent samples and ascertain whether they are above the rejection threshold or not. In Figure 11.2 we can see the powers of the nine methods introduced in this work. We denote by "Imean" the non-Gaussianity test when we are taking the mean of the differences of the negentropy over $\rho$ (Equation 10.7). "Emean" and "Emax" denote the methods when we use energy distance to compute the non-Gaussianity of the projections, taking the mean and the maximum of the differences respectively (Equation 10.8). In the same way, "MMDmean" and "MMDmax" denote the methods where MMD are used instead of negentropy (Equation 10.9).

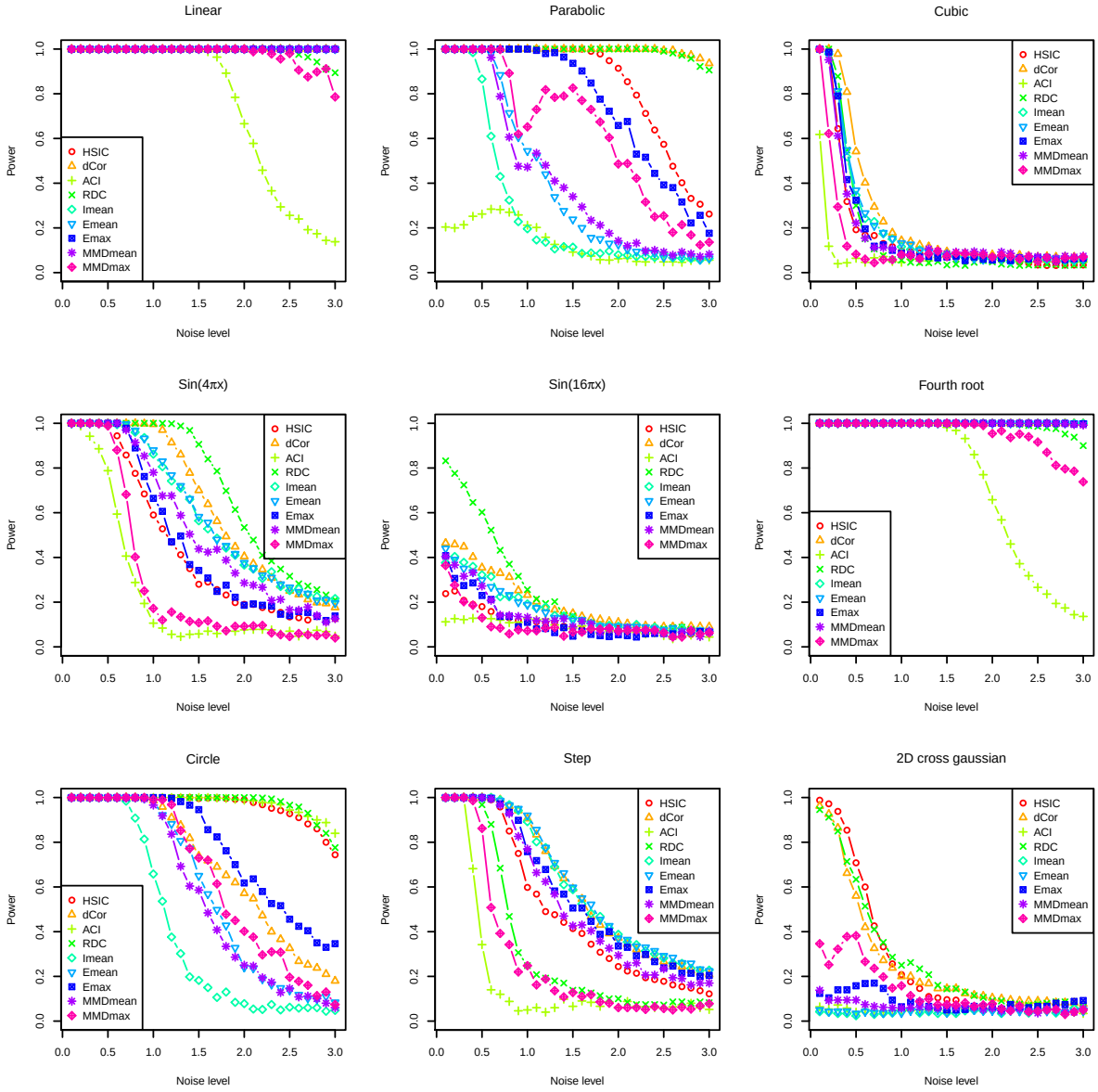



Figure 11.2: Power of the methods when increasing the noise level.

