UPC
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Departament de Teoria del Senyal i Comunicacions

DATA-DRIVEN INFORMATION-THEORETIC TOOLS UNDER A SECOND-ORDER STATISTICS PERSPECTIVE

Ph.D. Dissertation

Author: Ferran de Cabrera Estanyol
Advisor: Jaume Riba Sagarra

Barcelona, March 2023

A la Marta i la meva familia.

Abstract

A great deal is known about second-order statistics. For many years, most of signal processing problems have been addressed under the perspective of first and second-order moments of the data, known to be optimal under normal conditions. Nowadays data grow in complexity, rendering the second-order approach insufficient. To cope with this, descriptors of data that come from the field of information theory have been utilized in recent years as a substitute for classical second-order methods. Divergence, entropy, and information are the basis of information-theoretic methods, endowed with the capability of discovering the intricacies of underlying distributions. However, the estimation and utilization of information measures have also proven to be a difficult task, due to their nonlinear nature and the difficulty of estimating a density function.

Meanwhile, the field of machine learning has also advanced toward the analysis of data in a nonlinear fashion. Kernel methods are a primer example of this, procuring an apparently simple way of dealing with nonlinearities, called the kernel “trick”, but concealing a mathematically rigorous background that strengthens the method. A kernel performs linear operations in a possibly infinite-dimensional feature space without the requirement of explicitly operating in such space. Thanks to this, kernels have the potential to define nonparametric methods that may be unfeasible to address otherwise. This approach has been used to deal with the aforementioned information-theoretical measures, whose nonlinearity is easily manageable in the feature space. However, their strengths also become their disadvantages. To operate in an unvisited feature space is to lack interpretability. To operate in a high-dimensional feature space is an ingredient for sparsity, requiring a blind regularization. Furthermore, their data-driven approach often comes at the cost of a computational complexity that grows exponentially with the data size.

This dissertation deals with the analysis of complex phenomena embedded in large amounts of data by leveraging well-known second-order statistics tools. This task is performed from two different points of view. In the first part, and following a similar rationale to kernel methods, this dissertation develops a framework that is capable of dealing with nonlinearities in a linear fashion. To do so, the data is mapped into a feature space of higher dimensionality than the data space. However, this feature space is neither infinite-dimensional nor unknown, gaining not only in interpretability but also scalability for its use on large data sets. Correlation and covariance are measured in the feature space with the purpose of estimating measures of information, which constitutes the primary motivation of the mapping. While regularization is still needed in the proposed approach, a consequence of increasing the intrinsic dimensionality of the problem, the deterministic feature space allows the characterization of an appropriate regularization, which ends up exhibiting a strong duality with classical spectral estimation techniques.

The second part of the dissertation focuses on the use of information measures in problems that are typically solved through second-order statistics. Specifically, entropy is an uncertainty measure that provides better granularity of the underlying distribution than the variance. Entropy does not only retain the information of first and second-order statistics but also of higher-order statistics. The resulting methods gain in robustness, and at the same time, the information provided by the variance is still discerned in the entropy-based approach. A specific entropy estimator that derives into kernel methods is used for this task thanks to its relationship with U-statistics, which have the advantage of an asymptotic tendency to the sample variance. Consequently, the framework is again rotated, providing a unified rationale of information and second-order statistics.

Resum

Les estadístiques de segon ordre són una eina àmpliament coneguda i utilitzada gràcies al fet que les tècniques de processament de senyal clàssiques basades en aquestes són òptimes sota l'assumpció de Gaussianitat. No obstant això, avui dia el processat requereix tractar amb dades d'una complexitat incrementada, deixant obsoleta la formulació amb estadístiques de segon ordre. Últimament, ha crescut l'interès en la utilització de descriptors de dades que provenen del camp de la teoria de la informació, en substitució del primer i segon ordre. La divergència, l'entropia i la informació són la base d'aquest altre enfocament, dotats amb la capacitat de descobrir la riquesa de la distribució subjacent. Així i tot, la seva inherent no-linealitat i una estimació enrevessada dificulta la utilització d'aquestes mesures en la majoria dels casos.

Paral·lelament, el camp de l'aprenentatge automàtic també ha avançat en l'anàlisi no lineal de dades. Per exemple, els mètodes de nucli ("kernel") ofereixen una manera que és, aparentment, senzilla de tractar les no-linealitats. Aquest s'anomena el "truc" del nucli, però que amaga un fons matemàtic molt rigorós. Aquests nuclis permeten efectuar operacions lineals una vegada les dades són traslladades a un espai de característiques d'alta dimensió, però sense la necessitat d'explícitament operar en aquest espai. Gràcies a això, aquests tenen el potencial de ser utilitzats en mètodes no paramètrics que d'altra manera podrien esdevenir inviables, com per exemple l'estimació de les mesures esmentades anteriorment. Aquest plantejament, però, té també els seus problemes. Ja que l'espai de característiques no es visita, l'operació lineal es fa cegament. També són propicis a requerir una regularització, donat l'augment de dimensionalitat, la qual s'ha de fer, altre cop, cegament. I, a més, solen comportar una complexitat computacional elevada, car creixen exponencialment amb el nombre de dades.

Aquesta tesi desenvolupa les dues idees anteriors i les ajunta en un sol marc de treball per a l'anàlisi de fenòmens complexos presents en les dades, aprofitant també les estadístiques de segon ordre i el seu llarg recorregut en aquesta àrea de coneixement. Per a tal, ho analitzem de dues formes diferents. En primera instància, busquem una manera de tractar amb les no-linealitats de les dades de forma lineal, similar a com ho fan els mètodes de nucli, però limitant la dimensió de l'espai de característiques per tal de guanyar en ambdues interpretacions i complexitat computacional. Per a obtenir el millor dels dos mons, la dimensió és major a la de les dades, però menor que en mètodes de nucli. És llavors en aquest espai on calculem correlació i covariància, i ho traslladem a mesures d'informació. Si bé la regularització es manté necessària donat l'increment en la dimensió del problema, el fet que coneixem l'espai de característiques ens permet analitzar el procediment i proposar una regularització adequada, la qual acaba mostrant una forta connexió amb tècniques clàssiques d'estimació espectral.

La segona part de la tesi es centra en l'aplicació d'aquestes mesures d'informació, però en problemes que normalment es resolen amb la utilització de tècniques de segon ordre. En concret, utilitzem l'entropia, una mesura d'incertesa, per tal d'avaluar les propietats d'una funció de densitat de probabilitat que no pas la variància, ja que aquesta no només depèn de les estadístiques de primer i segon ordre, sinó també d'estadístiques d'ordre superior. Els mètodes que fan servir aquest plantejament guanyen en robustesa quan les dades no són Gaussianes, però sense perdre de vista la seva relació intrínseca amb la variància. Per a tal, fem ús d'un estimador concret que, mentre prové de mètodes de nucli, és una estadística no esbiaixada i manté una forta relació amb el moment de segon ordre. Amb això, l'anàlisi dona la volta sobre si mateix i permet el desenvolupament d'un plantejament comú entre informació i estadístiques de segon ordre.

Agraïments

Escric aquestes línies amb el dubte de la necessitat d'agrair a la gent que m'envolta i importa. Quatre paraules mal escrites en un paper, o teclat, sempre seran menys eficients que uns agraïments verdaders. Això és, però, un exercici de solemnitat: plasmo ara i aquí la meva gratitud a certa gent que ha estat present, i m'ha acompanyat, en una de les etapes més importants de la meva vida. Així, també, en podré passar vergonya anys més tard.

Agraeixo primer a la meva família, sobretot a la més pròxima. A ma germana, a ma mare, i a les seves respectives parelles, per fer-me costat i no dubtar de mi. Encara que sigui recent, també a la Gaia, per... néixer just a temps d'escriure la tesi. Et desitjo una vida feliç. A mon pare. Et trobo a faltar. Finalment, també a na Marta, per acompanyar-me i aguantar-me. Ara em toca a mi. I a la Tisi també, esclar.

També només tinc paraules d'admiració i gratitud per al director de la tesi, en Jaume. Ha sigut un camí llarg i difícil, però de veritat m'has ajudat a poder-lo caminar. En segona instància, a tots els membres de RODIN. Gràcies al Gregori, al Javier, al Francesc, a la Xell i al Josep. Tot i que jo a vegades sóc molt esquiu, no vol dir que no em senti benvingut, i això ha sigut gràcies a l'amabilitat de tots vosaltres. També voldria fer un petit agraïment als professors amb qui he tingut més contacte fent docència, al Sisco, al Josep i a la Marga, així com també al Sergio.

Per acabar, segurament tampoc seria aquí si no fos pels companys i companyes de doctorat que han patit, i pateixen, juntament amb mi. I això va tant per als "de dalt", en Jordi, en Francesc i en Sergi, per la seva companyonia i ajuda quan els he necessitat, com per als "de baix", el(s) Carlos, el Rachid, en Marc i l'Aniol, per aguantar-me. A aquests dos últims, particularment, també els desitjo molts de púdings en el que resta de doctorat. Mireia, entres pels pèls.

This dissertation has been supported by:

- Fellowship FI 2019 by the Secretary for University and Research of the Generalitat de Catalunya and the European Social Fund.
- Ministerio de Ciencia e Innovación, Gobierno de España, through projects WINTER (TEC2016-76409-C2-1-R) and RODIN (PID2019-105717RB-C22/MCIN/AEI/10.13039/501100011033).
- Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, under grants 2017 SGR 578 and 2021 SGR 01033.

Contents

Nomenclature

General notation

$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of real positive numbers
$\mathbb{C}$	Set of complex numbers
$\mathbb{N}$	Set of natural numbers with 0
$\mathbb{N}_*$	Set of natural numbers without 0
$\lceil \cdot \rceil$	Ceiling function
$\lfloor \cdot \rfloor$	Floor function
$\mathrm{Re}\left\{\right\}$	Real part
$\mathrm{Im}\left\{\right\}$	Imaginary part
$(\cdot)^*$	Conjugate
$\exp\left(\right)$	Exponential with base e
$\ln\left(\right)$	Natural logarithm

Vectors and matrices

$\mathbf{a}$	Column vector
$\mathbf{a} \in \mathbb{R}^N$	Real column vector with N elements
$\mathbf{a} \in \mathbb{C}^N$	Complex column vector with N elements
$[\mathbf{a}]_n$	n-th element of vector $\mathbf{a}$
$\mathbf{A}$	Matrix
$\mathbf{A} \in \mathbb{R}^{N \times M}$	Real matrix of dimension $N \times M$
$\mathbf{A} \in \mathbb{C}^{N \times M}$	Complex matrix of dimension $N \times M$
$[\mathbf{A}]_{n,m}$	Element at the n-th row and m-th column of matrix $\mathbf{A}$
$\mathbf{I}_N$	$N \times N$ identity matrix
$\mathbf{0}_N$	N-dimensional column vector that contains all zeros
$\mathbf{1}_N$	N-dimensional column vector that contains all ones

$\mathbf{0}_{N\times M}$	$N \times M$ matrix that contains all zeros
$\mathbf{1}_{N\times M}$	$N \times M$ matrix that contains all ones
$(\cdot)^T$	Transpose
$(\cdot)^H$	Conjugate transpose
$e^{\mathbf{a}}$	Element-wise exponential $\left[e^{\mathbf{a}}\right]_n = e^{a_n}$
$\mathbf{a}^{\alpha}$	Element-wise power of a vector
$\mathbf{A}^{1/2}$	Square-root of a positive semi-definite matrix such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$
diag $(\mathbf{a})$	Diagonal matrix whose diagonal is composed by $\mathbf{a}$
Toe $(\mathbf{a})$	Toeplitz matrix constructed from the vector $\mathbf{a}$
tr $(\mathbf{A})$	Trace of matrix $\mathbf{A}$
$\lambda_n(\mathbf{A})$	n-th eigenvalue of matrix $\mathbf{A}$
$\sigma_n(\mathbf{A})$	n-th singular value of matrix $\mathbf{A}$

Operators

$\lvert \cdot \rvert$	Absolute value or modulus of a complex number
$\lVert\cdot\rVert_p$	p-norm of a vector, matrix or function
$\lVert\cdot\rVert_{\mathrm{F}}$	Frobenius norm
$\lVert\cdot\rVert_{\mathrm{HS}}$	Hilbert-Schmidt norm
$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{X}}$	Inner product between $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{X}$, corresponding to $\sum_{\mathcal{X}} \mathbf{x}^T \mathbf{y}$
$\odot$	Schur, Hadamard, or element-wise product
$*$	Convolution operator
$\mathcal{F}\{\cdot\}$	Fourier transform operator
$\mathcal{F}^{-1}\{\cdot\}$	Inverse Fourier transform operator

Probability theory and random variables

Pr $\{a\}$	Probability of event a
$\lvert \mathcal{A} \rvert$	Cardinality of the set $\mathcal{A}$
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Proper complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathbb{E}_p\{\cdot\}$	Statistical expectation operator over the PMF or PDF p
Var $\{A\}$	Variance operator Var $\{A\} = \mathbb{E}_{p_A}\left\{(A - \mathbb{E}_{p_A}\{A\})^2\right\}$
Cov $\{A, B\}$	Covariance operator Cov $\{A, B\} = \mathbb{E}_{p_{A,B}}\{(A - \mathbb{E}_{p_A}\{A\})(B - \mathbb{E}_{p_B}\{B\})\}$
Corr $\{A, B\}$	Correlation operator Corr $\{A, B\} =$ Cov $\{A, B\}$ / (Var $\{A\}$ Var $\{B\}$)

Estimation theory

$\hat{a}$ An estimate of a

$\text{Bias}\left\{\hat{a}\right\}$ Expected difference between the expected value and the true value, defined as $\mathbb{E}_p\left\{\hat{a} - a\right\}$

$\text{NBias}\left\{\hat{a}\right\}$ Normalized Bias, defined as $\mathbb{E}_p\left\{\hat{a} - a\right\}/a$

$\text{Var}\left\{\hat{a}\right\}$ Expected value of the squared estimator deviation, defined as $\mathbb{E}_p\left\{\left(\hat{a} - \mathbb{E}_p\left\{\hat{a}\right\}\right)^2\right\}$

$\text{NVar}\left\{\hat{a}\right\}$ Normalized Var, defined as $\mathbb{E}_p\left\{\left(\hat{a} - \mathbb{E}_p\left\{\hat{a}\right\}\right)^2\right\}/a^2$

$\text{MSE}\left\{\hat{a}\right\}$ Mean squared error, defined as $\mathbb{E}_p\left\{\left(\hat{a} - a\right)^2\right\}$

$\text{NMSE}\left\{\hat{a}\right\}$ Normalized MSE, defined as $\mathbb{E}_p\left\{\left(\hat{a} - a\right)^2\right\}/a^2$

Other notation

$\Gamma\left(\cdot\right)$ Gamma function

$\mathbb{1}_{\{x \in \mathcal{A}\}}$ Indicator function that returns 1 if $x \in \mathcal{A}$, and 0 if $x \notin \mathcal{A}$

$\delta_{n,n'}$ Kronecker delta, which yields 1 if $n = n'$ and 0 otherwise

Acronyms and abbreviations

AP Adaptive Partitioning

APSK Amplitude and Phase-Shift Keying

AWGN Additive White Gaussian Noise

CCA Canonical Correlation Analysis

CDF Cumulative Distribution Function

CDM Canonical Dependence Matrix

CF Characteristic Function

COCO Constrained Covariance

CRB Cramér-Rao Bound

DA Data-Aided

DTM Divergence Transition Matrix

EM Expectation-Maximization

EVB Envelope-Based

FSO Frobenius Second-Order statistics

GC Generalized Coherence

GMM Gaussian Mixture Model

HS Hilbert-Schmidt

HGR Hirschfeld-Gebelein-Rényi

HSIC Hilbert-Schmidt Independence Criterion

ICA Independent Component Analysis

i.i.d. independent and identically distributed

ITL Information-Theoretic Learning

IP Information Potential

KCCA Kernel Canonical Correlation Analysis

KDE Kernel Density Estimate

KL KullbackLeibler

KNN k-Nearest Neighbors

KPCA Kernel Principal Component Analysis

LSMI Least-Squares Mutual Information

MI Mutual Information

MIP Modified Information Potential

MISE Mean Integrated Squared Error

MSC Magnitude-Squared Coherence

MSE Mean Squared Error

NCRB Normalized Cramér-Rao Bound

NDA Nondata-Aided

NLS Nonlinear Least-Squares

NMSE Normalized Mean Squared Error

NOCCO Normalized Cross-Covariance Operator

PCA Principal Component Analysis

PDF Probability Density Function

PMF Probability Mass Function

PSD Power Spectral Density

QAM Quadrature Amplitude Modulation

QPSK Quadrature Phase-Shift Keying

RKHS Reproducing Kernel Hilbert Space

SMI Squared-loss mutual information

SNR Signal-to-Noise Ratio

SP Sample Spacings

SVD Singular Value Decomposition

SVM Support Vector Machine

Chapter 1

Introduction

Information is, for those who are familiar with information theory, a measurable concept, a mathematical descriptor of the randomness of the source of such information. Claude E. Shannon leveraged this notion into the definition of entropy and Mutual Information (MI) [Sha48]. With them, he established the basis of the fundamental limits in data compression and data transmission [CT06; Gra11]. However, Shannon did not provide a unique definition for *information*. In fact, a proper definition of information is still nowadays a subject of discussion (see, for instance, [Pop17] Section 1.1). Inspired by the laws of thermodynamics, Shannon referred to information as the reduction of uncertainty, and, as such, deeply tied to Von Neumann's entropy [Von18]. Before that, the definition of information was discussed by Ralph Hartley [Har28], who referred to information as a "very elastic" term, discussing whether information carries a connotation of physiological or physical factors. In the end, Hartley determined that information is something that can be quantified, and, therefore, compared between two communication systems. Some years after, Alfred Rényi elucidated two main categories of information: *axiomatic* and *pragmatic*. Axiomatic refers to the definition that arises from a set of postulates that information must follow, as Shannon did in his seminal work. Pragmatic, on the other hand, is the definition more aligned with Hartley's. From the words of Rényi [Rén65]: "*This approach starts from certain particular problems of information theory and accepts as measures of the amount of information the quantities which present themselves in the solution. According to this point of view the real justification of some measure of information is that it does work.*". In other words, a measure of information is something that adds value to a problem, a quantity that is perceived for the purpose of assessing the information source, whatever that source may be, and it is the definition with which this dissertation is concerned.

From the pragmatic point of view, information gauges a variety of attributes of random variables and probability distributions. Entropy measures uncertainty, MI measures dependence (or independence), and more generally, the KullbackLeibler (KL) divergence (or relative entropy), measures discrimination [KL51]. In the last decades there has been an uprising interest on translating these measures of information to other fields beyond communications with the objective of characterizing the output of a signal or a system. The motivation behind this new paradigm is driven by the increasing nonlinearity of data models. As data grows in complexity, linear models may not hold in practice, and the classical approach of second-order statistics becomes skewed. In this framework, entropy and MI have been used to substitute, or to work in tandem, with the conventional statistical tools of variance and covariance. Thanks to their capability of grasping the intricacies of the underlying model, these measures are sought as more reliable descriptors of data. As a result, some researchers have used these measures in multiple different areas, such as data science, machine learning, neuroscience, economics, biology, language, and other experimental sciences (see [CT06; WKV09] and references therein).

A prominent example that benefits from this approach is the area of Information-Theoretic Learning (ITL), a term coined by Príncipe [Prí10; Erd02], which cuts across signal processing and machine learning by reviewing the learning process under the umbrella of information theory. In its core, ITL is concerned about employing information measures to improve the performance of adaptive systems. For example, a minimum error entropy criterion gains in robustness in front of the Mean Squared Error (MSE) for the identification of nonlinear dynamical systems [EP02], regression analysis [Hu+13], or data classification [SSA10]. In these cases, the advantage of entropy arises when the Gaussian assumption does not hold in practice, acknowledging that an information measure captures information beyond the mean and variance. As a result, information measures give an edge for more convoluted distributions, long tails, or faulty observations/outliers [HH18]. Nonetheless, this information-theoretic approach can be encountered in the classic signal processing literature. For example, it is devised in the Independent Component Analysis (ICA) method [NK11; Pha04], where MI is used to recede from Gaussianity, or in model-order selection problems [WK85; SS04], which employ the KL divergence for determining the discrepancy between the data and the modelled distribution.

One of the most challenging aspects of information-theoretic methods is to procure an accurate, if possible low complex, information estimation. These algorithms are required for the analysis of systems or data and posterior enhancement of a given application. Plug-in methods have been traditionally used for this purpose, which are based on estimating first the distributions of the observed data, and to measure the functional of interest in a second stage [MRL95; Par62]. However, these methods are generally susceptible to estimation errors when dealing with random variables with long tails in their distributions, and their rate of convergence may be too slow for real world case scenarios [HM93]. Other estimation approaches have been proposed, which surpass the capabilities of the base plug-in approach, although they are also, generally speaking, derivatives of plug-in estimates [WKV09; DV99; KSG04; BT11]. As a matter of fact, one of the key aspects for the proficiency of ITL is that most of its literature employs surrogates of Shannon's information measures. Generally speaking, the Rényi information measures are considered [Rén61; Rén65], which generalize the notion of Shannon's measures (and KL divergence in some cases) as the limit case of their additional parameter α, usually referred to as the entropic index. These surrogates have the advantage of being well-known in the field of information theory, where they have been characterized, discussed, and applied to many different problems [Csi95; Csi08; Ver15; EH14; BR78]. Sometimes, these surrogates are sufficient from an information-theoretic standpoint. The advantage and motivation for their use outside of information theory is that these surrogates may become "easier" to estimate [LPS08; PPS10; OUE08; Kim18; Sar16]. For instance, it has been shown that and estimate of Rényi's entropy converges faster than its Shannon counterpart [Ach+16], at least for discrete sources, and the case of $\alpha = 2$ (or second-order Rényi entropy) is shown to cope specifically well with the plug-in estimate, to the point that it can be written in closed-form expression [AH84; Joe89]. This last property is given thanks to the particular expression of the second-order Rényi entropy, which is a function of the L^2-norm of the probability function and resembles a second-order moment.

