UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

Detection of dependence based on kernel signal processing

Ferran de Cabrera Estanyol

A Master's Thesis
Submitted to the Faculty of the
Escola Tècnica d'Enyginyeria de Telecomunicació de
Barcelona
Universitat Politècnica de Catalunya

In partial fulfillment
of the requirements for the degree
MASTER IN TELECOMMUNICATIONS ENGINEERING

Advisor: Jaume Riba Sagarra

Barcelona, February 2017

Abstract

The measure of statistical dependence among two or more phenomena is an important task in a lot of scientific and engineering problems. Although Shannon's mutual information is a well measure of statistical dependence, its estimation from data usually becomes difficult due to the difficulty of estimating joint density functions. As a result, there is nowadays an increasing interest in defining alternative measures of dependence and, in parallel, a better understanding in which manner statistical dependence can be inferred directly from data samples.

Some alternative forms have been formulated based on kernel signal processing. It consolidates a powerful and applicable tool to many problems, specially for non linear approaches. Two of them will be reviewed in this work: the first, an alternative path of Shannon entropies through the generalization of them called Rényi entropies, and the second, that uses the kernel properties to build a Hilbert space allowing measures in high dimensional spaces.

This work intends to fill a gap on the understanding of the two proposals and their inherit properties, as well as proposing a more realistic and applicable case.

Abstracte

La mesura de dependència estadística entre dos o més successos és una tasca important en diversos camps científics i en la enginyeria. La mesura d'informació mútua definida per Claude Shannon és útil per mesurar-ho, però en casos reals es poden trobar adversitats donada la dificultat d'estimar la funció de densitat conjunta. Aquest fet ha portat que la cerca de mètodes alternatius estigui al alça, a més de la preocupació per caracteritzar aquesta dependència estadística a partir de unes mostres donades.

És per això que actualment existeixen diverses tècniques de processat basades en els anomenats kernels, que han esdevingut una eina adient per resoldre diversos problemes, especialment els no lineals. En aquest treball ens centrarem en dos d'ells. El primer és un camí alternatiu a l'entropia de Shannon, que serà precisament una generalització d'aquesta: les entropies de Rényi. El segon està basat en aprofitar les propietats dels kernels com a generadors d'espais Hilbertians, els quals permeten un àmbit de treball en espais infinits sense la necessitat de visitar-los.

Aquest treball té com a intenció millorar l'enteniment d'aquests mètodes així com també de proposar un cas real on aquestes mesures siguin aprofitables.

Acknowledgments

I must be first express my eternal gratitude to my advisor Jaume Riba for his patience with our infinite meetings, our never ending talks and to patiently and calmly responding my continuous concerns. He has opened a path to make this possible, leading me though the mysteries and labyrinthine field of the work.

I too wish to thanks my family, for their support during this course. We do know it has not been an easy academic year given all the past events, but your strong and unconditionally support has allowed me to be able to concentrate and pursue my objectives.

And finally, and in a special mention, to my father. I know you would be proud of me regardless of the outcome, and so this is for you.

Contents

List of Figures

Acronyms

AUC:	Area Under the Curve
AWGN:	Additive White Gaussian Noise
CCA:	Canonical Correlation Analysis
FIR:	Finite Impulse Response
FT:	Fourier Transform
GCC:	Generalized Cross-Correlation
HSIC:	Hilbert-Schmidt Independence Criterion
ICD:	Incomplete Cholesky Decomposition
I.I.D.:	Independent and Identically Distributed
IMSE:	Integrated Mean Squared Error
IT:	Information Theory
KDE:	Kernel Density Estimation
KICA:	Kernel Independence Component Analysis
KLD:	Kullback-Leibler Divergence
MMSE:	Minimum Mean Square Error
MVUE:	Minimum-Variance Unbiased Estimator
PDF:	Probability Density Function
RKHS:	Reproducing Kernel Hilbert Space
ROC:	Receiver Operating Characteristic

SNR:	Signal to Noise Ratio
TDE:	Time Delay Estimation
TDOA:	Time Difference Of Arrival
TOA:	Time Of Arival

Notation

a	Scalar
a^*	Complex conjugated of a
$\boldsymbol{a}$	Vector (column form)
$\mathbf{1}$	Column vector of ones
a_i	The i.th element of a vector $\boldsymbol{a}$
$\boldsymbol{a}_i$	Indexed vector

$\boldsymbol{A}$	Matrix with (i, j) entries A_{ij}
$\mathbf{I}$	Identity matrix
$\lvert\boldsymbol{A}\rvert$	Determinant of matrix $\boldsymbol{A}$
$tr\left(\boldsymbol{A}\right)$	Trace of matrix $\boldsymbol{A}$
$diag\left(\boldsymbol{A}\right)$	Diagonal of matrix $\boldsymbol{A}$
$\lVert\boldsymbol{A}\rVert_F$	Frobenius norm of matrix $\boldsymbol{A}$
$\boldsymbol{A}^T$	Transposed matrix $\boldsymbol{A}$
$\boldsymbol{A} \odot \boldsymbol{B}$	Schur-Hadamard or elementwise product

$\hat{b}$	Estimate of b
$E\left[b\right]$	Mathematical expectation of b
$\langle b, c\rangle$	Inner product of b and c

1 Introduction

The most known figure of merit to determine the degree of dependence is the mutual information defined by Shannon at 1948. His genuine and original work “A mathematical theory of communication” is still nowadays a reference for many of the current literature on Information Theory. In that time, the interest of the research was to determine the capacity of a noisy channel, but the result was on a higher scale, providing the basis of many further fields as probability, statistics, computer science and, in a special mention, communications. It is then unquestionable the importance of the figure of Claude E. Shannon on nowadays, and this work does not intends to be an exception.

The capacity of extracting knowledge from pure observations is one of the direct results of the mutual information measures. In a world of pure data there is an important concern on trying to get insights from it. In order to make it possible, there is an increasing seek of the appropriate statistical tools. The usual line of solutions to these kind of problems are found in Machine Learning, being it a neighboring field to the work. We are going to see that, like in Machine Learning, the interest relays on mapping the observations to an another space which is usually richer in information than the data space.

This Master’s Thesis objective is to define and analyze the presence of relation between two sets of data observations. To do so, the tool of choice will be the kernel. The notions of kernels appeared as a powerful field on Machine Learning allowing to solve non linear problems in an elegant manner. Its addition to signal processing has provided versatility and a suitable framework to this non linear perspective. In essence, the kernels are defined as functions that map the input data to a higher dimensional space based on the inner product of them. Its appealing definition and usage has promote the onset of kernel based techniques.

In respect of the matter at hand, there actually exist many techniques to obtain the desired degree of dependence. In this work we are going to review some of them, trying to scratching the core and emerging anew with a better understanding of what is happening. The objective is to fill a gap between the mathematical overview in the literature and a signal processing approach, and to fill a gap between the usually blind methodology of Machine Learning and the understanding on how the data is enhanced when kernel processing is used.

1.1 Thesis outline and organization

This Master's Thesis is organized in two blocks, being the second a consequence of the first. The first part will be entirely dedicated to the derivation and description of the methodologies that will be used as dependence detectors. Each detector is going to be reasonably obtained step by step and then analyzed from a perspective of goodness in the detection. The second part is focused on the derivation of the detectors to propose a latency estimator at the end of the work. Both parts will be highly correlated given the second is a practical approach to the first.

The general structure will be based on the ordered development of the methodologies. The first two chapters will be dedicated to the settlement of the problem as well as actual methodologies, while the rest of the work will be focused on the progression of the core of the work. Specifically, the structure is the following:

- In chapter 2 some generic expressions of detection theory are explained. It is going to be convenient to define the metrics and figures of merit that are going to be used in the rest of chapters, specially on the results presentation. Additionally, some background and definitions of kernel methodologies will be presented, which are going to be fundamental in the throughout of the work.

- In chapter 3 it is presented a state of the art in terms of detecting dependence. It begins with the correlation methods and finishes with a review on actual dependence detectors in the literature.

- Chapter 4 is the main core of the Master's Thesis. It is fully dedicated on the development of the detectors, its properties and the insights given by each one in memoryless cases. We are going to see four different approaches with the respective performance at the end of the chapter.

- In Chapter 5 it is going to be reviewed some channel model approaches in order to adapt the detectors to a more realistic case of transformations. We are going to see the disadvantage of the methodologies when memory is added and how to solve it.

- Chapter 6 will enclose the latency estimator proposal. The problem statement and its mapping to our case of dependence detectors will be reviewed here.
- Finally, chapter 7 is the conclusion of the work and such the final thoughts and some perspective to the future.

For any reader that founds it appropriate and/or interesting, the Matlab code used through the development of the work can be found in the following GitHub link:

https://github.com/FerranDeca/TFM-Detection-of-dependence-based-on-kernel-signal-processing

2 Problem statement and definitions

Before going on the main development, it is going to be useful to review the purpose of the work. The objective on doing so is manifold. First, the basis and objective of the work is defined and so the general workspace is built. It is also intended to define some tricks and gimmicks to build the detectors or to check their performance. It is important to remark that the equations and definition from this section will be widely used in the development of the work afterwards, and so the intention is nothing else that the reading and comprehension tempo of the Master's thesis would not be broken or interrupted.

Specifically, it is going to be seen that the problem in hands is highly related to detection theory procedures, additionally with some implications and figures of merit. Inside this part it is going to be reviewed the detection theory similarities that can be useful for enhancing the detection of dependence proposed in this work. Then, some performance indicators are going to be presented that in order to express as freely as possible the advantages and disadvantages of each detection. Lastly, an approach will be presented that will allow to advance faster once the estimators have been formulated.

To end up the chapter, we are going to review and define the kernel technique. Based on the title of the Master's Thesis, it is expected that the kernel methodologies conform the basis and will be widely present from now on.

2.1 Measuring and detection of statistical dependence

The general assumption of the work is that the knowledge about the statistics presented is null. The objective is then to determine the statistical dependence between two sets of observations in a blind way. The interest on this search of dependence are those applications on the understanding of the events is reduced, and so we are limited on raw observations. Objectively, the interest on building dependence detectors based on the lack of knowledge relies in applications that actually looks for the increase of the degree of cognition.

In general, we are going to define these observations as data samples $\{x_i\}$ and $\{y_i\}$ stored in vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Hence, the objective is clear: detect the dependence or independence of the two random variables based on these two vectors. We are going to see that this task of detection is not straightforward due to the implication of many factors.

2.1.1 Detection theory resemblance

It has to be pointed out the similarities of detection theory and the problem stated in this Master's thesis. Some definitions given to solve properly the detection problems are going to be useful to organize and test the solutions proposed.

The general framework is going to be a binary hypothesis based on a finite data set of samples. The objective is to define a dependence detector and so two hypothesis will be defined: hypothesis $\mathcal{H}_0$ will assume statistical independence while hypothesis $\mathcal{H}_1$ will assume dependence case. These hypothesis will be tested over a set of data $\boldsymbol{x} = x(1), x(1), ..., x(N)$ using a function called test statistic $T(X)$, and so the following function:

$$T(\boldsymbol{x}) = f\left(x(1), x(1), ..., x(N)\right) \tag{2.1}$$

The test statistic will be compared with a threshold to make the decision

$$T(\boldsymbol{x}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma \tag{2.2}$$

or equivalently, the space of observation will be segmented into disjoint decision regions $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$.

Using the test statistic and the trehshold, we can express the probability of success or failure with the following probabilities:

- Probability of false alarm or P_{FA}: Decide $\mathcal{H}_1$ when $\mathcal{H}_0$.
- Probability of detection or P_D: Decide $\mathcal{H}_1$ when $\mathcal{H}_1$.
- Probability of miss detection or P_{MD}: Decide $\mathcal{H}_0$ when $\mathcal{H}_1$.

Originally in detection theory, these probabilities are obtained through the integral over the region of detection hypothesis of the true pdf case. In real scenarios, it is a difficult task to measure them properly. In the case of the dependence detector, it is going to rely only on a finite set of data with unknown pdf. For these cases, the measure of P_{FA}, P_D and P_{MD} are more alike in machine learning classifiers, where the evaluation is done multiple times for different inputs and the measures of the probabilities are done based on the number of times a false alarm or miss detection events occurs.

For this purpose, L i.i.d samples for each hypothesis will be created, separated in M blocks of N samples, and so it will hold the following relation will be held: $L = M\,N$. From each block, a dependence measure from $T(\boldsymbol{x})$ will be extracted and compared to the threshold. Formally, it is going to be expressed as

$$T\left(\boldsymbol{X};\mathcal{H}_0\right) = \left[T\left(\boldsymbol{x}_1;\mathcal{H}_0\right)...T\left(\boldsymbol{x}_M;\mathcal{H}_0\right)\right] \tag{2.3}$$

for hypothesis of independence, and

$$T\left(\boldsymbol{X};\mathcal{H}_1\right) = \left[T\left(\boldsymbol{x}_1;\mathcal{H}_1\right)...T\left(\boldsymbol{x}_M;\mathcal{H}_1\right)\right] \tag{2.4}$$

for dependence, being $\boldsymbol{x}_i = x_{1,i}, ..., x_{N,i}$.

Then, for a fixed γ, the probabilities will have the following form :

$$P_{FA} = \sum_i \left(T\left(\boldsymbol{x}_i;\mathcal{H}_0\right) \geq \gamma\right)/M \tag{2.5}$$

$$P_D = \sum_i \left(T\left(\boldsymbol{x}_i;\mathcal{H}_1\right) \geq \gamma\right)/M \tag{2.6}$$

$$P_{MD} = \sum_i \left(T\left(\boldsymbol{x}_i;\mathcal{H}_1\right) < \gamma\right)/M = 1 - P_D \tag{2.7}$$

Once having tested a detector multiple times and the probabilities have been measured, the next natural step is to draw the called receiver operatic characteristic curve, or ROC curve. This curve represents the evolution of the probability of detection over the probability of false alarm for certain thresholds. So basically, it illustrates the performance of the detector as shown in figure 2.1.

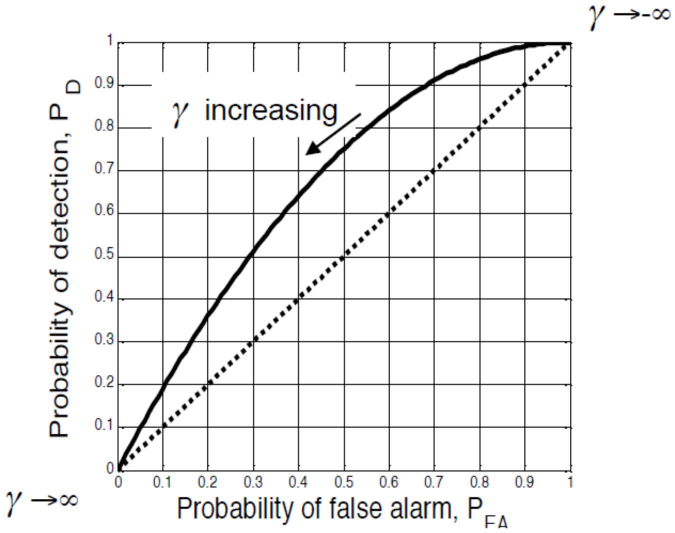

Figure 2.1: ROC curve example

The ROC curve allow to implement a visual performance indicator, but we can get a close expression for a more concrete indicator, explained at the next sub-section.

Asymptotic Behavior

Another useful characteristic of detection theory is the one given by the Stein's lemma. By this lemma, an asymptotic expression for the false alarm and error probabilities can be obtained. To develop it, it is necessary to define the pseudo-distance of Kullback-Leibler (KL) or the Kullback-Leibler Divergence (KLD).

Given two density functions f and g with $f, g \in \mathbb{S}$, the KL divergence is defined as the measure of the distance between the two probability distributions:

$$D\left(f \parallel g\right) = \int_S f(x) \log \frac{f(x)}{g(x)} dx \tag{2.8}$$

For the dependence estimators, the density functions for the estimators will be denoted as $p\left(X;\mathcal{H}_0\right)$ for the probability density under hypothesis $\mathcal{H}_0$, and $p\left(X;\mathcal{H}_1\right)$ under hypothesis $\mathcal{H}_1$. Then, considering these density functions, the Stein's lemma states an asymptotically behavior for the false alarm probability and probability of miss detection with the following expressions:

$$\lim_{N\to\infty} \frac{1}{N} \log P_{FA} = -D\left(p\left(X;\mathcal{H}_1\right) \parallel p\left(X;\mathcal{H}_0\right)\right) \tag{2.9}$$

$$\lim_{N\to\infty} \frac{1}{N} \log P_{MD} = -D\left(p\left(X;\mathcal{H}_0\right) \parallel p\left(X;\mathcal{H}_1\right)\right) \tag{2.10}$$

The demonstration can be reviewed on [7]. Do note that apart from demonstrating an asymptotic convergence it also has the implication that the false alarm and miss detection probability falls exponentially.

2.1.2 Detection performance indicators

At the end of the day, when the detectors are defined and explained, there will be a major need on testing its performance. It has been seen that the ROC curve is useful for this purpose, but it is also interesting to define more indicators to get comparable results.

The first one is the figure of merit called Area Under the Curve (AUC) [5]. It is a usual metric used to define the goodness of the ROC curves. Its principle is that the better the detector, the better is the area under the curve, and so more likely is the detector of making a good decision. The interpretation is that is comparing the average value of P_D for all the values of P_{FA}. The AUC will be 1 when the perfect detector is achieved but the metric that will be used in this work will be $1-AUC$ in the sense that, when better the detector the close to 0.

The second was given by the hands of Picinbono in [22], with the intention of providing a simple criteria to avoid the ROC curve calculation. As said, it provides an efficient solution that relies on the difference between the medium value of true dependent measures and the medium value of true independent measures. It is also usually called Signal to Noise ratio in the cases of detection with two hypothesis. To be more precise, it has the following form:

$$Deflection = \frac{\left(E\left[T(\boldsymbol{X};\mathcal{H}_1)\right] - E\left[T(\boldsymbol{X};\mathcal{H}_0)\right]\right)^2}{var\left[T(\boldsymbol{X};\mathcal{H}_0)\right]} \tag{2.11}$$

Contrary to the $1 - AUC$, the deflection provides a better detection when higher is the parameter, and so the mean values of the test statistics under the two hypothesis are more spaced between them.

By using these criteria, the system performance can be evaluated with only one number, and so they can be used to enhance the visual representation that provides the ROC. These two values will be specially useful for evaluating the performance of multiple detectors for different N to see its asymptotic behavior.

The threshold decision is also something to care about. The proposed metric is based on the perfect detection mapped into the ROC curve, that happens when we get a perfect probability of detection and a zero probability of false alarm. Then, the closer to the upper-left part of the figure the better. This can be measured by the euclidean distance between the (0,1) coordinates and the curve.