In general, the distance covariance method performs well, except for the circle relation. In

this case, the best method is RDC, which also has good performance in other cases. It seems that the circle dependence yields an unusual behaviour. In general, the tests proposed in this work perform reasonably well compared with the others, especially "Emax". For example in the linear case and in the fourth root. We can observe some fluctuations of the power when the maximums of the differences are used. This is due to the fact that the methods are quite sensitive to each particular sample when the maximums are taken and the scale constant affect more in these cases. We can also see that all the tests perform poorly for the sine of $16\pi$ and for the crossed Gaussian cases, so the dependences of these cases seems to be difficult to detect.

However, some of these examples have a strong linear dependence. Since we are interested in detecting non-linear dependences as well, we have repeated the same experiment but rotating the data to eliminate linear dependences. The data after the rotation (that is, the whitened data) can be seen in Figure 11.3. This whitening is made after scaling the data and before adding the noise. In figure 11.4 we can see the power of the methods for this whitened data.

In general, the power of the methods decreases. This indicates that the non-linear dependences are harder to detect. When whitening the linear relation, there are still non-linear dependences, but they are really difficult to detect. In fact, one of the components is equal to zero, and it exhibits a non-monotonic dependence when the noise level increases. The distance covariance can detect some dependences well, but its power has worsened. It seems that the RDC method is the less affected by the removing of the linear dependences. For example, for the circle dependence it is the best test, followed by two of the methods proposed in this work. It is interesting that our original method with the negentropy is able to detect some dependence in the $\sin(16\pi x)$ case, because all methods perform bad in this case when the linear dependences are present. This may be due to the fact that the chosen rotation increase the non-linear dependences that the negentropy detect better.

Next we will reproduce two different experiments taken from [30]. In the first one we will asses the power of the test as a function of the sample size. We will use a different set of data, also taken from [30]. Since the ACI method, proposed in that paper, assumes that the random variables are a mixture of Gaussians, all the variables are based on Gaussian distributions. The first one is a bivariate Gaussian with a correlation of 0.5. $(X, Y) \sim \mathcal{N}(0, \Sigma)$, where:

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

For the second example we generate a uniform random variable $Z \sim U[0, 2]$. The variables $X$ and $Y$ are the product of $Z$ and two independent standard Gaussian:

$$X = Z\widetilde{X} \text{ and } Y = Z\widetilde{Y},$$

where $\widetilde{X}, \widetilde{Y} \sim \mathcal{N}(0, 1)$. They are dependent because they share 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 $-0.8$, and probabilities 0.6, 0.2 and 0.2 respectively. The vector $(X, Y)$ has density:

$$0.6\mathcal{N}(0, \Sigma_1) + 0.2\mathcal{N}(0, \Sigma_2) + 0.2\mathcal{N}(0, \Sigma_3),$$