Another remarkable trend in the literature to deal with nonlinearities is the use of kernel methods. A kernel is a function that implicitly measures an inner product in an unvisited feature space, endowed with a solid mathematical background by the hands of James Mercer and Nachman Aronszajn [Mer09; Aro50]. Its characteristic mechanism is to map the data in a feature space of allegedly infinite dimension, conduct some linear operator, and return with just the result of such operation. Their approach is data-driven and universal, meaning that no model is assumed, and it is the data that leads to the appropriate solution. Thanks to the increase in the dimension with respect to the data space, it is usually stated that the problem can be solved in a linear fashion, either for classification following Covers' theorem [Cov65] or in regression problems. Its apparent simplicity, yet being a powerful tool, has made kernel methods a very appealing framework, being implemented for feature extraction and pattern recognition problems [SSM98; CS00; BJ02], in signal processing problems [Roj+18; PB04], communications theory [Din+13; Sav+15], information theory [Bac22], or ITL itself [Xu+08], among many others. For that matter, the use of kernel methods is not a stranger to the measure of information. The linearization in the feature space is also employed to measure information by means of covariance or correlation operators [SS00; Gre+05b; SSB+02]. The covariance and correlation operators in the feature space spanned by kernel methods can be seen, in fact, as a generalization of the selfsame operators in the Euclidean space [Mua+17]. Therefore, kernel methods provide a very interesting approach to the measure of information based on second-order statistics tools. As the methods that derive from measures of covariance and correlation are well-known, we found here an opportunity to strengthen both worlds by providing a unified approach. However, kernel methods are not without some caveats. Their mapping into a higher feature space is prone to sparsity in such space, driving a collection of problems for regularization of a space that is not visited. Furthermore, kernel methods do not learn nor question the required mapping dimension, since it is "let to decide" by the kernel. The result is that kernels tend to overshoot the minimum required problem dimension. This also translates as increasing the computational complexity, since the only other dimensionality left is the data, rendering the approach prohibitive for large data sets, and defying the sole purpose of procuring a data-driven approach. In fact, a lot of research has been conducted in an attempt to decrease the complexity of kernel methods. [RR07; Lim+15; LP20].

This dissertation tackles both the use of information measures to solve certain problems that are classically approached with second-order statistics, and the measurement of such information through covariance and correlation operators on a high-dimensional feature space. In essence, the objective is to provide a unified view of the intersection between information and second-order statistics. To perform such action, it is clear that some method of linearization is required, akin to kernel methods. However, while kernel methods are capable of solving very complex problems, they are also incapable of explaining the required process for learning or decision-making. This thesis opposes this point of view by going into detail on how data is processed, what is required for linearizing a nonlinear problem, and how information is measured. Afterwards, how to employ this knowledge to solve a problem with an information-theoretic perspective is also discussed.

1.1 Objectives

Following the overview of the information-theoretic philosophy, literature and methods, we then proceed to allocate this dissertation to a concrete framework with clear and detailed objectives. Generally speaking, we can divide the contents and associated philosophy of this work in two different topics. The first one is concerned about developing insightful tools for measuring meaningful indicators of the amount of information contained in raw data. The second topic addresses the use of information for the problem of parameter estimation. Regarding the first topic, the following are the primary focus with which the information estimators are derived:

- **Interpretability**: In machine learning, this term refers to the understanding of the learning or decision process, which has gained relevance in contemporary practical applications [Eld+17]. The framework and methods developed in this dissertation are based on an explicit feature map, given by the characteristic function, instead of an implicit one. By controlling the mapping itself, we gain insight, which can then be exploited to improve the estimation process itself.

- **Scalability**: Another relevant objective is to develop an estimation procedure whose computational complexity has an appropriate growth with the data size. This is partially done thanks to the explicit feature map, which allows focusing on the dimension of the feature space. By governing the dimensionality of the problem, we can provide a trade-off between complexity and accuracy, which is complemented by the interpretability of the problem.

- **Universality**: A particular emphasis is given to making the estimation procedure universal, that is, that the algorithm converges to the measure being estimated without any knowledge of the underlying probability distributions. Therefore, the proposed approach is data-driven, which implies that it makes decisions only based on the data observations, and a nonparametric estimate, i.e. it is distribution-free.

- **Second-order approach**: Finally, our main interest relies on the use of well-known techniques in classical signal processing problems. To be specific, we will leverage the Principal Component Analysis (PCA) and the Canonical Correlation Analysis (CCA) techniques for the estimation of information measures. This can be done thanks to characteristic function mapping, which links independence and uncorrelation. Due to this property, the method for estimating information with second-order statistics in the feature space arises naturally, providing an alternative path to kernel methods, which is similar but different in nature.

As a general remark, the interpretability, scalability, and universality of the approach can all be associated with the use of the characteristic function as a mapping and the posterior use of second-order statistics. On the one hand, it is a well-known tool in the field of probability theory with clear operational meaning. It can be shown that the covariance and correlation operators in the characteristic function space are fundamentally related to the same operators for the associated random variables. Furthermore, the regularization technique in this space will be strongly related to the contamination of sources with additive noise. As the contamination is deliberately introduced to regularize the problem, its distribution is known and its consequences to the characteristic function are measurable. On the other hand, the reduction of dimensionality will be provided by sampling the characteristic function. Thanks to this, the problem of estimating information is related to the estimation of covariance and correlation matrices, which is known to be consistent for a wide range of distributions, as well as the empiric estimate of the characteristic function. The proposed mapping and regularization with contamination help to the universality and scalability of the problem by controlling the overall characteristic function shape, admitting any kind of distribution.

The second topic, the information-theoretic approach to concrete problems, is more focused on the given applications at hand. Nevertheless, there is a general philosophy that accompanies how these problems are approached. In this sense, the objective is to contribute with a broad entropy-based viewpoint to some parameter estimation problem. The concrete objectives are the information-theoretic estimation of the variance of a random scalar sequence, the coherence of a random vector sequence, and the Signal-to-Noise Ratio (SNR) in a linearly modulated digital communications channel. The chosen information measure surrogate to perform such tasks is the Rényi entropy (with the entropic index equal to 2), which provides more granularity than the low-order moments. An accurate evaluation of the employed entropy estimator is provided. While still intrinsically under the kernel method standpoint, it is a particular formulation of kernel methods that yields an explicit feature map, contributing to the interpretability of the problem. Given that it is a kernel method, regularization is still required, although the overall effect of this regularization on the final estimate is known. Moreover, the entropy estimate is asymptotically related to the sample variance estimator, thus giving a joint vision to the topic of this dissertation.

1.2 Thesis outline and contributions

To end this chapter, the contents of the dissertation are elaborated, jointly with a list of the different research contributions for each of them. As mentioned above, most of the results that derive from this dissertation can be divided in two branches, where each one studies a different perspective of the relationship between information and second-order statistics. In the sequel, a brief explanation of every chapter is given:

- **Chapter 2** provides a general review of measures of information that are relevant to the contents of this thesis. Both linear-based (correlation and covariance) and nonlinear (information-theoretic) measures are addressed. The first will serve as the main second-order statistics methods on which the posterior framework is based, and the latter focuses on determining the appropriate information-theoretic surrogates for the purpose of estimation. While this chapter is generally composed of background concepts, it also provides a fresh and novel view of the difference between some surrogates of measures of information.

- **Chapter 3** unveils the relationship between kernel methods and some of the information-theoretical measures addressed in Chapter 2. First, the theory of reproducing kernel Hilbert spaces is concisely reviewed, and then some methods that take advantage of it are unravelled. All of the presented kernel methods are based on measuring covariance and correlation in the infinite-dimensional feature space, hence they are employed as a point of comparison for the method developed in the following chapter. Moreover, this chapter also deals with a specific entropy surrogate estimator that will be used in future derivations. Although this estimator is derived from the family of plug-in estimators, it is intrinsically related to kernel methods, and it possesses some properties that will be exploited in Chapter 5.

- **Chapter 4** develops the framework for the estimation of information measures in a finite-dimensional feature space, in contrast with kernel methods. On a first stage, the mapping is determined for discrete sources, emphasizing its implications. Then, the insights gained in the discrete case are employed for the continuous case. Numerical results are provided for the latter, focusing on the choice of the new hyper-parameters and the derived estimators are compared with existing methods in the literature. This chapter spans the first branch of the thesis, where information measures capitalize on the use of second-order statistics.

  The technical works comprised in this chapter are:

  - F. de Cabrera and J. Riba. “A novel formulation of Independence Detection based on the Sample Characteristic Function”. In: *26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018.
  - F. de Cabrera and J. Riba. “Squared-Loss Mutual Information via High-Dimension Coherence Matrix Estimation”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.
  - F. de Cabrera and J. Riba. “Regularized Estimation of Information via Canonical Correlation Analysis on a Finite-Dimensional Feature Space”. In: *IEEE Transactions on Information Theory*, Early Access, DOI: 10.1109/TIT.2023.3258182.

- **Chapter 5** tackles three different problems: the estimation of the determinant of a covariance matrix, the estimation of the magnitude-squared coherence, and the estimation of the SNR. All three applications share the same backbone by employing the entropy estimator addressed in Chapter 3. Thanks to the use of entropy, the estimators gain robustness in front of the lack of optimal conditions, whether they are faulty observations or a mismatch of the parameters of the channel. The chapter begins by analyzing the behaviour of the entropy estimator and then assesses the performances of the entropy-based applications separately. This chapter considers the second branch, where an information-theoretic measure is employed in exchange for second-order statistics.

  The technical works comprised in this chapter are:

  - F. de Cabrera, J. Riba and G. Vázquez. “Entropy-based covariance determinant estimation”. In: *IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Sapporo, Japan, 2017.
  - F. de Cabrera, J. Riba and G. Vázquez. “Robust estimation of the magnitude squared coherence based on kernel signal processing”. In: *51st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2017.

  - F. de Cabrera and J. Riba. “Entropy-Based Non-Data-Aided SNR Estimation”. In: *53rd Asilomar Conference on Signals, Systems, and Computers* , Pacific Grove, CA, USA, 2019.

- **Chapter 6** concludes the dissertation by outlining the contents and contributions of the thesis, as well as reflecting on possible future lines of research with the scope of expanding and enhancing the present work.

Other contributions: The following publications are related to some contents in this thesis. These are inside the general framework discussed here, albeit they are not part of the specific rationale of each chapter. Observations based on the relationship of these publications with the core contents of the thesis will be pointed out when necessary, thus widening the network of connections with the works done by the author.

- J. Riba and F. de Cabrera. “A Proof of de Bruijn Identity based on Generalized Price’s Theorem”. In: *IEEE International Symposium on Information Theory (ISIT)*, Paris, France, 2019.

- C. A. López, F. de Cabrera and J. Riba. “Estimation of Information in Parallel Gaussian Channels via Model Order Selection”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

The author has also contributed in:

- J. Riba, F. de Cabrera and J.M. Juan. “Multi-Satellite Cycle-Slip Detection and Exclusion Using the Noise Subspace of Residual Dynamics.” In: *26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018.

- A. Martí, J. Portell, D. Amblas, F. de Cabrera, M. Vilà, J. Riba, G. Mitchell, “Compression of Multibeam Echosounders Bathymetry and Water Column Data”. In *Remote Sensing*, 14(9):2063, 2022.

Chapter 2

A review of information measures

A fundamental task in this dissertation is to determine what information-theoretic measures are covered. There are many surrogates of the fundamental measures, which are Shannon's entropy, KL divergence, and mutual information, that coexist and provide different viewing angles to randomness, dissimilarity, and dependence, respectively. For instance, a generalization of the KL divergence [KL51] is the class of f-divergences, which, at the same time, include multiple other definitions, each one with its own set of properties and applications [LV06; EH14]. Shannon's entropy shares an identical situation, where it is usually generalized by the Rényi [Rén61] or the Tsallis entropies [Tsa88], and so on. However, not all these measures are equally suitable for the objective of this thesis. This chapter addresses the derivation and choice of the appropriate surrogates that may be more amicable with the data. Specifically, we sought those surrogates that have some kind of relationship with second-order statistics, since those are the ones that are, in principle, able to be estimated using well-known second-order techniques. For clarity in the exposition, these latter techniques are also addressed in this chapter.

The structure of this chapter is organized as follows. Section 2.1 provides a short exposition of the characteristic function and Szegös theorem. While the first is part of the essence of the mapping that will be proposed hereinafter, the second will be used to analyze asymptotic behaviours of the Toeplitz matrix that will appear due to the characteristic function mapping itself. Section 2.2 is a brief review of methods for measuring correlation, hence linear dependencies. In particular, we focus on the principal component and the canonical correlation analyses, as they will serve as the main tools for measuring information through second-order statistics. Lastly, in Section 2.3, several information-theoretic measures are studied. The rationale for choosing the appropriate surrogates is addressed, and some comparisons with their original counterparts are provided.

2.1 Preliminaries

2.1.1 A brief review of the characteristic function

The Characteristic Function (CF) is an important tool in the field of probability theory that serves for multitude of different applications [Luk63]. In its core, the CF has a unequivocal relationship with the Probability Density Function (PDF) of a given random variable, determined by the Fourier transform. This motivates the use of the CF as an equivalent descriptor of any random variable, which may be easier to work with. Since it is the Fourier pair of the PDF, it also carries some interesting properties that will be exploitable in order to solve complex problems. For instance, the CF can be used to find the moments of a random variable through the derivative, which may be useful for analyzing small-scale phenomena. Due to these reasons, the CF is an excellent tool, and it shall be used as a promoter of the contents of this thesis.

Definition 1. Let X be an absolutely continuous random variable with PDF $f_X(x)$ defined on the set $\mathcal{X}$. The CF of X is the function $\varphi : \mathbb{R} \to \mathbb{C}$ defined as follows:

$$\varphi_X(\omega) = \int_{\mathcal{X}} e^{j\omega x} f_X(x)\, \mathrm{d}x = \mathbb{E}_{f_X}\left\{e^{j\omega x}\right\}, \qquad \forall \omega \in \mathbb{R}. \tag{2.1}$$

The interpretation of the CF can be either the expectation over a complex exponential function or the Fourier transform with reverse sign in the complex exponential. Following this last interpretation, the subsequent properties are straightforward from the properties of the PDF:

1. Unity at origin: $\varphi_X(0) = 1$.
2. Bounded by one: $|\varphi_X(\omega)| \leq 1$.
3. Hermitian: $\varphi_X(-\omega) = \varphi_X^*(\omega)$.

Since the density function of a random variable is always absolutely integrable, the CF always exists. Furthermore, it can be shown that the CF is uniformly continuous for every $\omega \in \mathbb{R}$ [Luk70, Thm. 2.1.2]. The Fourier interpretation can also be used to determine the CF of the sum of random variables. For instance, let X_1 and X_2 be independent random variables, and let $Y = X_1 + X_2$ with PDF $f_Y(y)$. It is well-known that the distribution of Y is determined by the convolution of the distributions of X_1 and X_2:

$$f_Y(y) = \int_{\mathcal{X}} f_{X_1}(x)\, f_{X_2}(y - x)\, \mathrm{d}x. \tag{2.2}$$

Therefore, thanks to the convolution theorem of the Fourier transform, it is straightforward to see that their respective CF will be given by the product such that

$$\varphi_Y(\omega) = \varphi_{X_1}(\omega)\, \varphi_{X_2}(\omega). \tag{2.3}$$

Another relevant property of the CF is given by the uniqueness theorem, which states the following:

Theorem 2.1 (Uniqueness theorem [Luk70])**.** *Let $F_1(x)$ and $F_2(x)$ be two Cumulative Distribution Function (CDF)s. The functions $F_1(x)$ and $F_2(x)$ are identical if, and only if, their characteristic functions $\varphi_1(\omega)$ and $\varphi_2(\omega)$ are identical.*

However, this theorem only concerns CDFs, and it does not determine the inverse path, i.e. the CDF given a CF. For this, we need the following theorem:

Theorem 2.2 (Inversion theorem [Luk70])**.** *Let $\varphi_X(\omega)$ be the CF of the random variable X. If $\varphi_X(\omega)$ is absolutely integral over the whole real line, $\int_{\mathbb{R}} |\varphi_X(\omega)|\, d\omega < \infty$, then the following holds true:*

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \varphi_X(\omega)\, e^{-j\omega x} \mathrm{d}\omega, \tag{2.4}$$

with $f_X(x) = F_X'(x)$, and $f_X(x)$ being uniformly continuous for $x \in \mathbb{R}$.

Given the inversion theorem, the CF and the PDF are Fourier pairs provided that the CDF is differentiable. This last condition is relevant, since the uniqueness theorem governs over the CDF. The PDFs that fulfil this requirement are then leveraged by the bijective property of the Fourier transform, and there exist a one-to-one relationship between the CF and the PDF. However, these are only a subset (those whose derivative exists) of all possible PDFs, contrary to all the one-to-one relationship with all the CDFs.

The duality pair provided by the inverse theorem can also be used in applications where the estimation of probability functions is required, inherent, or directly main goal. For instance, since the PDF behaves like a Power Spectral Density (PSD), given both are nonnegative finite-area functions, it is possible to exploit well-known tools in the context of spectral estimation through the estimation the CF. This way, one can draw applications such that of estimation of the PDF [PL96], or the estimation of Shannon's entropy and the Kullback-Leibler divergence [BV00; Ram+09].

Next, we will proceed to the generalization of the CF for multivariate random variables. Let $X = \{X_1, ..., X_N\}$ be a N-variate random variable with PDF $f_X(\mathbf{x})$ defined on the set $\mathcal{X}$. The multivariate CF is then defined as the function $\varphi : \mathbb{R}^N \to \mathbb{C}$ such that

$$\varphi_X(\boldsymbol{\omega}) = \int_{\mathcal{X}} e^{j\boldsymbol{\omega}^T\mathbf{x}} f_X(\mathbf{x})\, \mathrm{d}\mathbf{x} \qquad \forall \boldsymbol{\omega} \in \mathbb{R}^N. \tag{2.5}$$

Given that the product between $\boldsymbol{\omega} = [\omega_1, ..., \omega_N]^T$ and $\mathbf{x} = [x_1, ..., x_N]^T$ leads to a sum, its exponential can be factorized:

$$\varphi_X(\omega_1, ..., \omega_N) = \int_{\mathcal{X}} e^{j\omega_1 x_1} \cdots e^{j\omega_N x_N} f_X(x_1, ..., x_N)\, \mathrm{d}x_1 \cdots \mathrm{d}x_N. \tag{2.6}$$

For independent random variables, i.e. $f_X(x_1, ..., x_N) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_N}(x_N)$, the separability of PDFs also entails the separability of CFs such that

$$\varphi_X(\omega_1, ..., \omega_N) = \int_{\mathcal{X}} e^{j\omega_1 x_1} \cdots e^{j\omega_N x_N} f_{X_1}(x_1) \cdots f_{X_N}(x_N)\, \mathrm{d}x_1 \cdots \mathrm{d}x_N \tag{2.7a}$$

$$= \int_{\mathcal{X}} e^{j\omega_1 x_1} f_{X_1}(x_1)\, \mathrm{d}x_1 \cdots \int_{\mathcal{X}} e^{j\omega_N x_N} f_{X_N}(x_N)\, \mathrm{d}x_N = \varphi_{X_1}(\omega_1) \cdots \varphi_{X_N}(\omega_N). \tag{2.7b}$$

Therefore, independence factorizes both PDFs and CFs equally.

Furthermore, by retrieving the definition of the CF as the expectation over some nonlinear function in (2.1), we can further elaborate on the implications of the separability. Without loss of generalization, let us fix $N = 2$ where $Z_1 = e^{j\omega_1 X_1}$ and $Z_2 = e^{j\omega_2 X_2}$ are the new complex random variables obtained from X_1 and X_2. The joint CF can be expressed as follows

$$\varphi_X(\omega_1, \omega_2) = \mathbb{E}_{f_X}\{Z_1 Z_2\}. \tag{2.8}$$

Clearly, if X_1 and X_2 are independent we can write

$$\varphi_X(\omega_1, \omega_2) = \mathbb{E}_{f_{X_1}}\{Z_1\} \mathbb{E}_{f_{X_2}}\{Z_2\} = \varphi_{X_1}(\omega_1) \varphi_{X_2}(\omega_2) \tag{2.9}$$

for $\omega_1, \omega_2 \in \mathbb{R}$. The implication is that we are moving an independence requirement to an uncorrelation one, thanks to the specific nonlinear mapping of the original random variables. Also note that, due to the uniqueness property of the CF, the converse is also true: if for any value of ω_1 and ω_2 these new variables are uncorrelated, then the original variables are independent. This implies that, if $\varphi_X(\omega_1, \omega_2) = \varphi_{X_1}(\omega_1) \varphi_{X_2}(\omega_2)$, then $f_X(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$, which is the converse statement from (2.7).

This property of the CF has been traditionally employed as a detector of independent random variables by testing the distance between the joint and the product of marginal CFs in [Csö85]. Moreover, the implication of the converse statement is that any form of dependence between X_1 and X_2 is revealed through correlation between the variables Z_1 and Z_2 for some values of ω_1 and ω_2. An independence detector was established in our conference paper [CR18] by detecting correlation, provided that a sufficient number of ω_1 and ω_2 are explored. In conclusion, this observation motivates the use of second-order statistics after some nonlinear mapping for measuring statistical dependency, and it will be leveraged to estimate measures of information in the following chapter.

2.1.1.1 Characteristic function and moments

We will now review the relationship between the CF and the k-th-order moment of the associated random variable. This one is another relevant property of the CF as it enables the study of local approximations as a function of mean, variance, skewness, etc. In particular, the moments of a random variable are linked to the CF through its derivatives at the origin, which is specifically useful for approximating a function through its Maclaurin series.

Consider the N-variate random variable $X = \{X_1, ..., X_N\}$ with CF $\varphi_X(\boldsymbol{\omega})$. By differentiating at the origin (see Appendix 7.1.1) we obtain

$$\left.\frac{\partial^{k_1+k_2+...+k_N}\varphi_X(\boldsymbol{\omega})}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}...\partial\omega_N^{k_N}}\right|_{\boldsymbol{\omega}=\mathbf{0}_N} = j^{k_1+k_2+...+k_N}\mathbb{E}_{f_X}\left\{X_1^{k_1}\cdots X_N^{k_N}\right\}, \tag{2.10}$$

where $k_n \in \mathbb{N}$ for $n = \{1, ..., N\}$. Here, the CF's derivative is expressed as general as possible, where k_n indicates the k_n-th derivative for the corresponding n-th random variable. However, it does not guarantee the existence of the derivative for all CFs. The following theorem determines the conditions for it to hold true:

Theorem 2.3 (Characteristic function and moments [Luk70])**.** *If the CF $\varphi_X(\boldsymbol{\omega})$ has a derivative of order k at $\boldsymbol{\omega} = \mathbf{0}_N$, then all the moments of X exist up to order k if k is even, or up to order $k-1$ if k is odd. It also follows that if the k-th moment of X exists, then the CF can be differentiated k times.*

Simply put, the existence of the moments of a random variable implies that these can be found by differentiating the CF at the origin. Now we are in terms of particularizing to the most relevant cases from the point of view of this thesis, which are the first and second-order moments.

Since k_n can be zero, then (2.10) includes the k_n-th partial derivative of the marginal CF $\varphi_{X_n}(\omega_n)$ such that

$$\left.\frac{\partial^{k_n}\varphi_{X_n}(\omega_n)}{\partial\omega_n^{k_n}}\right|_{\omega_n=0} = j^{k_1}\mathbb{E}_{f_X}\left\{X_n^{k_n}\right\}, \tag{2.11}$$

which is equivalent to determine the k-th derivative for the univariate case $N = 1$. From these, and focusing on the statistical expectation, the first and second-order moments correspond to

$$\mathbb{E}_{f_{X_n}}\{X_n\} = -j\left.\frac{\partial\varphi_{X_n}(\omega_n)}{\partial\omega_n}\right|_{\omega_n=0}, \qquad \mathbb{E}_{f_{X_n}}\{X_n^2\} = -\left.\frac{\partial^2\varphi_{X_n}(\omega_n)}{\partial\omega_n^2}\right|_{\omega_n=0}, \tag{2.12}$$

respectively. The particularization from (2.10) is as follows

$$\mathbb{E}_{f_X}\left\{X_1^{k_1}X_2^{k_2}\right\} = \frac{1}{j^{k_1+k_2}}\left.\frac{\partial^{k_1+k_2}\varphi(\omega_1,\omega_2)}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}}\right|_{\omega_1,\omega_2=0}, \tag{2.13}$$

which can be used to determine the appropriate required derivative for any $\{k_1, k_2\}$ combination.