Suppose a vector of T thresholds that provides the sweep of the ROC curve $\gamma = \gamma_1, ..., \gamma_T$, we can measure a P_{FA} and a P_D for each γ to obtain $\mathbb{P}_{\mathrm{FA}} = P_{FA}(\gamma_1), ..., P_{FA}(\gamma_T)$ and $\mathbb{P}_{\mathrm{D}} = P_D(\gamma_1), ..., P_D(\gamma_T)$, then the threshold is the one that obtain the close euclidean distance:

$$\gamma = \min_i \sqrt{P_{FA}^2(\gamma_i) + (1 - P_D(\gamma_i))^2} \tag{2.12}$$

2.1.3 Low SNR assumption

Being the detection of dependence the main focus on this work, the random values used as an input for the detectors will provide a low degree of dependence. The contrary, a big and visible dependence, would not be an interesting case as the result is kind of direct. Having this in mind, it is possible to make some approximations that will allow to make a more precise estimation and detection at nearly independence random vectors. The title of this section, low signal to noise ratio (SNR) assumption, is used because of its similarities with the proposed case.

Recalling the additive white Gaussian noise (AWGN), its channel capacity can be reduced to $C = B{\cdot}SNR$ when low SNR is accomplished. The removal of the logarithm allows to move from a logarithmic function to an approximation of a line when the signal level is near the noise level, or equivalently, when $SNR \approx 1$, and this is a very powerful trick to apply on the dependence detector.

For instance, we can recall Shannon's mutual information for continuous sources as:

$$I(X;Y) = \int\int f_{X,Y}(x,y) \log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right) dxdy \geq 0 \tag{2.13}$$

being X and Y two continuous random variables, $f_{X,Y}(x,y)$ its joint distribution and $f_X(x)$ and $f_Y(y)$ the marginal distributions. The inequality becomes equality when X and Y are independent. In the matter at hand, for random vectors close to independence, the inner part of the logarithm will be close to one given $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ at independence, and so it is immediate to recall the fundamental inequality for logarithms. This inequality defines an upper bound as $\log x \leq x - 1$ for $\forall x > 0$. The demonstration can be obtained from Taylor series:

$$\log(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} \leq (-1)^2 \frac{x^1}{1} + (-1)^3 \frac{x^2}{2} \leq x \tag{2.14}$$

The closer x to 1, the closer the equality, and then it is possible to proceed as follows:

$$\begin{aligned} I(X;Y) &\leq \int\int f_{X,Y}(x,y)\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}-1\right)dxdy \\ &= \int\int \frac{f^2_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy - \int\int f_{X,Y}(x,y)dxdy \\ &= \int\int \frac{f^2_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy - 1 \end{aligned} \tag{2.15}$$

Leading to this final inequality:

$$\int\int \frac{f^2_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy - 1 \geq I(X,Y) \geq 0 \tag{2.16}$$

Do note that this expression has a very high potential as a detector if an estimator of the first term can be formulated. As it is going to be seen, the pdf estimators will depend on the data samples and so a proper estimator from this expression will be actually possible. It is going to be seen in detail at Chapter 4.

2.2 Kernel definition and basics

The kernels will constitute the most important core of the work being all the estimators and detectors based on them. Thus, a proper definition and some basics on kernel processing will be useful to be described. The kernel on the field of signal processing is useful on many application. For example in Adaptive Filtering [17], in Fisher Discriminant Analysis [19], in comparing distributions by Maximum Mean Discrepancy [11] or even in biomedical engineering [43].

The kernel theory was proposed and consolidated in 1909 by the hands of Mercer [18], but for this work the notation and definitions will follow the ones by Aronszajn in [1], who defined the reproducing kernel theory. Another reference for kernel processing and its properties can be found in [14].

A kernel is defined as a continuous function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that operates in an input space $\mathcal{X}$. A kernel can also be defined as positive semi-definite when for any set of input data $\boldsymbol{x} \in \mathcal{X}$ of size N it satisfies:

$$\sum_{i,j}^{N} \alpha_i\alpha_j k\left(x_i, x_j\right) \geq 0 \quad \forall \alpha_i \in \mathbb{R} \tag{2.17}$$

A positive semi-definite kernel can be used to map the input values to an n-dimensional space where the variables live, called a feature space, through a function $\psi(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathcal{H}$ that assigns a kernel value $k(\boldsymbol{x}, \boldsymbol{y})$ to the input $\boldsymbol{y}$. Indeed, a feature space is associated with a positive semi-definite kernel through a inner product of the kernel, as demonstrated in [1]:

$$\langle k(\boldsymbol{x}, \cdot), k(\boldsymbol{y}, \cdot)\rangle = k(\boldsymbol{x}, \boldsymbol{y}) = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y})\rangle \tag{2.18}$$

which is called reproducing property, and for this the positive semi-definite kernels are commonly called reproducing kernels.

One interesting property of the inner product that defines a kernel, is that with a specific norm $\|g\| \equiv \sqrt{\langle g, g\rangle}$ it turns out into a Hilbert space. A Hilbert space $\mathcal{H}$, with function $f : \mathcal{X} \rightarrow \mathbb{R}$, is an inner product space that is complete, and so the Cauchy sequence $\left\{\boldsymbol{x}^{(k)}\right\}$ converges to an unique $\boldsymbol{x} \in \mathcal{H}$. In the cases of kernels, if there exists a function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defined by the norm, as kernels are, is then called reproducing kernel Hilbert space (RKHS). In fact, by the theorem of Moore-Aronszajn we can assure that any positive-definite function actually defines a RKHS, while a RKHS also defines a kernel by the means of the inner product which implies a positive-definite kernel. It is then a circle of definitions that provides mutual involvement.

It is also desirable to define the kernel matrix and its properties. For a set of N data values $x_1, ..., x_N$, the $N \times N$ kernel matrix $\boldsymbol{K}$ is composed by the elements $K_{i,j} = k(x_i, x_j)$, as represented in the equation that follows:

$$\boldsymbol{K}_x = \begin{bmatrix} k(x_1, x_1) & . \;\; . \;\; . & k(x_N, x_1) \\ . & . & . \\ . & . & . \\ . & . & . \\ k(x_1, x_N) & . \;\; . \;\; . & k(x_N, x_N) \end{bmatrix} \tag{2.19}$$

The properties of the kernels also defines the properties of the kernel matrix, for instance if the kernel is semi-definite positive, the kernel matrix will also be, fulfilling $\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \geq 0 \quad \forall \alpha_i \in \mathbb{R}$. Additionally, for Gaussian kernels we can also assure that the diagonal will be always a vector of ones, $diag(\boldsymbol{K}) = \mathbf{1}$, all its elements are positive, $(\boldsymbol{K})_{i,j} \geq 0$, and it conforms a Gram matrix, satisfying $K_{i,j} = K^*_{j,i}$.

The construction of the Kernel matrices is going to be fundamental as it is the first step for a mutual information estimator. The point on using the kernel matrix is its capability of manipulating kernels without considering the mapped feature space, and so it provides a link between raw data and the purpose of measure.

3 State of the art

To measure statistical dependence is not a novel problematic in statistics. In the current literature there exist many algorithms to address the measures. Some examples are the Maximal Information Coefficient (MIC) [26], the kernel Independent Component Analysis (KICA) [2], the distance of Brwonian Correlation (dCor) [34], the Mutual Information Dimension (MID) [33] or the Hilbert-Schmidt independence criteria (HSIC) [12] further developed to the Constrained Covariance (COCO) [13]. A part from these proceedings, the development of dependence measures in [23] and [41] will have a special mention due to its closeness with the firsts steps of the work.

Across all of these references, we do note a high relation between correlation and dependence, specially when the data is mapped into another space. Thus, first we are going to review some correlation methods that are interesting in the sense that will be useful afterwards. Then, we will take a look on some of the dependence detectors to obtain a better perspective of what is the general metric when dependence is sought.

3.1 Correlation detection methods and limitations

The correlation as a measure of dependence is a known and widely searched merit figure. The analysis of its existence has been studied for years and there are a lot of interesting techniques for many different purposes. In this work we are going to review three of them given its relation and implication throughout the development. The Pearson coefficient is a basic measure which provides a solid expression for the cross-correlation matrix estimate when the only knowledge about a random variable are data samples. The Frobenius norm provides an useful technique to estimate the degree of correlation for two vectors of data realizations, something that lacks in the Pearson coefficient. Finally, the Canonical Correlation analysis provides a measure of correlation that can be related to the mutual information under additive and white Gaussian channels.

3.1.1 Pearson coefficient

The most common and basic correlation detector for a population is the Pearson coefficient which is based on measuring the linear dependence of two random variables $x$and y, defined by:

$$\rho_{x,y} = E\left[(x - \bar{x})(y - \bar{y})\right] \tag{3.1}$$

being $\bar{x}$ and $\bar{y}$ the mean of the random variables x and y respectively. Do note the detector can also be written as the cross-covariance between the two random variables as $\rho_{x,y} = cov\,(x, y)$.

The coefficient results in $\rho_{x,y} = 0$ for uncorrelated random variables and $\rho_{x,y} \neq 0$ for correlated random variables. However, if the only knowledge of the random variables are random data samples $\boldsymbol{x} = \{x_1, ..., x_N\}$ and $\boldsymbol{y} = \{y_1, ..., y_N\}$, the coefficient can be estimated as

$$\hat{\rho}_{x,y} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x}_n)(y_n - \bar{y}_n) \tag{3.2}$$

where the factor $\frac{1}{N-1}$ is used to obtain an unbiased estimator and the sample mean $\bar{x}_n = \frac{1}{N} \sum\limits_{m=1}^{N} x_m$. As being an estimator, the detection of correlation has to be evaluated over a threshold γ in the following form:

$$\left|\hat{\rho}_{x,y}\right|^2 > \gamma \tag{3.3}$$

If we generalize the detection problem for multiple $\boldsymbol{x}$ and $\boldsymbol{y}$, the same principle can be used to measure the correlation between them. Suppose the data is organized in M_x and M_y vectors of N random samples over the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, with columns $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ containing the random data generated from the random variables $x, y \in \mathbb{R}$. Then, the sample correlation coefficient turns into

$$\hat{\boldsymbol{C}}_{x,y} = \frac{1}{N-1} \sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n)(\boldsymbol{y}_n - \bar{\boldsymbol{y}}_n)^H = \frac{1}{N-1} \left(\boldsymbol{X} - \bar{\boldsymbol{x}}\boldsymbol{1}^H\right)\left(\boldsymbol{Y} - \bar{\boldsymbol{y}}\boldsymbol{1}^H\right). \tag{3.4}$$

which is the sample cross-correlation matrix.

In order to simplify the expression, we can take $\left(\boldsymbol{X} - \bar{\boldsymbol{x}}\boldsymbol{1}^H\right)$ and define it as a centering of the signal. Thus, it is equivalent to project the signal into a new space that allows both signals to be compared independently of its original space. Formally, $\bar{\boldsymbol{x}}$ is the sample mean of the columns and so it can be expressed as $\bar{\boldsymbol{x}} = \frac{1}{N}\boldsymbol{X}\boldsymbol{1}$. Then we get

$$\left(\boldsymbol{X} - \bar{\boldsymbol{x}}\boldsymbol{1}^H\right) = \boldsymbol{X} - \frac{1}{N}\boldsymbol{X}\boldsymbol{1}\boldsymbol{1}^H = \boldsymbol{X}\left(\boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^H\right) = \boldsymbol{X}\boldsymbol{P} \tag{3.5}$$

being $\boldsymbol{P}$ the projector matrix to the orthogonal. It is known that the projector matrix have the form $\boldsymbol{A}\left(\boldsymbol{A}^H\boldsymbol{A}\right)^{-1}\boldsymbol{A}^H$, and so when $\boldsymbol{A} = \boldsymbol{1}$, then $\boldsymbol{P}$ results to be a projector to the dimension of ones.

The new reformulation of the sample cross-covariance matrix is

$$\hat{\boldsymbol{C}}_{x,y} = \frac{1}{N-1}\left(\boldsymbol{X}\boldsymbol{P}\right)\left(\boldsymbol{Y}\boldsymbol{P}\right)^H = \frac{1}{N-1}\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{P}^H\boldsymbol{Y}^H\right) = \frac{1}{N-1}\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{Y}^H\right) \tag{3.6}$$

in the sense that it tends to the real cross-covariance matrix in probability

$$\hat{\boldsymbol{C}}_{x,y} \underset{N\to\infty}{\longrightarrow} \boldsymbol{C}_{x,y} \tag{3.7}$$

This is known as the sample cross-correlation, and although is an efficient estimator of the cross-covariance based on raw data, the metric needed to determine the degree of correlation is not well defined. However, in the next sub-section it is going to be seen a procedure to exploit the information of this matrix in order to build a proper detector.

3.1.2 Frobenius norms

Given the Pearson coefficient approach there are many covariance degree estimators that are interesting and useful. Specifically, in a work by Santamaría et al., [27], it is demonstrated that under Gaussian data assumption and low correlation degree the optimal estimator is the squared Frobenius norm of the cross-correlation matrix. Hence, under an hypothesis of low covariance estimated from data, a degree of correlation can be formulated as

$$\left\|\hat{\boldsymbol{C}}_{x,x}\right\|_F^2 = tr\left(\hat{\boldsymbol{C}}_{x.x}^H\hat{\boldsymbol{C}}_{x.x}\right) \tag{3.8}$$

being $\hat{\boldsymbol{C}}_{x,x}$ normalized sample covariance matrix obtained in the previous Section, and so $\hat{\boldsymbol{C}}_{x,x} = \frac{1}{N-1}\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^T\right)$. This detector can be extensible by a cross-covariance by the means of $\left\|\hat{\boldsymbol{C}}_{x,y}\right\|_F^2$. The reason behind the Frobenius norm is to search over the covariance matrix for the largest singular value. If both hypothesis of uncorrelation and correlation are close enough, and the data is assumed Gaussian, a small change to the eigenvalues of the estimated cross-covariance matrix is visible and determinant to make the detection. It is actually a very powerful tool that allows to present the detector as follows:

$$\begin{aligned}\left\|\hat{\boldsymbol{C}}_{x,y}\right\|_F^2 &= tr\left(\hat{\boldsymbol{C}}_{x,x}^T\hat{\boldsymbol{C}}_{x,y}\right)\\ &= \left(\frac{1}{N-1}\right)^2 tr\left(\left(\boldsymbol{XPY}^H\right)^H\left(\boldsymbol{XPY}^H\right)\right)\\ &= \left(\frac{1}{N-1}\right)^2 tr\left(\boldsymbol{PX}^H\boldsymbol{XPY}^H\boldsymbol{Y}\right) \qquad (3.9)\end{aligned}$$

The implicit result on this methodology is that we are now interested on inner products, and so it does not require products between data of different sets nor outer products. Formally, it has begun with a sample cross-covariance matrix that gives attention to the data dimensionality M_x and M_y, called Primal form, and it has ended with a detector that prioritize the data sample length N, called Dual form. This approach differs from the usual procedure in signal processing, where the Primal form highlights. Then, it is interesting to stick on the Dual form given that M_x is generally larger than N. In fact, the matrices $\boldsymbol{X}^H\boldsymbol{X}$ and $\boldsymbol{Y}^H\boldsymbol{Y}$ are Gram matrices that later on will be called $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$.

The importance of this result will be seen later on, when the inner products will define a new space based on kernel processing and will allow flexibility over the future expressions.

3.1.3 Canonical Correlation analysis

The Canonical Correlation Analysis (CCA) is an another correlation detection widely used in the literature developed by Hotelling in 1936 [16]. It is based on finding the space formed by linear combinations between two sets of data $\boldsymbol{x} = \{x_1, ..., x_N\}$ and $\boldsymbol{y} = \{y_1, ..., y_N\}$ which provides maximum correlation. Additionally, an important remark on CCA is that it is invariant to affine transformations of the data.

The principle is similar to the one presented with the Pearson coefficient, but instead of projecting the matrices into the constant vector space, the projection is based on the direction vectors. Considering a pair of vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, the linear combinations from

$\boldsymbol{x}$ and $\boldsymbol{y}$ called canonical variates are $x = \boldsymbol{x}^H \boldsymbol{u}_i$ and $y = \boldsymbol{y}^H \boldsymbol{v}_i$, hence

$$\begin{aligned}\gamma_i &= \frac{E\left[xy\right]}{\sqrt{E\left[x^2\right]E\left[y^2\right]}} \\ &= \frac{E\left[\boldsymbol{u}_i^H \boldsymbol{x}\boldsymbol{y}^H \boldsymbol{v}_i\right]}{\sqrt{E\left[\boldsymbol{u}_i^H \boldsymbol{x}\boldsymbol{x}^H \boldsymbol{u}_i\right]E\left[\boldsymbol{v}_i^H \boldsymbol{y}\boldsymbol{y}^H \boldsymbol{v}_i\right]}} \\ &= \frac{E\left[\boldsymbol{u}^H \boldsymbol{C}_{x,y} \boldsymbol{v}\right]}{\sqrt{E\left[\boldsymbol{u}_i^H \boldsymbol{C}_{x,x} \boldsymbol{u}_i\right]E\left[\boldsymbol{v}_i^H \boldsymbol{C}_{y,y} \boldsymbol{v}_i\right]}}\end{aligned} \tag{3.10}$$

where maximum canonical correlation is the maximum γ_i with respect to different $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, which is an indicator to how strong a relationship is.

There is also a relation between the canonical correlation and mutual information which relies in the property of mutual information of being additive for independent variables. If the independence is accomplished, then the mutual information is the sum of mutual information between x_i and y_i. For Gaussian variables this means:

$$I\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{1}{2}\sum_i \log\left(\frac{1}{1-\gamma_i^2}\right) \tag{3.11}$$

The implication behind is that it is only possible when the correlation does not depend on higher order statistic dependencies other than correlation. Additionally, for a low correlation degree we can linearize the logarithm around $log\left(1\right)$, as it is going to be explained at Section 2.1.3, and so we get

$$\begin{aligned}I\left(\boldsymbol{x}, \boldsymbol{y}\right) &\leq \frac{1}{2}\sum_i \left(\frac{1}{\left(1-\gamma_i^2\right)} - 1\right) \\ &= \frac{1}{2}\sum_i \left(\frac{\gamma_i^2}{1-\gamma_i^2}\right)\end{aligned} \tag{3.12}$$

which actually tends to the sum of γ_i^2. This way, it can be compared with 3.9 in the sense that the Frobenius norm is the sum of the squared elements of the matrices of data $\boldsymbol{X}$ and $\boldsymbol{Y}$, and so it looks on all directions instead of focusing where the information lies in the search of the degree of correlation. In this case but, we do only sum the degree of correlations given $\boldsymbol{u}$ and $\boldsymbol{v}$, but at the cost of finding them first.