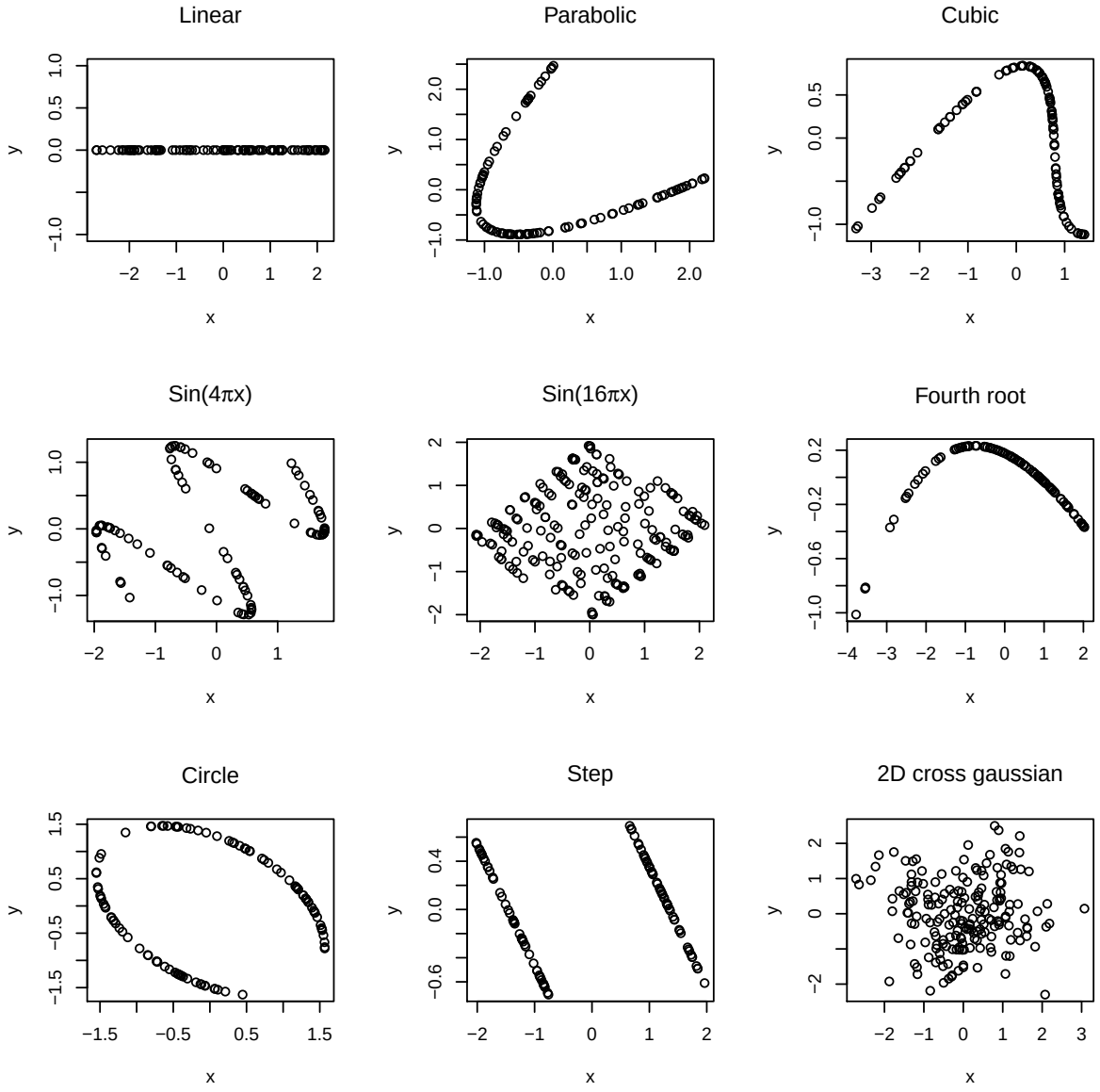



Figure 11.3: Whitened data dependences for the first experiment.

where

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}.$$

The variables of the last one are generated as a Gaussian random variable with correlation coefficient 0.8 and then multiply each variable with white Gaussian noise:

$$(X, Y) = Z\varepsilon, \text{ where } Z \sim \mathcal{N}(0, \Sigma_2) \text{ and } \varepsilon \sim \mathcal{N}(0, \Sigma_1).$$

These relations can be seen in Figure 11.5.

The power is measured for sample sizes $25, 50, 100, 150$ and 200. The results of these experiments are presented in Figure 11.6. These results differs from the ones in [30], where a value $\alpha = 0.1$ was used.