In conclusion, the CF provides a preliminary link between statistical dependence and the moments of a random variable. We are particularly interested on the cases in which the CF can be approximated by the mean and variance of X, since these are the ones that piece (2.8) and (2.13) together. For instance, if we approximate (2.8) by its Maclaurin decomposition up to the second order, we can see that the correlation between Z_1 and Z_2 is indeed related to the correlation between the original variables X_1 and X_2, albeit approximately. Although it is reduced to an asymptotic tendency, e.g. when the CF can be approximated by a constant around the origin, it provides an analysis tool for the posterior derivations based on the CF.

2.1.2 Limit theorems of Toeplitz matrices

We will encounter multiple Toeplitz matrices over the course of the dissertation. Most times, we are interested in operating with these matrices to achieve some concrete goal, let it be to estimate some measure or to characterize a distribution. Furthermore, we are particularly interested in the characterizations of these matrices for an increasing matrix dimension and how is the limit behaviour of their eigenvalues. This subsection deals with the limit theorems of the behaviour of Toeplitz matrices, which will prove to be useful when handling various types of Toeplitz matrices.

To begin with, we will define both Toeplitz and circulant matrices:

Definition 2. A Toeplitz matrix $\mathbf{T} \in \mathbb{C}^{N \times N}$ is matrix with constant-valued diagonals

$$[\mathbf{T}]_{n,n'} = [\mathbf{T}]_{n-n',0} = t_{n-n'}, \tag{2.14}$$

for $n, n' = 0, ..., N-1$. In matrix form, that is

$$\mathbf{T} = \begin{bmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{-N+1} \\ t_1 & t_0 & t_{-1} & \ddots & \vdots \\ t_2 & t_1 & \ddots & \ddots & t_{-2} \\ \vdots & \ddots & \ddots & t_0 & t_{-1} \\ t_{N-1} & \cdots & t_2 & t_1 & t_0 \end{bmatrix}. \tag{2.15}$$

Generally speaking, all Toeplitz matrices that will arise along the thesis will also be Hermitian, which endows them with the structure $t_n = t^*_{-n}$. Consequently, and for simplicity, we will construct a Hermitian-Toeplitz from the vector $\mathbf{t} = [t_0, t_1, \cdots, t_{N-1}]^T$, denoted by $\mathbf{T} = \mathrm{Toe}(\mathbf{t})$.

Definition 3. A circulant matrix $\mathbf{C}$ is a special case of Toeplitz matrix where every row of the matrix is a right/left circular shift of the row above/below:

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{-1} & c_{-2} & \cdots & c_{-N+1} \\ c_{-N+1} & c_0 & c_{-1} & \ddots & \vdots \\ c_{-N+2} & c_{-N+1} & \ddots & \ddots & c-2 \\ \vdots & \ddots & \ddots & c_0 & c_{-1} \\ c_{-1} & \cdots & c_{-N+2} & c_{-N+1} & c_0 \end{bmatrix}. \tag{2.16}$$

Both Toeplitz and circulant matrices, and their properties, are well-known in the literature [GS58; Dav79]. In the context of limit behaviours, it is of particular interest the property that circulant matrices are diagonalizable by the unitary discrete Fourier matrix $\mathbf{W}$. In particular, any circulant matrix satisfies

$$\mathbf{C} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^H, \tag{2.17}$$

where

$$\mathbf{W} = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{-j2\pi/N} & e^{-j2\pi 2/N} & \cdots & e^{-j2\pi(N-1)/N} \\ 1 & e^{-j2\pi 2/N} & e^{-j2\pi 4/N} & \cdots & e^{-j2\pi 2(N-1)/N} \\ \vdots & \vdots & \vdots & \ddots & e^{-j2\pi 3(N-1)/N} \\ 1 & e^{-j2\pi(N-1)/N} & e^{-j2\pi 2(N-1)/N} & e^{-j2\pi 3(N-1)/N} & e^{-j2\pi(N-1)(N-1)/N} \end{bmatrix}, \tag{2.18}$$

and $\boldsymbol{\Lambda}$ is a diagonal matrix that contains the eigenvalues of $\mathbf{C}$, which correspond to the samples of the discrete Fourier transform of the sequence c_{-n} (or directly the sequence c_n depending on how the circulant matrix has been defined). Toeplitz matrices have also a similar property with Vandermonde matrices [YXS16], whose columns are composed by sampled complex exponential functions. However, our interest relies on the fact that Toeplitz matrices are asymptotically circulant matrices, whose properties, and particularly the eigenvalue behaviour, can then be easily exploited through the Fourier transform. In this sense, the following theorems will be used to improve the analysis and performance of the proposed estimation techniques based on operating with Toeplitz matrices.

Theorem 2.4 (Szegö's theorem on asymptotic eigenvalues behaviour [GS58]). *Let t_n be the sequence that determines a Hermitian-Toeplitz matrix $\mathbf{T}$ for $n = 0, ..., N-1$. Consider $f(x)$ a bounded and real-valued continuous function defined by the Fourier series with coefficients t_n, related by*

$$t_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-jnx} \mathrm{d}x. \tag{2.19}$$

Moreover, let $F(\lambda)$ be a continuous function in the same interval as $f(x)$. Then, Szegö's theorem states that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} F(\lambda_n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(f(x)) \mathrm{d}x, \tag{2.20}$$

where λ_n are the eigenvalues of $\mathbf{T}$.

Note that, by construction of the Fourier series, it is inherent that $f(x)$ is considered to be a periodic function with period 2π, hence the normalization and limits of the integral.

Theorem 2.5 (Szegö's theorem on asymptotic circulant matrices [Gra+06]). *Let t_n be the sequence that determines a Hermitian-Toeplitz matrix $\mathbf{T}$ for $n = 0, ..., N-1$, and that it possesses a limited number of M nonzero entries in its diagonals with $M < N$. Consider a circulant matrix $\mathbf{C}$ composed by the sequence c_n with*

$$c_n = \begin{cases} t_{-n} & n = 0, 1, ..., M-1 \\ t_{k-n} & n = k - M + 1, ..., k-1 \\ 0 & \textit{otherwise} \end{cases} \tag{2.21}$$

for $k = 0, ..., N-1$. Then, $\mathbf{T}$ is asymptotically $\mathbf{C}$ such that

$$\lim_{N \to \infty} \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} \sum_{n'=0}^{N-1} \left| [\mathbf{T}]_{n,n'} - [\mathbf{C}]_{n,n'} \right|^2} = 0. \tag{2.22}$$

To put it simply, the previous theorem states that a *banded* Toeplitz matrix is asymptotically a circulant matrix. However, this is still a restrictive constraint that may not be applicable with most Toeplitz matrices. It is possible, however, to relax the condition with a milder assumption. In particular, the most general and relaxed assumption that guarantees the behaviour described in the previous theorem is that the sequence t_n is square-integrable for $N \to \infty$. This is usually referred as the weak conditions of Szegö's theorem [Bin12]. In that case, a Toeplitz matrix can still be asymptotically approximated by an equivalent circulant matrix.

The relevance of this limit case can be encountered in the behaviour of their respective eigenvalues. In particular, the following Corollary to Theorem 2.5 encompasses the desired property for asymptotically large Toeplitz matrices:

Corollary 2.5.1. [Gra+06, Lemma 4.3] *Let $\mathbf{T}$ and $\mathbf{C}$ be the Toeplitz and circulant matrices from Theorem 2.5, and let $\lambda_n(\mathbf{T})$ and $\lambda_n(\mathbf{C})$ be their respective eigenvalues. Then*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \left(\lambda_n^{\beta}(\mathbf{T}) - \lambda_n^{\beta}(\mathbf{C}) \right) = 0 \tag{2.23}$$

for any positive integer β. For a finite N, then

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} \left(\lambda_n^{\beta}(\mathbf{T}) - \lambda_n^{\beta}(\mathbf{C}) \right) \right| \leq \gamma N^{-1/2}, \tag{2.24}$$

where γ is independent from N.

As a result, not only the matrices are asymptotically the same, but also their eigenvalues, hence any Toeplitz matrix with the aforementioned assumptions can be processed as a circulant matrix. We are purposely interested in this property given the diagonalization of circulant matrices. In particular, to diagonalize a Toeplitz matrix by the unitary Fourier matrix will serve as a tool for simplifying matrix inversions.

2.2 Measures of linear dependence

Before delving into the core content of the thesis, we will briefly review the analyses of linear dependence that will be used to leverage the estimation of information measures and their surrogates. Albeit well-known, the contents of this section play a very important role in future parts of the thesis. It is then within reason to review some of the most important methods that measure linear dependence through second-order statistics. To be concrete, we will be focusing on the PCA and on the CCA.

2.2.1 Principal component analysis

PCA is generally regarded as a dimensionality reduction algorithm that computes the direction of maximum variance of a set of multivariate data [Jol02]. Its principle is based on finding the orthogonal weight vectors that maximize data variation. These weight vectors, called the principal components, are then used to linearly transform the original data to new uncorrelated variables. The dimensionality reduction comes from the property that the variances of these new variables are ordered, which serves as good descriptor of the most relevant directions and similarity among the data.

Consider the N-variate zero-mean random variable $X = \{X_1, .., X_N\} \in \mathcal{X}$ defined on the set[1] $\mathcal{X} \subseteq \mathbb{C}^N$ with PDF $f_X(\mathbf{x})$. Let $\mathbf{x}(i) \in \mathbb{C}^N$ be L independent and identically distributed (i.i.d.) samples from X, where $i = 0, ..., L-1$ and $\mathbf{x}(i) = [x_1(i), ..., x_N(i)]^T$. The data will be processed in the form of a data matrix $\mathbf{X} \in \mathbb{C}^{N\times L}$ composed by the column vectors $\mathbf{X} = [\mathbf{x}(0), ..., \mathbf{x}(L-1)]$. PCA seeks for the linear combinations $z = \mathbf{u}^H\mathbf{x}$ that maximize data variation, where $\mathbf{u} \in \mathbb{C}^N$. PCA is then expressed as:

$$\rho_{\text{PCA}} = \max_{\mathbf{u}} \frac{\mathbb{E}_{f_X}\left\{|z|^2\right\}}{\mathbf{u}^H\mathbf{u}} = \max_{\mathbf{u}} \frac{\mathbf{u}^H\mathbb{E}_{f_X}\left\{\mathbf{x}\mathbf{x}^H\right\}\mathbf{u}}{\mathbf{u}^H\mathbf{u}} = \max_{\mathbf{u}} \frac{\mathbf{u}^H\mathbf{R}_x\mathbf{u}}{\mathbf{u}^H\mathbf{u}}, \tag{2.25}$$

where $\mathbf{R}_x$ is the autocorrelation matrix and whose empirical estimate is

$$\hat{\mathbf{R}}_x = \frac{1}{L}\sum_{i=0}^{L-1}\mathbf{x}(i)\mathbf{x}^H(i) = \frac{1}{L}\mathbf{X}\mathbf{X}^H. \tag{2.26}$$

Note that, due to X being a zero-mean variable, the autocorrelation and autocovariance matrices coincide. In the opposite case, the mean value of each column must be subtracted, or the autocovariance matrix must be computed with

$$\hat{\mathbf{C}}_x = \frac{1}{L-1}\sum_{i=0}^{L-1}\left(\mathbf{x}(i) - \frac{1}{L}\sum_{j=0}^{L-1}\mathbf{x}(j)\right)\left(\mathbf{x}(i) - \frac{1}{L}\sum_{j=0}^{L-1}\mathbf{x}(j)\right)^H = \frac{1}{L-1}\mathbf{X}\mathbf{P}\mathbf{X}^H, \tag{2.27}$$

where $\mathbf{P}_{\mathbf{1}}^{\perp} = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}_N^T/L$ is the projection matrix onto the orthogonal space spanned by $\mathbf{1}_N$, which is a real symmetric $\mathbf{P}_{\mathbf{1}}^{\perp} = \left(\mathbf{P}_{\mathbf{1}}^{\perp}\right)^T$ and idempotent $\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{P}_{\mathbf{1}}^{\perp} = \mathbf{P}_{\mathbf{1}}^{\perp}$ matrix, and the $L-1$ term

[1] Complex-valued variables are defined in preparation for the next sections, where a nonlinear mapping to a complex space will be provided, hence the data will also be complex-valued.

is given by the Bessel correction. Then the empirical PCA just becomes

$$\hat{\rho}_{\text{PCA}} = \max_{\mathbf{u}} \frac{\mathbf{u}^H \hat{\mathbf{C}}_x \mathbf{u}}{\mathbf{u}^H \mathbf{u}}. \tag{2.28}$$

It is also worth noting that the denominator $\mathbf{u}^H\mathbf{u}$ constrains $\mathbf{u}$ to not fall into the trivial solution $\|\mathbf{u}\|_2^2 = \mathbf{u}^H\mathbf{u} \to \infty$. Otherwise, without this constraint, $\mathbf{u} \in \mathbb{C}^N$ is usually defined as a unit-length weight vector.

The previous maximization problem is actually related to the problem of finding the points of zero derivative of the Rayleigh quotient, endowing the analysis with physical interpretations [Bor98]. The Rayleigh quotient is defined as

$$\rho = \frac{\mathbf{u}^H \mathbf{A} \mathbf{u}}{\mathbf{u}^H \mathbf{B} \mathbf{u}}, \tag{2.29}$$

where $\mathbf{A}$ and $\mathbf{B}$ are Hermitian (symmetric for the real-valued case) matrices and $\mathbf{B}$ is positive semi-definite. It can be shown that the zero-derivative points of this quotient can be solved by means of the generalized eigenvalue problem [GKC19]

$$\mathbf{A}\mathbf{U} = \mathbf{B}\mathbf{U}\boldsymbol{\Lambda}, \tag{2.30}$$

where $\mathbf{U} \in \mathbb{C}^{N\times N}$ is a square unitary matrix whose columns are the eigenvectors $\mathbf{u}_n$ for $1 \leq n \leq N$, and $\boldsymbol{\Lambda} \in \mathbb{C}^{N\times N}$ is a diagonal matrix containing the eigenvalues λ_n associated to the eigenvectors $\mathbf{u}_n$. The eigenvalue decomposition is then

$$\mathbf{B}^{-1}\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}. \tag{2.31}$$

This perspective allows to unify multiple related algorithms to a common framework of solving the generalized eigenvalue problem [BLK97]. In the case of PCA, it is equivalent to fix $\mathbf{A} = \hat{\mathbf{C}}_x$ and $\mathbf{B} = \mathbf{I}_N$, where $\mathbf{U} \in \mathbb{C}^{N\times N}$ is composed by the column vectors $\mathbf{u}_n$. Also, since $\mathbf{B}^{-1}\mathbf{A} = \mathbf{I}_N\mathbf{A} = \mathbf{A}$ is a symmetric matrix, then we have $\mathbf{U}^{-1} = \mathbf{U}^H$. Note that, while in (2.28) only one vector $\mathbf{u}_1$ and the largest principal component are assessed, in (2.31) all the subsequent principal components with associated weight vectors $\mathbf{u}_n$, for $n = 2, .., N$, are also computed. Therefore, the empirical $\hat{\rho}_{\text{PCA}}$ corresponds to the largest eigenvalue λ_1 with eigenvector $\mathbf{u} = \mathbf{u}_1$. All the remaining principal components are ordered inside of matrix $\boldsymbol{\Lambda}$.

The estimated principal component can also be expressed as

$$\hat{\rho}_{\text{PCA}} = \|\hat{\mathbf{C}}_x\|_2, \tag{2.32}$$

where $\|\cdot\|_2$ denotes the spectral norm in the case of matrices, and corresponds to the largest singular value such that

$$\|\hat{\mathbf{C}}_x\|_2 = \sigma_1\left(\hat{\mathbf{C}}_x\right) = \sqrt{\lambda_1\left(\hat{\mathbf{C}}_x^H \hat{\mathbf{C}}_x\right)}. \tag{2.33}$$

In this case, given that $\hat{\mathbf{C}}_x$ is a Hermitian, thus normal, matrix with $\hat{\mathbf{C}}_x\hat{\mathbf{C}}_x^H = \hat{\mathbf{C}}_x^H\hat{\mathbf{C}}_x$, we can also express the largest singular value as $\sigma_1(\hat{\mathbf{C}}_x) = |\lambda_1(\hat{\mathbf{C}}_x)|$.

2.2.2 Canonical correlation analysis

Consider now two multivariate zero-mean random variables

$$X = \{X_1, .., X_{N_x}\} \in \mathcal{X}, \qquad Y = \left\{Y_1, .., Y_{N_y}\right\} \in \mathcal{Y}, \tag{2.34}$$

defined on the sets $\mathcal{X} \subseteq \mathbb{C}^{N_x}$ and $\mathcal{Y} \subseteq \mathbb{C}^{N_y}$, respectively, with joint PDF $f_{XY}\left(\mathbf{x}, \mathbf{y}\right)$, and marginal PDFs $f_X\left(\mathbf{x}\right)$ and $f_Y\left(\mathbf{y}\right)$. Let $\{\mathbf{x}\left(i\right), \mathbf{y}\left(i\right)\}$ be L i.i.d. observations with $i = 0, ..., L-1$. The corresponding data matrices are $\mathbf{X} \in \mathbb{C}^{N_x\times L}$ and $\mathbf{Y} \in \mathbb{C}^{N_y\times L}$. CCA [Hot36] finds the new variables, denoted as the canonical variables $z_x = \mathbf{u}_x^H\mathbf{x}$ and $z_y = \mathbf{u}_y^H\mathbf{y}$ for $\mathbf{u}_x \in \mathbb{C}^{N_x}$ and $\mathbf{u}_y \in \mathbb{C}^{N_y}$, that maximize the correlation coefficient ρ_{CCA}, called the canonical correlation.

Unlike PCA, CCA identifies linear combinations between two different variables while determining the directions in which their correlation is stronger. The greater canonical correlation is then obtained by solving the following maximization problem:

$$\rho_{\text{CCA}} = \max_{\mathbf{u}_x,\mathbf{u}_y} \frac{\mathbb{E}_{f_{X,Y}}\left\{z_x z_y^*\right\}}{\sqrt{\mathbb{E}_{f_X}\left\{|z_x|^2\right\}\mathbb{E}_{f_Y}\left\{|z_y|^2\right\}}} \tag{2.35a}$$

$$= \max_{\mathbf{u}_x,\mathbf{u}_y} \frac{\mathbf{u}_x^H\mathbb{E}_{f_{X,Y}}\left\{\mathbf{x}\mathbf{y}^H\right\}\mathbf{u}_y}{\sqrt{\left(\mathbf{u}_x^H\mathbb{E}_{f_X}\left\{\mathbf{x}\mathbf{x}^H\right\}\mathbf{u}_x\right)\left(\mathbf{u}_y^H\mathbb{E}_{f_Y}\left\{\mathbf{y}\mathbf{y}^H\right\}\mathbf{u}_y\right)}} \tag{2.35b}$$

$$= \max_{\mathbf{u}_x,\mathbf{v}} \frac{\mathbf{u}^H\mathbf{R}_{xy}\mathbf{u}_y}{\sqrt{\mathbf{u}_x^H\mathbf{R}_x\mathbf{u}_x}\sqrt{\mathbf{u}_y^H\mathbf{R}_y\mathbf{u}_y}}, \tag{2.35c}$$

where $\mathbf{R}_{xy}$ is the cross-correlation matrix, whose estimate is

$$\hat{\mathbf{R}}_{xy} = \frac{1}{L}\sum_{i=0}^{L-1}\mathbf{x}(i)\mathbf{y}^H(i) = \frac{1}{L}\mathbf{X}\mathbf{Y}^H, \tag{2.36}$$

and $\mathbf{R}_x$ and $\mathbf{R}_y$ are the autocorrelation matrices as in (2.26). Similarly as with PCA, if nonzero-mean random variables are used then covariance matrices are used instead, being the cross-correlation matrix

$$\hat{\mathbf{C}}_{xy} = \frac{1}{L-1}\sum_{i=0}^{L-1}\left(\mathbf{x}(i) - \frac{1}{L}\sum_{j=0}^{L-1}\mathbf{x}(j)\right)\left(\mathbf{y}(i) - \frac{1}{L}\sum_{j=0}^{L-1}\mathbf{y}(j)\right)^H = \frac{1}{L-1}\mathbf{X}\mathbf{P}\mathbf{Y}^H, \tag{2.37}$$

and the autocovariance matrices as in (2.27). CCA can also be solved through the generalized eigenvalue problem [BLK97]. This time, the matrices from (2.31) are

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}_{N_x\times N_x} & \hat{\mathbf{C}}_{xy} \\ \hat{\mathbf{C}}_{yx} & \mathbf{0}_{N_y\times N_y} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \hat{\mathbf{C}}_x & \mathbf{0}_{N_x\times N_y} \\ \mathbf{0}_{N_y\times N_x} & \hat{\mathbf{C}}_y \end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix} \mathbf{U}_x \\ \mathbf{U}_y \end{bmatrix}, \tag{2.38}$$

where $\mathbf{U}_x \in \mathbb{C}^{N_x\times N}$, $\mathbf{U}_y \in \mathbb{C}^{N_y\times N}$, and $N = \min\{N_x, N_y\}$.

The eigenvalue decomposition can then be written by two different equations

$$\begin{cases} \hat{\mathbf{C}}_x^{-1}\hat{\mathbf{C}}_{xy}\hat{\mathbf{C}}_y^{-1}\hat{\mathbf{C}}_{yx}\mathbf{U}_x = \mathbf{\Lambda}_x^2\mathbf{U}_x \\ \hat{\mathbf{C}}_y^{-1}\hat{\mathbf{C}}_{yx}\hat{\mathbf{C}}_x^{-1}\hat{\mathbf{C}}_{xy}\mathbf{U}_y = \mathbf{\Lambda}_y^2\mathbf{U}_y \end{cases}. \tag{2.39}$$

In this case, $\mathbf{\Lambda}_x^2 \in \mathbb{R}^{N_x\times N_x}$ and $\mathbf{\Lambda}_y^2 \in \mathbb{R}^{N_y\times N_y}$ contain the same squared canonical correlations (up to the N-th, the rest are zero-valued), and the largest corresponds to $\hat{\rho}_{\text{CCA}}^2$. Equivalently, one can perform the Singular Value Decomposition (SVD) $\hat{\mathbf{C}} = \mathbf{F}\mathbf{\Sigma}\mathbf{G}^H$, or $\mathbf{\Sigma} = \mathbf{F}^H\hat{\mathbf{C}}\mathbf{G}$, where

$$\hat{\mathbf{C}} = \hat{\mathbf{C}}_x^{-1/2}\hat{\mathbf{C}}_{xy}\left(\hat{\mathbf{C}}_y^{-1/2}\right)^H \tag{2.40}$$

is called the empirical *coherence* matrix [SM00], and the diagonal matrix $\mathbf{\Sigma}$ contains all the canonical correlations. The empirical maximum canonical correlation can then be expressed as

$$\hat{\rho}_{\text{CCA}} = \|\hat{\mathbf{C}}\|_2. \tag{2.41}$$

While CCA is usually portrayed as a measure of the maximum correlation, the assessment of all canonical correlations (contained in $\mathbf{\Lambda}_x$ or $\mathbf{\Lambda}_y$) have other multiple applications, e.g. in model-order selection [RS15]. A similar scenario was studied in our conference paper [LCR20], where the coherence matrix is used to determine the mutual information (due to Gaussian assumption, CCA and information are closely related) by detecting the number of correlated signals in a channel.