3.2 Current dependence detectors

The current existing detectors of dependence based on kernel signal processing are many and manifold. Apart from the ones that are going to be further developed in the throughout of the work, some interesting approaches are the works by Gretton and Smola et al., providing the Hilbert-Schmidt independence criterion (HSIC). The idea is to define two feature maps for each variable $\boldsymbol{x}$ and $\boldsymbol{y}$ through the kernel functions, and to measure the correlation in the Hilbert space that these kernels reproduce. The detection of correlation in that space is shown to be highly related with the dependence in the data space. The main idea is adjacent with the idea behind the sampling of the Characteristic function in section 4.4. Actually, we will see that it results in a very similar expression to the HSIC one:

$$HSIC = \frac{1}{N^2} tr\left(\boldsymbol{KPLP}\right) \tag{3.13}$$

being $\boldsymbol{K}$ and $\boldsymbol{L}$ Gram kernel based matrices. Also do note its resemblance with 3.9. It is given because the Hilbert spaces will be seen as an extension of the Frobenius norm when kernels are applied. These relations are going to be seen in detail in Chapter 4.

The extension of the CCA for measuring independence through kernel methods, called Kernelized Correlation Component Analysis (KICA) [2] do also follows the seek of the degree of correlation in the infinite space. In this case, the canonical correlation is given by

$$\gamma_i = \frac{\boldsymbol{\alpha}_1^T \boldsymbol{K}_x \boldsymbol{K}_y \boldsymbol{\alpha}_2}{\sqrt{\boldsymbol{\alpha}_1^T \boldsymbol{K}_x \boldsymbol{\alpha}_1}\sqrt{\boldsymbol{\alpha}_2^T \boldsymbol{K}_y \boldsymbol{\alpha}_2}} \tag{3.14}$$

being $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ the Gram matrices associated with a kernel function and $\boldsymbol{\alpha}$ direction vectors of the space spanned also by the kernel function $\psi(\boldsymbol{x}) : \ \mathcal{X} \rightarrow \mathcal{H}$. The idea behind is the same as in CCA, but instead of looking for linear relations in the data space, these relations are looked in a Hilbert space.

Generally speaking, in the analysis of dependence between two vectors of data observations it is going to be seen that there is no good reason to stay in the data space. The addition of kernels and so the mapping to a higher space provides an information not present in the original space. Kernels signal processing will then be the key on detecting dependence.

4 Dependence detection in memoryless systems

The next natural step is to move to the main development. This chapter consolidates the core of the work in the sense that we derive the detectors of dependence and we analyze them by its functionality and characteristics.

In this chapter we are firstly reviewing the estimation of the pdf given N observations in order to, then, estimate the entropy associated with them. Within this, we are going to naturally derive the kernel processing from the estimation of the pdf, and so we will relate the kernel processing with the information theory. Based on this relation, three detectors will be extracted. The first is directly obtained from the density estimate, a process that can be reviewed in [23] and in [41]. The second and the third will be derived from the properties of the Rényi entropies and from the U-statistics respectively. The appropriate step will then be to review the insights of the detectors and to analyze what is happening in the kernel processing applied. The last detector will be developed from the characteristic function, an alternative path to densities to characterize a random process, and from the Hilbert-Schmidt norm, an extension of the Frobenius norm when kernels are used. Finally, the detectors will be tested to determine the performance with a synthetic model.

4.1 Parzen density estimation

In order to estimate the mutual information of multiple random variables $I(X_1, ..., X_N)$, it is therefore needed an estimation of entropy $H(X_1, ..., X_N)$. The first step is to define an estimator of the probability density function that works on samples from a random variable, and so do Kernel Density Estimation (KDE) techniques. The most important definitions for KDE were given by Emanuel Parzen (1962) and Murray Rosenblatt (1956), being the Parzen window estimate the preferred choice in this work [21].

Let us assume a sequence of i.i.d. random samples $x_1, x_2, ..., x_M$, then the estimation of its pdf is given by

$$\hat{f}_x(x) = \frac{1}{N}\sum_{i=1}^{N} g\left(\frac{x - x_i}{\sigma}\right) \tag{4.1}$$

The proposal from Parzen is a plug-in algorithm that relays on the summation of evaluated functions, also called kernels, depending on the data samples. The interesting point on the proposed estimator was the capability of reconstructing the pdf given a finite amount of random values, and so a non-parametric model to set the problem. It does act as a local builder, and so the outliers have minor impact at the reconstruction. Also, do note this estimator acts as an average over the samples using a kernel g, which is actually a semi-definite kernel, but whose definition differs from the kernels on section 2.2 in the sense that the metrics of inner products are not necessarily defined there.

It is coherent to think that for most distributions a Gaussian kernel is the most appropriate, but for non-negative distributions it may cause a non desirable tail effect on the negative part of the estimate. At any case, at the matter in hand this little caveat will not make any point but other properties.

Most important, for a proper estimation of the pdf, the kernel g must have the following characteristics:

- $g(x) \geq 0$ for any x
- $lim_{x\to\infty} |xg(x)| = 0$
- $\int_{-\infty}^{\infty} g(x)dx = 1$

To fulfill these properties, the Gaussian function g is used:

$$g\left(\frac{\boldsymbol{x} - \boldsymbol{y}}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\parallel \boldsymbol{x} - \boldsymbol{y} \parallel^2}{2\sigma^2}\right). \tag{4.2}$$

One of the reasons on the choice of Gaussian form is because the parameter σ, which can be obtained through an optimal like rule-of-thumb if the true pdf has Gaussian form and Gaussian kernels are used. At any case, it is proven that generally speaking the Gaussian kernel provides the best estimation of the true pdf.

The selection of σ, user oriented, is one of the major constrains of the work as well as a turning point given its impact on the pdf estimation. A low bandwidth will result on spiky estimate and a high bandwidth into an oversmoothing estimate.The first thought is to sweep for a range of σ until some cost function is accomplished, although it is not a practical solution when multiple estimations are needed as the computational time could be a drawback.

Silverman showed in 1986 that for Gaussian densities the optimal kernel bandwidth is a thumb rule based on the standard deviation of the data [32]. The proposal for a univariate Gaussian distribution was the following one:

$$\sigma_k = \hat{\sigma}_{data} \left(\frac{4}{3N} \right)^{\frac{1}{5}} \cong 1.06\hat{\sigma}_{data} N^{-1/5} \tag{4.3}$$

For multivariate normal density, the formula generalizes as

$$\sigma_k = \hat{\sigma}_{data} \left(\frac{4}{N\left(2d+1\right)} \right)^{\frac{1}{4+d}} \tag{4.4}$$

The scope under these assumptions was to reduce at minimum the integrated mean squared error (IMSE) assuming true normal distributions. The IMSE cost function is specially used on KDE mainly because the bias of the integral of the pdf estimation leads to zero unlike the mean squared error (MSE) which is biased. The IMSE is defined by:

$$E\left[\int_{\mathbb{R}} \left(\hat{f_x}(x) - f_x\left(x\right)\right)^2 dx\right] \tag{4.5}$$

It has to be pointed out that for univariate kernels the optimal bandwidth is of order $O\left(N^{-1/5}\right)$ while for multidimensional kernels it is of order $O\left(N^{-1/(4+d)}\right)$. The point is that for multidimensional estimates larger bandwidths are needed. If we think in a two dimensional pdf it has some sense to use wider kernels to fill the spread data properly. Another point of view is to map this expression to the multivariate normal distribution, where the determinant of the covariance matrix Σ determines the volume of the data spread.

For kernel density estimator there is also another rule-of-thumb given by Scott in 1979 for a more general true densities [28]. The proposed for univariate variables was $\sigma_k = 3.49\hat{\sigma}_{data} N^{-1/3}$. Although Scott gives a good approximation of pdf, this work will maintain the Gaussian assumption for the reasons explained previously.

Back to pdf estimation, for multiple random variables the joint probability density function estimator extends to:

$$\hat{f_X}(X) = \frac{1}{N} \sum_{i=1}^{N} g\left(\frac{X - X_i}{\sigma_i} \right) \tag{4.6}$$

It can also be written with a multivariate kernel $g(x_1, ..., x_N)$ as following:

$$\hat{f}_{X_1,...,X_N}(x_1, ..., x_N) = \frac{1}{N}\sum_{i=1}^{N} g\left(\frac{x_1 - x_{1,i}}{\sigma_1}, ..., \frac{x_N - x_{N,i}}{\sigma_N}\right) \tag{4.7}$$

The multivariate pdf estimation have some more insights than the marginal estimation because of the σ variability. For different kernel bandwidth the equation transforms into a weighted summation kernels, being the marginal kernel bandwidth the weight for each kernel. The purpose is to add control to the multidimensional pdf estimation and to adapt every random variable to its own standard deviation.

At any case, the focus of this work is to measure independence between two random variables X and Y and so from now on a more maneuverable expression is going to be used. Do note that these expressions can always be generalized to multiple random variables.

Then, rewriting the previous expressions leads to

$$\hat{f}_{X,Y}(x, y) = \frac{1}{N}\sum_{i=1}^{N} g\left(\frac{x - x_i}{\sigma_x}, \frac{y - y_i}{\sigma_y}\right) \tag{4.8}$$

where $g(x, y)$ fulfills

- $g(x, y) \geq 0$ for any x and y
- $\int\int g(x, y)dxdy = 1$

Another reason to use Gaussian kernels is that under independence of the data, it is possible to apply separability on the joint density estimator, leading to

$$g(x, y) = g(x)g(y) \tag{4.9}$$

and the joint probability function becomes

$$\hat{f}_{X,Y}(x, y) = \frac{1}{N}\sum_{i=1}^{N} g\left(\frac{x - x_i}{\sigma_x}\right) g\left(\frac{y - y_i}{\sigma_y}\right) \tag{4.10}$$

4.2 Rényi entropy and mutual information

It is widely known that entropy $H(X)$ is a quantification of information or uncertainty of a random variable X, defined as

$$H(X) = -\sum_{x \in X} P_X(x) \log P_X(x) \tag{4.11}$$

and being X a discrete random variable, P_X its distribution and $P_X(x_i)$ the probability of observing x_i.

This approach was first introduced by Shannon [31] in 1948. This equation consolidates the information theory background given its usefulness. From there, other important statements can be derived as mutual information $I\left(X;Y\right)$, conditional entropy $H\left(X|Y\right)$, joint entropy $H\left(X;Y\right)$, Kullback-Leibler divergence $D\left(X||Y\right)$, and so on. So it is clear that Information Theory (IT) is a key stone on many applications and studies, by now and by the introduction.

The extension to continuous sources, called differential entropy $h(x)$, appeared the same year by both Shannon and Wiener [42] as

$$h(X) = -\int_{\mathbb{S}} f_X(x) \log\left(f_X(x)\right) dx \tag{4.12}$$

being X a continuous random variable with a pdf f_X defined on $\mathbb{S}$.

From here and recalling equation 4.1, the new estimate for marginal differential entropy becomes

$$\hat{h}(x) = -\int_{\mathbb{S}} \hat{f}_x(x) \log\left(\hat{f}_x(x)\right) dx \tag{4.13}$$

which can be rewritten as

$$\hat{h}(x) = -\frac{1}{N^2}\left(\int_{-\infty}^{\infty} \sum_{i=1}^{N} g\left(\frac{x-x_i}{\sigma_x}\right) \log\left(\sum_{i=1}^{N} g\left(\frac{x-x_i}{\sigma_x}\right)\right) dx\right) \tag{4.14}$$

At this point the difficulty of solving the equation has increased significantly due to the logarithm inside the integral. The inner part of it could be forced to be close to one in order to apply the fundamental inequality, but this would only force to stick to that regimen and it can be limiting afterwards. Luckily, Shannon entropy is only a particular case of a major family of entropies. This family consists in a generalization form that allows versatility and a more straightforward approach to the entropy estimator. This generalization came from Rényi at nearly 60s in the following form [24]:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{x \in X} P_X^\alpha(x) \right) \tag{4.15}$$

with the following properties:

- $\alpha > 0$
- Non-negative, concave and bounded.
- $H_\alpha(X)$ is a continuous function with probabilities $P_X(x)$, $x \in X$
- It converts to Shannon entropy when $\alpha \to 1$: $H(X) = \lim_{\alpha \to 1} H_\alpha(X)$
- For $\alpha \to 0$ it recalls Max or Hartley's entropy: $\log |X| = \lim_{\alpha \to 0} H_\alpha(X)$
- For $\alpha \to \infty$ it recalls Min or Chebyshev entropy: $-\log \max_i p_i = \lim_{\alpha \to \infty} H_\alpha(X)$
- $H_0 \geq H_1 \geq H_2 \geq ... \geq H_\infty$
- Additivity property remains as $H_\alpha(X,Y) = H_\alpha(X) + H_\alpha(Y)$ at independence, but not sub-additivity.
- Other Shannon entropy properties also remain, as permutationally symmetric, recursivity and monotonicity.

Note that α-entropy, or α-Rényi entropy, behaves like a group of functions for measuring uncertainty with different meanings for each α, being Shannon entropy a characterization of the generalized formula, with the addition that most of the properties that Shannon entropy holds are also fulfilled with the generalization of entropy. For instance, the Shannon entropy can be seen as a subset of the α entropy.

Although all the properties that inherits from Shannon entropy, the sub-additivity property is not fulfilled with Rényi entropies, as exposed in [8] and in [36], only case $\alpha \to 1$ and $\alpha \to 0$. This implication means that, under assumption of independence, it is no possible to bound the joint entropy like Shannon's $H(X,Y) < H(X) + H(Y)$ and so a possible detector of dependence cannot rely on it. On section 4.2.4 the estimator will be derived based on an alternative path.

Its extension to continuous sources is then formulated as

$$h_\alpha(X) = \frac{1}{1-\alpha} \log \left(\int_{\mathbb{S}} f_X^\alpha(x) dx \right) \tag{4.16}$$

This work will be focused in the case $\alpha = 2$, which is the called 2-Rényi entropy or Collision entropy. The interest in H_2 is multiple. First, as being a lower bound of Shannon entropy, it might be more efficient than Shannon's for entropy maximization, as explained in [23]. Additionally, the inner part of the logarithm is called Information potential V_2, which formally has the form

$$V_2 = \int_{\mathbb{S}} f_X^2(x)dx \tag{4.17}$$

and it can be described as monotonic decreasing function that can be estimated non-parametrically from pairwise sample differences as it is going to be seen, which is one of the main focus in this work. From its definition it can be seen that works as a 2 power norm of the pdf, which properties can come in handy lately.

At this point, the equation 4.13 can be reformulated to fit the Rényi entropy as

$$\hat{h_2}(x) = -\log\left(\int_{\mathbb{S}} \hat{f}_X^2(x)dx\right) = -\log(\hat{V_2}(x)) \tag{4.18}$$

and therefore

$$\begin{aligned}
\hat{h_2}(x) &= -\log\frac{1}{N^2}\int\limits_{-\infty}^{\infty}\left(\sum_{i=1}^{N} g\left(\frac{x-x_i}{\sigma_x}\right)\right)^2 dx \\
&= -\log\frac{1}{N^2}\int\limits_{-\infty}^{\infty}\sum_{i=1}^{N}\sum_{j=1}^{N} g\left(\frac{x-x_i}{\sigma_x}\right) g\left(\frac{x-x_j}{\sigma_x}\right) dx \\
&= -\log\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int\limits_{-\infty}^{\infty} g\left(\frac{x-x_i}{\sigma_x}\right) g\left(\frac{x-x_j}{\sigma_x}\right) dx \\
&= -\log\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k_{\sigma_x}\left(x_j - x_i\right)
\end{aligned} \tag{4.19}$$

Then, the information potential results in:

$$\hat{V_2}(x) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k_{\sigma_x}\left(x_j - x_i\right) \tag{4.20}$$

Do note that the result is only applicable with certain kernels. As previously pointed, Gaussian kernels are assumed, being then

$$k_\sigma\left(x_j - x_i\right) = \int\limits_{-\infty}^{\infty} g\left(\frac{x - x_i}{\sigma_x}\right) g\left(\frac{x - x_j}{\sigma_x}\right) dx = \exp\left(-\frac{\parallel x_j - x_i \parallel^2}{2\sigma_x^2}\right). \tag{4.21}$$

From equation 4.19 there is an additional step that is intrinsic on its definition, which is the matrix form expression. With this replacement, the formula transforms into a more intuitive form like the summation of all the elements is. Thus, the new expression transforms into

$$\hat{h}_2(x) = -\log\frac{1}{N^2}\left(\mathbf{1}^T \boldsymbol{K}_x \mathbf{1}\right) = -log\left(\hat{V_2}(x)\right). \tag{4.22}$$

The same process can be applied for joint 2-Rényi entropy. It is a necessary step as it is needed for measuring 2-Rényi mutual information and it will give some insights in its matrix form. For two random variables, the joint Rényi entropy is defined as:

$$\begin{aligned}
\hat{h_2}(x,y) &= -\log\left(\int\int \hat{f}^2_{X,Y}\left(x,y\right)dx\right) \\
&= -\log\left(\int\int\left(\frac{1}{N}\sum_{i=1}^{N} g\left(\frac{x-x_i}{\sigma_x}\right) g\left(\frac{y-y_i}{\sigma_y}\right)\right)\times\right. \\
&\qquad \left.\left(\frac{1}{N}\sum_{j=1}^{N} g\left(\frac{x-x_j}{\sigma_x}\right) g\left(\frac{y-y_j}{\sigma_y}\right)\right) dxdy\right) \\
&= -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int\int g\left(\frac{x-x_i}{\sigma_x}\right) g\left(\frac{y-y_i}{\sigma_y}\right)\times\right. \\
&\qquad \left.\left(g\left(\frac{x-x_j}{\sigma_x}\right) g\left(\frac{y-y_j}{\sigma_y}\right)\right) dxdy\right) \\
&= -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k_{\sigma_x}\left(x_j - x_i\right) k_{\sigma_y}\left(y_j - y_i\right)\right) \\
&= -\log\left(\frac{1}{N^2}\left(\mathbf{1}^T\left(\boldsymbol{K}_x \odot \boldsymbol{K}_y\right)\mathbf{1}\right)\right) \qquad (4.23) \\
&= -\log\left(\hat{V_2}\left(x;y\right)\right)
\end{aligned}$$

The final expression of the equation leads to think about a joint Rényi entropy as a summation of point to point comparison between two matrices built on pairwise kernel measures. At independence, and from additivity property, the expected result is the same as adding both separate kernel matrices summation, so it leads to think about this equation as a searcher of divergence between random vectors. At dependence it is trickier to make an analogy because the sub-additivity property is not fulfilled.

Equation 4.19 and equation 4.23 will be the pivoting cores to the future expressions for defining independence detectors. It reflexes a proper way to quantify the 2-Rényi entropy from an average of kernels evaluated to random samples differences.