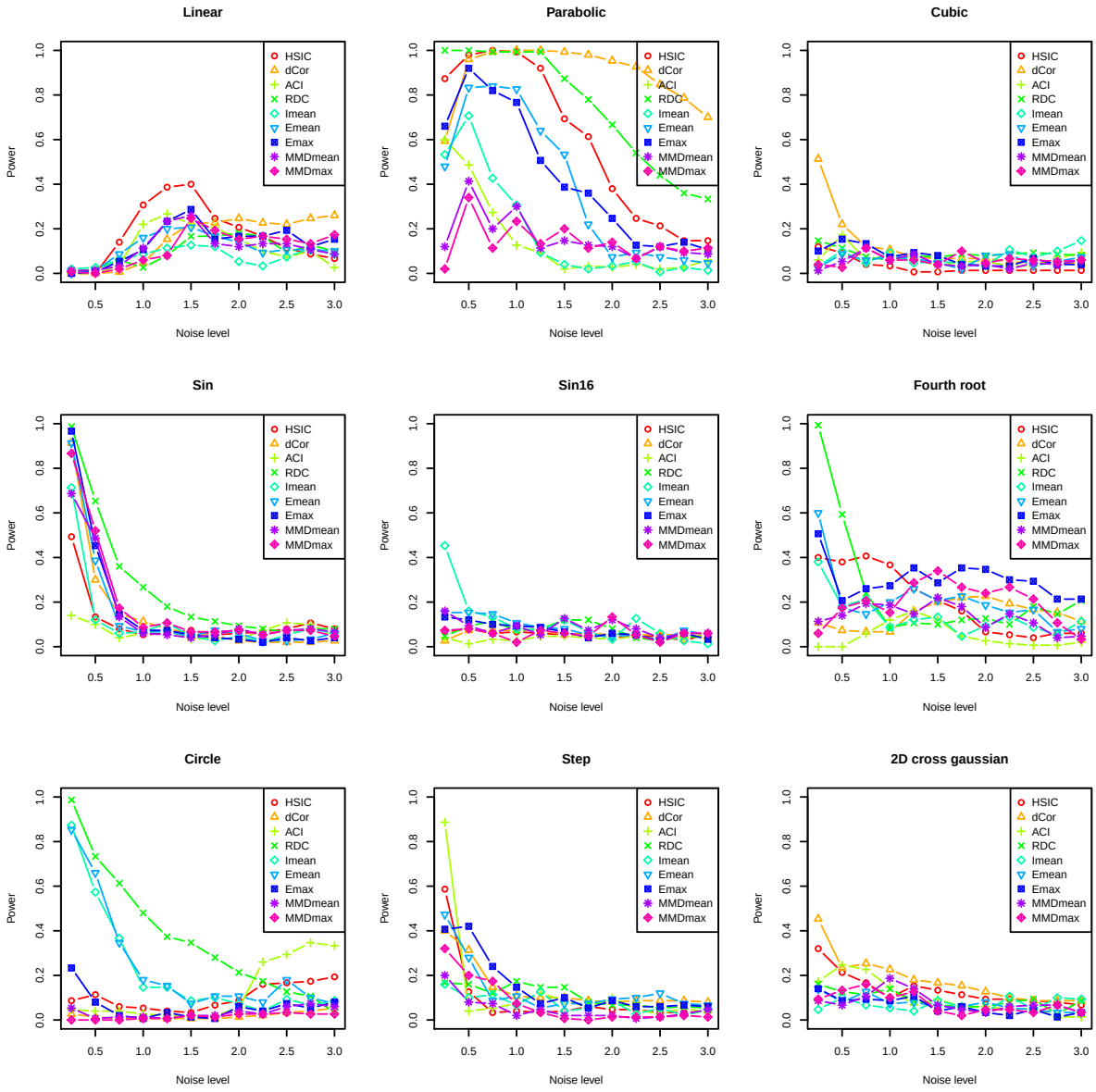



Figure 11.4: Power of the methods for whitened data when increasing the noise level.

It is clear that ACI method performs better for these examples than for the previous ones. The measures of non-Gaussianity in terms of the negentropy and the energy distance perform very well in the first example, but they perform extremely poor in the mixture. In the other two examples the performances of these methods are not so bad. The best overall results are obtained using the energy distance of the maximum.

We have carried out the experiments after removing the non-linear dependences by whitening the data, as in the previous experiments. In this case, all methods perform poorly an no conclusion can be drawn from the results.

Finally we have performed a final experiment following [30]. Two independent random variables, $X$ and $Y$, both having zero mean and unit variance. $X$ is a uniform random variable, $X \sim U[-\sqrt{3}, \sqrt{3}]$, whereas $Y$ is a combination of two uniform random variables each having

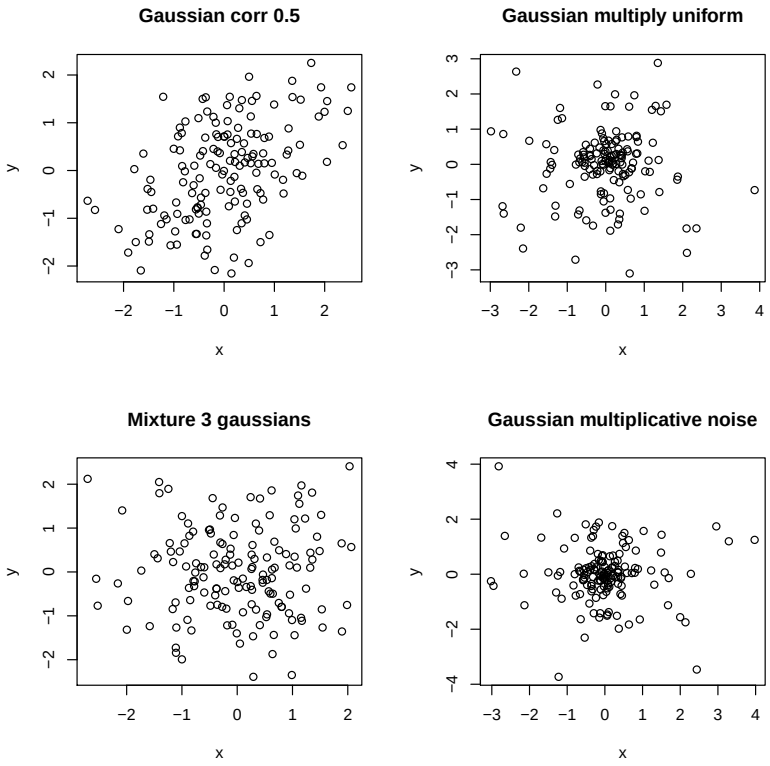



Figure 11.5: Data dependences for the second experiment.

equal probability of occurrence on disjoint support. $Y$ has density:

$$0.5U[-1,-0.5] + 0.5U[0.5,1].$$

We generate a new pair of random variables by rotating this random variable pair $(X,Y)$. The covariance matrix does not change and, thus, the correlation between the new variables stays zero. However, the dependence between them change. The new variables are independent if and only if the angle of rotation is zero and dependent otherwise. Therefore, we will test the power of the methods when increasing the rotation angle of the variables. We will use a sample size of 100. In Figure 11.7 we can see the rotated variables with an angle of $\frac{\pi}{20}$.