A common problem with CCA is that its performance greatly depends on the available number of samples. The measurement of the canonical correlations is prone to errors if the variables are of high dimensionality but reduced sample size [Bao+19]. This behavior follows from the requirement of estimating sample covariance matrices, which may also be flawed if the sample size is small. In order to overcome this problem, some algorithms have been designed that minimize the overall dimensionality of the involved matrices [Pez+04; AN17]. This issue is relevant because, as we employ the CCA to estimate information, the dimensionality of the data will be translated to the dimensionality of some feature space. Therefore, the choice of the dimensionality of such feature space will become an important issue to monitor.

2.3 Measures of information

Once measures of linear dependence have been reviewed, we then provide a review of information-theoretic measures that are pivotal within the objective of measuring information through classical signal processing techniques based on second-order statistics. Generally speaking, these measures originate from Shannon's entropy, the KL divergence and MI. While these surrogates succeed on providing an assessment of the quantity of uncertainty, dissimilarity and dependence, respectively, their properties may or not may be different from their original measures. These surrogates share a common property however, which is the capability of being expressed as the squared Euclidean norm of some functional based on probability distributions. For this reason, we refer to these surrogates as "quadratic measures of information". Although the focus of this thesis is on continuous sources, the discrete case will also be addressed as a guiding line to better understand the origin, properties and uses. Lastly, the following measures of information will be based on univariate random variables, given that the generalizations to multivariate random variables can be easily done if required.

2.3.1 Uncertainty measures

Consider the discrete random variable X with Probability Mass Function (PMF) $p_X(x)$ defined on the set $\mathcal{X}$. Shannon's entropy (in *nats*) is defined as

$$H(X) = -\sum_{\mathcal{X}} p_X(x) \ln p_X(x) = -\mathbb{E}_{p_X}\{\ln p_X(x)\}. \tag{2.42}$$

This measure was born from Shannon in the pursuit of quantifying the average information of the outcome of a given random variable [Sha48]. Shannon's entropy gives a measure of randomness of a random variable, which is maximized by the uniform distribution. The importance of this measure is well-known in the fields of communications and information theory, and its use has since then escalated tremendously by being used in very disparate fields (the reader is referred to [WKV09; CT06], and references therein, for a review of areas and applications where entropy arises in a natural manner).

In an attempt to define the uncertainty of a continuous random variable, Shannon proposed the so called *differential* entropy. Given a continuous random variable $X \in \mathcal{X}$ with PDF $f_X(x)$, it is defined as follows:

$$h(X) = -\int_{\mathcal{X}} f_X(x) \ln f_X(x)\,\mathrm{d}x = -\mathbb{E}_{f_X}\{\ln f_X(x)\}. \tag{2.43}$$

In contrast to its discrete counterpart, this definition of entropy can attain negative values, which may be counter-intuitive for the purpose of quantifying information. This is due to the fact that differential entropy is not the limit case of Shannon's entropy [CT06, Sec. 8.3]. Nevertheless, outside the basic definition of uncertainty, differential entropy can still be used in other settings in which a negative outcome does not, by itself, denote an improper generalization from the discrete case (although it is).

In terms of a random variable, entropy is sensitive to the probability of the outcomes and not to their values, which would be the case of the mean or the variance. Thus, the capability of entropy for describing a random variable differs from that of the classic second-order statistics. As such, entropy may be a better data descriptor that measures local particularities of the PDF, while variance measures its global spread. This idea will be particularly addressed on Chapter 5.

Regarding the scope of this thesis, these measures of information carry some inherent problems to is own definition. First, the estimation of differential entropy from a set of observations is not an easy task. Although there are numerous estimators in the literature with varying grade of success (see [WKV09] for a review of estimators and their applications), most of them suffer from a great computational complexity for big data-sets, especially for nonparametric estimators. Secondly, the logarithm involved in its definition can be bothersome for some distributions. Generally speaking, to cope with the estimation of differential entropy, most estimators rely on plug-in methods. These follow the methodology that first estimates $f_X(x)$ and then they plug it into the desired functional. This means that, if a given density is not easily estimated, e.g. distributions with long tails, then this approach may lead to bigger errors. A potential solution can be encountered in the method of *importance sampling* [TK10], where the expectation (as the one given in (2.43)) is estimated by a weighted average of i.i.d. observations. However, this approach concerns a target distribution, from which the expectation is computed, while we are interested in developing an universal approach, capable of estimating the entropy given any underlying distribution. A more straightforward approach is to alleviate the estimation error by increasing the sample size, but this would mean to increase further the computational complexity. For these reasons, it may be interesting to study other forms of entropy that may convey a better procedure of estimation, and that also describe the quantity of information given by the outcome of a random variable.

In light of the work of Shannon, there was an upraise in the pursuit of knowledge regarding the meaning, use and significance of these new measures of information. In 1961, a mathematician called Alfréd Rényi proposed a generalization of Shannon's entropy. In [Rén61] he introduced a class of parameterized entropies that preserved most of its axioms, but relaxing the strong additivity axiom with a weaker postulate (additivity) [Csi08; ABH18]. Concretely, the additivity property states that the joint entropy of various independent random variables is equal to the sum of the entropies of these random variables. It does not, however, hold any inequality as with Shannon's entropy.

In the case of a discrete random variable X, it is defined as follows:

$$H_\alpha(X) = \frac{1}{1-\alpha} \ln \sum_{\mathcal{X}} p_X^\alpha(x) = \frac{1}{1-\alpha} \ln \mathbb{E}_{p_X}\left\{p_X^{\alpha-1}(x)\right\}. \tag{2.44}$$

We will refer to the parameter $\alpha \in \mathbb{R}^+ \backslash \{1\}$ as the entropic index, and it is used to determine the order of Rényi's entropy. While this parameter opens the possibility for infinite representations of Rényi's entropy, generally speaking the most used and studied values are 0,1,2 and ∞. For $\alpha \to 1$ this measure adopts the special case of being defined only by the limit, which can be shown that is indeed Shannon's entropy through Hôpital's rule [Rén61]:

$$\lim_{\alpha \to 1} H_\alpha(X) = H(X). \tag{2.45}$$

The implications are that, not only does Rényi's entropy provide an alternative measure of randomness, but it also generalizes Shannon's entropy. Regarding the other cases, these are enumerated below:

1. *Max / Hartley entropy*:

$$H_0 = \ln |\mathcal{X}|, \tag{2.46}$$

where $|\cdot|$ indicates the cardinality of the set, and assuming that all probabilities are nonzero.

2. *Collision entropy*:
$$H_2\left(X\right) = -\ln \left\|p_X\left(x\right)\right\|_2^2, \tag{2.47}$$
where $\left\|\mathbf{a}\right\|_p = \left(\sum_{i=1}^n \left|a_i\right|^p\right)^{1/p}$ is the p-norm of the vector $\mathbf{a}$ for $p \in [1, \infty)$.

3. *Min entropy*:
$$H_\infty\left(X\right) = \min_i \left\{-\ln p_X\left(x_i\right)\right\}, \tag{2.48}$$
which is another case of convergence for $\alpha \to \infty$ in order to preserve the continuity.

Furthermore, it is know that the Rényi entropies are monotonically decreasing with respect to the entropic index. In particular, $H_\alpha\left(X\right)$ satisfies

$$H_\alpha\left(X\right) \geq H_\beta\left(X\right) \quad s.t.\, \alpha \leq \beta. \tag{2.49}$$

Therefore, we can establish a clear relation between the particular cases:

$$H_0\left(X\right) \geq H\left(X\right) \geq H_2\left(X\right) \geq ... \geq H_\infty\left(X\right), \tag{2.50}$$

where equality holds for uniform distributions.

From these, we are particularly interested in the case $\alpha = 2$, given that it provides a measure of information based on the Euclidean norm of a given PMF. The name, collision entropy, indicates that this particular surrogate measures the probability of two of the same sample coinciding. However, this notion will be lost when moving to the continuous case. It is also worth noting that this surrogate can be naturally obtained by applying Jensen's inequality thanks to the concavity of the natural logarithm function:

$$H\left(X\right) = -\mathbb{E}_{p_X}\left\{\ln p_X\left(x\right)\right\} \geq -\ln \mathbb{E}_{p_X}\left\{p_X\left(x\right)\right\} = H_2\left(X\right). \tag{2.51}$$

While being conceived as a generalization of Shannon's entropy, the family of Rényi entropies has found applications on a very varied scientific fields. Some examples are encountered in statistics [BR78], coding [Cam65; Csi95], cryptography [Skó15], thermodynamics [Bae22], or physics [Don16]. However, Rényi's entropy suffers for some of its limitations due to the relaxation of its properties. For instance, the definition of conditional Rényi's entropy is not so well-defined as the conditional entropy described by Shannon [FB14], which translates to a challenging transition to the Rényi mutual information [Csi95; LP19a]. The definition used in this thesis will be addressed further on, when describing the surrogates of dependence measures in Subsection 2.3.3.

In the case of continuous variables, Rényi's entropy is generalized in a similar fashion to that of differential entropy:

$$h_\alpha\left(X\right) = \frac{1}{1-\alpha} \ln \int_{\mathcal{X}} f_X^\alpha\left(x\right) \mathrm{d}x = \frac{1}{1-\alpha} \ln \mathbb{E}_{f_X}\left\{f_X^{\alpha-1}\left(x\right)\right\}. \tag{2.52}$$

Similarly to the bridge between Shannon's and differential entropies, the differential Rényi entropy is not the limit case of Rényi's entropy [TBA16], and can also be negative.

The differential Rényi entropy of order 2, or directly the second-order Rényi entropy, is then

$$h_2\left(X\right) = -\ln \int_{\mathcal{X}} f_X^2\left(x\right) \mathrm{d}x = -\ln \mathbb{E}_{f_X}\left\{f_X\left(x\right)\right\}. \tag{2.53}$$

Alternatively, one can generalize the p-norm to a function space L^p and express the second-order Rényi entropy as

$$h_2\left(X\right) = -\ln \left\|f_X\left(x\right)\right\|_2^2, \tag{2.54}$$

since the PDF, and its p-th power, are Lebesgue integrable with

$$\left\|f_X\left(x\right)\right\|_p = \left(\int_{\mathcal{X}} \left|f_X\left(x\right)\right|^p \mathrm{d}x\right)^{1/p}. \tag{2.55}$$

As a result, the second-order Rényi entropy can be seen as analogous to the second-order moment. However, the expectation of the PDF is evaluated instead of the possible outcomes of the random variable. This perspective allows us to move from the classical view of second-order statistics, to the second-order information measures. Moreover, note that no logarithm is involved in the measure itself, since it can be added a posteriori, unlike with Shannon's entropy. This property suggests that it may be less complex to estimate this surrogate of entropy rather than Shannon's entropy. As a matter of fact, this property has been explored for some authors in the same pursuit of estimating information. In [Prí10, Sec. 2.3] an estimator is derived by taking advantage of this attribute, which will be addressed in Section 3.3 due to its close relation with kernel methods. An equivalent result is observed in [OUE08] by blending the kernel density estimate with the sample-spacing method.

Finally, let us briefly address other families of entropy. In particular, the Tsallis entropy is a widely spread measure in the field of statistical physics, born from the work of Tsallis to generalize the Boltzmann-Gibbs entropy [Tsa88]. However, here we will focus on the information-theoretic aspect of it, that is, that entropy is a function of probability distributions. For a discrete random variable X, Tsallis entropy is defined as

$$S_\alpha\left(X\right) = \frac{1}{1-\alpha}\left(\sum_{\mathcal{X}} p_X^\alpha\left(x\right) - 1\right) = \frac{1}{1-\alpha}\left(\mathbb{E}_{p_X}\left\{p_X^{\alpha-1}\left(x\right)\right\} - 1\right), \tag{2.56}$$

whereas in the continuous case it is

$$s_\alpha\left(X\right) = \frac{1}{1-\alpha}\left(\int_{\mathcal{X}} f_X^\alpha\left(x\right)\mathrm{d}x - 1\right) = \frac{1}{1-\alpha}\left(\mathbb{E}_{f_X}\left\{f_X^{\alpha-1}\left(x\right)\right\} - 1\right). \tag{2.57}$$

The Tsallis entropy exhibits similar properties to Rényi's entropy given its close proximity. In fact, one can define Rényi's entropy from Tsallis entropy with

$$h_\alpha\left(X\right) = \frac{1}{1-\alpha}\ln\left(1 + \left(1-\alpha\right)s_\alpha\left(X\right)\right). \tag{2.58}$$

Similar to Rényi's entropy, the Tsallis entropy also tends to Shannon's entropy for $\alpha \to 1$. Despite this, Tsallis entropy is not a common measure encountered in information theory problems. This is due to the lack of logarithm in its definition, since it is the logarithm function that allows the quantification of information (in bits, nats, etc). Although for $\alpha = 2$ it presents the same properties than the second-order Rényi entropy, such as the capability of being expressed as a second-order statistic (note that they share the integral of the α-th power of the PDF), we will not focus our attention with this entropy measure. However, it is relevant to note that by estimating (or employing) Rényi's entropy, one can always recover the Tsallis entropy, if required per application.

2.3.2 Dissimilarity measures

Next, we move to divergences. Jointly with entropy, these are considered the fundamental measures of information, being the MI a particular case. The procedure to identify the surrogate will follow a similar structure to those of the uncertainty measures.

Consider now two PMFs $p_X\left(x\right)$ and $p_Y\left(x\right)$ defined on the same set $\mathcal{X}$. The KL divergence [KL51] is given by:

$$D\left(p_X \parallel p_Y\right) = \sum_{\mathcal{X}} p_X\left(x\right)\ln\frac{p_X\left(x\right)}{p_Y\left(x\right)} = \mathbb{E}_{p_X}\left\{\ln\frac{p_X\left(x\right)}{p_Y\left(x\right)}\right\}. \tag{2.59}$$

The KL divergence is nonegative, and zero if and only if $p_X\left(x\right) = p_Y\left(x\right)$. It is also well-known that is not a proper distance metric, given that it is not symmetric with respect to its arguments. For this reason, it may be more appropriate to categorize the divergences as dissimilarity measures, which increase as the disparity increases. To avoid limit cases and inconsistencies with its definition, the convention is to assume $0 \ln 0/0 = 0$, $0 \ln 0/a = 0$ and $a \ln a/0 = \infty$, such that $D\left(p_X \parallel p_Y\right) \geq 0$.

Regarding continuous distributions, consider $f_X\left(x\right)$ and $f_Y\left(x\right)$, and assume absolute continuity with respect to each other. The KL divergence is then

$$D\left(f_X \parallel f_Y\right) = \int_{\mathcal{X}} f_X\left(x\right) \ln\left(\frac{f_X\left(x\right)}{f_Y\left(x\right)}\right) \mathrm{d}x = \mathbb{E}_{f_X}\left\{\ln\left(\frac{f_X\left(x\right)}{f_Y\left(x\right)}\right)\right\}. \tag{2.60}$$

The translation to continuous random variables is well-defined in the case of the KL divergence, given that its properties are preserved. Notoriously, $D\left(f_X \parallel f_Y\right)$ is also nonnegative for any $f_X\left(x\right)$ and $f_Y\left(x\right)$, unlike with the leap between Shannon's and differential entropy.

The KL divergence belongs to the class of f-divergences, which are a set of dissimilarity measures between probability distributions. The divergences that pertain in this class satisfy a set of common properties, notoriously the nonnegativity and the data processing inequality [LV06]. The most commonly used divergences and relationships among them can be found in [GS02; GSS14; SV16]. From these we are particularly interested in the Rényi divergence, D_α, and the χ^2-divergence, D_{χ^2}, since their definition is akin to a measure based on second-order statistics.

Similarly to the case of Shannon's entropy, the Rényi divergence [Rén61] generalizes the notion of the KL divergence with

$$D_\alpha\left(f_X \parallel f_Y\right) = \frac{1}{\alpha-1} \ln \int_{\mathcal{X}} f_X^\alpha\left(x\right) f_Y^{1-\alpha}\left(x\right) \mathrm{d}x = \frac{1}{\alpha-1} \ln \mathbb{E}_{f_X}\left\{\left(\frac{f_X\left(x\right)}{f_Y\left(x\right)}\right)^{\alpha-1}\right\}, \tag{2.61}$$

where for $\alpha \to 1$ this measure is asymptotically the KL divergence. For an extensive review of its operational characterization, the reader is referred to [EH14]. Also note that D_α has been directly expressed for continuous distributions. Unlike with entropy measures, where some properties are lost in the transition from variables with finite support to an infinite one, the case of Rényi divergence does not need any specific clarification, and the extension from finite spaces is the same for any value of α.

Following the identification of an appropriate surrogate in the case of entropy, we will directly focus our attention to the case of $\alpha = 2$. However, note that many other values of α resolve into well-known divergences in the literature. For example, for $\alpha = 1/2$ the resulting divergence is a function of both Bhattacharyya and Hellinger distances [Bha46; GS02].

For $\alpha = 2$, then (2.61) resolves into

$$D_2\left(f_X \parallel f_Y\right) = \ln \int_{\mathcal{X}} \frac{f_X^2\left(x\right)}{f_Y\left(x\right)} \mathrm{d}x = \ln \mathbb{E}_{f_X}\left\{\frac{f_X\left(x\right)}{f_Y\left(x\right)}\right\}, \tag{2.62}$$

which we refer to as the second-order Rényi divergence. Again, this surrogate can be immediately obtained by applying Jensen's inequality:

$$D\left(f_X \parallel f_Y\right) = \mathbb{E}_{f_X}\left\{\ln \frac{f_X\left(x\right)}{f_Y\left(x\right)}\right\} \leq \ln \mathbb{E}_{f_X}\left\{\frac{f_X\left(x\right)}{f_Y\left(x\right)}\right\} = D_2\left(f_X \parallel f_Y\right), \tag{2.63}$$

where equality is hold if and only if $p_X\left(x\right) = p_Y\left(x\right)$, which denotes 0 divergence. Otherwise, the strict inequality is held by the strict concavity of the natural logarithm.

The case of second-order Rényi divergence is particularly interesting since it shares similar properties to that of the second-order Rényi entropy. In particular, one can express this divergence as a function of an Euclidean norm such that

$$D_2\left(f_X \parallel f_Y\right) = \ln \left\| \frac{f_X\left(x\right)}{\sqrt{f_Y\left(x\right)}} \right\|_2^2, \tag{2.64}$$

showing that the capability of being expressed as a second-order moment of Rényi information measures is shared among them for $\alpha = 2$.

Alternatively, the second-order Rényi entropy can also be expressed as follows:

$$D_2\left(f_X \parallel f_Y\right) = \ln\left(1 + D_{\chi^2}\left(f_X \parallel f_Y\right)\right), \tag{2.65}$$

where $D_{\chi^2}\left(f_X \parallel f_Y\right)$ denotes the Pearson χ^2-divergence

$$D_{\chi^2}\left(f_X \parallel f_Y\right) = \int_{\mathcal{X}} \frac{f_X^2\left(x\right)}{f_Y\left(x\right)} \mathrm{d}x - 1 = \mathbb{E}_{f_X}\left\{\frac{f_X\left(x\right)}{f_Y\left(x\right)}\right\} - 1 = \left\|\frac{f_X\left(x\right)}{\sqrt{f_Y\left(x\right)}}\right\|_2^2 - 1. \tag{2.66}$$

By doing some manipulations (see Appendix 7.1.2), one can also express (2.66) in its more well-known form:

$$D_{\chi^2}\left(f_X \parallel f_Y\right) = \int_{\mathcal{X}} \frac{\left(f_X\left(x\right) - f_Y\left(x\right)\right)^2}{f_Y\left(x\right)} \mathrm{d}x = \mathbb{E}_{f_X}\left\{\frac{\left(f_X\left(x\right) - f_Y\left(x\right)\right)^2}{f_X\left(x\right) f_Y\left(x\right)}\right\} = \left\|\frac{f_X\left(x\right) - f_Y\left(x\right)}{\sqrt{f_Y\left(x\right)}}\right\|_2^2. \tag{2.67}$$

From the fundamental logarithm and Jensen's inequalities ($\ln\left(1+x\right) \leq x$), we can state the following bounds:

$$D\left(f_X \parallel f_Y\right) \leq D_2\left(f_X \parallel f_Y\right) \leq D_{\chi^2}\left(f_X \parallel f_Y\right). \tag{2.68}$$

The equality holds for $f_X\left(x\right) = f_Y\left(x\right)$, otherwise the inequality is strict for nonzero divergence. The relationship between the χ^2 and KL divergences can also be tackled by assuming closeness between the two probability functions. To be concrete, as the divergence measures approximate to zero, we can consider $f_Y\left(x\right) = f_X\left(x\right) + \epsilon\Delta\left(x\right)$ where ϵ is an arbitrarily small number and $\Delta\left(x\right)$ is a function with null area. Then, we can write

$$D\left(f_X \| f_X + \epsilon\Delta\right) = \frac{1}{2} D_{\chi^2}\left(f_X \| f_X + \epsilon\Delta\right) + O(\epsilon^3), \tag{2.69}$$

as can be seen in Appendix 7.1.3. As a result, one can see the χ^2-divergence as a local approximation of the KL divergence for close distributions. Moreover, the analysis shows that $\frac{1}{2}D_{\chi^2}$ is not a bound, but an approximation. The rationale behind this local approximation can also be encountered in [HSZ15], where their interest relies on translating information theory problems to linear algebra problems, a similar philosophy than the one of this thesis. Likewise, the relation between MI and SMI under local approximations has also been expressed in [Hua+19, Eq. (61)] (and corresponding footnote), although more focused on providing an insightful measure of local information geometry.

The χ^2-divergence is a recurrent measure of dissimilarity in the literature. Its use can be traced back to Pearson, in the study of associations among categorical data [Pea04], where it was referred to as the mean-square contingency. In the past decades, the χ^2-test has been particularly used for the goodness of fit method for parameter estimation [CL54]. The relationship between D_{χ^2} and D_2 is conveyed in a similar fashion to that of between the Tsallis entropy s_α and Rényi's entropy h_α. However, while in the case of s_α the lack of logarithm was detrimental in the sense of measuring uncertainty, here the dissimilarity measure does not lose its meaning even without the logarithm. Nonetheless, D_{χ^2} looses the additivity property with respect to independent (i.e. multiplicative) components in $f_X\left(x\right)$ or $f_Y\left(x\right)$, which is a property satisfied by both KL and Rényi divergences.