But there is an stronger implication within these expressions. At this point we have arrived to the first important result in the work. It turns out that, given the Gaussian kernel $k_\sigma (x_j - x_i)$ is a semi-definite function defined by the inner product of kernels $\int g(x - x_i)\, g(x - x_j)\, dx$, the 2-Rényi entropy is then obtained through a kernel as defined in section 2.2. In fact, is the equation 4.21 that assures the semi-definite positive property of the kernel k. Hence, the matrices $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ are kernel matrices that maps the input data to a feature space. Actually, if we rebuild the joint information potential as

$$\hat{V_2}(x;y) = \frac{1}{N^2} tr\left(\boldsymbol{K}_x^T \boldsymbol{K}_y\right) \tag{4.24}$$

it is obtained a similar expression to the one in equation 3.9 which also looked for a relation even if it was the correlation. The point is, in order to measure the 2-Rényi entropy, we are looking for a relation in a higher dimensional space. When the detectors will be defined this implication will be important to analyze them.

Finally, to conclude the subsection, a formal definition for Rényi mutual information is needed. Although Rényi did not define the α mutual information, he actually defined a divergence measure for a general α based on the Kullback-Leibler divergence $D_{KL}(f||g)$, and from there it is possible to define the mutual information. In [25] it can be seen the α divergence as

$$D_\alpha (f \parallel g) = \frac{1}{1-\alpha} log \int f(x) \left(\frac{f(x)}{g(x)}\right)^{\alpha - 1} dx. \tag{4.25}$$

Its properties are the following ones:

- $D_\alpha (f||g) \geq 0 \ , \forall f, g \ , \alpha > 0$
- $D_\alpha (f||g) = 0$ iif $f(x) = g(x)\ \forall x \in \mathbb{R}$
- $\lim\limits_{\alpha \to 1} D_\alpha (f||g) = D_{KL}(f||g)$

which can be proven as in [23].

From there, a definition for the α mutual information can be built based on the property of Shannon's mutual information that fulfills

$$I(x) = D_{KL}(f_x(x_1, ..., x_n)||f_x(x_1)f_x(x_2)...f_x(x_n)) \tag{4.26}$$

considering a continuous n-dimensional random variable X, with marginal distributions $f_X(x_o)$ and joint distribution $f_X(x_1, ..., x_n)$. Then it is possible to use this property and equation 4.25 to built the mutual information as follows:

$$I_\alpha(x) = \frac{1}{1-\alpha} log \int ... \int \frac{f_x^\alpha(x_1, ..., x_n)}{\prod\limits_{o=1}^{n} f_x^{\alpha-1}(x_o)} dx_1 dx_2 ... dx_n \tag{4.27}$$

In the case of two continuous random variables used in this work, being called x and y, and assuming the $\alpha = 2$ case or the called 2-Rényi mutual information, the equation simplifies to

$$I_2(x; y) = -log \int \int \frac{f_{x,y}^2(x, y)}{f_x(x) f_x(y)} dxdy \tag{4.28}$$

Do note this expression is similar to the used to explain the Low SNR regimen, which will provide an effective form to estimate the Shannon's mutual information through the estimation of the pdf's. This is in fact what is going to be done in the subsection 4.2.2.

4.2.1 Bias, variance and kernel bandwidth

At this point it is important to make a pause before deriving the detectors in order to analyze the 2−Rényi estimate by the means of it expected value and variance. The interesting point is that both of them are highly correlated with the called kernel bandwidth σ_k, and so its selection will consolidate as an important task throughout the rest of the work.

Let us first take the marginal 2−Rényi entropy defined by

$$\hat{h}_2(x) = -\log \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} k_\sigma(x_j - x_i) \tag{4.29}$$

Being the logarithm a monotonic function it will not provide any contribution, alongside with the constant, so we will not count them. Additionally, we can also remove the non-informative elements, and so the cases $i = j$, which only provides a bias to the estimator. The $i > j$ cases are repetitive, and so they will also not provide information. Hence, lets express the estimator as

$$\hat{\hat{h}}_2(x) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N}^{N} k_\sigma(x_j - x_i) \tag{4.30}$$

Within this bias removal, and being the sum contained by i.i.d. elements, we have now a called Unbiased statistic, and so it is interesting to review it.

The U-Statistics are defined as a class of statistics that allows to derive a minimum-variance unbiased estimator (MVUE) from an unbiased estimator of a parametric function $\theta = \theta(\boldsymbol{x})$. The introduction to the U-Statistics were first given by Hoeffding at 1948 within [15]. A more general discussion of U-Statistics can be found in [29], and for specifically on non-parametric models at [9].

The interest on U-Statistics for the non-parametric estimator desired lies in the possibility of building an estimator close to the asymptotic mean or variance of the random samples. In general, from an estimator built from samples $x_1, ..., x_N$ defined by the distribution $f(x)$, the estimator may be represented as:

$$\theta(\boldsymbol{x}) = E\left[h\left(x_1, ..., x_N\right)\right] = \int ... \int h\left(x_1, ..., x_N\right) df(x_1)...df(x_N) \tag{4.31}$$

given a generic kernel function $h\left(x_1, ..., x_N\right)$. Then, the corresponding U-statistic for estimate θ on a data of $M \leq N$ samples is obtained through averaging the kernel symmetrically over observations:

$$U_N\left(x_1, ..., x_N\right) = \frac{1}{\begin{pmatrix} N \\ M \end{pmatrix}} \sum_c h\left(x_{i_1}, ..., x_{i_M}\right) \tag{4.32}$$

with c denoting the the subset of M random samples obtained from the total N.

The richness of using U-statistics in our case is the processing of pairwise data that defines Serfling in his book. He was trying to evaluate the statistics from pairs of data of $2, 3, 4...$ and to extend to the case where M were taken. From the pairwise perspective, we would have a U-statistic of $M = 2$.

For instance, if we do an estimation of the variance of the data samples $\theta = \sigma_x^2 = var\left(\boldsymbol{x}\right)$, we can express the variance of a pair of data, and so $M = 2$, as

$$\sigma_{i,j}^2 = \frac{var\left(x_i\right) + var\left(x_j\right)}{2} = E\left[\frac{\left(x_i - x_j\right)^2}{2}\right] \tag{4.33}$$

Then, defining a pairwise kernel that holds the data as the pairwise variance from the previous equation $h\left(x_i, x_j\right) = \left(x_i - x_j\right)^2/2$, the U-statistic holds that

$$U_N = \frac{2}{N\left(N-1\right)} \sum_{1 \leq i < j \leq N}^{N} \frac{\left(x_i - x_j\right)^2}{2} = \frac{1}{N\left(N-1\right)} \left(\sum_{i=1}^{N} x_i^2 - N\mu^2\right) = \sigma_x^2 \tag{4.34}$$

The point is, when estimating the variance through a pairwise kernel, the result is the variance of the data. Do note that it has a great similarity with equation 4.30, with the exception of the kernel used. If we now consider that the kernel bandwidth is sufficiently large, for instance $\sigma_k \to \infty$, then the pairwise measures of the Gaussian kernel leads to measures on the parabolic regime of the Gaussian, and so we can state

$$\frac{2}{N\left(N-1\right)}\sum_{1\leq i<j\leq N}^{N} k_{\sigma}\left(x_j - x_i\right) \underset{\sigma_k\to\infty}{\longrightarrow} \frac{2}{N\left(N-1\right)}\sum_{1\leq i<j\leq N}^{N}\left(1-\frac{\left(x_j-x_i\right)^2}{\alpha}\right) \tag{4.35}$$

being $\alpha \in \mathbb{R}$ a constant with $\alpha > 0$. Finally, by the sense of the U-statistics, we can state that the expectation of the unbiased estimator tends to $1 - \sigma_x^2/\alpha$. The interpretation is that the $2-$Rényi entropy tends to be affine to a second-order statistic of the U-statistics when using wide kernels. The expectation of the opposite case, and so $\sigma_k \to 0$, would lead to pairwise measures of high narrow band functions, and so to 0.

Another implication is more related with the change of the kernel function used to determine the expectation. On Section 4.3 we will see that, for any semi-definite positive kernel, the called kernel trick allows versatility in the sense that we can change the first kernel to a second one also semi-definite positive and the results does not alter. In this case the change of kernel implies that, even using a Gaussian kernel, on the feature space we are measuring the a metric proportional to the variance of the data.

4.2.2 Mutual information estimation

Once having a proper definition for the Rényi entropy estimation and Rényi mutual information, the next natural step is to introduce an estimation of this last. From equation 4.28, it could be pretty intuitive to replace the joint pdf and marginal pdf with the estimation, but it is actually more clever to use the structure of the Rényi entropy and the expectation theorem. If we define the 2-Rényi mutual information for continuous sources as the following one:

$$I_2(x;y) = \log(C(x;y)) \tag{4.36}$$

where

$$C(x;y) = \int\int \frac{f_{x,y}^2(x,y)}{f_x(x)f_y(y)}dxdy = \int\int f_{x,y}(x,y)\frac{f_{x,y}(x,y)}{f_x(x)f_y(y)}dxdy \tag{4.37}$$

Then it is possible to express $C(x, y)$ as an expectation like

$$C(x; y) = E\left[\frac{f_{x,y}(x, y)}{f_x(x) f_y(y)}\right]. \tag{4.38}$$

Now, if we assume a sufficiently large number of random values for each X and Y it is possible to recall the law of large numbers, where the sample average converges to the expectation, we get the following expression:

$$\hat{C}(x; y) = \frac{1}{N}\sum_{j=1}^{N} \frac{f_{x,y}(x_j, y_j)}{f_x(x_j) f_y(y_j)} \tag{4.39}$$

At this point the integral has been removed and we can process freely so substitute this equation with 4.6 and 4.10, resulting in

$$\hat{C}(x; y) = \frac{1}{N}\sum_{j=1}^{N} \frac{\frac{1}{N}\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right) g\left(\frac{y_j - y_i}{\sigma_y}\right)}{\left(\frac{1}{N}\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right)\right)\left(\frac{1}{N}\sum\limits_{i=1}^{N} g\left(\frac{y_j - y_i}{\sigma_y}\right)\right)} \tag{4.40}$$

Finally, the first approach to Rényi mutual information estimator is the resulting by combining the previous expression and the definition in 4.36:

$$\hat{I}_2(x; y) = \log\left(\sum_{j=1}^{N} \frac{\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right) g\left(\frac{y_j - y_i}{\sigma_y}\right)}{\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right) \sum\limits_{i=1}^{N} g\left(\frac{y_j - y_i}{\sigma_y}\right)}\right) \tag{4.41}$$

This is the first attempt on estimating the Rényi mutual information. It has an intuitive form like Shannon's mutual information where joint pdf is confronted with the product of marginal pdf. At independence, the joint pdf should result in a measure similar to the product of marginal pdf and so a close to zero measure at its whole. For a dependence hypothesis, it is not completely reliable as because it is not possible to assure the independence assumption solution. We do need to bound this estimation in order to consolidate a detector.

Then, an alternative path has to be found. If we stick on the low SNR assumption, it is possible to write the following inequality:

$$C(x; y) - 1 \geq I(x; y) \geq 0, \tag{4.42}$$

being $I(X,Y)$ the Shannon mutual information. This inequality can be obtained from 2.16, and it makes our detector the following:

$$\hat{C}(x;y) - 1 = \sum_{j=1}^{N} \frac{\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right) g\left(\frac{y_j - y_i}{\sigma_y}\right)}{\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right) \sum\limits_{i=1}^{N} g\left(\frac{y_j - y_i}{\sigma_y}\right)} - 1 > \gamma \tag{4.43}$$

Note that, as the logarithm is monotonic, the previous expression is equivalent to this one:

$$T_1\left(\boldsymbol{x}, \boldsymbol{y}\right) = \log\left(\sum_{j=1}^{N} \frac{\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right) g\left(\frac{y_j - y_i}{\sigma_y}\right)}{\sum\limits_{i=1}^{N} g\left(\frac{x_j - x_i}{\sigma_x}\right) \sum\limits_{i=1}^{N} g\left(\frac{y_j - y_i}{\sigma_y}\right)}\right) > \gamma^{'} \tag{4.44}$$

resulting in the first independence detector of this work.

As viewed, this detector is obtained through the search of estimating the 2-Rényi entropy, and we have noticed that we can use it as a detector of independence through the low SNR regimen. Although it has some potential to be described as a kernel matrix, its own structure makes difficult to do so, so we are going to stick under the Shannon's mutual information resemblance.

4.2.3 Mutual information estimation based on non-additivity of joint Rényi entropy

The sub-additivity property of Shannon's joint entropy gives a very interesting approach on detecting dependence, but lets not forget that Rényi entropies does not fulfill this property. At any case, Rényi entropies does fulfill the additivity properties, so even if it is not possible to assert that the joint 2-Rényi entropy is always above the sum of the marginals at dependence, it is possible to assume that the value of the joint will be close to that value.

In this case, the detector can be formulated based on the lack of equality between the joint 2-Rényi entropies and the marginal ones:

$$\mid H_2\left(X;Y\right) - H_2\left(X\right) - H_2\left(Y\right) \mid > 0 \tag{4.45}$$

The proposed detector for continuous random variables is:

$$\begin{aligned}
&\left|\hat{h}_2\left(x;y\right)-\hat{h}_2\left(x\right)-\hat{h}_2\left(y\right)\right|>\gamma\\
&\left|\log\hat{V}_2\left(x;y\right)-\log\hat{V}_2\left(x\right)-\log\hat{V}_2\left(y\right)\right|>\gamma\\
&\left|\log\left(\frac{\hat{V}_2\left(x;y\right)}{\hat{V}_2\left(x\right)\hat{V}_2\left(y\right)}\right)\right|>\gamma
\end{aligned}\tag{4.46}$$

which can be simplified by the assumption of near independence, or low SNR assumption, as follows:

$$\left|\frac{\hat{V}_2\left(x;y\right)}{\hat{V}_2\left(x\right)\hat{V}_2\left(y\right)}-1\right|>\gamma\tag{4.47}$$

Finally, from equation 4.19 and 4.23, we get the following final expression for the second detector proposed:

$$\begin{aligned}
T_2\left(\boldsymbol{x},\boldsymbol{y}\right)&=\left|N^2\frac{\left(\mathbf{1}^T\left(\boldsymbol{K}_x\odot\boldsymbol{K}_y\right)\mathbf{1}\right)}{\left(\mathbf{1}^T\boldsymbol{K}_x\mathbf{1}\right)\left(\mathbf{1}^T\boldsymbol{K}_y\mathbf{1}\right)}-1\right|\\
&=\left|N^2\frac{tr\left(\boldsymbol{K}_x\boldsymbol{K}_y\right)}{\left(\mathbf{1}^T\boldsymbol{K}_x\mathbf{1}\right)\left(\mathbf{1}^T\boldsymbol{K}_y\mathbf{1}\right)}-1\right|>\gamma
\end{aligned}\tag{4.48}$$

The most important implication in the second detector is that we have naturally derived a properly kernel processing as pointed out in the information potential estimator. Given its structure, the detector looks for dependence relations in the new feature space and compare it with the marginal ones. It actually is intrinsically the same metric of what we are doing in the first detector, but instead of doing it column by column, this new detector does it for the data at its whole.

4.2.4 Mutual information estimation based on Cauchy-Schwartz inequality approach

The lack of sub-additivity of the joint Rényi entropy suggest a slight modification of the previous detector, but this time based on U-Statistics and the Cauchy-Schwartz inequality.

In order to derive the estimator into a U-statistics, it is needed to remove the bias from it. As known from Section 4.2.1, the joint information potential $\hat{V}_2(X,Y)$ does not provide information over the evaluation of the kernel with the the samples with themselves additionally with the repeated terms. Then, it is desirable to remove them:

$$\begin{aligned}\hat{V}_2(x;y) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k_{\sigma x}\left(x_j - x_i\right) k_{\sigma y}\left(y_j - y_i\right) \\ &= \frac{1}{N} + \frac{2}{N\left(N-1\right)}\sum_{1\le i<j\le N}^{N} k_{\sigma x}\left(x_j - x_i\right) k_{\sigma y}\left(y_j - y_i\right) \qquad (4.49)\end{aligned}$$

Considering the informative second term

$$\hat{\hat{V}}_2(x;y) = \frac{2}{N\left(N-1\right)}\sum_{1\le i<j\le N}^{N} k_{\sigma x}\left(x_j - x_i\right) k_{\sigma y}\left(y_j - y_i\right) \qquad (4.50)$$

it is now a U-statistic $\hat{\hat{V}}_2(X,Y) = U\left(x_1, ..., x_M\right)$. As explained, it can be seen as an unbiased estimate of the statistical mean of the kernel function based on the samples like:

$$E\left[\hat{\hat{V}}_2(x;y)\right] = E\left[k_{\sigma x}\left(x_a - x_b\right) k_{\sigma y}\left(y_\mathrm{a} - y_b\right)\right] \qquad (4.51)$$

Now, using the Cauchy-Schwartz inequality, we can state that

$$\left(E\left[k_{\sigma x}\left(x_a - x_b\right) k_{\sigma y}\left(y_\mathrm{a} - y_b\right)\right]\right)^2 \le E\left[k_{\sigma x}^2\left(x_a - x_b\right)\right] E\left[k_{\sigma y}^2\left(y_\mathrm{a} - y_b\right)\right] \qquad (4.52)$$

with equality in case of dependence. Therefore the following inequality can be proposed:

$$\frac{E\left[k_{\sigma x}^2\left(x_a - x_b\right)\right] E\left[k_{\sigma y}^2\left(y_\mathrm{a} - y_b\right)\right]}{\left(E\left[k_{\sigma x}\left(x_a - x_b\right) k_{\sigma y}\left(y_\mathrm{a} - y_b\right)\right]\right)^2} - 1 > 0 \qquad (4.53)$$

for finally, reversing the U-statistic transform and leading into the third proposed detector:

$$\frac{\sum\limits_{1\le i<j\le N}^{N} k_{\sigma x}^2\left(x_j - x_i\right) \sum\limits_{1\le i<j\le N}^{N} k_{\sigma y}^2\left(y_j - y_i\right)}{\left(\sum\limits_{1\le i<j\le N}^{N} k_{\sigma x}\left(x_j - x_i\right) k_{\sigma y}\left(y_j - y_i\right)\right)^2} - 1 > \gamma \qquad (4.54)$$

To express in matrix form, the matrix $\boldsymbol{L}$ will be defined as a $N \times N$ upper triangular matrix containing $N\left(N-1\right)/2$ ones, with the following structure:

$$\boldsymbol{L}=\begin{bmatrix} 0 & 1 & . & . & 1 \\ & . & . & . & . \\ & . & & . & . & . \\ & . & & & . & 1 \\ 0 & . & . & . & 0 \end{bmatrix} \tag{4.55}$$

Then, if we define $\tilde{\boldsymbol{K}}_x=\boldsymbol{L}\odot\boldsymbol{K}_x$, the third detector is expressed as

$$\begin{aligned} T_3\left(\boldsymbol{x},\boldsymbol{y}\right) &= \frac{\left(\boldsymbol{1}^T\left(\tilde{\boldsymbol{K}}_x\odot\tilde{\boldsymbol{K}}_x\right)\boldsymbol{1}\right)\left(\boldsymbol{1}^T\left(\tilde{\boldsymbol{K}}_y\odot\tilde{\boldsymbol{K}}_y\right)\boldsymbol{1}\right)}{\left(\boldsymbol{1}^T\left(\tilde{\boldsymbol{K}}_x\odot\tilde{\boldsymbol{K}}_y\right)\boldsymbol{1}\right)^2}-1 \\ &= \frac{tr\left(\tilde{\boldsymbol{K}}_x\tilde{\boldsymbol{K}}_x\right)tr\left(\tilde{\boldsymbol{K}}_y\tilde{\boldsymbol{K}}_y\right)}{tr\left(\tilde{\boldsymbol{K}}_x\tilde{\boldsymbol{K}}_y\right)^2}-1>\gamma \end{aligned} \tag{4.56}$$

Despite the structure of this detector is similar to the previous one, specially on the denominator, the final expression leads to different meanings. This time we are removing the non-informative part of the kernel matrix, and so the repeated values, and we are evaluating the data spammed in the infinite space of the informative pairs. This means that we are reducing the data compared in the unknown space and so it is expected for the detector to be less sensitive in terms of bias-variance trade-off. Even though the kernel bandwidth still affects in this affair.