The procedure will be the same as for the previous experiments. We will check the power for angles $0, \frac{\pi}{70}, \frac{\pi}{40}, \frac{\pi}{20}, \frac{\pi}{12}$ and $\frac{\pi}{10}$. We have added an extra point with respect to the original experiment of the paper. The powers of the methods can be seen in Figure 11.8.

The best methods in this case are our tests with energy distance and MMD when taking the maximum of the differences. ACI method do not perform well, as well as distance covariance for small angles. It has no sense to whiten the data in this case, since whitening is rotating the data, and then we would break the aim of the experiment.

To sum up, it seems that distance covariance is a good method for general dependences, although it fails in some ones. However, although we have checked several examples, all of them are similar. For the first experiment all the variables $Y$ are a function of a uniform, and in the second and third experiments all the variables are variations of Gaussians. We have not

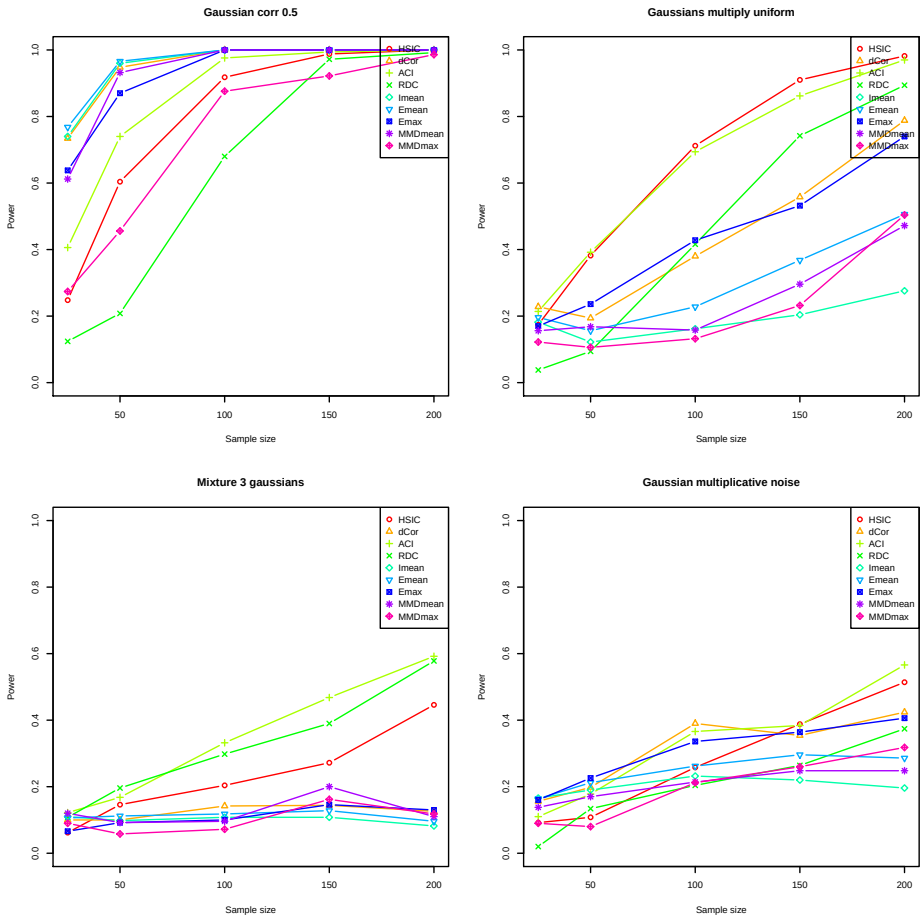



Figure 11.6: Power of the methods when increasing the sample size.

checked, for example, heavy tails variables when the dependences are in the tails. Maybe we have checked the tests, unconsciously, using a set of experiments that favours distance covariance. As for other methods, when there are some evidence that the data is a mixture of Gaussian distributions, ACI method or equivalents performs well. Besides, our methods perform well generally. They are clearly the best in the last example, when the dependences of the data are purely non-linear. They perform also well in the first experiment when the linear dependences are removed, specially in a difficult case as the circle, when most of the methods fail. When the linear dependences are present, they are on the average. However, we can not decide from this experiments if some of the proposed methods is better than the others.

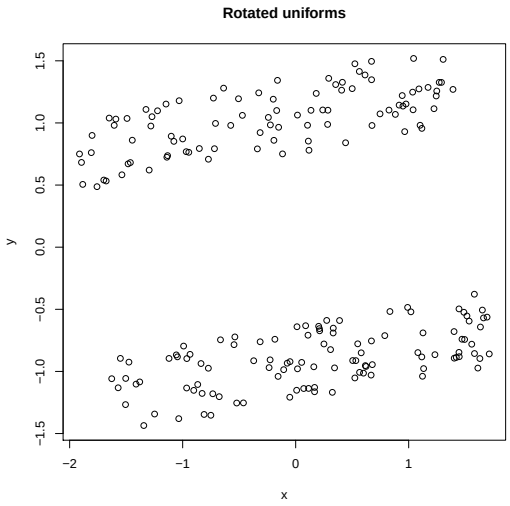



Figure 11.7: Data for the third experiment with a rotation angle of $\pi/20$.