The particularly expression provided in (2.67) is also of interest since it differs from the ones in both (2.64) and (2.66), thus providing an alternative measure to a similar approach. As usual, we are particularly interested in information measures that convey a similar structure to that of second-order statistics. Hence, while both (2.66) and (2.67) satisfy this condition, the latter will be unveiled to portray a very specific structure in the case of measuring dependence.

2.3.3 Dependence measures

Finally, we will briefly address the special case of measuring dependence. Since the measure of dependence is strongly tied to the measure of dissimilarity, the properties and analysis are identical to the ones in previous subsection. In short, one just needs to measure the dissimilarity between the joint probability distribution and the product of marginal ones. However, although dependence can be seen as a particular case, it is still relevant due to its well-known interpretation and applications.

We first begin with the KL divergence. Consider now the joint PDF $f_{X,Y}(x,y)$ defined on the set $\mathcal{X}\times\mathcal{Y}$, and the product of marginal distributions $f_X(x)\,f_Y(y)$ also defined on $\mathcal{X}\times\mathcal{Y}$. The KL divergence between these distributions is as follows

$$D\left(f_{X,Y}\parallel f_Xf_Y\right)=\int_{\mathcal{X}\times\mathcal{Y}}f_{XY}(x,y)\ln\frac{f_{X,Y}(x,y)}{f_X(x)\,f_Y(y)}\mathrm{d}x\mathrm{d}y=\mathbb{E}_{f_{X,Y}}\left\{\ln\left(\frac{f_{X,Y}(x,y)}{f_X(x)\,f_Y(y)}\right)\right\}, \tag{2.70}$$

which is classically referred to as the MI with notation $I(X;Y)$. This measure is nonnegative with $I(X;Y)\geq 0$ and symmetric with respect to its arguments $I(X;Y)=I(Y,X)$.

This measure was born from Shannon in the same mathematical analysis in which Shannon's entropy was defined [Sha48]. In its classical paper, $I(X;Y)$ determines the theoretical limits of the capacity of the channel between the transmitter, associated with the random variable X, and the receiver, associated with the random variable Y. Although the characterization of the MI in the field of communications runs much more deeper, in terms of this thesis we are just interested in its information-theoretic interpretation as a functional of probability distributions. In this sense, our interest is generalized for many other applications with a similar point of view.

Outside of the theory of communications, the MI is "simply" regarded as a general descriptor of the relationship between random variables. Statistics and machine learning are two fields where the MI has found a variety of applications. Particularly, it can be encountered in feature selection [GE03], independent component analysis [HO00], and in the study of neural networks [TZ15], among others. In these applications, the most straightforward description of $I(X;Y)$ is that it "tests" for independence, that is, if $I(X;Y)=0$, then X is independent to Y due to $f_{XY}(x,y)=f_X(x)\,f_Y(y)$ in the independent case. That is why, in the case of its estimation, it is usually referred to as an independence measure, and not the opposite. Another way to interpret this measure is to trace en equivalence to correlation. Both measures indicate a relationship between random variables with a given degree of dependence. However, correlation only determines linear relationships, while MI is generally regarded as a nonlinear descriptor of statistical dependence. This notion will become pivotal in the development of strategies for estimating information.

Similarly to the derivation of the MI from the KL divergence, one can particularize the Rényi divergence with

$$I_\alpha(X;Y)=D_\alpha\left(f_{X,Y}\parallel f_Xf_Y\right)=\frac{1}{\alpha-1}\ln\int_{\mathcal{X}\times\mathcal{Y}}\frac{f^\alpha_{X,Y}(x,y)}{\left(f_X(x)\,f_Y(y)\right)^{\alpha-1}}\mathrm{d}x\mathrm{d}y \tag{2.71a}$$

$$=\frac{1}{\alpha-1}\ln\mathbb{E}_{f_{X,Y}}\left\{\left(\frac{f_{X,Y}(x)}{f_X(x)\,f_Y(y)}\right)^{\alpha-1}\right\}, \tag{2.71b}$$

where $I_\alpha(X;Y)$ is called the Rényi α-information. It is important to note that other definitions may be given to the Rényi α-information [Ver15; LP19a]. The most relevant are the ones given by Arimoto [Ari77], Csiszár [Csi95] and Sibson [Sib69], where their interest mostly relied on the definition of Rényi α-information through the conditional Rényi entropy of a discrete channel. Instead, the definition from (2.71) is directly derived from D_α, as it appears in [Rén07] , and is in agreement with [TZI15]. Either way, the properties of (2.71) are still shared with the ones of (2.61), and it is asymptotically equivalent to the MI for $\alpha\to 1$.

For $\alpha=2$ we have the second-order Rényi information:

$$I_2(X;Y)=\ln\int_{\mathcal{X}\times\mathcal{Y}}\frac{f^2_{X,Y}(x,y)}{f_X(x)\,f_Y(y)}\mathrm{d}x\mathrm{d}y=\ln\mathbb{E}_{f_{X,Y}}\left\{\frac{f_{X,Y}(x,y)}{f_X(x)\,f_Y(y)}\right\} \tag{2.72a}$$

$$= \ln \left\| \frac{f_{X,Y}(x,y)}{\sqrt{f_X(x)\, f_Y(x)}} \right\|_2^2 . \tag{2.72b}$$

Following the same steps as in (2.65), we can express

$$I_2(X;Y) = \ln\left(1 + I_{\mathrm{s}}(X;Y)\right), \tag{2.73}$$

where

$$I_{\mathrm{s}}(X;Y) = D_{\chi^2}\left(f_{X,Y} \parallel f_X f_Y\right) = \mathbb{E}_{f_{X,Y}}\left\{\frac{f_{X,Y}(x,y)}{f_X(x)\, f_Y(y)}\right\} = \left\| \frac{f_{X,Y}(x,y)}{\sqrt{f_X(x)\, f_Y(y)}} \right\|_2^2 - 1 \tag{2.74a}$$

$$= \mathbb{E}_{f_{X,Y}}\left\{\frac{\left(f_{X,Y}(x,y) - f_X(x)\, f_Y(y)\right)^2}{f_{X,Y}(x,y)\, f_X(x)\, f_Y(y)}\right\} = \left\| \frac{f_{X,Y}(x,y) - f_X(x)\, f_Y(y)}{\sqrt{f_X(x)\, f_Y(y)}} \right\|_2^2 - 1. \tag{2.74b}$$

The dependence measure $I_{\mathrm{s}}(X;Y)$ will be called the Squared-loss mutual information (SMI), coined in [Suz+09] for the purpose of feature selection. However, other authors have referred to this measure as the mean-square contingency [Hir35], a term characterized by Pearson in [Pea04], and also studied by Rényi as a measure of dependence in [Rén59]. It should also be noted that in [Suz+09] the term refers to half of the magnitude of (2.74). The reason behind this constant is to approximate $I_{\mathrm{s}}(X;Y)$ to $I(X;Y)$ in the low dependence regime (as it follows from (2.69)). However, here we will strictly define the SMI as it resolves from the χ^2-divergence.

Following the bound from the dissimilarity measures in (2.68), the measures of dependence follow the same rules with

$$I(X;Y) \le I_2(X;Y) \le I_{\mathrm{s}}(X;Y). \tag{2.75}$$

In this case, the inequalities are strict for dependent random variables, and the equality is to zero for independent random variables with $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$. It also follows from (2.69) that, for small values of dependence with $f_X(x) f_Y(y) = f_{X,Y}(x,y) + \epsilon\Delta(x,y)$, where ϵ is an arbitrarily small value and $\Delta(x,y)$ is defined on the set $\mathcal{X} \times \mathcal{Y}$ and is constrained to have null area, we can then express

$$I(X;Y) = \frac{1}{2} I_{\mathrm{s}}(X;Y) + O(\epsilon^3). \tag{2.76}$$

A similar observation can be encountered in [GN18] on the topic of co-clustering contingency tables, and in [Hua+19] in the study of local measures of information from the perspective of information geometry. In order to illustrate this approximation, Figure 2.1 shows the closeness for the small dependence scenario and univariate random variables. The distributions shown are the Gaussian, the Pareto with location parameter $\theta = 1$ for both marginal distributions [DV00], the Student's t-distribution with $\nu = 10$ degrees of freedom [KN04] and a Gaussian Mixture Model (GMM) distributed as $(X,Y) \sim \mathcal{N}(0, \mathbf{\Sigma}_\rho)/2 + \mathcal{N}(0, \mathbf{\Sigma}_{-\rho})/2$, and the marginal random variables distributed as $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$, where $\mathbf{\Sigma}_a = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$ and $\rho \in [0,1)$, as it is encountered in [Res+11] under the name of "X" model. Note that the Gaussian variable has MI $I(X;Y) = -0.5\ln\left(1-\rho^2\right)$ and SMI $I_{\mathrm{s}}(X;Y) = \rho^2/\left(1-\rho^2\right)$, which are equal up to the first order approximation $f(\rho) \approx f(0) + f'(0)\rho$ with

$$I(X;Y) = -0.5\ln\left(1-\rho^2\right) \approx \frac{\rho^2}{1-\rho^2} = I_{\mathrm{s}}(X;Y) \tag{2.77}$$

given that $\frac{d}{d\rho}\ln\left(1-\rho^2\right) = \frac{-2\rho}{1-\rho^2}$. Due to this, the Gaussian is more easily approximated and its curve approximates the equality faster. Unfortunately, the other distributions are not so easily computed (either the MI or the SMI). Its relation, however, seems to be of a higher-order approximation since they are further from the equality line.

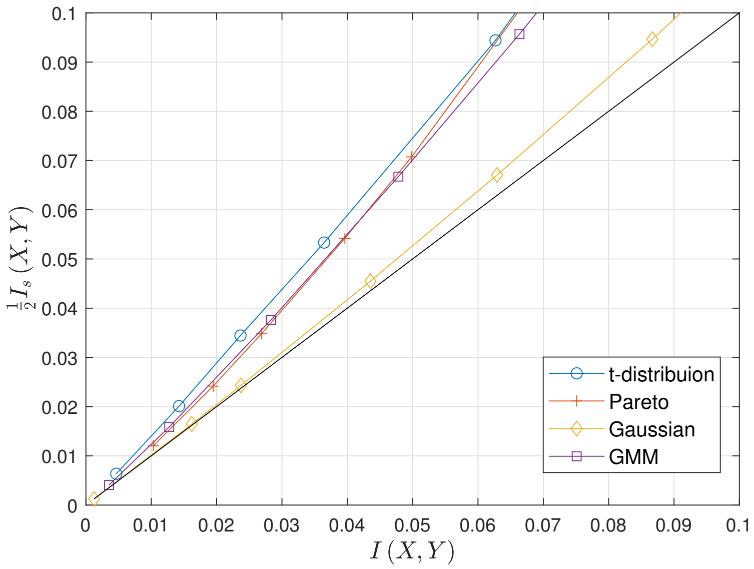


Figure 2.1: Half of the SMI versus the MI for different distributions.

2.3.4 Other dependence measures

Apart from the classical derivations from the KL divergence and the MI, there are other measures of dependence that may prove to be of interest during the course of the thesis. Generally speaking, the rest of this subsection addresses *in*dependence measures rather than measuring dependence from the point of view of information. For this reason, the empirical estimation of the following measures usually serves as independence detectors instead of conveying a tangible measure, e.g. the channel capacity in the case of the MI. However, the following measures do convey, in a way or another, a connection with measures from the previous subsections, although it may not be apparent.

2.3.4.1 Hirschfeld-Gebelein-Rényi maximal correlation coefficient

The Hirschfeld-Gebelein-Rényi (HGR) coefficient is a generalization of the well-known Pearson coefficient. In particular, the HGR coefficient determines the maximum correlation coefficient between two random variables after a given mapping, which is usually nonlinear. Let X and Y be two random variables with PDFs $f_X(x)$ and $f_Y(y)$ defined on the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, with joint probability distribution $f_{X,Y}(x,y)$ defined on the set $\mathcal{X}\times\mathcal{Y}$ and let $g_X:\mathcal{X}\to\mathbb{R}$ and $g_Y:\mathcal{Y}\to\mathbb{R}$ be Borel-measurable functions. The HGR coefficient [Hir35; Rén59] is then

$$\rho_{\text{HGR}}(X;Y)=\sup_{g_X,g_Y}\mathbb{E}_{f_{X,Y}}\{g_X(X)\,g_Y(Y)\}\ \text{ s.t.}\begin{cases}\mathbb{E}_{f_X}\{g_X(X)\}=\mathbb{E}_{f_Y}\{g_Y(Y)\}=0\\ \mathbb{E}_{f_X}\{g_X^2(X)\}=\mathbb{E}_{f_Y}\{g_Y^2(Y)\}=1\end{cases}\tag{2.78}$$

Due to the constraints, the expectation is equivalent to measure the correlation between $g_X(X)$ and $g_Y(Y)$. Consequently, the HGR coefficient may also be written as

$$\rho_{\text{HGR}}(X;Y)=\sup_{g_X,g_Y}\text{Corr}\{g_X(X),g_Y(Y)\}\ \text{ s.t.}\begin{cases}\mathbb{E}_{f_X}\{g_X(X)\}=\mathbb{E}_{f_Y}\{g_Y(Y)\}=0\\ \mathbb{E}_{f_X}\{g_X^2(X)\}=\mathbb{E}_{f_Y}\{g_Y^2(Y)\}=1\end{cases}\tag{2.79a}$$

$$=\sup_{g_X,g_Y}\frac{\text{Cov}\{g_X(X),g_Y(Y)\}}{\sqrt{\text{Var}\{g_X(X)\}}\sqrt{\text{Var}\{g_Y(Y)\}}}\ \text{ s.t.}\begin{cases}\mathbb{E}_{f_X}\{g_X(X)\}=\mathbb{E}_{f_Y}\{g_Y(Y)\}=0\\ \mathbb{E}_{f_X}\{g_X^2(X)\}=\mathbb{E}_{f_Y}\{g_Y^2(Y)\}=1\end{cases}\tag{2.79b}$$

$$= \sup_{g_X, g_Y} \frac{\mathbb{E}_{f_{X,Y}} \left\{ \left(g_X(X) - \mathbb{E}_{f_X}\{g_X(X)\} \right) \left(g_Y(Y) - \mathbb{E}_{f_Y}\{g_Y(Y)\} \right) \right\}}{\sqrt{\mathbb{E}_{f_X}\left\{g_X^2(X)\right\} - \mathbb{E}_{f_X}^2\{g_X(X)\}} \sqrt{\mathbb{E}_{f_Y}\left\{g_Y^2(Y)\right\} - \mathbb{E}_{f_Y}^2\{g_Y(Y)\}}}$$
$$\text{s.t.} \begin{cases} \mathbb{E}_{f_X}\{g_X(X)\} = \mathbb{E}_{f_Y}\{g_Y(Y)\} = 0 \\ \mathbb{E}_{f_X}\left\{g_X^2(X)\right\} = \mathbb{E}_{f_Y}\left\{g_Y^2(Y)\right\} = 1 \end{cases}. \tag{2.79c}$$

Particularly, in [Rén59], Rényi was primarily interested in characterizing the dependency between two random variables through a given statistic with a list of predetermined desirable properties. In terms of the HGR coefficient, these are:

1. $\rho_{\text{HGR}}(X;Y)$ is defined for a pair of nonconstant random variables X and Y.
2. It is symmetric:
$$\rho_{\text{HGR}}(X;Y) = \rho_{\text{HGR}}(Y;X). \tag{2.80}$$
3. It is bounded:
$$0 \leq \rho_{\text{HGR}}(X;Y) \leq 1. \tag{2.81}$$
4. $\rho_{\text{HGR}}(X;Y) = 0$ only for independent random variables, and $\rho_{\text{HGR}}(X;Y) = 1$ only for a strict dependence and Borel-measurable functions.
5. If $g_X, g_Y : \mathbb{R} \to \mathbb{R}$ are bijective Borel-measurable functions, then
$$\rho_{\text{HGR}}(X;Y) = \rho_{\text{HGR}}(g_X(X); g_Y(Y)). \tag{2.82}$$
6. If X and Y are correlated Gaussian variables, then
$$\rho_{\text{HGR}}(X;Y) = |\rho(X;Y)|, \tag{2.83}$$
where
$$\rho(X;Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \tag{2.84}$$
is the Pearson correlation coefficient.

While these properties are purposely desirable for determining statistical dependency, other well-known dependence measures do not necessarily satisfy them. For instance, the MI does not meet some of them, as it is notoriously not upper-bounded nor it is strictly the Pearson coefficient for Gaussian variables (although it is a function of it). It is also worth noting that in [Rén59] the SMI is also reviewed under the point of view of these properties. Similarly to the MI, the SMI also does not satisfy some of them, although a transformation can be applied to bound the SMI between zero and one.

Generally speaking, the supremum is taken over an infinite-dimensional space for continuous random variables, usually deeming the HGR coefficient as computationally not feasible [LHS13]. However, it is possible to obtain a computationally tractable implementation thanks to kernel methods. This implementation will be addressed in Section 3.2.

2.3.4.2 Quadratic measure of dependence

The rest of independence measures focus on the requirement of being zero if and only if the random variables are independent. For this purpose, it is usually enough to test either the condition $f_{X,Y}(x,y) = f_X(x)\,f_Y(y)$ (hypothesis $\mathcal{H}_0$) or $f_{X,Y}(x,y) \neq f_X(x)\,f_Y(y)$ (hypothesis $\mathcal{H}_1$). This indicator (see [Set+11, Eq. (4)]) can be portrayed as follows:

$$\xi(X;Y) = \mathbb{E}_{f_{X,Y}} \left\{ \frac{(f_{X,Y}(x,y) - f_X(x) f_Y(y))^2}{f_{X,Y}(x,y)} \right\} = \|f_{X,Y}(x,y) - f_X(x) f_Y(y)\|_2^2. \tag{2.85}$$

As can be seen, the previous expression is just the squared norm of a difference of densities, which portrays the distance between the joint PDF and the product of the marginal ones. The simplicity in its formulation comes with the cost of lacking a connection with the MI, since neither inequalities or local behavior can be stated in the way that have been established for the SMI.

2.3.4.3 Distance covariance and correlation

Finally, we will briefly address the distance covariance and distance correlation. These are statistical tests of independence for random vectors that do not necessarily have the same dimension. Again, both tests focus on the property of being zero only for independent random variables, and they possess a close link with the previous measure $\xi(X;Y)$.

Let $\varphi_X(\boldsymbol{\omega}_1)$, $\varphi_X(\boldsymbol{\omega}_2)$ and $\varphi_{X,Y}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)$ be the CFs, defined in (2.5), of the random variables X, Y and $X \times Y$, respectively, with $\boldsymbol{\omega}_1 \in \mathbb{R}^N$, $\boldsymbol{\omega}_2 \in \mathbb{R}^M$. Then, the distance covariance [SRB07] is as follows:

$$\mathcal{V}^2(X;Y,\upsilon) = \left\| \varphi_{X,Y}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2) - \varphi_X(\boldsymbol{\omega}_1)\,\varphi_X(\boldsymbol{\omega}_2) \right\|_\upsilon^2, \tag{2.86}$$

where the norm $\|\cdot\|_\upsilon^2$ is defined as

$$\left\| g_{X,Y}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2) \right\|_\upsilon^2 = \int_{\mathbb{R}^N}\int_{\mathbb{R}^M} \left| g_{X,Y}(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2) \right|^2 \upsilon(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2)\, \mathrm{d}\boldsymbol{\omega}_1 \mathrm{d}\boldsymbol{\omega}_2. \tag{2.87}$$

Here υ acts as a weight function so that the integral exists and the distance covariance is scale and rotation invariant. The proposed weight function in [SRB07; SR09] is as follows:

$$\upsilon(\boldsymbol{\omega}_1,\boldsymbol{\omega}_2) = \frac{1}{c_N c_M \left\|\boldsymbol{\omega}_1\right\|_2^{1+N} \left\|\boldsymbol{\omega}_2\right\|_2^{1+M}}, \tag{2.88}$$

where $c_P = \pi^{(1+P)/2}\Gamma\left((1+P)/2\right)$.

Unlike the quadratic measures of dependence, this measure is the distance between the joint and the product of the marginal characteristic functions but defined on the space of the weighted L^2-norm. However, since the PDF and the characteristic function are Fourier pairs, the distance covariance is closely related to the test from (2.85). For an appropriate choice of weight function, the $\xi(X;Y)$ and the $\mathcal{V}^2(X;Y,\upsilon)$ tests are equivalent thanks to the Plancherel-Parseval theorem.

One of the main advantages of this measure is that by expanding the norm and operating with the characteristic functions, one can determine an estimate that only depends on pairwise differences among the random variables. Specifically, in [SRB07] it is shown that, for finite first-order moments of the random variables X and Y, the distance covariance coefficient is equivalent to the following expression:

$$\begin{aligned}\mathcal{V}^2(X;Y) = {} & \mathbb{E}_{f_{X,Y}}\left\{\mathbb{E}_{f_{X',Y'}}\left\{\left\|X-X'\right\|_2\left\|Y-Y'\right\|_2\right\}\right\} \\ & + \mathbb{E}_{f_X}\left\{\mathbb{E}_{f_{X'}}\left\{\left\|X-X'\right\|_2\right\}\right\}\mathbb{E}_{f_Y}\left\{\mathbb{E}_{f_{Y'}}\left\{\left\|Y-Y'\right\|_2\right\}\right\} \\ & - 2\mathbb{E}_{f_{X,Y}}\left\{\mathbb{E}_{f_X}\left\{\left\|X-X'\right\|_2\right\}\mathbb{E}_{f_Y}\left\{\left\|Y-Y'\right\|_2\right\}\right\},\end{aligned} \tag{2.89}$$

where $f_{X,Y} = f_{X',Y'}$, $f_X = f_{X'}$ and $f_Y = f_{Y'}$. From this expression, one can easily determine an empirical estimate of the distance covariance by substituting the expectations with sample means, as it appears in [SRB07].

Finally, the distance correlation extends the notion of covariance to a one similar to the Pearson coefficient, and is defined as

$$\mathcal{R}^2(X;Y,\upsilon) = \begin{cases} \frac{\mathcal{V}^2(X;Y,\upsilon)}{\sqrt{\mathcal{V}^2(X,\upsilon)}\sqrt{\mathcal{V}^2(Y,\upsilon)}} & \mathcal{V}^2(X,\upsilon)\mathcal{V}^2(Y,\upsilon) > 0 \\ 0 & \mathcal{V}^2(X,\upsilon)\mathcal{V}^2(Y,\upsilon) = 0 \end{cases}, \tag{2.90}$$

for $\mathcal{V}^2(Z,\upsilon) = \mathcal{V}^2(Z;Z,\upsilon)$. These measures are of particular interest since they can be related to a kernel measure of independence, as will be seen in the following section. However, similar to $\xi(X;Y)$, they do not convey any significance with respect to other information measures other than resembling a correlation coefficient and following the desirable properties for any dependence measure from Subsection 2.3.4.1.