The approach is interesting in the sense that what is desired is the unbiased estimator instead of a more general expression, and the Cauchy-Schwartz inequality provides a metric in order to compare the dependence relation.

4.3 The kernel idea

Given the last two detectors proposed we can get some insights of what is happening when measuring the dependence. Remembering that these detectors have been drawn from the Parzen window estimate and the properties of the 2−Rényi entropy, it turns out that they are actually based on kernel mapping. This is a little of an interesting result given the fact that Parzen window estimate uses a kernel as a mass-particle and not as a kernel as defined in section 2.2. As a matter of fact, it is possible to think about some implications:

- The evaluation of the data among the integral operators of the function g has been derived as a kernel of point to point differences, moving from the general description of a kernel $k\left(x_i, x_j\right)$ to a more specific for our case $k\left(x_j - x_i\right)$. Hence we have moved from the inner product to pairwise differences among the data samples.

- In order to get a comparable detection for $\boldsymbol{x}$ and $\boldsymbol{y}$, the kernels that maps the data should be the same, and for this we can impose $k_{\sigma x}(x, y) = k_{\sigma y}(x, y)$. The same happens when using different hypothesis testing, different kernel bandwidth for each hypothesis would not lead to a proper detection as the bias-variance from each case should be similar to be compared.

Being the second important for the detectors built, the first has some important insights. In kernel processing in general we are interested on the inner products of the data, and so we look about linear implications. Within the use of pairwise Gaussian kernels, it is maintained the property of measuring these inner products in a infinite-dimensional space through the kernel function in the data input space. This implies that, if an algorithm is constructed based on semi-definite positive kernels it is possible to replace that kernel by another semi-definite positive kernel and the measures in the feature space will remain. That's why Gaussian kernels, among others, are said to be universal [38]. This is known as kernel trick [37] and it consolidates the basis of the kernel signal processing due to the advantages explained.

Another implication that will be useful is given by the structures of the kernel matrices. Even though we don't know about the feature map provided by the kernels, we do know there is some type of relation in that space, reflected in the measurement of the data through the kernel matrices. Being them semi-definite positive and Gram matrices, it is possible to exploit the given structure to reduce the computational complexity derived from the measurement of the data at the feature map. Doing so, the computational complexity of the detectors can be reduced. We will come back to this at Section 4.5.

4.4 Hilbert-Schmidt norms

We have seen until now the properties of a kernel and its extension to a RKHS. The interest now is to use these properties to build a detector of dependence through a cross-covariance detector. The bridge that allow the link between the two of them is going to be, in fact, the Hilbert space. The mapping function to the feature space $f : \mathcal{X} \to \mathbb{R}$ that defines a Hilbert space $\mathcal{H}$ allows to constrict the dimension of measures, or in another words, it makes possible to get insights of a much superior or infinite-dimension without the need of visiting it. A good reference for this property are the works by Smola and Gretton et al. ([13, 12]).

To make it possible, the Hilbert-Schmidt norm has to be defined. Consider A a linear operator that maps between two RKHS defined by the function f and g, with the corresponding Hilbert Spaces $\mathcal{F}$ and $\mathcal{G}$, and so $A : \mathcal{F} \rightarrow G$. If both Hilbert spaces are separable, so they contain an orthonormal system defined by u and v orthonormal bases, then the norm of the operator A is defined as

$$\|A\|_{HS}^2 \triangleq \sum_{i,j} \langle Au_i, v_j \rangle_2^2 \tag{4.57}$$

Do note that for matrix operators it corresponds to the Frobenius norm, and so a Hilbert-Schmidt norms are extensions of the Frobenius norm for Hilbert spaces. The interesting approach to this norm is that, in fact, we can use it to measure a cross-covariance operator. Recovering the equation 3.9, we can state that, if matrcesx $\boldsymbol{X}$ and $\boldsymbol{Y}$ are kernel based, and so we have used a semi-definite positive kernel to map raw data to a Hilbert space, then the Frobenius norm of the cross-covariance will lead to a Hilbert-Schmidt norm of the cross-covariance. The reasoning behind is to us these metrics to measure a Frobenius norm of an unknown space through a Hilbert-Schmidt norm, and so avoid functional analysis.

Despite the convenience of the Hilbert-Schmidt norm, it is still needed a mathematical reason to formulate a dependence detector based on this norm.

4.4.1 Characteristic function uniform sampling interpretation

Given a random variable x with density probability function $f_x(x)$, then the characteristic function that defines it is the Fourier transform of the pdf:

$$\phi_x(u) = E\left[e^{jux}\right] = \int_{\mathbb{R}} f_x(x)e^{jux}du \quad u \in \mathbb{R} \tag{4.58}$$

The characteristic function can always be obtained if there exist a pdf of the random variable, as well as the pdf always exist if a random variable can be expressed by its characteristic function through the inverse Fourier transform (FT). The existence of the FT is a necessary condition for the existence of the characteristic function, while the summability of the squared pdf is a sufficient condition. For a joint pdf, the characteristic function extends to

$$\phi_{x,y}(u;v) = \int \int_{\mathbb{R}} f_{x,y}(x,y) e^{jux+jvy} dxdy = E\left[e^{jux+jvy}\right] = E\left[e^{jux}e^{jvy}\right] \quad u,v \in \mathbb{R} \tag{4.59}$$

Do note that it is a special case of a FT. While the usual FT is evaluated through a temporal domain, this case do evaluate it for another domain: the values of the pdf of the original random variable. It is actually pretty direct considering that the pdf is a function of likeliness given a determined pdf's value. A good reference for other works that measures FT as "amplitude" transforms can be found in [20].

Another useful implication under the characteristic function is that, for a true pdf of Gaussian shape, the characteristic function is also Gaussian. For instance, if we define $f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$, its characteristic function will be $\phi_x(u) = e^{-\sigma^2 u^2/2}$, providing some interest in being stick under Gaussian assumption. Then, given its exponential and expectation form, one interesting inherent property is the separability when x and y are independents. Hence the following relation is accomplished:

$$\phi_{x,y}(u;v) = E\left[e^{jux}\right] E\left[e^{jvy}\right] \tag{4.60}$$

The expression of the characteristic function allows to express it as $z_1 = e^{jux}$, which can be seen as function that transforms x from its own space to another space or feature space. If we stick in independence assumption, the function z_x and z_y defined from the variables x and y hold that

$$E\left[z_x z_y^*\right] = E\left[z_x\right] E^*\left[z_y\right] \tag{4.61}$$

or equivalently

$$E\left[\left(z_x - \bar{z}_x\right)\left(z_y - \bar{z}_y\right)^*\right] = 0 \tag{4.62}$$

which results in the same structure as equation 3.1 and so it consolidates an uncorrelation detector under independence assumption. This is a very simple way to obtain an independence detector through the correlation of z_x and z_y for any real values of u and v. The drawback is that the uncorrelation property has to be verified for infinite number of pairs $\{u,v\}$, and so the dimension of the feature space is infinite. This leads to a search of methodologies that maintain the properties of infinite spaces but in finite spaces in order to build the desired detector.

A first approach can be obtained if the feature space is constrained by assuming finite support of the joint characteristic function, thus finite support of the marginal characteristic functions for $-U > u > U$ and for $-V > v > V$ as follows:

$$\phi_{x,y}(u;v) = 0 \quad \phi_x(u) = \phi_y(v) = 0 \quad -U > u > U,\, -V > v > V \tag{4.63}$$

The constrain over the characteristic function has implicit that the pdf of the random variable has also to be finite. It is an extension of the classical time-frequency trade-off. This is a strange case in terms of continuous random variables, being at most cases sufficiently large or infinite like in the Gaussian case due is tails. Then, the constraint can be relaxed by admitting an arbitrary small number instead of a zero:

$$|\phi_{x,y}(u;v)| < \varepsilon \tag{4.64}$$

Although this is an interesting approach, it is not going to be further developed because of it would open an another path different from kernel processing, distracting of the main focus of the work. However, the interest with characteristic function is not ended.

This time we will limit the dimension of the mapped space by doing a sample of the discordant element u by $n\triangle$, being $n = 1, ..., N_z$. This way, the vector form of the samples characteristic function is

$$\boldsymbol{z}_{x,i} = \left\{ e^{jn\Delta x_i} \right\}_{n=1...N_z} \tag{4.65}$$

$$\boldsymbol{z}_{y,i} = \left\{ e^{jn\Delta y_i} \right\}_{n=1...N_z}, \quad i = 1...N \tag{4.66}$$

being a $\boldsymbol{x}$ a sequence of scalar random samples $\{x_i\}_{i=1,..,N}$, which follows the assumption of lack of knowledge about the random variables but random data samples. With this, the constraint of infinite space has been relaxed and we are in terms of analyzing the correlation in order to study its implication with mutual information. Additionally, do note that N_x and N_y defines the dimensionality of the new feature space.

The covariance matrix is then the following:

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{z}_1^H \boldsymbol{z}_1 & . & . & . & \boldsymbol{z}_1^H \boldsymbol{z}_N \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ \boldsymbol{z}_N^H \boldsymbol{z}_1 & . & . & . & \boldsymbol{z}_N^H \boldsymbol{z}_N \end{bmatrix}$$

being the (i, j).th element of the correlation matrix $R_{i,j} = \sum_{n=1}^{N_z} e^{-jn\Delta x_i} e^{jn\Delta x_j}$. At this point it can be questionable to use this sampling metric as a good approximation of the real mapping of the characteristic function, specially because of the lack of well-defined scalar products. In concrete, we are interested in the case $\sum_{n=1}^{N_z} e^{-jn\Delta x_i} e^{jn\Delta x_j}$ when $N_z \to \infty$ in order to get a well-defined scalar product. To do so, the mapping can be modified to assure finite norm functions.

To avoid infinite sampling, let us modify the mapping by rewriting the vectors as functions of λ as

$$\boldsymbol{z}_{x,i} = z_{x,i}(\lambda) = e^{j\lambda x_i} G(\lambda) \tag{4.67}$$

being $G(\lambda)$ a window with $\int |G(\lambda)|^2 \, d\lambda = 1$ in order to maintain the inner products. This way, the function has unit norm

$$\int_{\infty}^{\infty} |z_{x,j}(\lambda) \, d\lambda|^2 \, d\lambda = 1$$

and the scalar product is defined:

$$\langle \boldsymbol{z}_{x,i}, \boldsymbol{z}_{x,j} \rangle \; = \int_{-\infty}^{\infty} z_{x,i}^*(\lambda) \, z_{x,j}(\lambda) \, d\lambda \; = \int_{-\infty}^{\infty} e^{j\lambda(x_j - x_i)} |G(\lambda)|^2 \, d\lambda \tag{4.68}$$

which is actually related with the product between the vectors, and so we are trying to avoid an infinite sampling by adding the window.

Do note that λ defines the values of u that are a priori relevant. With so, we are again in the same constrain as in equation 4.63. However, we are now more interested on reducing the values for which $|G(\lambda)|^2 \approx 0$, and so we are assuming the true pdf to decay rapidly.

If the expression is looked carefully, we note that by substituting the pairwise measures $x_j - x_i$ with another variable α, we can express it as a function as follows

$$k(\alpha) \; = \int_{-\infty}^{\infty} e^{j\lambda a} |G(\lambda)|^2 \, d\lambda \; = \int_{-\infty}^{\infty} g(\alpha - \beta) \, g^*(\beta) \, d\beta \tag{4.69}$$

which turns out that it corresponds to the inverse Fourier transform of the window $G(\lambda)$. The gimmick is to define the window as Gaussian, which does not imply any restriction, with unitary area:

$$|G(\lambda)|^2 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\lambda^2/2\sigma^2} \tag{4.70}$$

and to define $k(\alpha)$ as a finite-energy signal $|k(\alpha)| \leq k(0) = 1$, symmetric $k(\alpha) = k(-\alpha)$, real, and asymptotically zero when $\alpha \to \pm\infty$, and so a valid autocorrelation function.

Then, the function k can be expressed as

$$k\left(\alpha\right) = e^{-\alpha^2\sigma^2/2} = e^{-\left(x_j - x_i\right)^2\sigma^2/2} \tag{4.71}$$

resulting in a kernel function of pairwise measures that has been used until now. Do note that the parameter σ is still controversial due to the fact that is not the same as the kernel bandwidth σ_k described until now, but it is highly related to it. In this case σ provides, as explained, the bandwidth of $G\left(\lambda\right)$ and so which values of the characteristic function are supposed irrelevant. It is important to assure the independence case for any of the values of the origin space, to assure 4.60 to be still valid, and so to assure that characteristic functions are still separable. To do so it is correct to assume the convergence of $G\left(\lambda\right)$ as necessary to provide this link of independence, removing importance from the tails and centering where the information must be evaluated.

At any case, the value of σ is related to the kernel bandwidth in the sense that it will highly depend one from the another. A higher σ provides a more sensitive evaluation but a more restrictive kernel bandwidth, and vice-versa. To sum up, it is a trade off between sensitivity and variance.

Recalling the steps followed, we have started from a characteristic function uniform sampling, we have seen that its correlation can be obtained from pairwise measures, and the inner product results in a Gaussian kernel function. Within these steps, it is intrinsically defined a proper kernel with well defined inner products, and so it is possible to state that it provides the mapping to an infinite feature space. The correlation matrix is now defined as a kernel matrix with elements $K_{i,j} = e^{-\left(x_j - x_i\right)^2\sigma^2/2}$, which is actually a Gram matrix related to the correlation from the characteristic function. Now we can use this knowledge to build a cross-correlation detector, which result to be the Frobenius norm of the cross-covariance matrix estimation as in equation 3.9:

$$\left\|\hat{\boldsymbol{C}}_{x,y}\right\|_F^2 = \left(\frac{1}{N-1}\right)^2 tr\left(\boldsymbol{P}\boldsymbol{K}_x\boldsymbol{P}\boldsymbol{K}_y\right) \tag{4.72}$$

But what is more is that, given the Gram matrices are kernel based, the measure of this norm is actually a Hilbert-Schmidt norm as defined in equation 4.57, and so this detector is actually measuring the correlation in a infinite space. Remembering that the condition of independence with the characteristic functions was to assure the separability in a infinite space, we have now obtained a independence detector by using the Hilbert-Schmidt norm that looks for that independence in the desired space. Hence, the fourth detector proposed is the following:

$$T_4\left(\boldsymbol{x}, \boldsymbol{y}\right) = \left(\frac{1}{N-1}\right)^2 tr\left(\boldsymbol{P}\boldsymbol{K}_x\boldsymbol{P}\boldsymbol{K}_y\right) > \gamma \tag{4.73}$$

The insight is that we are now capable of measuring independence by measuring correlation on the feature space mapped by the kernels, which differs from other detectors in the sense that this time it is correlation based. Generally speaking, the interesting approach of this detector is that it can relate independence and incorrelation instead of separating them, even if this relation appears on an infinite space.

4.5 Fast computation through Incomplete Cholesky Decomposition

All the detectors presented until now have a polynomial computational complexity of the order $O\left(N^2\right)$ due to the kernel matrices dimensions. This is a direct result of deriving the detectors from a Dual form expression. It is clear that is desirable to reduce the polynomial rate to another relation. For this purpose, the first thought is to decompose the matrix in order to reduce its complexity. The most common used in these cases is the QR decomposition, but given the property of the kernel matrix of being definite semi-positive, the particular case is called Cholesky decomposition. A reference for this decomposition can be found in [10] and [35].

Given a $N \times N$ symmetric and positive definite matrix $\boldsymbol{K}$, the matrix can be decomposed as $\boldsymbol{K} = \boldsymbol{G}^T\boldsymbol{G}$, where matrix $\boldsymbol{G}$ corresponds to an $N \times N$ upper triangular matrix. This decomposition can be obtained through the reproducing property of the positive semi-definite kernels. Given a kernel k, its function ψ that maps into a feature space and a the matrix $\boldsymbol{X}$ defined in section 3.1.1 of dimension $M_x \times N$ whose columns are projections of a data set, a QR decomposition can be applied in the feature space such that

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{G} \tag{4.74}$$

being $\boldsymbol{Q}$ an $M_x \times N$ orthogonal matrix containing the orthonormal basis derived from the Gram-Schmidt method. Recalling that we are working on a regimen of $M_x >> N$, which is an atypical situation in signal processing. Then $\boldsymbol{Q}$ holds that

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I} \tag{4.75}$$

Now we are settled to define the decomposition of the matrix:

$$\boldsymbol{K} = \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{G}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{G} = \boldsymbol{G}^T\boldsymbol{G} \tag{4.76}$$

The (i,j) element of the matrix $\boldsymbol{G}$ is obtained by the Gram-Schmidt decomposition as follows:

$$G_{i,j} = \frac{1}{\sqrt{d_j}}\left(K_{i,j} - \sum_{k=1}^{j-1} G_{k,j}G_{k,i}\right) \quad i = j+1, ..., N \tag{4.77}$$

which corresponds to evaluate the inner product $\langle\psi(x_j), \psi(x_i)\rangle = K_{i,j}$ and the basis vector $\boldsymbol{q}_j$ for $i > j$, that corresponds to the component lying in the feature space by the basis vectors to $j-1$: $\sum\limits_{k=1}^{j-1} \langle\boldsymbol{q}_k, \psi(x_j)\rangle \langle\boldsymbol{q}_k, \psi(x_i)\rangle$. The vector $\boldsymbol{d}$ stores the squared residual norm of the orthogonal vectors. It is initialized with $\boldsymbol{d} = diag(\boldsymbol{K})$, which leads to $\mathbf{1}_N$ for Gaussians, and updated by $d_i \to d_i - G_{j,i}^2$.