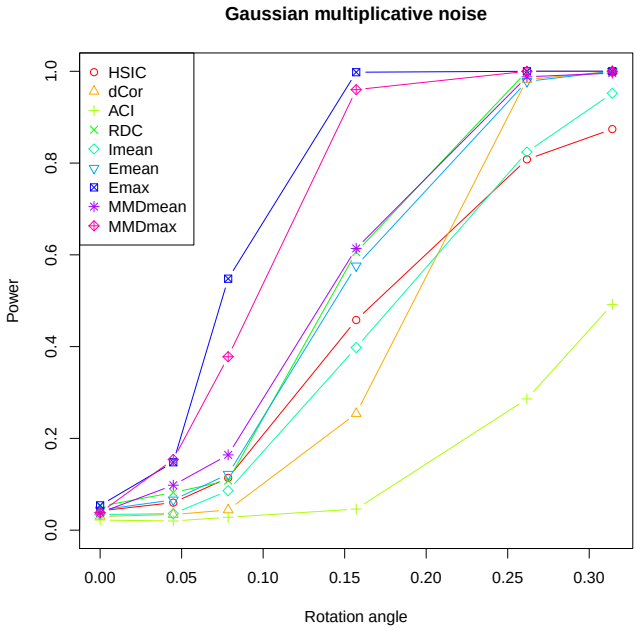



Figure 11.8: Power of the methods when increasing the rotation angle.

# Chapter 12

# Conclusions and future work

One of the aims of this work was to order and summarize the published research on homogeneity tests based on embeddings of probability distributions in an RKHS's and tests based on the energy distance. We have introduced and analysed the properties of different homogeneity and independence tests. Some of these independence test are novel in the literature. One of the main results presented is that both families of methods are actually the same. In fact, energy distance is a particular case of probability embedding test with a uncommon kernel. This particular embedding can be interpreted as the limit of embeddings using Laplacian kernels with large width.

A lot of efforts have been dedicated in the literature to choose the RKHS for the embedding, although it does not affect significantly the results of the test for general data sets. We have seen empirically that distance covariance performs better than HSIC in most of the cases that we have proven. Distance covariance is obtained with the same procedure as energy distance, so the connection between HSIC and distance covariance is similar to the one between MMD and energy distance. It is remarkably that the kernel that allows to write energy distance as a kernel embedding has not been widely studied, despite the empirical results.

Our main goal when developing the novel tests described was to detect non-linear dependences, which is still a difficult problem of practical interest. We have confirmed this fact, since most of the state-of-the-art tests fail when the linear dependences of the data are removed. We have seen that all the methods analysed in Chapters 3 to 7 are closely related. It seems empirically that our methods also present a similitude with the original versions of energy distance and MMD, however this is still an open question.

In the last part of the work we have assessed the power of all the methods using different dependences and procedures. The main conclusion of these experiments is that the performance of the tests is variable and strongly depends on the dependence structure of the data. One of the consequences of these results is that each of the novel tests presented in Chapter 10 is more powerful for some particular structures of the data. However, we can affirm that some of them perform better than the others when the dependence of the data is purely non-linear, which was the aim of their design.

There are mainly three research lines open at this time. The first one is to complete and settle down the meaning of the original version of the energy distance as a kernel embedding method. These relation can be also established between distance covariance and HSIC methods, although it is almost unexplored yet.

The other two lines are related with the novel independence tests. On the one hand, we have mentioned that the theoretical justification of these tests is given only for the case of Gaussian marginals. Therefore, it is necessary to provide full mathematical support for general marginal distributions. Some results have been obtained in this line, as Theorem 18, but there are still work to do. On the other hand, we have observed that random projections can be applied to improve and develop independence tests. Most of the results related to random projections have been formulated from the point of view of functional data. Therefore, we propose to explore this type of results to reuse them to characterize independence of random variables.