2.4 Summary

This chapter has reviewed some information-theoretic measures with advantageous properties for estimation. In exchange for the classical measures of information classically utilized in the literature, which are those that derive from the KL divergence, the surrogates gain relevance thanks to their particular expressions. While there are some concessions in terms of properties, these new measures provide a fresh view on how to measure uncertainty, dissimilarity, and information, while at the same time being well-known figures of merit in other areas outside the signal processing and information theory fields. In particular, the second-order Rényi entropy and the SMI are relevant surrogates with plenty of potential different applications, but are still partially unknown in digital communications, applied statistics, machine learning, etc. Moreover, a local approximation analysis for the SMI has also been conducted. The outcome is that the SMI is an upper-bound except for the small dependence regime, where it behaves as an approximate estimation of twice the MI.

From the point of view of this thesis, the interest in these surrogates mainly relies upon their capability of being expressed as a second-order moment. From this observation, the proposal is to develop methods that are capable of exploiting these expressions by leveraging second-order statistics. These last techniques have a very long and rich history and much is known of their capabilities, so it becomes natural to try to unify the two apparently different standpoints. The gap, however, remains in the nonlinearity presented by an information measure.

In view of this objective, we move to the next chapter to review the definition and uses of kernel methods. Kernels are particularly designed to cope with nonlinear problems in a linear fashion. It is then within reason to benefit from the structure of kernels to merge the nonlinear information measures with well-known techniques that measure correlation, such as PCA and CCA, addressed at the beginning of this chapter. We will see that, in fact, there are existing techniques that capitalize on measuring linear dependencies using kernel methods, and we will contextualize them within the thesis.

Chapter 3

Information-theoretic kernel methods

Kernel methods are powerful techniques that allow solving complex nonlinear problems in the field of machine learning. While the origins of a proper kernel function trace back to the beginnings of the last century [Mer09], its popularity exponentially grew upon its use for binary classification problems [BGV92], which then was particularized to be known as Support Vector Machine (SVM). In SVM, the classification is done by optimizing the hyperplane that linearly separates the classes, defined by the inner product between a given support vector and the data. However, this approach, referred to as the primal model, can not provide linearly separable classes with nonlinear data. Instead, the kernels allow for a nonlinear mapping and posterior measurements of inner products in a high-dimensional feature space, possibly infinite-dimensional, where the problem can be solved in a linear fashion with high probability. This statement follows directly from Cover's theorem [Cov65]. The alternative optimization problem cast by the kernel methods is referred to as the dual model [CS00]. In the dual model the inner product in the feature space is done implicitly, that is, the inner product is not explicitly calculated, but directly obtained thanks to the kernel function. This property constitutes the main advantage of kernels, as computations in a infinite-dimensional space are prohibitively complex. The advantages provided by the kernels are not limited to SVM, but can be generalized by any method that is defined by an inner product and may require a high-dimensional mapping. Thanks to this, kernel methods are widely used in a plethora of nonlinear problems, and their applications have been spread in the machine learning and signal processing fields [HSS08; Roj+18; PLH11]. In many of theses cases, we may also refer to the use of kernels as "kernel signal processing". On the other hand, the cost of embracing kernel methods is usually a high computational complexity. Although it may vary case by case, while avoiding explicitly visiting the feature space, the complexity grows exponentially with the data size.

This chapter is structured as follows. First, the concept of a kernel is disseminated through an introduction to the kernel function in Section 3.1. The objective is to give a brief background on what constitutes a kernel and how to use it as a mapping to infinite-dimensional feature spaces for the purpose of solving linear problems after a nonlinear transformation. In 3.2 we provide a unified review on kernel methods employed for estimating information measures. We focus on these techniques that are linked with the measures addressed in the previous chapter. Lastly, in Section 3.3, a particular case is drawn, where the kernel method arises naturally rather than from the ground up.

3.1 Reproducing kernel Hilbert spaces

While the theory of kernels was consolidated by Mercer in [Mer09], the following definitions will follow the ones provided by Aronszajn in [Aro50]. The exposition of the theory of reproducing kernel Hilbert spaces will also follow a similar structure as the one given in [Vae10].

Definition 4. A Hilbert space $\mathcal{H}$ is a complete inner product space in the sense of Cauchy, i.e. every Cauchy sequence converges inside the space.

Example 4.1. The L^2-space with inner product

$$\langle f, g \rangle = \int f(x) \overline{g(x)} \mathrm{d}x \tag{3.1}$$

is a Hilbert space. From all the L^p-spaces, only the L^2 is a norm induced by a inner product, thus a Hilbert space.

In terms of the signal processing field, Hilbert spaces allows to deal with signals with infinite length, endowing them with the constrain of finite energy. As a consequence, we say that the signal has finite norm, thus its scalar product is well-defined. Note that, while there are finite-dimensional Hilbert spaces (e.g. the Euclidean space), here we are interested in the generalization to infinite-dimensional feature spaces.

Let us now proceed by defining a kernel, which will then be put in relation to a Hilbert space:

Definition 5. A kernel is a continuous, symmetric and positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies

$$\sum_{i,j=0}^{L-1} \alpha_i \alpha_j k(\mathbf{x}(i), \mathbf{x}(j)) \geq 0 \quad \forall \alpha_i, \alpha_j \in \mathbb{R}, \tag{3.2}$$

for any set of data $\{\mathbf{x}(i)\}_{i=0,..,L-1} \in \mathcal{X}$.

Example 5.1. The polynomial function

$$k(\mathbf{x}(i), \mathbf{x}(j)) = \left(\mathbf{x}^H(i)\mathbf{x}(j) + a\right)^b \tag{3.3}$$

with $a \geq 0$, is a kernel. Note that the dot product is included in this family $\langle \mathbf{x}(i), \mathbf{x}(j) \rangle = \mathbf{x}^H(i)\mathbf{x}(j)$, and therefore it is also a kernel.

Example 5.2. The radial basis (or Gaussian) function

$$g(\mathbf{x}(i), \mathbf{x}(j)) = \exp\left(-\frac{\|\mathbf{x}(i) - \mathbf{x}(j)\|^2}{2\sigma^2}\right) \tag{3.4}$$

is a kernel.

The kernel function can also be seen as a similarity measure. For instance, the dot product from Example 5.1 is well-known in this regard, since it intrinsically computes the cosine of the angle between $\mathbf{x}(i)$ and $\mathbf{x}(j)$, and is equal to zero for orthogonal vectors. In the Gaussian kernel case, the similarity measure is clearly seen in the subtraction between two vectors, which is equal to one for $\mathbf{x}(i) = \mathbf{x}(j)$, and it tends to zero as the dissimilarity between $\mathbf{x}(i)$ and $\mathbf{x}(j)$ increases. By computing similarities of two vectors and returning a single real value, kernels are sought as descriptors of data.

Let us now consider a function $\phi : \mathcal{X} \to \mathcal{H}$ defined on the nonempty set $\mathcal{X}$ that provides the feature map given an input $\mathbf{x}(i)$ onto the Hilbert space $\mathcal{H}$.

Definition 6. The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of the Hilbert space $\mathcal{H}$ if

1. For $\forall \mathbf{x}(i) \in \mathcal{X}$, every function $k(\mathbf{x}(i), \cdot)$ belongs to $\mathcal{H}$.
2. For $\forall \mathbf{x}(i) \in \mathcal{X}$ and $\forall \phi \in \mathcal{H}$, it satisfies the reproducing property:

$$\langle \phi, k(\mathbf{x}(i), \cdot) \rangle_{\mathcal{H}} = \phi(\mathbf{x}(i)) \tag{3.5}$$

Particularly, the reproducing property also states that

$$\langle k\left(\mathbf{x}(i), \cdot\right), k\left(\mathbf{x}(j), \cdot\right)\rangle_{\mathcal{H}} = \langle \phi\left(\mathbf{x}(i)\right), \phi\left(\mathbf{x}(j)\right)\rangle_{\mathcal{H}} = k\left(\mathbf{x}(i), \mathbf{x}(j)\right). \tag{3.6}$$

If a reproducing kernel exists, then $\mathcal{H}$ is called a Reproducing Kernel Hilbert Space (RKHS). It also follows that a kernel is positive definite only if it is a reproducing kernel. Intuitively, (3.6) shows that the inner product of the mapped data through the function ϕ can be directly obtained as the evaluation of a kernel function. In other words, we can avoid the computation of the inner product in the Hilbert space $\mathcal{H}$ thanks to the reproducing kernel.

It may also be useful to define the kernel matrix:

Definition 7. A kernel matrix $\mathbf{K} \in \mathbb{R}^{L\times L}$ is a positive definite matrix $\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \geq 0$ for $\forall\boldsymbol{\alpha} \in \mathbb{R}^L$, with elements $[\mathbf{K}]_{i,j} = k\left(\mathbf{x}(i), \mathbf{x}(j)\right)$ for $\{\mathbf{x}(i)\}_{i=0,..,L-1} \in \mathcal{X}$.

Due to the definition of a kernel, the kernel matrix is also a Gram matrix since its entries are given by an inner product $\{\mathbf{K}\}_{i,j} = \langle \phi\left(\mathbf{x}(i)\right), \phi\left(\mathbf{x}(j)\right)\rangle_{\mathcal{H}}$, satisfying $\mathbf{K} = \mathbf{K}^T$. In some ways, kernel matrices have many properties in common with correlation matrices. Both matrices are Gram matrices, but what differentiates them is in which space the inner product occurs.

Finally, we can state the following theorem:

Theorem 3.1 (Moore-Aronszajn theorem [Aro50])**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite function. A unique RKHS exists with reproducing kernel k.*

Remark 3.1.1. While a RKHS determines a unique reproducing kernel k, the converse is not true [SSB+02, Sec. 2.2.3].

The relevance of the Moore-Aronszajn theorem is that it states that any positive definite function $k\left(\mathbf{x}(i), \mathbf{x}(j)\right)$ is a reproducing kernel. Moreover, since every kernel is positive definite, then every reproducing kernel is a kernel, which is at the same time a positive definite function. In this way, this theorem closes a circle of definitions that revolve around the theory of RKHS.

One of the most powerful techniques within RKHS is the famously called “kernel trick”. This *trick* refers to the capability of interchanging kernel functions to solve the same problem [SSB+02, Remark 2.8]. The only restriction is that we work with positive definite kernel functions, which endows them with the theory of RKHS. In essence, the *trick* is to actually choose a given kernel function, and to not worry about the RKHS that its inner product spans. Furthermore, from (3.6), one does not even worry about computing the inner product. The most exemplary use of the *kernel trick* is in the application of SVM, where the dual model is precisely taking advantage of it [SAP10].

Kernel methods can then be described as a simple approach to complex problems. It is precisely the flexibility provided by the reproducing property and the *kernel trick* what makes this approach attractive for a plethora of different problems. Despite this, a blind mapping onto a infinite-dimensional feature space may entail other problems that may be avoidable with lower-dimensional representations of the data. Specifically, as the dimensionality increases, the embedded data onto the feature space tends to become sparse. The result is that some problems may tend to be ill-conditioned. This effect is generally referred to as “the curse of dimensionality” [BN06]. This problem can be solved in two ways. The first is to exponentially increase the number of data available, in hopes that the sparsity is reduced in the high-dimensional space. Since the amount of data may vary per application, this solution is often disregarded. The second and more common solution is regularization, which is usually based on an ad hoc approach. Regardless of the *curse of dimensionality*, many kernel-based algorithms enjoy from multiple benefits of its intrinsic high-dimensional feature space. In most algorithms, the sole capability of solving nonlinear problems outweighs, for example, the need for regularization.

3.2 Kernel measures of information

In the last decades, the theory of RKHS has been harnessed for numerous applications in the fields of probability and statistics, particularly in statistical learning and inference [BT11]. This last group includes the estimation of measures of information [Xu+08; ZC06]. Since information measures can be seen as nonlinear descriptors of random variables, it should be within reason the use of nonlinear techniques such as kernel methods. The rationale endowed to these problems is shared among them: a nonlinear mapping provided by the kernel functions should help to solve the problem in a linear fashion. These linear solutions are frequently found in measures of correlation and/or covariance, thus second-order statistics. This is particularly interesting as it is a first approach to relate information theory measures with second-order statistics, and they shall be used as a basis for the thesis. However, as will be noted along the rest of this section, the study of correlation and covariance operators in a possibly infinite-dimensional feature space comes with a cost. While it is true that we can relate probabilistic measures to the dissimilarity measures spanned by the kernels [ZC06], the analysis of these operators in the Hilbert space is often a difficult task. These operators need to be reduced in a matrix form through a lower-dimensional representation of the Hilbert space, while still needing regularization techniques to avoid ill-conditioned problems, or matrices in this case.

In the sequel, some of the most important approaches to the use of RKHS to measure information will be provided. In each case, the focus will become how such measures become manageable from the point of view of an empirical estimate, and how these estimates relate to the original information measure.

3.2.1 Kernel principal component analysis

As a starting point, we will address the *kernelized* version of the PCA. While in its original formulation it is considered a measure of linear dependence (within the components of the same data-set), the kernel method promotes this dependence to a nonlinear one. Simply put, Kernel Principal Component Analysis (KPCA) performs PCA in the feature space. Consequently, if the kernel function used to map the data onto the feature space is the dot product, then both methods are equivalent.

For now, the KPCA does not hold any apparent relationship with any of the previously covered information measures. However, we will see in Chapter 4 how the KPCA is actually related to an entropy measure. Therefore, it is still relevant to discuss how the PCA can be handled within the theory of RKHS, and to serve as a basis for the *kernelized* version of the CCA, which does directly relate to an information measure.

Let $g_X \in \mathcal{K}$, where $\mathcal{K}$ is a RKHS with associated kernel k. Consider the feature map

$$\phi_X\left(X\right) = k\left(X, \cdot\right) \qquad \forall \phi_X \in \mathcal{K} \quad \forall X \in \mathcal{X}. \tag{3.7}$$

Thanks to the reproducing property from (3.5), we can express

$$g_X\left(X\right) = \left\langle \phi_X\left(X\right), g_X \right\rangle_{\mathcal{K}}. \tag{3.8}$$

The equivalent formulation of the PCA from (2.25) is then

$$\rho_{\text{KPCA}}\left(X\right) = \max_{g_X} \ \text{Cov}\left\{\left\langle \phi_X\left(X\right), g_X\right\rangle_{\mathcal{K}}, \left\langle \phi_X\left(X\right), g_X\right\rangle_{\mathcal{K}}\right\}. \tag{3.9}$$

This formulation takes advantage of the kernel *trick* in order to define the KPCA as the covariance of the inner product of a given feature space, following similar steps as in [BJ02]. Next, we will consider a finite-dimensional subspace of the RKHS in order to formulate an empirical estimate of this measure.

For now, assume that the data is centered in the feature space, i.e. $\mathbb{E}_{f_X}\left\{g_X\left(X\right)\right\} = 0$. Let $\mathbf{x}(i) \in \mathbb{R}^N$ be L i.i.d. observations with $0 \leq i \leq L-1$, whose images in the feature space are $\boldsymbol{\Phi}_X = \left[\boldsymbol{\phi}_X\left(\mathbf{x}(0)\right), ..., \boldsymbol{\phi}_X\left(\mathbf{x}(L-1)\right)\right]$. Then, the empirical representation of the autocovariance is expressed as

$$\hat{\text{Cov}}\left\{\left\langle \phi_X\left(X\right), g_X\right\rangle_{\mathcal{K}}, \left\langle \phi_X\left(X\right), g_X\right\rangle_{\mathcal{K}}\right\} = \frac{1}{L}\sum_{i=0}^{L-1} \left\langle \boldsymbol{\phi}_X\left(\mathbf{x}(i)\right), g_X\right\rangle_{\mathcal{K}} \left\langle \boldsymbol{\phi}_X\left(\mathbf{x}(i)\right), g_X\right\rangle_{\mathcal{K}}^{H}. \tag{3.10}$$

Equivalently, g_X can be linearly decomposed by the images of the feature space such that $\mathbf{g}_X = \sum_{k=0}^{L-1} u_k \boldsymbol{\phi}_X(\mathbf{x}(k))$. This linear combination serves as the principal component that lies in the feature space [KG03]. Then, the autocovariance from (3.10) becomes

$$\begin{aligned}
&\hat{\mathrm{Cov}}\left\{\langle \boldsymbol{\phi}_X(\mathbf{x}), \mathbf{g}_X\rangle_{\mathcal{K}}, \langle \boldsymbol{\phi}_X(\mathbf{x}), \mathbf{g}_X\rangle_{\mathcal{K}}\right\} && (3.11a)\\
&= \frac{1}{L}\sum_{i=0}^{L-1}\left\langle \boldsymbol{\phi}_X(\mathbf{x}(i)), \sum_{k=0}^{L-1} u_k \boldsymbol{\phi}_X(\mathbf{x}(k))\right\rangle_{\mathcal{K}}^{H}\left\langle \boldsymbol{\phi}_X(\mathbf{x}(i)), \sum_{l=0}^{L-1} u_l \boldsymbol{\phi}_X(\mathbf{x}(l))\right\rangle_{\mathcal{K}} && (3.11b)\\
&= \frac{1}{L}\sum_{i=0}^{L-1}\sum_{k=0}^{L-1}\sum_{l=0}^{L-1} u_k^* k(\mathbf{x}(k), \mathbf{x}(i))\, k(\mathbf{x}(l), \mathbf{x}(i))\, u_l && (3.11c)\\
&= \frac{1}{L}\mathbf{u}^H\mathbf{K}\mathbf{K}\mathbf{u}. && (3.11d)
\end{aligned}$$

By substituting the previous expression into (3.9), the supremum then translates to the maximum over the vector $\mathbf{u}$. However, similar to the trivial solution to the original PCA problem in (2.25), we either constrain the principal components with

$$\mathbf{g}_X^H \mathbf{g}_X = \sum_{k=0}^{L-1}\sum_{l=0}^{L-1} u_k^* \boldsymbol{\phi}_X^H(\mathbf{x}(k))\, \boldsymbol{\phi}_X(\mathbf{x}(l))\, u_l = \mathbf{u}^H\mathbf{K}\mathbf{u} = 1, \tag{3.12}$$

or perform the equivalent maximization problem with $\mathbf{u}^H\mathbf{K}\mathbf{u}$ in the denominator. Lastly, we will consider the case in which the data is not centered in the feature space. Although this is an apparent difficult task, it can be shown [SSM98] that the kernel matrix of the centered variables in the feature space corresponds to

$$\bar{\mathbf{K}} = \mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{K}\mathbf{P}_{\mathbf{1}}^{\perp}, \tag{3.13}$$

where $\mathbf{P}_{\mathbf{1}}^{\perp}$ is the projection matrix from (2.27). By gathering all previous considerations, the equivalent problem formulation is the following:

$$\hat{\rho}_{\text{KPCA}}(X) = \max_{\mathbf{u}} \frac{1}{L}\frac{\mathbf{u}^H\bar{\mathbf{K}}\bar{\mathbf{K}}\mathbf{u}}{\mathbf{u}^H\bar{\mathbf{K}}\mathbf{u}}. \tag{3.14}$$

Following the Rayleigh quotient (2.29), the solution to the maximization problem is equivalent to finding the maximum eigenvalue of the generalized eigenvalue problem

$$\frac{1}{L}\bar{\mathbf{K}}\bar{\mathbf{K}}\mathbf{U} = \bar{\mathbf{K}}\mathbf{U}\boldsymbol{\Lambda}. \tag{3.15}$$

Since $\bar{\mathbf{K}}$ is a symmetric (Gram) matrix, the previous eigenvalue problem is equivalent to solving

$$\frac{1}{L}\bar{\mathbf{K}}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}, \tag{3.16}$$

whose largest eigenvalue leads to the following empirical estimate

$$\hat{\rho}_{\text{KPCA}}(X) = \frac{1}{L}\|\bar{\mathbf{K}}\|_2. \tag{3.17}$$

3.2.2 Kernel canonical correlation analysis

Following the KPCA, we will now discuss the tractable version of the HGR coefficient from Subsection 2.3.4.1. Thanks to the RKHS framework, it is actually possible to determine an empirical estimate of the HGR by letting the functions g_X and g_Y be elements of RKHSs. This implementation is referred to as the Kernel Canonical Correlation Analysis (KCCA) [BJ02], since its empirical estimate is obtained by solving the CCA problem through kernel matrices, similarly as with the KPCA.