For a direct measure of the columns of $\boldsymbol{G}$, it can be done by generalizing the previous expression:

$$\boldsymbol{g}_j = \frac{1}{\sqrt{d_j}}\left(\boldsymbol{k}_j - \sum_{k=1}^{j-1} G_{k,j}\boldsymbol{g}_k\right) \quad i = 1, ..., N \tag{4.78}$$

The Cholesky decomposition does not give the desired advantage on computational time but, it turns out that if the eigenvalues of $\boldsymbol{K}$ decays rapidly it is possible to reduce its spacial dimension by allowing a certain error on the reconstruction of the matrix ([3]). This is the called Incomplete Cholesky Decomposition, or ICD, and it allows to express the kernel matrix as $\boldsymbol{K} \approx \tilde{\boldsymbol{G}}^T\tilde{\boldsymbol{G}}$, being $\tilde{\boldsymbol{G}}$ a $D \times N$ matrix with $D \leq N$, when the eigenvalues of $\boldsymbol{K}$ stored at $\boldsymbol{d}$ decay rapidly. The error ε is an arbitrary small positive number that can be quantified as follows:

$$\left\|\boldsymbol{K} - \tilde{\boldsymbol{G}}^T\tilde{\boldsymbol{G}}\right\| < \varepsilon \tag{4.79}$$

In order to reduce at maximum the computational complexity, from $O(ND^2)$ to $O(ND)$, the evaluation of $\tilde{\boldsymbol{G}}$ is done by pivoting the columns of $\boldsymbol{K}$ for which the error is minimized and stopping when it is reached. The pivot is selected by tracking the vector $\boldsymbol{d}$ and by measuring the values from the actual step to the end $\boldsymbol{d}_{i:N}$. The update has the following form

$$\boldsymbol{d}_{i:N} = tr\left(\boldsymbol{K}_{i:N} - \boldsymbol{G}_i^T\boldsymbol{G}_i\right) = \mathbf{1}_{N-i+1} - tr\left(\boldsymbol{G}_i^T\boldsymbol{G}_i\right) \tag{4.80}$$

where $\boldsymbol{G}_i$ corresponds to the sub-matrix of $\boldsymbol{G}$ with rows from i to N, or what we want to measure, and the columns from 1 to $i-1$, or what we have already measured. Finally, the new pivoting column of $\boldsymbol{K}$ corresponds to that with the maximum singular value which is the position of max value in $\boldsymbol{d}$ from i to N.

In terms of algorithmia, the error parameter is user oriented in terms of making a decision on the reduction of D. The less the error allowed, the higher the output dimension and vice versa. The normal premise of the work will lead to a narrow difference between independent and dependent estimations, so it is not very recommended on allowing a high error, but the contrary will lead to a poor dimensionality reduction and so it becomes a trade-off.

The algorithm, extracted from [30] with slight changes adapted for Gaussian kernels, is the following:

Algorithm 4.1 Incomplete Cholesky Decomposition with pivoting

```
Input: $\{x_i\}_{1 \leq i \leq N}$, $k$, $\varepsilon$
$\boldsymbol{d} = \mathbf{1}_N$, $\boldsymbol{p} = [1, 2, ..., N]^T$
$\boldsymbol{G}(:, 1) = [k(x_1 - x_1), k(x_1 - x_2), ..., k(x_1 - x_N)]$
for $i = 1 \ : \ N$
    if $i \neq 1$
        $\boldsymbol{d}\left(i \ : \ N\right) = \mathbf{1}_{N-i+1} - \left(\boldsymbol{G}\left(i : N, 1 : i-1\right) \odot \boldsymbol{G}\left(i : N, 1 : i-1\right)\right) \mathbf{1}_{i-1}$
    end if
    if $\sum \boldsymbol{d}\left(i \ : \ end\right) < \varepsilon$
        $Break$
    end if
    $j* = \arg\max \boldsymbol{d}\left(i \ : \ end\right)$
    $\boldsymbol{p}\left(i\right) \leftrightarrow \boldsymbol{p}\left(j*\right)$
    $\boldsymbol{G}\left(i, 1 : i-1\right) \leftrightarrow \boldsymbol{G}\left(j*, 1 : i-1\right)$
    $\boldsymbol{G}\left(i, i\right) \leftrightarrow \sqrt{\boldsymbol{d}\left(j*\right)}$
    for $j = i+1 \ : \ N$
        $pivot\left(j-i\right) = k\left(x_{p(i)} - x_{p(j)}\right)$
    end for
    $\boldsymbol{G}\left(i+1 : N, i\right) = \left(\boldsymbol{pivot} - \boldsymbol{G}\left(i+1 : N, 1 : i-1\right) \boldsymbol{G}\left(i, 1 : i-1\right)^T\right) / \boldsymbol{G}\left(i, i\right)$
end for
Sort rows of $\boldsymbol{G}$ according to $\boldsymbol{p}$
```

In summary, the ICD provides a change in computational complexity from $O\left(N^2\right)$ to $O\left(ND\right)$, which can make a huge difference on computational time. To avoid the caveat of measuring the whole $\boldsymbol{K}$, matrix $\tilde{\boldsymbol{G}}$ is constructed through pivoting certain columns of $\boldsymbol{K}$ and stopping when the error is obtained. This allows to do not compute

$\boldsymbol{K}$ at its whole but only pairwise measures until the algorithm detects that is sufficiently representative.

With the preconditioning that the ICD provides, it is not only possible to accelerate the process of evaluating the data with a kernel to build $\boldsymbol{K}$ but also to reduce the expressions of some of the detectors presented. In fact, the most interesting simplifications are

$$\mathbf{1}^T \boldsymbol{K} \mathbf{1} = \left\| \mathbf{1}^T \tilde{\boldsymbol{G}}^T \right\|_F^2 \tag{4.81}$$

and

$$\begin{aligned} \mathbf{1}^T \left(\boldsymbol{K}_x \odot \boldsymbol{K}_y \right) \mathbf{1} &= tr\left(\boldsymbol{K}_x^T \boldsymbol{K}_y \right) \\ &= tr\left(\tilde{\boldsymbol{G}}_x^T \tilde{\boldsymbol{G}}_x \tilde{\boldsymbol{G}}_y^T \tilde{\boldsymbol{G}}_y \right) \\ &= tr\left(\tilde{\boldsymbol{G}}_y \tilde{\boldsymbol{G}}_x^T \tilde{\boldsymbol{G}}_x \tilde{\boldsymbol{G}}_y^T \right) \\ &= \left\| \tilde{\boldsymbol{G}}_y \tilde{\boldsymbol{G}}_x^T \right\|_F^2 \end{aligned} \tag{4.82}$$

Do note that with this reduction we are still measuring inner products but with the advantage of being a $D \times N$. This implies that we are now measuring matrices of $D \times D$ elements, which differs from the usual $N \times N$. Additionally, the kernel matrices $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ will provide different rank, and so different dimension D_x and D_y, highlighting the importance of obtaining the expressions in this form.

4.6 Detectors performance

To end up with the chapter, the detectors will be evaluated for a given presented problem. The proposal is to prepare a model where the correlation is not a reliable measure but there still exist dependence, so we are forcing to detect independence over correlation. Following the notation until now, M blocks of N random data samples will be generated from a bivariate Gaussian model $\mathcal{N} \sim (0, \boldsymbol{\Sigma})$ with the covariance matrix

$$\Sigma = \begin{bmatrix} \sqrt{1-\rho^2} & \rho \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \qquad 0 \leq \rho \leq \sqrt{2}/2 \tag{4.83}$$

The parameter ρ allows to constrict or relax the correlation, being $\rho = 0$ leading to two vectors of independent Gaussian random samples and $\rho = \sqrt{2}/2$ leading to two completely dependent and correlated vectors of random data samples. Then, to decorrelate the signal, a non linear operation will be used. To be more precise, the generated vector $\boldsymbol{x}$ will be randomly multiplied a random vector $\boldsymbol{w}$ satisfying

$$w_n = \begin{cases} 1 & p = 1/2 \\ -1 & p = 1/2 \end{cases} \tag{4.84}$$

This is one of the procedures used in [26], which will lead to a cross alike form as shown in figure 4.1.

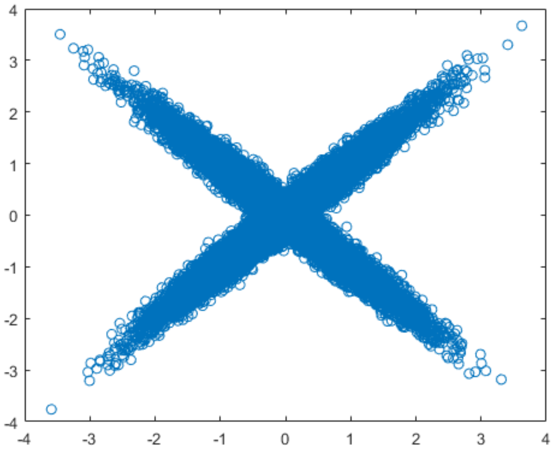

Figure 4.1: Model for dependent and uncorrelated random values, $\rho = 0.7$

The proceeding will be to use the data generated and decorrelated as hypothesis of dependence $\mathcal{H}_1$ and to generate two complete random vectors of data samples for the independence hypothesis $\mathcal{H}_0$. Another consideration is the criteria selection of the threshold γ that will classify the output of the sufficient statistics, or detectors, as dependent or independent. In general, the ideal situation is the one with an universal threshold that is good enough for all the detectors, but it is not possible to assure an universal floor value. This is a direct result from constructing each detector with its own metrics and so each one will result in a bias-variance trade off adapted by its own. This caveat has some similarity with the Normalized Least Mean Square algorithm (NLMS), where the data is scaled depending on its power. In the NMLS case, the algorithm is not supposed to be universal but adaptive. Thus, the same idea will be used in order to compare the efficiency of each detector. The threshold will be based on the floor value provided by the measures under hypothesis of independence, and so the mean.

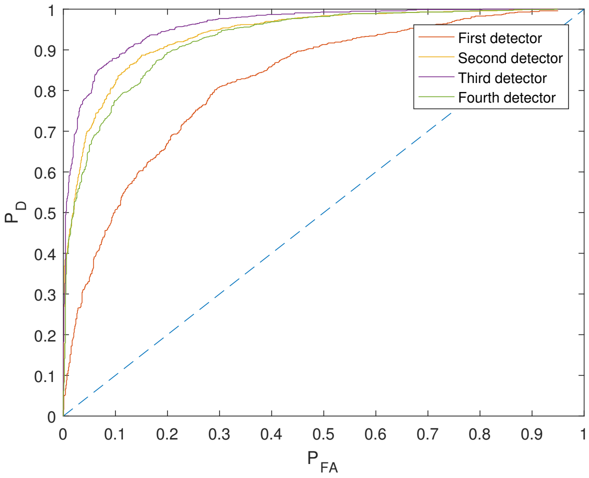

Figure 4.2: Detectors ROC curve for $\rho = 0.3, N = 1000$

A first review of the detectors are found in figure 4.2. It can be seen that the first detector derived directly from the KDE provides the worst performance while the third detector, derived from the u-statistics, provides the best, closely followed by the second and fourth. The parameter ρ provides a difficulty degree to the detection in the sense that higher data size is needed for more restrictive ρ and vice versa. At any case, the number of samples N should be usually high enough in the pro of the detection. In order to enhance the results, the figure 4.3 shows the logarithmic scale for the ROC curve with a higher dependency but fewer points to see in a closer way the detectors performance. In this case, the probability of miss detection is used instead of the probability of detection to get a better detector when closer to zero.

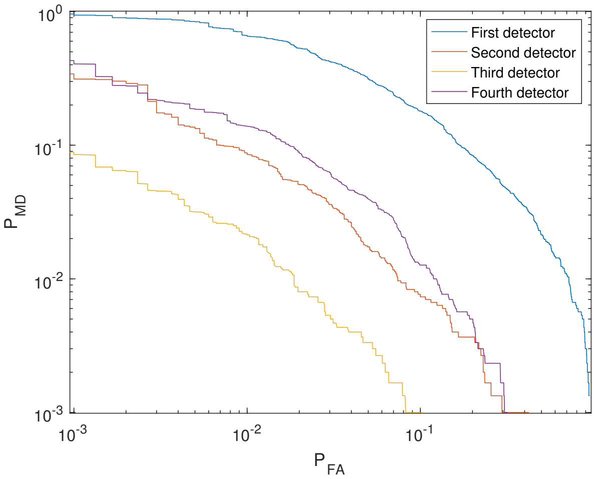

Figure 4.3: Log scale of ROC for $\rho = 0.5, N = 300$

In order to see the asymptotic behavior of the detectors, a sweep of the dimension of the data N is done. It is expected that, in terms of $1 - AUC$, the detectors tends to zero and so the perfect discrimination between independent and dependent hypothesis. In figure 4.4 can be seen that the last three detectors do tend to the perfect detection in exception of the first one. This is mainly caused because the first detector provides a high variance in terms of test statistics measures.

Taking a close look on the first detector in 4.44, it can be seen that its format of measuring mutual information row by row, instead of the whole matrix, provides a more sensitive output when the mean of a row is close to zero, and so we observe outliers. In figure 4.5 we can see the output of the test statistic of the detectors for both hypothesis. Apart from the outliers of the first detector, the third and the fourth provides the minimum variance over all the detectors but at the cost of introducing bias, specially the fourth. In these cases, the hypothesis $\mathcal{H}_0$ is getting far from the expected close to zero output. On the other hand, in the second detector the hypothesis are close to zero but at the cost of a higher bias. Within this, it is consolidated that the choice of detectors become a trade-off of bias variance as expected.

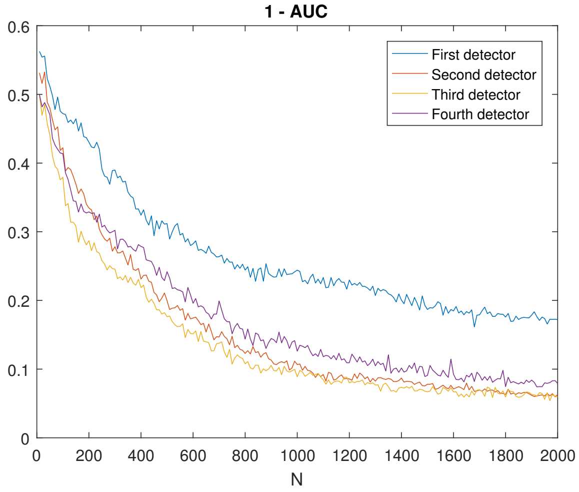

Figure 4.4: Detectors $1 - AUC$ for $\rho = 0.3$, N from 50 to 2000

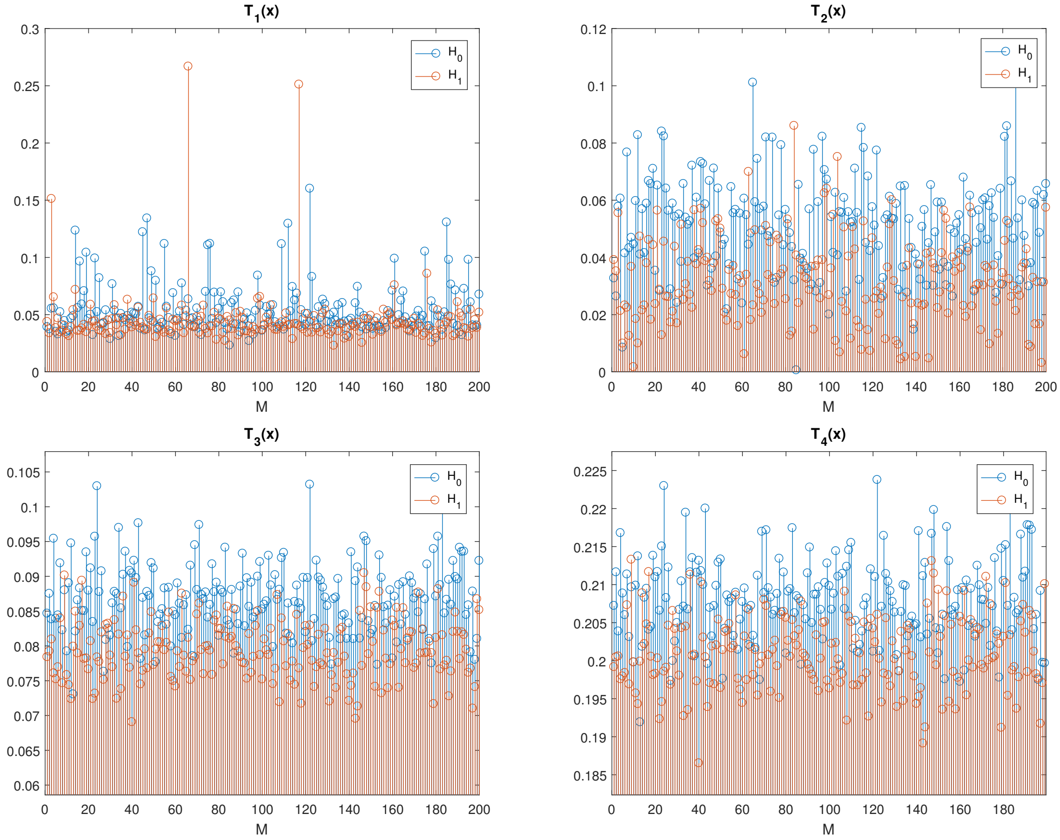

Figure 4.5: Test statistic outputs for both hypothesis, $\rho = 0.3, N = 500$

If we take a look on he deflection, we can see that it increases with a better detection as seen in figure 4.6. This is in fact a result of both hypothesis separating naturally when N increases. The reasoning is that with more samples the kernel methods get more representative data from the input, and so the information of them in the mapped space is richer. Then, the detector is capable of separating better both hypothesis resulting in a better detection.

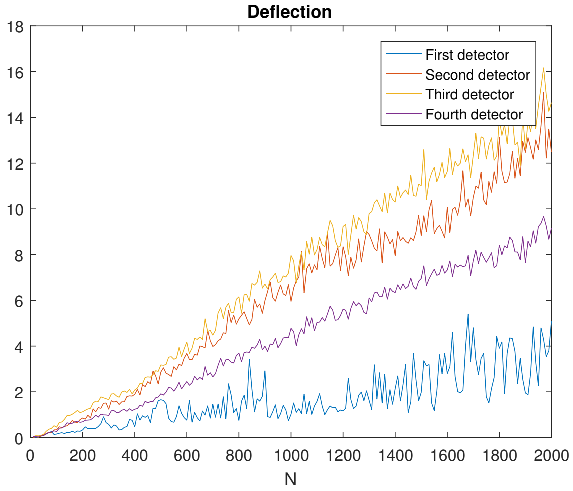

Figure 4.6: Detectors Deflection for $\rho = 0.3$, N from 50 to 2000

5 Dependence detection in Hammerstein systems

When applying the problem of dependence detection on reals cases, some sort of distortions are expected from a given channel. Noise addition, attenuation or multipath are only some of the expected results on communication through a channel. In the signal processing field it is usual to assume a given model for a channel in order to study its affects and how they can be solved. In order to be capable of extending the detectors into a more general situation, and preparing for the latency estimation, this chapter will provide some channel characterizations with solutions to the degrade of the signal.