# Glossary

In general, the page number refers to the first occurrence of the term.

| TERM | DEFINITION | PAGE |
|---|---|---|
| Complete space. | Metric space where every Cauchy sequence has a limit in the space. | 8 |
| Pre-Hilbert space. | A non complete complex vector space on which there is an inner product. | 8 |
| Hilbert space. | A complex vector space on which there is an inner product and complete with respect to the distance function induced by this inner product. | 1 |
| Metric space | A space for which distances between all members of the space are defined | 8 |
| Topological space | A set $\mathcal{X}$ together with a collection of subsets of $\mathcal{X}$ such that (1) the empty set and $\mathcal{X}$ are open, (2) any union of open sets are open and (3) the intersection of any finite number of open sets is open. | 17 |
| $\mathcal{L}^p(\mathcal{X}), 1 < p < \infty$. | Functions such that $\|f\|_p \equiv \left(\int_{\mathcal{X}} \|f(x)\|^p \mathrm{d}x\right)^{\frac{1}{p}} < \infty$. | 5 |
| $\mathcal{L}^\infty(\mathcal{X})$. | Functions such that $\|f\|_\infty \equiv \sup_{x\in\mathcal{X}} \|f(x)\| < \infty$. (Supremum norm) | 16 |
| $\mathcal{C}(\mathcal{X})$. | Continuous functions with support on $\mathcal{X}$ | 11 |
| Mutually singular | Two measures $\mu$ and $\nu$ are called mutually singular if there exist a set $A$ such that $\mu(A) = 1$ and $\nu(A) = 0$. | 17 |
| Signed measure. | Generalization of measure by allowing it to have negative values. | 17 |

| | | |
|---|---|---|
| Lebesgue measure. | An extension of the classical notions of length and area to more complicated sets. Given an open set $S \subset \mathbb{R}$ formed by $k$ disjoint intervals $(a_k, b_k)$, i.e. $S = \bigcup_k (a_k, b_k)$, the Lebesgue measure of $S$ is $\lambda(A) = \sum_k \lvert b_k - a_k \rvert$. | 5 |
| Borel measure. | Measure on a topological space that is defined on all open sets (and thus on all Borel sets). | 17 |
| Dominated measure | A measure $\mu$ is dominated by another measure $\nu$ if for every measurable set $A$, $\nu(A) = 0$ implies $\mu(A) = 0$. This is written as $\mu \ll \nu$. | 41 |
| Borel set. | Set in a topological space that can be formed from open sets (or, equivalently, from closed sets) through the operations of countable union, countable intersection, and relative complement. | 52 |
| Dense set. | A subset $S$ of a topological space $\mathcal{X}$ such that every point $x \in \mathcal{X}$ either belongs to $S$ or is a limit point of $S$. | 16 |
| Proper subset | A proper subset $S'$ of a set $S$, denoted $S' \subset S$, is a subset that is strictly contained in $S$. | 59 |
| Compact space | Space $\mathcal{X}$ such that each of its open covers (Collection of open sets whose union contains $\mathcal{X}$) has a finite subcover.<br>$A \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded. | 16 |
| Relatively compact subspace | A subspace $S$ of a topological space $\mathcal{X}$ whose closure (all points in $S$ plus the limit points of $S$) is compact. | 110 |
| Compact operator. | An operator between Hilbert spaces, $T : \mathcal{H} \to \mathcal{G}$, such that the image under $T$ of any bounded subset of $\mathcal{H}$ is a relatively compact subset of $\mathcal{G}$. | 24 |
| Bounded operator. | An operator between normed vector spaces, $T : \mathcal{X} \to \mathcal{Y}$, such that $\exists M > 0$ such that $\lVert T(v) \rVert_{\mathcal{Y}} \leq M \lVert v \rVert_{\mathcal{X}}$ $\forall v \in \mathcal{X}$. The operator norm, $\lVert T \rVert_{op}$, is the smallest such $M$. | 9 |
| Isometric embedding | A map between metric spaces $\phi : \mathcal{X} \to \mathcal{Y}$ with metrics $d_{\mathcal{X}}$ and $\mathcal{Y}$ such that $d_{\mathcal{Y}}(\phi(x), \phi(y)) = d_{\mathcal{X}}(x, y)$. | 8 |
| Bilinear | A function of two variables that is linear with respect to each of its variables. | 8 |

| | | |
|---|---|---|
| Measurable | A function between measurable spaces $f : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$, meaning that $\mathcal{X}$ and $\mathcal{Y}$ are sets equipped with respective sigma algebras $\mathcal{A}$ and $\mathcal{B}$, such that $f^{-1}(E) \equiv \{x \in \mathcal{X} \mid f(x) \in E\} \in \mathcal{A}, \forall E \in \mathcal{B}$. | 12 |
| Absolutely continuous | A function $f$ such that $\forall \varepsilon > 0\ \exists \delta > 0$ such that whenever a finite sequence of pairwise disjoint intervals $(x_k, y_k)$ satisfies $\sum_k (y_k - x_k) < \delta$ then $\sum_k \lvert f(y_k) - f(x_k) \rvert < \varepsilon$. | 5 |
| Almost everywhere | A property holds almost everywhere if the set for which the property does not hold has measure zero. | 5 |
| Almost surely. | An event happens almost surely (a.s.) if it happens with probability one. | 40 |
| Almost surely convergence. | A sequence $X_n$ converges almost surely (a.s.) towards $X$ if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$. | 24 |
| Norm convergence. | The sequence $f_n$ converges in the norm $\lVert \cdot \rVert$ to $f$ if $\forall \varepsilon > 0$ $\exists N$ such that $\lVert f_n - f \rVert < \varepsilon\ \forall n \geq N$. | 24 |
| Distribution convergence | A sequence of random variables $X_n$ converges in distribution to a random variable $X$ if $\lim_{n\to\infty} F_n(x) = F(x)\ \forall x$, where $F_n$ and $F$ are the distribution functions of the variables. | 23 |