Let $g_X \in \mathcal{K}$ and $g_Y \in \mathcal{L}$, where $\mathcal{K}$ and $\mathcal{L}$ are RKHSs with associated kernels k and l, respectively. Consider the feature maps

$$\phi_X(X) = k(X,\cdot), \quad \phi_Y(Y) = l(Y,\cdot) \qquad \forall \phi_X \in \mathcal{K} \quad \forall \phi_Y \in \mathcal{L} \quad \forall X \in \mathcal{X}. \tag{3.18}$$

Again, following the application of the reproducing property as in (3.11), we can express

$$\begin{aligned} g_X(X) &= \langle \phi_X(X), g_X \rangle_{\mathcal{K}}, \\ g_Y(Y) &= \langle \phi_Y(Y), g_Y \rangle_{\mathcal{L}}. \end{aligned} \tag{3.19}$$

The equivalent formulation of the HGR from (2.78) is then

$$\rho_{\text{KCCA}}(X;Y) = \sup_{g_X, g_Y} \text{ Corr}\left\{ \langle \phi_X(X), g_X \rangle_{\mathcal{K}}, \langle \phi_Y(Y), g_Y \rangle_{\mathcal{L}} \right\} \tag{3.20a}$$

$$= \sup_{g_X, g_Y} \frac{\text{Cov}\left\{ \langle \phi_X(x), g_X \rangle_{\mathcal{K}}, \langle \phi_Y(y), g_Y \rangle_{\mathcal{L}} \right\}}{\sqrt{\text{Var}\left\{ \langle \phi_X(x), g_X \rangle_{\mathcal{K}} \right\} \text{Var}\left\{ \langle \phi_Y(y), g_Y \rangle_{\mathcal{L}} \right\}}}. \tag{3.20b}$$

Again, let us assume, for now, that the data is centered in the feature space, i.e. $\mathbb{E}_{f_X}\{g_X(X)\} = \mathbb{E}_{f_Y}\{g_Y(Y)\} = 0$. Let $\mathbf{x}(i) \in \mathbb{R}^{N_x}$ and $\mathbf{y}(i) \in \mathbb{R}^{N_y}$ be L i.i.d. observations from the random variables X and Y with $0 \leq i \leq L-1$, whose images in the feature space are $\mathbf{\Phi}_X = [\boldsymbol{\phi}_X(\mathbf{x}(0)), ..., \boldsymbol{\phi}_X(\mathbf{x}(L-1))]$ and $\mathbf{\Phi}_Y = [\boldsymbol{\phi}_Y(\mathbf{y}(0)), ..., \boldsymbol{\phi}_Y(\mathbf{y}(L-1))]$, respectively. The empirical representation of the correlation is then

$$\hat{\text{Corr}}\left\{ \langle \phi_X(\mathbf{x}), g_X \rangle_{\mathcal{K}}, \langle \phi_Y(\mathbf{y}), g_Y \rangle_{\mathcal{L}} \right\} = \frac{\frac{1}{L}\sum_{i=0}^{L-1} \langle \boldsymbol{\phi}_X(\mathbf{x}(i)), g_X \rangle_{\mathcal{K}} \langle \boldsymbol{\phi}_Y(\mathbf{y}(i)), g_Y \rangle_{\mathcal{L}}^H}{\sqrt{\frac{1}{L}\sum_{i=0}^{L-1} \langle \boldsymbol{\phi}_X(\mathbf{x}(i)), g_X \rangle_{\mathcal{K}} \langle \boldsymbol{\phi}_X(\mathbf{x}(i)), g_X \rangle_{\mathcal{K}}^H} \sqrt{\frac{1}{L}\sum_{i=0}^{L-1} \langle \boldsymbol{\phi}_Y(\mathbf{y}(i)), g_Y \rangle_{\mathcal{L}} \langle \boldsymbol{\phi}_Y(\mathbf{y}(i)), g_Y \rangle_{\mathcal{L}}^H}}. \tag{3.21}$$

The functions g_X and g_Y can be linearly decomposed by the images of the feature space such that $\mathbf{g}_X = \sum_{k=0}^{L-1} u_k \phi_X(\mathbf{x}(k))$ and $\mathbf{g}_Y = \sum_{l=0}^{L-1} v_l \phi_Y(\mathbf{y}(l))$. These linear combinations serve as the canonical variates that lie in the feature space [KG03]. Then, the covariance from (3.21) becomes

$$\hat{\text{Cov}}\left\{ \langle \boldsymbol{\phi}_X(\mathbf{x}), g_X \rangle_{\mathcal{K}}, \langle \boldsymbol{\phi}_Y(\mathbf{y}), g_Y \rangle_{\mathcal{L}} \right\} \tag{3.22a}$$

$$= \frac{1}{L}\sum_{i=0}^{L-1} \left\langle \boldsymbol{\phi}_X(\mathbf{x}(i)), \sum_{k=0}^{L-1} u_k \boldsymbol{\phi}_X(\mathbf{x}(k)) \right\rangle_{\mathcal{K}}^H \left\langle \boldsymbol{\phi}_Y(\mathbf{y}(i)), \sum_{l=0}^{L-1} v_l \boldsymbol{\phi}_Y(\mathbf{y}(l)) \right\rangle_{\mathcal{L}} \tag{3.22b}$$

$$= \frac{1}{L}\sum_{i=0}^{L-1}\sum_{k=0}^{L-1}\sum_{l=0}^{L-1} u_k^* k(\mathbf{x}(k), \mathbf{x}(i))\, l(\mathbf{y}(l), \mathbf{y}(i))\, v_l \tag{3.22c}$$

$$= \frac{1}{L}\mathbf{u}^H \mathbf{K}\mathbf{L}\mathbf{v}, \tag{3.22d}$$

where $\mathbf{u}$ and $\mathbf{v}$ are L-dimensional column vectors, and $\mathbf{K} \in \mathbb{R}^{L\times L}$ and $\mathbf{L} \in \mathbb{R}^{L\times L}$ are the kernel matrices from Definition 7 with entries $[\mathbf{K}]_{i,j} = k(\mathbf{x}(i), \mathbf{x}(j))$ and $[\mathbf{L}]_{i,j} = l(\mathbf{y}(i), \mathbf{y}(j))$. The variances correspond to the autocovariance from (3.11) with

$$\hat{\text{Var}}\left\{ \langle \phi_X(\mathbf{x}), g_X \rangle_{\mathcal{K}} \right\} = \frac{1}{L}\mathbf{u}^H \mathbf{K}\mathbf{K}\mathbf{u} \tag{3.23}$$

$$\hat{\text{Var}}\left\{ \langle \phi_Y(\mathbf{y}), g_Y \rangle_{\mathcal{L}} \right\} = \frac{1}{L}\mathbf{v}^H \mathbf{L}\mathbf{L}\mathbf{v}. \tag{3.24}$$

By gathering all previous derivations, and considering that the kernel matrices have been centered in the feature space (3.13), we then have

$$\hat{\rho}_{\text{KCCA}}(X;Y) = \max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^H \bar{\mathbf{K}}\bar{\mathbf{L}}\mathbf{v}}{\sqrt{\mathbf{u}^H \bar{\mathbf{K}}\bar{\mathbf{K}}\mathbf{u}}\sqrt{\mathbf{v}^H \bar{\mathbf{L}}\bar{\mathbf{L}}\mathbf{v}}}. \tag{3.25}$$

Clearly, since we have moved from functions to vectors, the supremum is now taken as the $\mathbf{u}$ and $\mathbf{v}$ that maximizes the canonical correlate, or in this case, the estimate of the HGR coefficient. The equivalent generalized eigenvalue problem $\mathbf{AW} = \mathbf{BW\Lambda}$ can now be determined similarly as in the original CCA problem (2.38) with

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \bar{\mathbf{K}}\bar{\mathbf{L}} \\ \bar{\mathbf{L}}\bar{\mathbf{K}} & \mathbf{0} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \bar{\mathbf{K}}\bar{\mathbf{K}} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{L}}\bar{\mathbf{L}} \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}. \tag{3.26}$$

Since the HGR coefficient is specifically the maximum above all the possible solutions, its estimation corresponds to only the largest eigenvalue.

Nonetheless, this tractable computation of the HGR coefficient carries some problems. On the one hand, the problem yields $\hat{\rho}_{\text{KCCA}}(X;Y) = 1$ for $\mathbf{u}$, and $\mathbf{v}$ being columns of $\mathbf{L}^{-1}\mathbf{K}$, which occurs when the kernel matrices are invertible [HSS04]. Or, in other words, finding the correlation coefficient becomes trivial. This is prone to happen if the dimensionality of the feature space is much higher than the one of the input space, a characteristic indicator of overfitting [Vae10]. For instance, the classic Gaussian kernel spans a feature space of infinite dimensionality, enticing the need of regularization. Either way, this issue is in line with the common problems of kernel methods addressed at the end of Section 3.1. To avoid overfitting, the usual proposed solution [BJ02] is to regularize such that the matrix (3.25) becomes

$$\hat{\rho}_{\text{KCCA}}(X;Y) = \max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^H\bar{\mathbf{K}}\bar{\mathbf{L}}\mathbf{v}}{\sqrt{\mathbf{u}^H\bar{\mathbf{K}}\bar{\mathbf{K}}\mathbf{u} + \varepsilon\mathbf{u}^H\bar{\mathbf{K}}\mathbf{u}}\sqrt{\mathbf{v}^H\bar{\mathbf{L}}\bar{\mathbf{L}}\mathbf{v} + \varepsilon\mathbf{v}^H\bar{\mathbf{L}}\mathbf{v}}}, \tag{3.27}$$

where $\varepsilon > 0$ is the regularization constant. This procedure can be seen as an equivalent to performing the Tikhonov regularization [BLG08], a commonly method used for regularizing overfitted problems with high number of parameters.

On the other hand, not any kernel function preserves the proprieties of the HGR coefficient. As shown in [Gre+05b], $\rho_{\text{KCCA}}(X;Y) = 0$ for independent random variables if $\mathcal{K}$ and $\mathcal{L}$ are the sets of bounded and continuous functions, i.e. the kernels k and l are continuous and bounded. It is also worth noting that in [BJ02] the kernels were required to be smooth functions, which is a stronger condition than continuous functions. Specifically, the KCCA usually requires the kernel functions to be smooth [BJ02] and bounded functions [Gre+05b]. If these two conditions are met, then we can ensure that $\rho_{\text{KCCA}}(X;Y) = 0$ for independent random variables.

Finally, by proceeding similarly as with the CCA [HSS04], we can determine that the HGR coefficient can be estimated by solving either of the two following eigenvalue problems:

$$\begin{cases} \left(\bar{\mathbf{K}} + \varepsilon\mathbf{I}\right)^{-1}\bar{\mathbf{L}}\left(\bar{\mathbf{L}} + \varepsilon\mathbf{I}\right)^{-1}\bar{\mathbf{K}}\mathbf{U} = \mathbf{\Lambda}_k^2\mathbf{U} \\ \left(\bar{\mathbf{L}} + \varepsilon\mathbf{I}\right)^{-1}\bar{\mathbf{K}}\left(\bar{\mathbf{K}} + \varepsilon\mathbf{I}\right)^{-1}\bar{\mathbf{L}}\mathbf{V} = \mathbf{\Lambda}_l^2\mathbf{V} \end{cases}, \tag{3.28}$$

where the diagonal matrices $\mathbf{\Lambda}_k^2$ and $\mathbf{\Lambda}_l^2$ contain all the squared kernelized canonical correlations. This time, since $\bar{\mathbf{K}}$ and $\bar{\mathbf{L}}$ have the same dimensionality, $\mathbf{\Lambda}_k^2$ and $\mathbf{\Lambda}_l^2$ are equivalent and contain all the same eigenvalues. Therefore, we can express the estimated KCCA coefficient as

$$\hat{\rho}_{\text{KCCA}}(X;Y) = \sqrt{\lambda_1\left(\left(\bar{\mathbf{K}} + \varepsilon\mathbf{I}\right)^{-1}\bar{\mathbf{L}}\left(\bar{\mathbf{L}} + \varepsilon\mathbf{I}\right)^{-1}\bar{\mathbf{K}}\right)}. \tag{3.29}$$

This expression can be further simplified by operating with the largest eigenvalue such that $\lambda_1(\mathbf{AB}) = \lambda_1(\mathbf{BA})$ (which can be derived from [HJ12], Theorem 1.3.22). Then, we can finally express

$$\hat{\rho}_{\text{KCCA}}(X;Y) = \sqrt{\lambda_1(\mathbf{\Gamma}_k\mathbf{\Gamma}_l)} \tag{3.30}$$

for $\mathbf{\Gamma}_k = \bar{\mathbf{K}}\left(\bar{\mathbf{K}} + \varepsilon\mathbf{I}\right)^{-1}$ and $\mathbf{\Gamma}_l = \bar{\mathbf{L}}\left(\bar{\mathbf{L}} + \varepsilon\mathbf{I}\right)^{-1}$.

3.2.3 Constrained covariance

The regularization problem of KCCA is, indeed, its rigid form that follows from the HGR coefficient and its properties. While the whitening process of the correlation operator ensures a full-fledged dependence measure, it ends up being detrimental to the estimation process. This observation led to Gretton et al. [Gre+05c] to explore a different RKHS based methods to explore less restrictive dependence measures. Specifically, they proposed to switch the correlation operator to a covariance operator, referred to as the Constrained Covariance (COCO). This interchangeability can be done at the expense of dropping some of the properties from Subsection 2.3.4.1, mainly the equality to one for deterministic dependence. The advantage is that the covariance and correlation operators behave identically under independent random variables. Therefore, the focus in [Gre+05c] shifts to a test of independence while providing a more stable measure without the need for regularization.

The ultimate goal of the COCO is to determine the maximum cross-covariance between the spaces spanned by the functions g_X and g_Y. Following the same terms used for defining the HGR coefficient, the COCO is defined as

$$\rho_{\mathrm{COCO}}\left(X;Y\right) = \sup_{g_X, g_Y} \ \mathrm{Cov}\left\{g_X\left(X\right), g_Y\left(Y\right)\right\}. \tag{3.31}$$

Again, let $g_X \in \mathcal{K}$ and $g_Y \in \mathcal{L}$ be functions that map the data onto the RKHSs $\mathcal{K}$ and $\mathcal{L}$, and let k and l be the kernels associated to these RKHSs. Following (3.20), we can determine

$$\rho_{\mathrm{COCO}}\left(X;Y\right) = \sup_{g_X, g_Y} \ \mathrm{Cov}\left\{\left\langle\phi_X\left(X\right), g_X\right\rangle_{\mathcal{K}}, \left\langle\phi_Y\left(Y\right), g_Y\right\rangle_{\mathcal{L}}\right\}, \tag{3.32}$$

where ϕ_X and ϕ_Y are the feature maps as in (3.18). Then, we can outright express the covariance operator as it appears in (3.22):

$$\hat{\mathrm{Cov}}\left\{\left\langle\phi_X\left(\mathbf{x}\right), g_X\right\rangle_{\mathcal{K}}, \left\langle\phi_Y\left(\mathbf{y}\right), g_Y\right\rangle_{\mathcal{L}}\right\} = \frac{1}{L}\mathbf{u}^H\bar{\mathbf{K}}\bar{\mathbf{L}}\mathbf{v}, \tag{3.33}$$

where the data has been centered in the feature space. The empirical COCO then becomes

$$\hat{\rho}_{\mathrm{COCO}}\left(X;Y\right) = \max_{\mathbf{u},\mathbf{v}} \ \frac{1}{L}\mathbf{u}^H\bar{\mathbf{K}}\bar{\mathbf{L}}\mathbf{v}. \tag{3.34}$$

We can also express $\hat{\rho}_{\mathrm{COCO}}\left(X;Y\right)$ in terms of the generalized eigenvalue problem [Gre+05b] with

$$\mathbf{A} = \frac{1}{L}\begin{bmatrix} \mathbf{0} & \bar{\mathbf{L}} \\ \bar{\mathbf{K}} & \mathbf{0} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}, \tag{3.35}$$

whose maximum eigenvalue leads to the following empirical estimate:

$$\hat{\rho}_{\mathrm{COCO}}\left(X;Y\right) = \frac{1}{L}\|\bar{\mathbf{K}}^{1/2}\bar{\mathbf{L}}^{1/2}\|_2 = \frac{1}{L}\|\mathbf{K}^{1/2}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}^{1/2}\|_2, \tag{3.36}$$

where the square root of a positive semi-definite matrix (which kernel matrices are) $\mathbf{M}^{1/2}$ denotes $\mathbf{M}^{1/2}\mathbf{M}^{1/2} = \mathbf{M}$.

Similarly to the KCCA, the COCO also requires some conditions on the kernels. While the set of continuous and bounded functions is sufficient for $\rho_{\mathrm{COCO}}\left(X;Y\right) = 0$ under independent random variables, in [Gre+05b] it is argued that these conditions provide a rich choice of functions that may not guarantee good convergence properties of the empirical estimate $\hat{\rho}_{\mathrm{COCO}}\left(X;Y\right)$. Therefore, they propose to limit k and l to be *universal* kernels, that is, kernel functions whose RKHS is dense in the space of continuous functions [MXZ06]. Either way, the most well-known and used kernel, the Gaussian kernel from (3.4), meets all of the previous properties.

3.2.4 Hilbert-Schmidt Independence Criterion

The Hilbert-Schmidt Independence Criterion (HSIC) [Gre+05a] is an extension of the COCO that does not compute only the largest covariance coefficient but all the covariance operator spectrum. This way, this generalization accomplishes two things: it does not require regularization due to measuring covariance, and drops the supremum carried over from the HGR coefficient, which allows for an easier measurement both statistically and in computational complexity. At independence, it behaves identically to the COCO. Hence, the key point of developing an independence measure is preserved.

In order to see the HSIC, let us begin by defining the Hilbert-Schmidt (HS) norm:

Definition 8. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with othonormal bases f_n and g_m, respectively, and let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator. The HS norm of A is then defined as

$$\|A\|_{\text{HS}}^2 = \sum_{n,m} \left|\langle f_n, A g_m\rangle_{\mathcal{H}_2}\right|^2 . \tag{3.37}$$

In the case of finite-dimensional Euclidean spaces, the HS norm corresponds to the Frobenius norm. For instance, let $\mathbf{A} \in \mathbb{R}^{N\times M}$ be a matrix operator $\mathbf{A} : \mathbb{R}^M \to \mathbb{R}^N$ with left and right singular vectors correspond to $\{\mathbf{f}_n\}_{n=1,\dots,N}$ and $\{\mathbf{g}_m\}_{m=1,\dots,M}$, respectively. On the one hand, the Frobenius norm is defined as

$$\|\mathbf{A}\|_{\text{F}} = \sqrt{\sum_{n,m} \left|[\mathbf{A}]_{n,m}\right|^2} = \sqrt{\sum_{i}^{\min(N,M)} \sigma_i^2(\mathbf{A})} = \text{tr}\left(\mathbf{A}\mathbf{A}^T\right) = \text{tr}\left(\mathbf{A}^T\mathbf{A}\right). \tag{3.38}$$

On the other hand, the squared HS norm from (3.37) becomes

$$\|\mathbf{A}\|_{\text{HS}}^2 = \sum_{n,m} \left|\langle \mathbf{A}^T\mathbf{f}_n, \mathbf{g}_m\rangle\right|^2 = \sum_{n,m} \left|\mathbf{f}_n^H\mathbf{A}\mathbf{g}_m\right|^2 = \sum_{n,m} \left|[\boldsymbol{\Sigma}]_{n,m}\right|^2 = \sum_{i}^{\min\{N,M\}} \sigma_i^2(\mathbf{A}) = \|\mathbf{A}\|_{\text{F}}^2 , \tag{3.39}$$

where $\boldsymbol{\Sigma} = \mathbf{F}^H\mathbf{A}\mathbf{G}$ is the diagonal matrix that contains the singular values σ_i, and $\mathbf{F}$ and $\mathbf{G}$ are the matrices that contains the left and right singular vectors. It is also straightforward to see that

$$\|\mathbf{A}\|_{\text{HS}}^2 = \sum_{i}^{\min\{N,M\}} \sigma_i^2(\mathbf{A}) \geq \sigma_1^2(\mathbf{A}) = \|\mathbf{A}\|_2^2 , \tag{3.40}$$

or, equivalently, $\|\mathbf{A}\|_{\text{HS}} \geq \|\mathbf{A}\|_2$, where equality holds for $N = M = 1$ or for zero-valued norms.

Consider now the RKHSs $\mathcal{K}$ and $\mathcal{L}$ with mapping functions $g_X \in \mathcal{K}$ and $g_X \in \mathcal{L}$. The HSIC is then similar to the COCO but, instead of taking only the supremum, it measures the HS norm of the covariance such that

$$\rho_{\text{HSIC}}(X;Y) = \|\text{Cov}\left\{\langle \phi_X(X), g_X\rangle_{\mathcal{K}}, \langle \phi_Y(Y), g_Y\rangle_{\mathcal{L}}\right\}\|_{\text{HS}}^2 . \tag{3.41}$$

The empirical measure of the HSIC can then be directly obtained by substituting the spectral norm from (3.36) with the squared HS norm. However, as we have spanned the data onto a finite dimensional subspace of an infinite-dimensional Hilbert space, the HS norm becomes the Frobenius norm, thus becoming easily attainable. The estimator is then expressed as follows

$$\hat{\rho}_{\text{HSIC}}(X;Y) = \left(\frac{1}{L}\left\|\mathbf{K}^{1/2}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}^{1/2}\right\|_{\text{HS}}\right)^2 = \frac{1}{L^2}\text{tr}\left(\mathbf{L}^{1/2}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{K}^{1/2}\mathbf{K}^{1/2}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}^{1/2}\right) \tag{3.42a}$$

$$= \frac{1}{L^2}\text{tr}\left(\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{K}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}\right), \tag{3.42b}$$

where the final expression is obtained thanks to the circularity of the trace. It also should be noted that in [Gre+05a] the HS norm is averaged over $(L-1)^2$ elements instead of L^2 as in the previous expression. While $(L-1)^2$ is in line with the unbiased sample covariance, due to the Bessel correction, here we will express the empirical HSIC with L^2 to maintain its direct derivation from the COCO.

Due to the inequality from (3.40), the HSIC upper-bounds the COCO such that

$$\rho_{\text{HSIC}}(X;Y) \geq \rho_{\text{COCO}}^2(X;Y), \tag{3.43}$$

where equality holds for $\rho_{\text{HSIC}}(X;Y) = \rho_{\text{COCO}}(X;Y) = 0$. Therefore, both COCO and HSIC are valid options to detect independence. However, to compute the trace of a matrix is computationally less complex than performing the complete SVD. Since the dimensionality of kernels matrices increases exponentially with the data-set length, this reduced complexity advantage becomes significant rapidly.

Moreover, the HSIC has a strong relationship with the distance covariance from Subsection 2.3.4.3. To see this connection, let us express the HSIC in terms of expectations over kernel functions. From [Gre+05a], the HSIC becomes

$$\begin{aligned}\rho_{\text{HSIC}}(X;Y) = \; & \mathbb{E}_{f_{X,Y}}\left\{\mathbb{E}_{f_{X',Y'}}\left\{k(X,X')\, l(Y,Y')\right\}\right\} \\ & + \mathbb{E}_{f_X}\left\{\mathbb{E}_{f_{X'}}\left\{k(X,X')\right\}\right\}\mathbb{E}_{f_Y}\left\{\mathbb{E}_{f_{Y'}}\left\{l(Y,Y')\right\}\right\} \\ & - 2\mathbb{E}_{f_{X,Y}}\left\{\mathbb{E}_{f_X}\left\{k(X,X')\right\}\mathbb{E}_{f_Y}\left\{l(Y,Y')\right\}\right\}.\end{aligned} \tag{3.44}$$

It is worth noting that the previous expression can be directly obtained by expanding the covariance from (3.41) with $\langle\phi_X(X), \phi_X(X')\rangle_{\mathcal{K}} = k(X,X')$ and $\langle\phi_Y(Y), \phi_Y(Y')\rangle_{\mathcal{L}} = l(Y,Y')$. From here, the connection between the expansion of the distance covariance in (2.89) and (3.44) becomes straightforward. For a complete proof through the use of semimetrics, the reader is referred to [Sej+13]. However, by realizing that the Euclidean norm is a form of a linear kernel function, the equivalence becomes a clear case of the kernel "trick". Here, the expression in (3.44) generalizes the concept by letting k and l be any kernel function, but the underlying measure remains equivalent. In essence, the following expression states a generalized distance correlation measure:

$$\mathcal{R}^2(X;Y) = \frac{\left\|\text{Cov}\left\{\langle\phi_X(X), g_X\rangle_{\mathcal{K}}, \langle\phi_Y(Y), g_Y\rangle_{\mathcal{L}}\right\}\right\|_{\text{HS}}^2}{\left\|\text{Var}\left\{\langle\phi_X(X), g_X\rangle_{\mathcal{K}}\right\}\right\|_{\text{HS}}\left\|\text{Var}\left\{\langle\phi_Y(Y), g_Y\rangle_{\mathcal{L}}\right\}\right\|_{\text{HS}}}, \tag{3.45}$$

where the numerator is directly the HSIC, and the original measure of distance correlation can be recovered with the right choice of kernel functions. By following the same steps as in (3.42), we can write

$$\hat{\mathcal{R}}^2(X;Y) = \frac{\frac{1}{L^2}\left\|\mathbf{K}^{1/2}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}^{1/2}\right\|_{\text{HS}}^2}{\frac{1}{L}\left\|\mathbf{K}^{1/2}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{K}^{1/2}\right\|_{\text{HS}}\frac{1}{L}\left\|\mathbf{L}^{1/2}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}^{1/2}\right\|_{\text{HS}}} = \frac{\text{tr}\left(\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{K}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}\right)}{\sqrt{\text{tr}\left(\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{K}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{K}\right)\text{tr}\left(\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{L}\right)}}. \tag{3.46}$$

3.2.5 Normalized cross-covariance operator

Lastly, we will finish the kernel methods for measuring information with the Normalized Cross-Covariance Operator (NOCCO). The core idea of the NOCCO is to provide an alternative to the KCCA with better, and known, consistency [Fuk+07; FBG07]. Similarly with the HSIC, the NOCCO can also be seen as the generalization to the HS norm of the KCCA. Thanks to this, the NOCCO becomes a measure of information rather than independence, with a close link with other well-known measures in the literature.