Firstly, some channel models will be proposed and its extension to the detectors yet developed. For any new addition, the detectors will be assessed to verify in which point the information is lost. The main progress is going to be the addition of the memory in the channel, and so the memory into the data samples. We will see in which degree this implication affects the detection and how to solve it by adding memory to the detectors.

5.1 Hammerstein systems and applications

The Hammerstein system is a specific configuration of a nonlinear model, usually used to characterize or evaluate different channels jointly with the Wiener system. Its scope is to define a channel by the means of the input $x\left[n\right]$ and the output $y\left[n\right]$ through a memory channel composed by a memoryless nonlinear transformation $f\left(\cdot\right)$ and a finite impulse response (FIR) filter $h\left(\cdot\right)$. It is actually a specific case of the Volterra nonlinear systems [40], described by

$$y\left[n\right] = h_0 + \sum_{i=1}^{\infty} h_i\left(x\left[n\right]\right) \tag{5.1}$$

being h_i polynomial integrals operators with $h_i\left(x\left[n\right]\right) = \sum\limits_{n_1} ... \sum\limits_{n_i} h_{n_1 ... n_i} x\left[n - n_1\right] ... x\left[n - n_i\right]$ for the discrete case. The expression results in a very wide possibilities of channel characterization but with increasing estimation difficulty when the data input and the nonlinearity grows, and so it us usual to select predefined models that are sufficient to evaluate certain channels behaviors. One of them is the reduced model called Wiener model proposed in [4], conformed by a non-linear dynamic function followed by a linear non-dynamic one:

$$y\left[n\right] = f\left(\sum_{i=0}^{M-1} h_i x\left[n-i\right]\right) \tag{5.2}$$

with the corresponding block diagram from [39]:

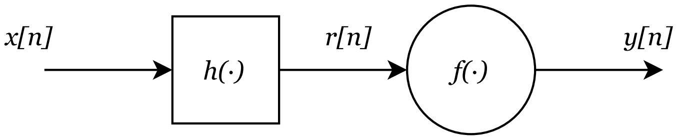

Figure 5.1: Wiener model diagram

To define the Hammerstein system, it is only needed to inverse the blocks from the Wiener system, resulting in a linear block followed by a non-linear dynamic:

$$y\left[n\right] = \sum_{i=0}^{M-1} h_i f\left(x\left[n-i\right]\right) \tag{5.3}$$

with a block diagram like the following one:

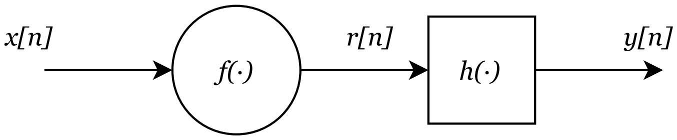

Figure 5.2: Hammerstein model diagram

The Hammerstein-Wiener models have many applications in many engineering problems. One of the main uses in the literature is the non-linear process identification of many kinds, specially to determine the parameters of the model in order to characterize and define them for multiple objectives.

In the case of this work, the Hammerstein model will have a special interest on the current problematic due to the fact that is a natural extension of the non-linear and one to one transformation used to decorrelate the data samples. This way, the addition of the filter will not invalidate the previous analysis of the detectors but adding more complexity to them.

The next step is to use this model to look on how far transformations due to the channel are distorting the original signal, and so what can be done in order to detect some dependence between the original sequence and the sequence at the output of the channel. To do so, the input $x[i]$ will be related to the metrics used until now in the sense that $\boldsymbol{x} = \{x_i\}_{i=1,..,N}$ and so $x[i] = x_i$. To make it more realistic, an additive white Gaussian noise will be added to the channel drawn from a zero-mean normal distribution with variance σ: $\boldsymbol{n} \sim \mathcal{N}(0, \sigma)$.

At this point, the resemblance with an AWGN channel is direct, and so the objective of the detector will be to look for dependence over the input $\boldsymbol{x}$ and the output $\boldsymbol{y}$.

As a first evaluation, and to define a starting point, the basic model will be composed by the non-linear operator and the addition of noise, as in the figure 5.3.

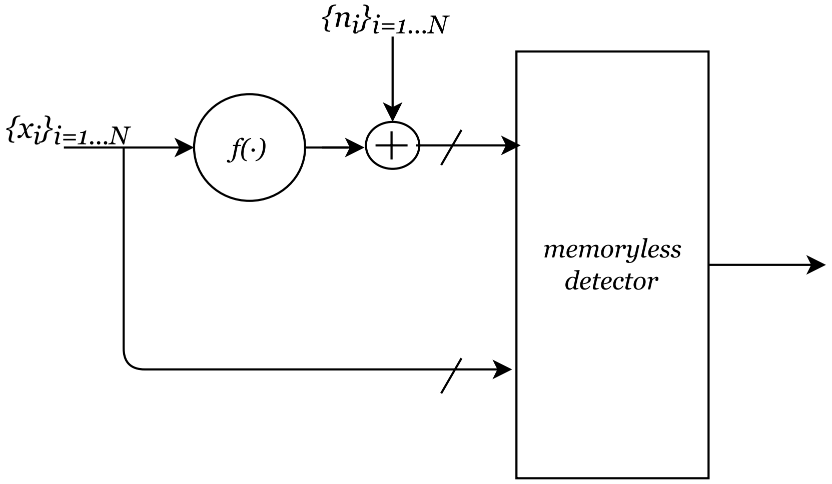

Figure 5.3: Memoryless channel detector proposal

It is expected that the noise power respect to the power of the signal, and so the SNR, has a great impact on the detection task. In figure 5.4 the impact of multiple noise levels can be reviewed. It can be seen that, for an equal signal and noise power, the detection task is in fact not deteriorated, but when the power of the noise augments, the real signal gets masked and the task becomes more difficult as expected. At any case, the noise addition does not provide a great impact on the final output if being at levels close to the signal.

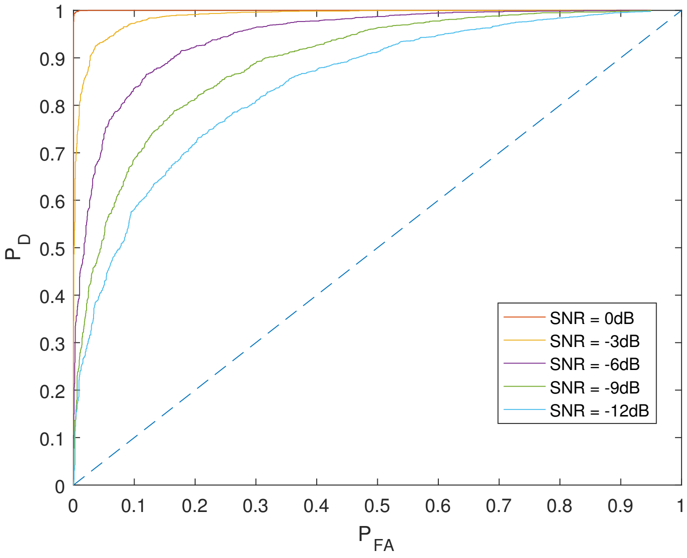

Figure 5.4: Noise impact on detection for $N = 500$

Given the memoryless performance, we can now move to introduce some memory type transformation. As a matter of fact, it is only needed to apply the Hammerstein model plus the addition of noise, as in figure 5.5. The used linear filter to test the channel was a unit-energy triangular one with memory depth of 5. Although is a very simple implementation, it is sufficient to see the effects on the detection task.

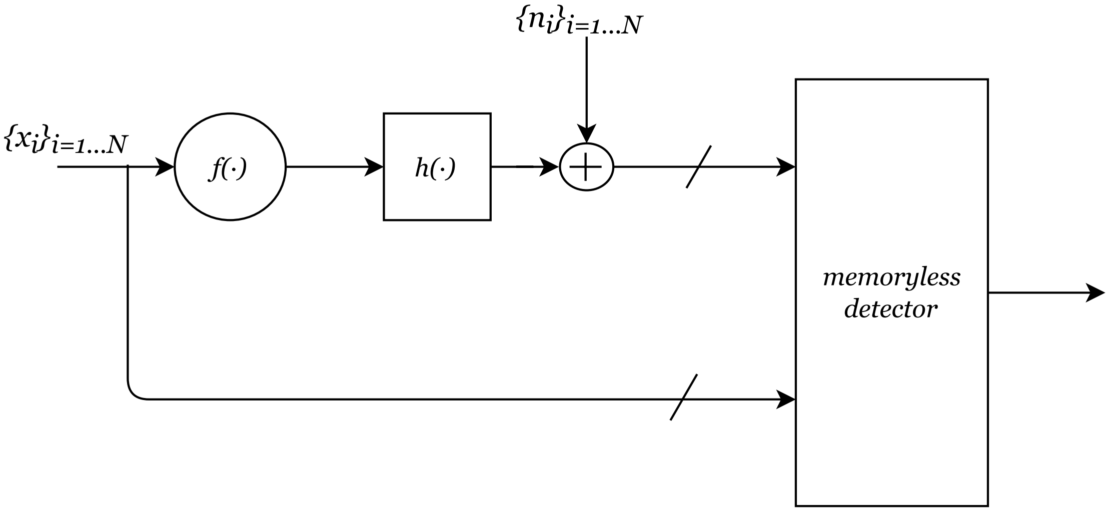

Figure 5.5: Hammerstein channel detector proposal

The addition of the filter provides some time-varying transformation that is expected to difficult the task of detection. As can be seen in figure 5.6 the pairwise measures have lost the instant information with a dynamic system, and so with memory. The result implies that, in order to recover the relation between pairs, it is needed to add memory in the detector in some way.

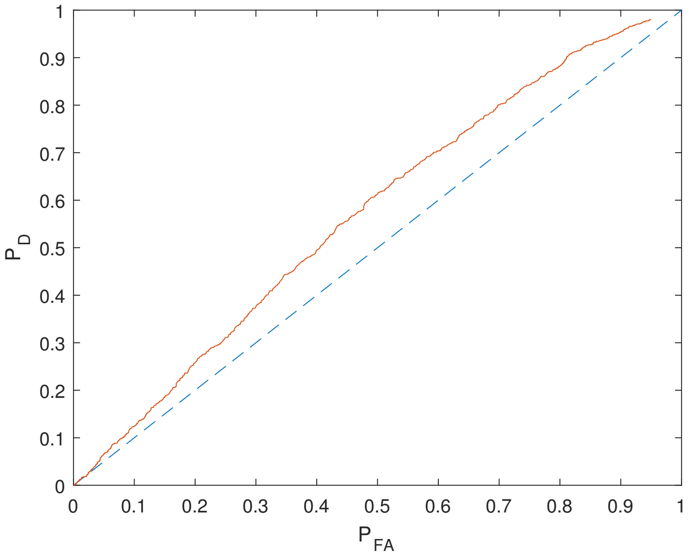

Figure 5.6: Hammerstein model detection performance for $N = 500$

A possible solution to recover the information lost could be to use Matched filter. But this kind of filtering is thought to improve the SNR in terms of stochastic noise addition, so it actually does not provide signal regeneration to the memory problematic. In this case, another solution has to be applied.

5.2 Adding memory as input kernel dimensionality

In order to recover the data after it has been distorted by the filter, the proposal is to use the kernel definitions to introduce some kind of memory inside the detectors as in figure 5.7. The objective is to not only provide information of the sample being measured with a kernel but to provide information of the sample itself and the near ones in the sense of indexing.

To do so, we will take advantage of the structure of the kernel by evaluating the kernel with vectors instead of scalars and so we are going to move from kernel pairwise measure

$k(x_j, x_i)$ to $k(\boldsymbol{x}_j, \boldsymbol{x}_i)$. Within this change, we are mapping two vectors into an scalar, and so the feature space is now group oriented instead of pair oriented. Given the filter is lineal, the purpose is to analyze the lineal combinations in the infinite space given two input vectors. As reviewed in Section 4.3, the lineal products are intrinsically measured by the kernel trick when using a Gaussian kernel.

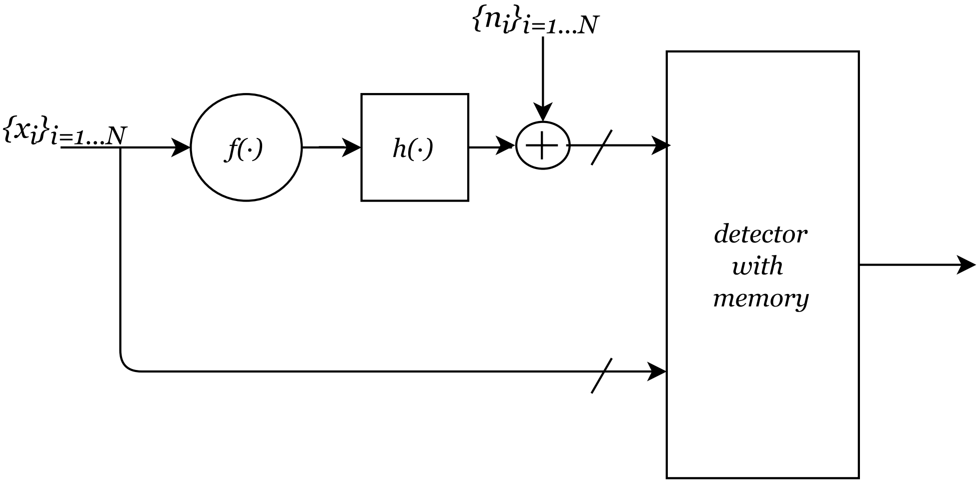

Figure 5.7: Hammerstein channel detector with memory proposal

Consider an input signal of random data realizations $\boldsymbol{x} = \{x_i\}_{i=1,..,N}$, the kernel input is going to be re-ordinate as a $d \times L$ matrix with $L = N - d + 1$ and d the dimensionality introduced to the kernel. The matrix $\boldsymbol{X}$ will then be

$$\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 & . \ . \ . & x_{N-d+1} \\ x_2 & x_3 & . \ . \ . & x_{N-d+2} \\ . & . & & . \\ . & . & & . \\ x_d & x_{d+1} & . \ . \ . & x_N \end{bmatrix} \tag{5.4}$$

with $\boldsymbol{x}_i$ the i.th column vector of the matrix. This is a common reorganization of the data in signal processing for time variant schemes. The vectors $\boldsymbol{x}_i$ are usually called run-time signal, being each column a different time instant. The kernel matrix is then constructed with elements

$$K_{i,j} = e^{-\|\boldsymbol{x}_j - \boldsymbol{x}_i\|^2 / 2\sigma_k^2} \tag{5.5}$$

which is similar to evaluate a multivariate Gaussian function.

An important addition to the new kernel measure is the change of the kernel bandwidth σ_k. We have reviewed in equation 4.4 that for multivariate density estimations the optimal standard deviation has a dimensionality parameter included. As the detectors does not imply a different derivation for a multidimensional case, and so they are generalizable, the kernel bandwidth only requires to be tuned by its dimensionality. The reasoning is that we are now spanning the data that lies in the infinite space, and so the kernel bandwidth requires a higher value to evaluate that space properly.

To make a comparison, in figure 5.8 it is viewed the effect of the kernel bandwidth in both memory and memoryless detection using the third detector. For the memoryless detector, the pairwise kernel provides similar performance when a kernel bandwidth near the one obtained from Silverman rule. On the counterpart, we can see that the performance improves using a higher σ_k than the kernel bandwidth of the memoryless detector σ_1. In this case, the detector with memory is more sensitive to the change of kernel bandwidth.

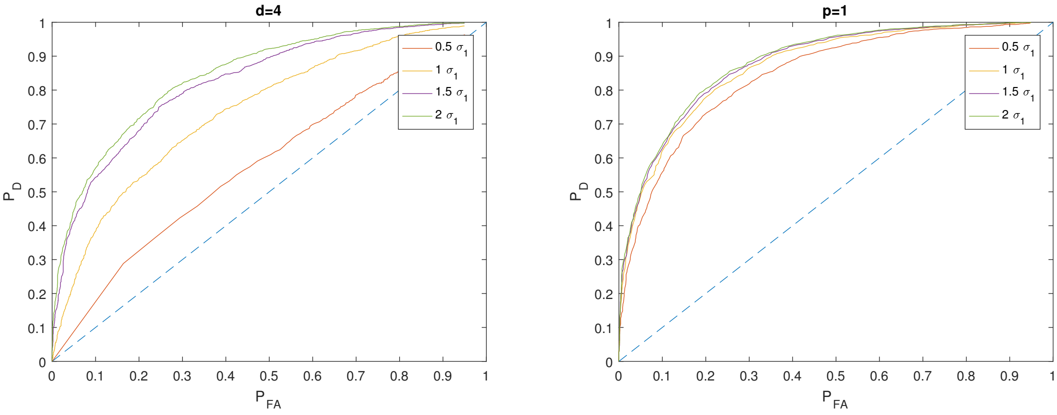

Figure 5.8: Performance based on bandwidth value for $\rho = 0.3, N = 500$

Within the spanning of the data there is another implication. Being the kernel matrices a representation of the so called infinite space, it is supposed that the matrix becomes rank complete or close to it in order to enclose the spanning in that space. The result is that the Incomplete Cholesky decomposition does not reduce the computational complexity when dimensionality is added, and so there is no gain to use it. In figure 5.9 is shown a comparison between the reduction of D that provides the Incomplete Cholesky Decomposition when we add dimensionality to the kernel versus when the kernel is pairwise. The general implication between input and kernel dimensionality is an open issue that deserves further attention, but it is out of the scope of this work.

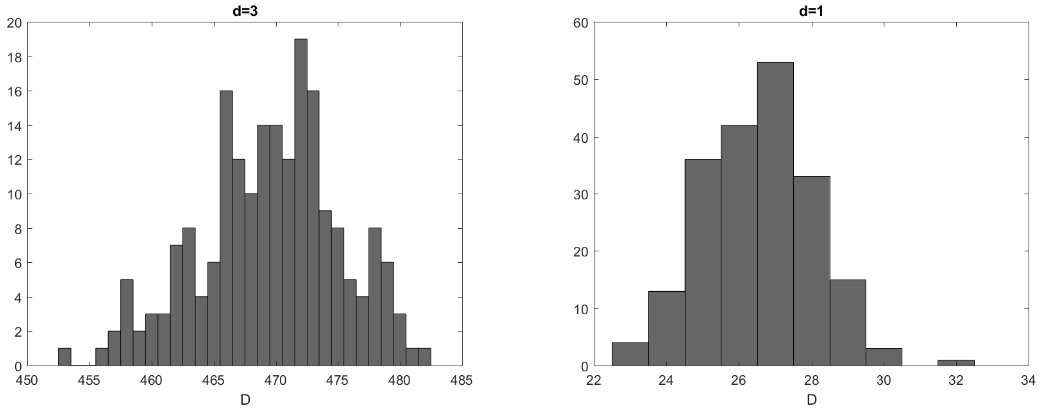

Figure 5.9: Dimension D from ICD for $d = 3$ and $d = 1$ with $N = 500$

To finish this chapter, figure 5.10 is provided. This figure represents the detection performance using the Hammerstein model when $d = 4$ versus the same performance obtained in figure 5.6 and so when $d = 1$. It is clear that for a given d and using the Hammerstein model, the detector can discern between dependence and independence, detecting the linear operator applied by the filter properly. Do note the detector is blind in terms of knowledge of the channel, and so no additional information is needed but the depth of the memory added by the channel.