# Bibliography


[1] Gretton A., Borgwardt K.M., Rasch M.J anod Schölkopf B., and Smola A.J. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[2] Berlinet A. and Thomas-Agnan C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science + Business Media, New York, 2004.

[3] Gretton A. Introduction to RKHS, and some simple kernel algorithms. Advanced Topics in Machine Learning. Lecture conducted from University College London, 2013.

[4] Steinwart I. and Christmann A. *Support Vector Machines. Information Science and Statistics.* Springer, 2008.

[5] Borgwardt K.M., Gretton A., Rasch M.J., Kriegel H.P., Schölkopf B., and Smola A.J. Integrating structured biological data by kernel Maximum Mean Discrepancy. *Bioinformatics*, 21(14):e49–e57, 2006.

[6] Sriperumbudur B.K., Gretton A., Fukumizu K., Schölkopf B., and Lanckriet G.R.G. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[7] Bingham M.S. and Parthasarathy K.R. A probabilistic proof of Bochner's theorem on positive definite functions. *Journal of London Mathematical Society*, s1-43(1):626–632, 1968.

[8] Wendland H. *Scattered Data Approximation.* Cambridge University Press, 2005.

[9] Smola A.J., Gretton A., Song L., and Schölkopf B. A Hilbert space embedding for distributions. In *Proceedings of the Conference on Algorithmic Learning Theory (ALT) 4754*, pages 13–31. Springer, 2007.

[10] R.J. Serfling. *Approximation theorems of mathematical statistics.* John Wiley & Sons, Inc., 2008.

[11] Reed M. and Simon B. *Methods of modern mathematical physics. Vol. 1: Functional Analysis.* Academic Press, San Diego, 1980.

[12] Gretton A., Bousquet O., Smola A., and Schölkopf B. *Measuring Statistical Dependence with Hilbert-Schimdt Norms.* Max Planck Institute for Biological Cybernetics, 2005.

[13] Gretton A., Fukumizu K., Teo C.H., Song L., Schoelkopf B., and Smola S. A kernel statistical test of independence. *NIPS*, 21, 2007.

[14] Sejdinovic D., Sriperumbudur B., Gretton A., and Fukumizu K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[15] Székely G.J. and Rizzo M.L. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013.

[16] Rizzo M.L. and Székely G.J. DISCO analysis: A non parametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010.

[17] Székely G.J. and Rizzo M.L. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

[18] Székely G.J. and Rizzo M.L. On the uniqueness of distance covariance. *Statistics and Probability Letters*, 82:2278–2282, 2012.

[19] Székely G.J. and Rizzo M.L. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

[20] Berrendero J.R., Cuevas A., and Torrecilla J.L. Variable selection in functional data classification: a maxima-hunting proposal. Unpublished, 2015.

[21] Székely G.J. and Rizzo M.L. Testing for equal distributions in high dimension. *InterStat*, 5, November 2004.

[22] Székely G.J., Rizzo M.L., and Bakirov N.K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[23] Berg C., Christensen J.P.R., and Ressel P. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, volume 100 of *Graduate Texts in Mathematics*. Springer, New York, 1984.

[24] Maa J.F., Pearl D.K., and Bartoszynski R. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *The Annals of Statistics*, 24(3):1069–1074, 1996.

[25] Lyons R. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

[26] Gretton A., Fukumizu K., and Sriperumbudur B.K. Discussion of: Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1285–1294, 2009.

[27] Cardoso J.F. Dependence, correlation and Gaussianity in Independent Component Analysis. *Journal of Machine Learning Research*, 4:177–1203, 2003.

[28] Cuesta-Albertos J.A., Fraiman R., and Ransford T. A sharp form of the Cramer–Wold theorem. *Journal of Theoretical Probability*, 20:201–209, 2007.

[29] Lopez paz D., Hennig P., and Schölkopf B. The Randomized Dependence Coefficient. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1–9. 2013.

[30] Rao M., Seth S., Xu J., Chen Y., Tagare H., and Príncipe J.C. A test of independence based on a generalized correlation function. *Signal Processing*, 91:15–27, 2011.