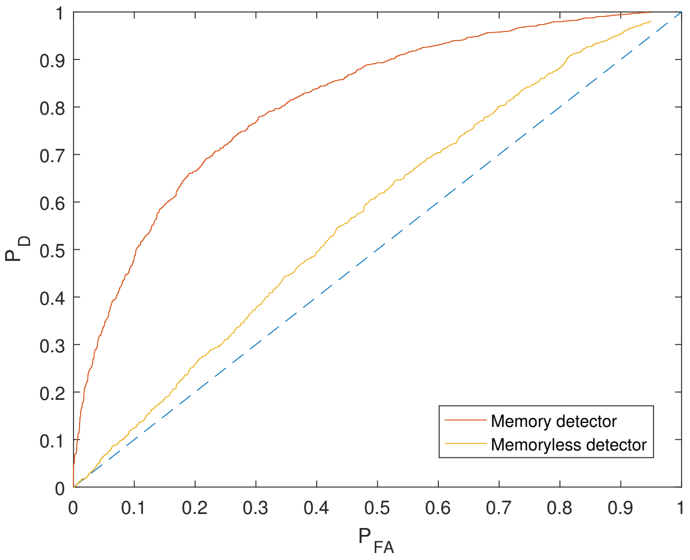

Figure 5.10: Hammerstein detector with memory ROC curve, $N = 500$, $d = 4, SNR = 0dB$

6 Latency estimation

Once the channel has been defined and analyzed, the natural step is to use the knowledge acquired to construct a more practical model for an approach to real problematic. Given the natural step of the detectors of measuring the dependence between two vectors of random values, the proposal of this work is to use them to estimate the latency. The objective of this estimation is manifold, i.e. determine at what epoch a given information arrives to its destination, at the same time of detecting at which node it arrives.

To do so, the Hammerstein model will be used to characterize the channel and the detector input will be based on a sliding window that search over the output sequence of the channel. The latency estimator will result in the window instant when the dependence is maximum.

6.1 Problem statement and applications

The Time delay estimation (TDE), or latency estimation, is an area with a high number of interest due to its multiple applications. Usual uses of TDE cover diverse fields as radar, seismology, geophysics or communications. When specifically looking on time delay in signal processing it is usually also described as Time Difference of Arrival (TDOA) estimation. The reason is that it is used as a metric to define the lag or propagation between two known points instead of detecting the Time of Arrival (TOA), which defines the time of forward and backward propagation. Common applications of the TDOA includes array of sensor approaches, submarine telecommunications or phone localization techniques, among others.

The general definition of the time difference of arrival considers a model like

$$r\left[i\right] = \alpha s\left[i-\tau\right] + n\left[i\right] \quad i = 0, 1, ..., N-1 \tag{6.1}$$

being $r[i]$ the received signal and $n\left[i\right]$ the noise at the time instant i, $s\left[i-\tau\right]$ the signal of interest and α the attenuation of the channel. The scope is then to estimate the delay τ.

Among multiple techniques, the idea of the generalized cross-correlation (GCC) algorithm will be interesting due to its closeness on what we are going to do with the dependence detector ([6]). For random signals, lets consider another signal

$$r_2\left[i\right] = s\left[i\right] + n_2\left[i\right] \tag{6.2}$$

with $s\left[i\right]$ and $n_2\left[i\right]$ being zero-mean and white processes. The scope is to minimize a cost function through the minimum mean square error (MMSE):

$$MMSE\left(\tilde{\alpha}, \tilde{\tau}\right) = E\left[\left(r_2\left[i\right] - \tilde{\alpha} r\left[i - \tilde{\tau}\right]\right)^2\right] \tag{6.3}$$

Without the need of entering into details, it can be seen that the estimation of the delay corresponds to the value of $\tilde{\tau}$ which provides maximum correlation between $r_2\left[i\right]$ and $r\left[i - \tilde{\tau}\right]$, and so

$$\hat{\tau} = \underset{\tilde{\tau}}{argmax}\left(E\left[r_2\left[i\right] r\left[i - \tilde{\tau}\right]\right]\right) \tag{6.4}$$

To sum up, in order to find the delay we just need to compute a cross-correlation method between the observed signal and signal generated with similar metrics and same noise power. It is direct that, as with the kernel methods the correlation is observed, by measuring the dependence between the output of the channel and a realization of the original signal it is like applying the cross-correlation method here exposed.

6.2 Extending the classical time difference of arrival estimation problem

Lets assume $\boldsymbol{y}$ a vector that contains realizations of a unknown random process at the output of a Hammerstein channel with $\boldsymbol{y} = \{y_i\}_{i=1,..,N}$. Consider that at a time instant D, we do observe L values from $\mathbf{y}$, namely $\boldsymbol{y}_D = y_{1+D}, ..., y_{L+D}$ with $L \leq N - D$ and $0 \leq D \leq N - L$. If the input of the channel is known with $\boldsymbol{x} = \{x_i\}_{i=1,..,N}$, then the objective is to estimate the time instant $\hat{D}$ in which the channel output $\boldsymbol{y}_D$ corresponds to the input $\boldsymbol{x}_D$. As we have seen, by GCC the solution is the instant when $\boldsymbol{y}_D$ is maximum correlated with $\boldsymbol{x}_D$.

Reformulating 6.3 for the Hammerstein channel, we have that the new cost of function is:

$$MMSE\left(\tilde{h}, \tilde{\tau}\right) = E\left[\left(r_2\left[i\right] - \tilde{h} f\left(r\left[i - \tilde{\tau}\right]\right)\right)^2\right] \tag{6.5}$$

This way, we do reformulate the channel as an attenuation to incorporate the Hammerstein channel. The new latency estimator will also be based on measuring the cross-correlation, but this time it will be done by the means of kernel processing.

To extend the dependence detectors as latency estimators the proposal is based on generating N data samples as channel input, and to crop L of them at a random instant D, namely $\boldsymbol{x}_L$. This way, the latency estimation is based on comparing the cropped sequence with the output of the channel for multiple time instant. Then, the cross-correlation method would be

$$\hat{\tau} = \underset{D}{argmax} \left(E \left[(\boldsymbol{x}_L - \mu_x) \left(\boldsymbol{y}_D^T - \mu_y \right) \right] \right) = \underset{D}{argmax} \boldsymbol{C}_{x,y_D} \tag{6.6}$$

By using the detectors, we know that it is being measured the correlation on the feature space mapped by the kernel, and so it is being applied a cross-correlation method on an infinite space by the means of dependence on the data space. For example, the third detector for latency estimator is the following:

$$\hat{\tau} = \underset{D}{argmax} \left(\frac{tr\left(\tilde{\boldsymbol{K}}_x \tilde{\boldsymbol{K}}_x\right) tr\left(\tilde{\boldsymbol{K}}_{y_D} \tilde{\boldsymbol{K}}_{y_D}\right)}{tr\left(\tilde{\boldsymbol{K}}_x \tilde{\boldsymbol{K}}_{y_D}\right)^2} - 1 \right) \tag{6.7}$$

being $\tilde{\boldsymbol{K}}_{y_D}$ and $\tilde{\boldsymbol{K}}_x$ kernel matrices of $L \times L$.

The figure 6.1 shows the detector output for multiple D in multiple realizations. The latency estimator was tested with $N = 1000$ and adding memory to the detector in order to detect the dependence close to D after the Hammerstein channel. Also do not the effect of the memory addition to the detector. Around the peak of the detector, it is reflected the behavior of the run-time vectors near the real one. Within this, we can assure that the kernel with dimensionality is correctly detecting the data redundancy of the input matrix 5.4, providing a more robust result to the general estimation.

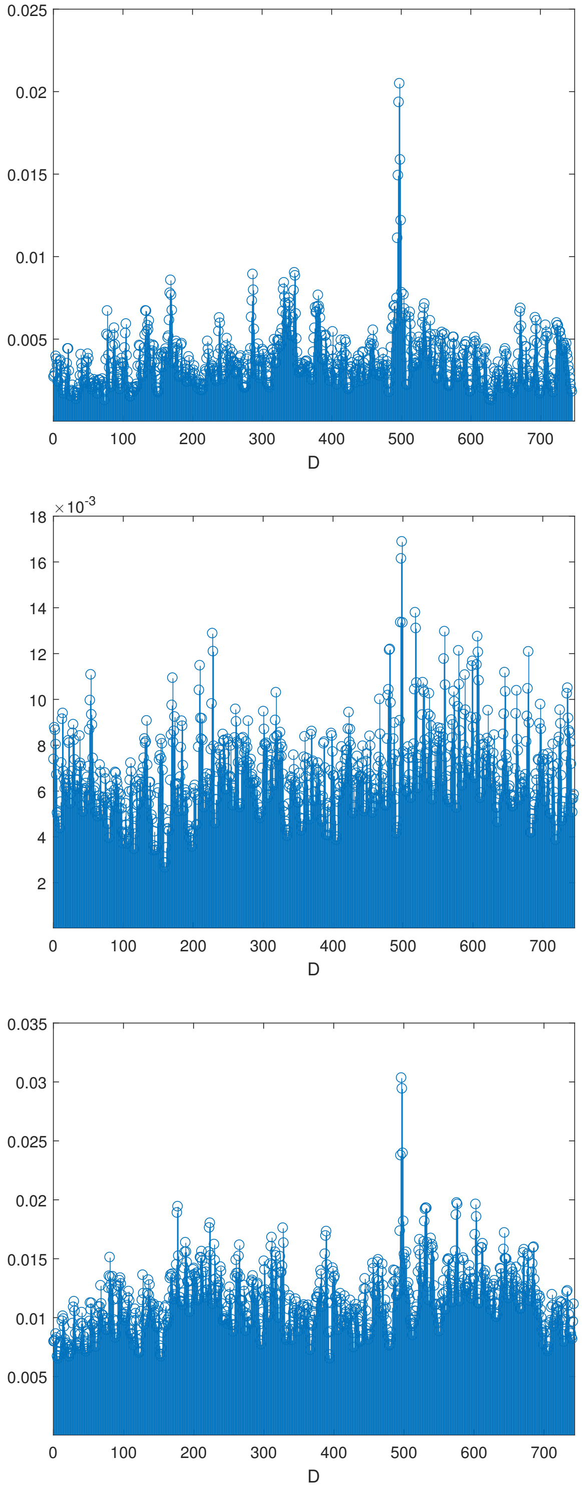

Figure 6.1: Latency estimation output for multiple delays, $D = 500$, $L = 250$

Recalling the threshold setting problem, it has to be pointed out that we do now have access to constant observations under independence hypothesis. Then, it is direct to make the threshold to be calibrated automatically by the floor level. Thanks to these constant observations, the bias of the detector does not make any implication on latency estimation due to the fact that we can define different levels of threshold depending on the detector used.

7 Conclusion

The kernel signal processing has provided an attractive framework for non linear signal processing. We have seen that with the addition of kernel functions, we are now capable of using a broaden methodology that allows to solve elegantly the problems proposed. Additionally, by the universality of the kernels, it turns out to be a very versatile and powerful tool capable of fitting a wide range of problems.

Reviewing the work, in this Master's Thesis the kernel methodology has been studied in terms of detecting dependency between two sets of observations. We have derived a kernel environment from the Parzen window estimate and the Rényi entropies based on the pairwise measures and we have seen its properties. The estimation of the Rényi entropy has been extended into a kernel processing problem, studying the tendency of the estimator depending on the kernel bandwidth used and deriving three detectors from it, based on the bound that the $2-$Rényi mutual information provides.

It has been reviewed the importance of the reproducing property, pointing out the advantages of being in a Hilbert space through the Hilbert-Schmidt norm. In that space, the relation between the correlation and the dependency emerges, so we took advantage of it. By studying the characteristic function of a given pdf, we have seen how to define a correlation metric that depends of the independence assumption on an infinite space. From there, we have adapted it to the framework of data observations and how to derive it in order to avoid the infinite space problematic, by the means of kernel processing.

To wrap the obtained detectors into the dependence detection context, we have evaluated its performance for a given data set dimension and its asymptotic convergence when this dimension is increased.

In the final part of the work, a more general detection model has been viewed by the means of characterizing a channel. We have seen the effects of non linear transformations, the addition of noise and the effects of a dynamic system. The increase of the kernel dimensionality has provided a solution to recover the spatial information by moving from pairwise to pairs of run-time vectors.

Finally, we have applied the knowledge obtained with the Hammerstein model to apply the detectors to a more suitable case. We have seen that it is possible to detect the point to point latency by looking at the correlation in the Hilbert space.

7.1 Future work

Although the metrics of the work are generally well-defined, the scope of the methodology remains to be seen. A latency approach has been proposed to give a real framework to the problem, but the final purpose can actually be diverse. The work intended to give a more general comprehension on the kernel methodology for detecting dependence. But as has been told, the applicable fields are based on the ones with a lack of knowledge and so the possibilities nowadays are fullness.

Specifically in the work, there is a need on a better understanding on the kernel procedure when dimensionality of the data is added. We have seen the effects on recovering dynamic information from a detector point of view. But the study on how this dimensionality increase affects the kernel itself is something that needs an study for a better comprehension.

From the detection point of view, two natural extensions can be drawn:

- The direct estimation of mutual information instead of pure detection. Although an estimator has been derived, there is a lack on focusing in the pure estimator at its own. We have derived the detectors in a bounded point of view, looking at being close to zero near independence and assuming a low SNR regime. The extension to the general purpose of estimating can provide even more insights on the study of dependence, expanding the scope possibilities.

- The direction of information. It is an ambitious point of view that has to be reviewed carefully. The presence of dependence and the estimation of the degree of dependence is just the first step on determining the cause of the dependence and the direction of it. A better understanding of the direction can be used to draw relations between nodes in a lack of information framework. A good scope for this evaluation is, for example, the better understanding of unknown networks topologies and so the study of network tomography.

Beginning with known structures for detecting dependence and ending with an applicable case, there is a something certain: this research field is still big, open, and his future has yet to be drawn.

Bibliography

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, pages 337-404, 1950.

[2] F.R. Bach and M.I. Jordan. Kernel independent component analysis. *Journal of Machine Learning*, 3, pages 1-48, 2002.

[3] F.R. Bach and M.I. Jordan. Predictive low-rank decomposition for kernel methods. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, pages 33–40, 2005.

[4] S.A. Billings and S.Y. Fakhouri. Identification of nonlinear systems using the wiener model. *Electronics Letters*, 13, Issue 17, 1977.

[5] A.P. Bradley. The use of the area under the curve in the evaluation of mmachine learning algorithms. *Pattern*, 30, 1996.

[6] Jingdong Chen, Yiteng Huang, and Jacob Benesty. *Time Delay Estimation*, pages 197–227. Springer US, Boston, MA, 2004.

[7] T.M. Cover and J.A Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[8] S. Fehr and S. Berens. On the conditional rényi entropy. *IEEE transaction on information theory*, 60(11), 2014.

[9] D.A.S Fraser. *Nonparametric methods in statistics.* D.A.S Fraser, 1957.

[10] G.H. Golub and C.F. Van Loan. *Matrix Computations, Third edition.* Johns Hopkins University Press, 1996.

[11] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research 13*, 2012.

[12] A. Gretton, A. Smola, O. Bousquet, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2005.

[13] A. Gretton, A. Smola, R. Herbrich, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 2005.

[14] M. Hein and O. Bousquet. Kernels, associated structures and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, 2004.

[15] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19, pages 293-325, 1948.

[16] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28, 1936.

[17] W. Liu, J.C. Príncpie, and S. Haykin. *Kernel Adaptive Filtering.* Wiley, 2010.

[18] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209, pages 415-446, 1909.

[19] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers. Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48, August 1999.

[20] Alba Pagès Zamora. *Transformada de Fourier en Processament no Lineal del Senyal.* PhD thesis, Universitat Politècnica de Catalunya, 1996.

[21] E. Parzen. *On a Estimation of a Probability Density Function and Mode*, volume 33, 1065-1067. Ann. Math. Statist., 1962.

[22] B. Picinbono. On deflection as a performance criterion in a detection. *IEEE Transactions on Aerospace and Electronic Systems*, 31, pages 1072-1081, 1995.

[23] J.C. Príncipe. *Information Theoretic Leraning.* Springer, 2010.

[24] A. Rényi. On measures of entropy and information. *Mathematical Institue Hungarian Academy of Sciences*, Volume 1, 1961.

[25] A. Rényi. *Probability Theory.* Mineola, 2007.

[26] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, December 2011.

[27] I. Santamaría, D. Ramírez, J. Vía, and L.L. Scharf. Locally most powerful invariant tests for correlation and sphericity of gaussian vectors. *IEEE Transaction on Information Theory*, 59, 2012.

[28] D.W. Scott. On optimal and data-based histograms. *Biometrika*, 66, pages 605-610, 1979.

[29] R.J. Serfling. *Approximation Theorems of Mathematical Statistics.* John Willey & Sons, 1980.

[30] S. Seth and J.C. Príncipe. On speeding up computation in information theoretic learning. In *International Joint Conference on Neural Networks, IJCNN*, 2009.

[31] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, pages 379–423, 623–656, 1948.

[32] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability, 1986.

[33] M. Sugiyama and K.M Borgwardt. Measuring statistical dependence via the mutual information dimension. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, 2013.

[34] G.J Székely, M.L Rizzo, and N.K Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 2007.

[35] J.S. Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[36] A. Teixeira, A. Matos, and L. Antunes. Conditional rényi entropies. *IEEE transaction on information theory*, 58(7), 2012.

[37] S. Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective.* Elsevier, 2015.

[38] S. Theodoridis, K. Salavakis, and P. Bouboulis. Online learning in reproducing kernel hilbert spaces. *Academic Press Library in Signal Processing*, 1, pages 883-987, 2013.

[39] S.V Vaerenbergh. *Kernel Methods for Nonlinear Identification, Equalization and Separation of Signals*. PhD thesis, University of Cantabria, Department of Communications Engineering, 2009.

[40] V. Volterra. Sopra le funzioni che dipendono da altre funzioni. nota ii. *Rendiconti della Reale Accademia dei Lincei*, 3, pages 141-146, 1887.

[41] Q. Wang, S.R.Kulkarni, and S.Verdú. *Universal Estimation of Information Measures for Analog Sources*. Now publishers, 2009.

[42] N. Wiener. *Cybernetics: Control and Communication in the Animal and the Machine*. The MIT Press, 1948.

[43] H. Xiong and X. Chen. Kernel-based distance metric learning for microarray data classification. *BMC Bioinformatics*, 2006.