The NOCCO is defined as follows:

$$\rho_{\text{NOCCO}}(X;Y) = \left\|\text{Corr}\left\{\langle\phi_X(X), g_X\rangle_{\mathcal{K}}, \langle\phi_Y(Y), g_Y\rangle_{\mathcal{L}}\right\}\right\|_{\text{HS}}^2$$

$$= \left\| \frac{\text{Cov}\left\{\left\langle \phi_X\left(x\right), g_X\right\rangle_{\mathcal{K}}, \left\langle \phi_Y\left(y\right), g_Y\right\rangle_{\mathcal{L}}\right\}}{\sqrt{\text{Var}\left\{\left\langle \phi_X\left(x\right), g_X\right\rangle_{\mathcal{K}}\right\} \text{Var}\left\{\left\langle \phi_Y\left(y\right), g_Y\right\rangle_{\mathcal{L}}\right\}}} \right\|_{\text{HS}}^2 . \tag{3.47}$$

Note that this expression is equivalent to (3.20), but with the HS norm instead of the supremum. By operating with the empirical matrices obtained by the finite-dimensional representations of the covariance operators [Fuk+07], similarly as in (3.42), the empirical NOCCO is then

$$\hat{\rho}_{\text{NOCCO}}\left(X;Y\right) = \text{tr}\left(\mathbf{\Gamma}_k \mathbf{\Gamma}_l\right), \tag{3.48}$$

where $\mathbf{\Gamma}_k = \bar{\mathbf{K}}\left(\bar{\mathbf{K}} + \varepsilon\mathbf{I}\right)^{-1}$ and $\mathbf{\Gamma}_l = \bar{\mathbf{L}}\left(\bar{\mathbf{L}} + \varepsilon\mathbf{I}\right)^{-1}$ as in (3.30). The NOCCO can then be seen as a generalization of the KCCA where all the spectrum of some expression with kernel matrices is computed instead of only the largest eigenvalue. Also note that regularization is still needed in order to properly operate with the inverses.

Moreover, it is shown in [FBG07] that the NOCCO is corresponds to the SMI, as it is defined in Subsection 2.3.3. The proof of this relationship is rather complex and out of the scope of this subsection, thus it will not be detailed. However, in [Fuk+07, Thm. 4] the statement is proved through the eigenvalue decomposition of the correlation and covariance operators, and whose result is referred to as the kernel-free integral expression. In particular, the result is that $\rho_{\text{NOCCO}}\left(X;Y\right) = I_{\text{s}}\left(X;Y\right)$ under some assumptions. Furthermore, they also show that the empirical NOCCO is consistent with

$$\lim_{L\to\infty} \hat{\rho}_{\text{NOCCO}}\left(X;Y\right) = \rho_{\text{NOCCO}}\left(X;Y\right). \tag{3.49}$$

Therefore, the SMI becomes the measure of information that is estimated by (3.48). This observation is important since it directly relates a kernel measure with an expression of some probability density functions. While the correlation and covariance in the data space only capture the first moments of a random variable, the correlation and covariance in the infinite-dimensional feature space are also capable of representing the higher-order moments that are required to define complex functions such as a PDF. This will become a core aspect of the thesis as we develop other measure of information based on a mapping of the data onto other, more controlled, feature spaces.

3.2.6 Unified kernel map

Once all the kernel measures of information are defined, we can determine a map of connections between all of them and their significance in relation to other measures of information. Figure 3.1 shows the path followed by the rationale provided in this subsection. Apart from defining the KPCA, which serves to generalize to the KCCA, we have begun by trying to estimate the HGR coefficient by means of kernel methods. This has required to determine a kernelized version of the CCA, based on measuring the maximum correlation (spectral norm) after some mapping to the feature space. Then, the COCO tries to avoid the computational problems associated to the matrix inverses required in KCCA by measuring only the covariance in the feature space. The HSIC generalizes this last concept by measuring and summing up all the covariance coefficients instead of only the largest (HS norm). By doing this, the HSIC gains in granularity and becomes more sensible to potential small covariance coefficients that also denote dependence, an information lost by measuring only the largest one. And finally, we realize that, in order to asses a proper measure of information, the normalization with the marginal variances is required, leading to the NOCCO. Thanks to both measuring all the spectrum and being a proper measure of information, the NOCCO can be uniquely determined as an estimator of the SMI. All the same, we realize that the SMI is intrinsically related to measuring second-order in a high dimensional space. Generally speaking, here we can observe a trend between these statistical methods: only those that are designed as measures of correlation have an information measure counterpart. While a similar relationship has been done with the HSIC and a quadratic measure of information, namely the distance covariance, this last one lacks a strong connection with the MI or its surrogates. As a result, this joint rationale

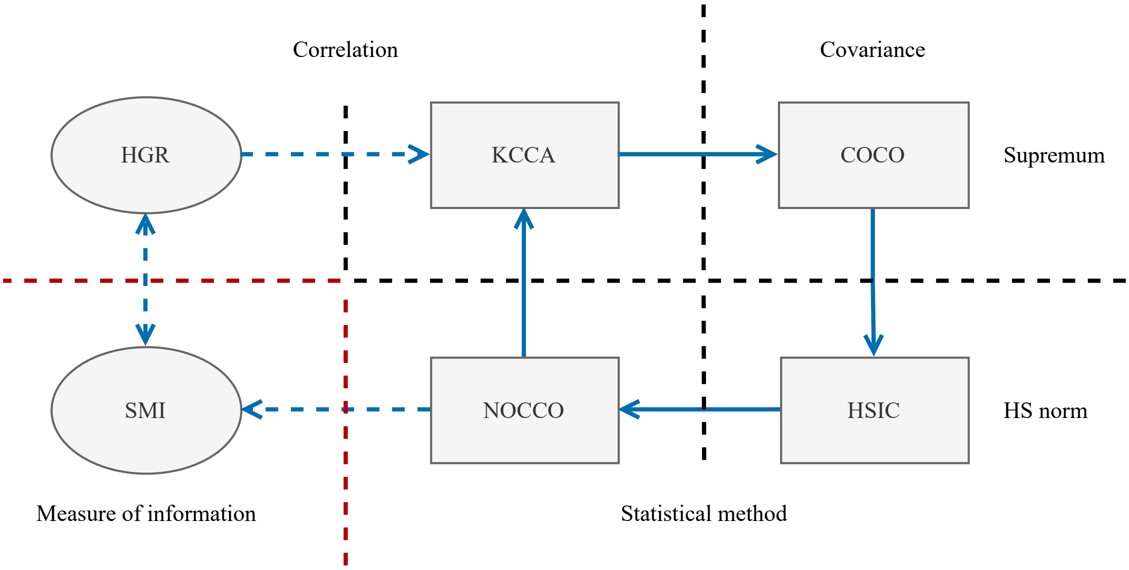


Figure 3.1: Block diagram of the relationship between different kernel measures and their measures of information counterparts.

further suggests that the SMI is the suitable measure if we want to estimate information through second-order statistics.

On an additional note, the connection between the HGR and SMI has not been completely addressed in this subsection since it would require an additional step to all the previous formulations. Specifically, it is neither based on a correlation measure by itself nor it is strictly defined as the HS norm of some operator. The red dashed line indicates these additional required steps, although it can be crossed under some mild assumptions. This issue will be addressed in the following chapter, jointly with a finite-dimensional representation of the feature space. However, the relationship between these two measures can be surmised by realizing that, under some specific matrix operator, the HGR becomes the largest eigenvalue (3.30) and the SMI the sum of all of them (3.48).

3.3 The second-order Rényi entropy

So far, some estimators that rely on the RKHS have been reviewed. Their common goal is to measure some covariance/correlation operator in a Hilbert space of unknown dimensionality thanks to kernel methods. Next, we will see the opposite path. Instead of relying on kernels from the beginning, some special cases lean into kernel methods in a natural manner. In this case, we recover the second-order Rényi entropy due to its expression as the L^2-norm of a PDF, and we will see how this property leads to kernel methods naturally. The derivation of the estimator presented in this section first appeared in [AH84], and the rationale will be based on the work of Príncipe et al. in [Prí10], although we will introduce further concepts in order to broaden the discussion, such as its connection with unbiased statistics. For this reason, we begin by introducing these concepts to enhance the afterward exposition of the estimation of second-order Rényi entropy.

3.3.1 Parzen-Rosenblatt window estimate

A direct procedure for the estimation of information measures is to first estimate the PDFs, and then to estimate the desired expression [MRL95; Bei+97]. These kind of methods are called "plug-in" estimators, as they plug the estimate of the probability function in the functional of interest. The most well-known plug-in estimator is the one referred to as the Kernel Density Estimate (KDE).

However, the definition of a kernel in this algorithm is different to the one given in Definition 5. For this reason, and to not mistake terminology, this estimation methodology will be denominated here as the Parzen-Rosenblatt window estimate [Par62; Ros56]. For clarity, let us first begin by univariate random variables, and generalize to multivariate random variables afterwards.

Definition 9. A window $g_h : \mathbb{R} \to \mathbb{R}$ with bandwidth h is a function that satisfies

1. It is nonnegative $g_h(x) \geq 0, \qquad \forall x \in \mathbb{R}$.
2. It has unit area $\int_{\mathbb{R}} g_h(x)\, \mathrm{d}x = 1$.
3. It is finite with $\lim\limits_{|x| \to \infty} g_h(x) = 0$.

Generally speaking, window functions are considered to be zero-valued outside of some interval. Particularly, it is the bandwidth of the function that determines the interval in which the function is not zero-valued. While this interval may vary from function to function in terms of h, the bandwidth is integrated in the definition as a general approach to bandwidth-limited functions.

Example 9.1. The Gaussian window is defined as

$$g_h(x) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{x^2}{2h^2}\right), \tag{3.50}$$

where the bandwidth coincides with the standard deviation of the Gaussian PDF.

Many more window functions could be defined as examples, but we will limit to the Gaussian window as it is the one that will be mainly used in the derivation of estimators of information.

Consider L i.i.d. samples $\{x(i)\}_{i=0,\ldots,L-1}$ drawn from the random variable $X \in \mathbb{R}$ with probability distribution $f_X(x)$. An estimate of $f_X(x)$ is given by

$$\hat{f}_X(x) = \frac{1}{L} \sum_{i=0}^{L-1} g_h(x - x(i)). \tag{3.51}$$

As can be seen, the estimate is a uniformly weighted average of the window function centered at the samples $x(i)$. For $\hat{f}_X(x)$ to be a PDF, the first two properties from Definition 9 must be ensured. It is also worth noting that we are employing a different definition of the estimator than in [Par62], which is the usual depiction of the Parzen-Rosenblatt window estimate in the literature. The reason to include the bandwidth parameter inside the window function, instead to (3.51), is just for clarity of exposition in forthcoming expressions.

In [Par62] the bias and variance of the estimator are also evaluated, which correspond to

$$\text{Bias}\left\{\hat{f}_X(x)\right\} = \mathbb{E}_{f_X}\left\{\hat{f}_X(x)\right\} - f_X(x) = \frac{h^2}{2}\frac{\partial^2 f_X(x)}{\partial x^2}\int_{\mathbb{R}} z^2 g_1(z)\, \mathrm{d}z + O\left(h^2\right), \tag{3.52}$$

$$\text{Var}\left\{\hat{f}_X(x)\right\} = \mathbb{E}_{f_X}\left\{\left|\hat{f}_X(x) - f_X(x)\right|^2\right\} = \frac{1}{hL} f_X(x) \int_{\mathbb{R}} g_1^2(z)\, \mathrm{d}z + O\left(\frac{1}{hL}\right). \tag{3.53}$$

As can be seen, this is a biased estimator for $h > 0$, as any window function with a given bandwidth smooths the underlying PDF. In particular, it is well-known that the expected estimated probability density is just the convolution of the target density with the window function [Sil86]. To see that relation, let $X_0, \ldots, X_{L-1}$ be L independent random variables that share the same distribution $f_{X_i}(x) = f_X(x)$. Then we have

$$\mathbb{E}_{f_{X_i}}\left\{\hat{f}_X(x)\right\} = \frac{1}{L}\sum_{i=0}^{L-1} \mathbb{E}_{f_{X_i}}\left\{g_h(x - X_i)\right\} = \mathbb{E}_{f_{X_i}}\left\{g_h(x - X_i)\right\} \tag{3.54a}$$

$$= \int_{\mathbb{R}} g_h\left(x - X_i\right) f_X\left(X_i\right) \mathrm{d}X_i = \left(g_h * f_X\right)(x)\,. \tag{3.54b}$$

The expected PDF then corresponds to the true one $f_X(x)$ for $h = 0$, since (3.54) is then equivalent to the convolution with a Dirac delta function $\delta(x)$ and given that $(\delta * f_X)(x) = f_X(x)$. This result is in line with what Parzen wanted to achieve in [Par62], which was an asymptotically unbiased estimator for $L \to \infty$ at the same time as $h \to 0$. However, in practical applications, h cannot be a null (or too small) value, since the number of available data is limited, which implies that Var$\{\hat{f}_X(x)\}$ would increase infinitely. As a result, the bias-variance relationship is a trade-off that depends primarily on the window bandwidth. On the positive side, Bias$\{\hat{f}_X(x)\}$ tends to zero quadratically with h for $h \to 0$, while Var$\{\hat{f}_X(x)\}$ increases linearly, hence small values of h are still preferred. Numerous attempts to solve this conundrum have been provided in the literature [Che15]. The most used and accepted is possibly the Silverman's rule of thumb [Sil86]. Concretely, this rule states that, for Gaussian data with a Gaussian kernel, the optimal choice of kernel bandwidth is

$$h_\text{S} = \hat{\sigma}\left(\frac{4}{3L}\right)^{1/5}, \tag{3.55}$$

where $\hat{\sigma}$ is the empirical standard deviation of the data $\{x(i)\}_{i=0,\ldots,L-1}$. This bandwidth is optimized by minimizing the Mean Integrated Squared Error (MISE), defined as

$$\text{MISE}\left\{\hat{f}_X(x)\right\} = \int_{\mathbb{R}} \left(\hat{f}_X(x) - f_X(x)\right)^2 \mathrm{d}x. \tag{3.56}$$

Although this optimal bandwidth is derived under the Gaussian assumption, it is generally used as a recurrent bandwidth with generally good performance for Gaussian windows. For that, it is a rule of thumb that avoids complex computations for different underlying distributions.

For N-variate random variables $X \in \mathcal{X}$ with $\mathcal{X} \subseteq \mathbb{R}^N$, one has just to take multivariate window functions such that

$$\hat{f}_X(\mathbf{x}) = \frac{1}{L}\sum_{i=0}^{L-1} g_\mathbf{H}\left(\mathbf{x} - \mathbf{x}(i)\right) \tag{3.57}$$

for $\mathbf{x}(i) = [x_1(i), ..., x_N(i)]^T$ and $\mathbf{H} \in \mathbb{R}^{N\times N}$ being the symmetric and positive definite bandwidth matrix. In this case, the multivariate window function can also be easily defined by generalizing the properties from Definition 9 to a multivariate function. Generally speaking, isotropic Gaussian windows are typically used. These are of the form

$$g_\mathbf{H}(\mathbf{z}) = \frac{1}{(2\pi)^{N/2}\left|\mathbf{H}\right|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{H}^{-1}\mathbf{z}\right), \tag{3.58}$$

where $\mathbf{H} = h^2\mathbf{I}_N$, which can be simplified to

$$g_h(\mathbf{z}) = \frac{1}{(2\pi)^{N/2} h^N} \exp\left(-\frac{\|\mathbf{z}\|^2}{2h^2}\right). \tag{3.59}$$

In the multivariate case, the relationship between the kernel bandwidth and the bias is maintained as in (3.52), but the variance decreases exponentially with N [Sil86, Sec. 4.3] such that

$$\text{Var}\left\{\hat{f}_X(\mathbf{x})\right\} = \frac{1}{h^N L} f_X(\mathbf{x}) \int_{\mathbb{R}^N} g_\mathbf{H}^2(\mathbf{z})\,\mathrm{d}\mathbf{z} + O\left(\frac{1}{h^N L}\right), \tag{3.60}$$

which maintains the desired bias-variance trade-off. Lastly, by following similar steps to those in (3.56), Silverman's rule is then as follows:

$$h_\text{S} = \hat{\sigma}\left(\frac{4}{(2N+1)\,L}\right)^{\frac{1}{N+4}}, \tag{3.61}$$

where $\hat{\sigma} = \sum_{n=1}^{N} \hat{\sigma}_n / N$ is the average marginal standard deviation, with $\hat{\sigma}_n$ being the sample estimator of the n-th variable. These notions will become useful for defining estimators of information of multivariate random variables.

3.3.2 Unbiased statistics

Another important concept that will naturally emerge from the estimator of entropy is the class of unbiased statistics or, in short, U-statistics. This class of statistics was introduced by Hoeffding in [Hoe48], where he was concerned about characterizing the statistics that are asymptotically normal under a given structure. From a general point of view, this class is a generalization of the sample average that allows for the derivation of minimum-variance unbiased estimators. For this reason, this class of statistics is usually employed in nonparametric estimation [Ser09].

Consider L i.i.d. observations $x\left(i\right)$ of the random variable X with density function $f_X\left(x\right)$ for $i=0,...,L-1$, and consider the parameter θ for which there is an unbiased estimator of the form

$$\theta=\mathbb{E}_{f_X}\left\{g\left(x\left(i_1\right),...,x\left(i_N\right)\right)\right\} \tag{3.62}$$

for some symmetric function g [1], that is

$$g\left(z_1,...,z_N\right)=\frac{1}{N!}\sum_{P_N}g\left(z_{p_1},...,z_{p_N}\right), \tag{3.63}$$

where the summation is over the set P_N of all permutations of the N-dimensional vector in the argument. Again, this notion can be generalized to multivariate random variables, but we will remain with the univariate case for clarity of exposition.

Definition 10. Let $\left\{x\left(i\right)\right\}_{i=0,..,L-1}$ have the same distribution, and consider a symmetric function $g\left(z_1,...,z_N\right)$ with $N\leq L$. The U-statistic of the parameter θ is given by the average of the function g over all the observations:

$$U=\frac{1}{\binom{L}{N}}\sum_{P_{L,N}}g\left(x\left(i_1\right),...,x\left(i_N\right)\right), \tag{3.64}$$

where the summation is over the set $P_{L,N}$ of all possible combinations over the binomial coefficient $\binom{L}{N}$.

Example 10.1. (Sample mean) Consider $N=1$ and the function $g\left(z\right)=z$. The U-statistic of g is the sample mean:

$$U=\frac{1}{L}\sum_{i=0}^{L-1}x(i). \tag{3.65}$$

Example 10.2. (Sample variance [Ser09]) Consider $N=2$ and $g\left(z_1,z_2\right)=\frac{1}{2}\left(z_1-z_2\right)^2$. The U-statistic of g is the sample variance:

$$U=\frac{2}{L(L-1)}\sum_{i_1=0}^{L-1}\sum_{\substack{i_2=1\\i_2>i_1}}^{L-1}\frac{1}{2}\left(x(i_1)-x(i_2)\right)^2=\frac{1}{L(L-1)}\sum_{0\leq i_1<i_2\leq L-1}\left(x(i_1)-x(i_2)\right)^2 \tag{3.66a}$$

$$=\frac{1}{L-1}\sum_{0\leq i_1\leq L-1}\left(x(i_1)-\frac{1}{L}\sum_{0\leq i_2\leq L-1}x(i_2)\right)^2. \tag{3.66b}$$

While it can be shown the suitability of these statistics in terms of their asymptotic Gaussian distribution [Ser09], we are just interested in the fact that they are, indeed, unbiased estimates. In

[1] Typically, g is also referred to as a kernel. In order to not overlap definitions and properties, here we will not give g a proper name since it will be easily recognizable under a given context. Nonetheless, any kernel can be used as a function g for any U-statistic, specially due to its symmetry property.

this sense, this definition will be used for identifying appropriate estimators that have a similar structure, and to exploit their properties.

For instance, one of its main properties comes in the form of the "incomplete" U-statistics. The *incomplete* adjective refers to the action of sampling fewer terms than the complete version of U-statistic. Let us define the incomplete U-statistics as follows:

Definition 11. Let U be a U-statistic with symmetric function $g\left(z_1, ..., z_N\right)$ over L i.i.d. observations $\left\{x\left(i\right)\right\}_{i=0,..,L-1}$:

$$U = \frac{1}{\binom{L}{N}} \sum_{P_{L,N}} g\left(x\left(i_1\right), ..., x\left(i_N\right)\right) \tag{3.67}$$

An incomplete U-statistic, denoted U_{i}, with the same symmetric function g, is given by

$$U_{\mathrm{i}} = \frac{1}{\binom{M}{N}} \sum_{P_{M,N}} g\left(x\left(i_1\right), ..., x\left(i_N\right)\right), \tag{3.68}$$

where $N \leq M < L$.

As can be deduced, the incomplete U-statistics are also unbiased, and the cost of fewer samples is an increase of variance with $\mathrm{Var}\left\{U_{\mathrm{i}}\right\} \geq \mathrm{Var}\left\{U\right\}$. Particularly, for $\mathrm{Var}\left\{g\right\} = \sigma_g^2$, if the subsamples are selected randomly from the pool of all available samples, then [Blo76]

$$\mathrm{Var}\left\{U_{\mathrm{i}}\right\} = \frac{\sigma_g^2}{M} + \left(1 - \frac{1}{M}\right) \mathrm{Var}\left\{U\right\}. \tag{3.69}$$

Generally speaking, the main advantage of U_{i} is the reduced computational complexity, as all the combinations $P = \binom{L}{N}$ increase the number of operations exponentially with $O\left(L^N\right)$ [CK19]. However, these kind of statistics may also shine for nonrandomized sampling. A potential case of use of this approach is a random process that is not i.i.d., since it then becomes sensitive to a random subsampling. The asymptotic behaviour in these cases has also been studied in [Jan84]. While these notions may not directly influence the estimation of information from this subsection, their utility will become apparent when employing the entropy estimator in some particular cases in Section 5.4.

3.3.3 Information potential

Let us take the second-order Rényi entropy from (2.53) and express it as

$$h_2\left(X\right) = -\ln \int_{\mathcal{X}} f_X^2\left(x\right) \mathrm{d}x = -\ln V\left(X\right). \tag{3.70}$$

The argument of the logarithm is called the Information Potential (IP) [Prí10], and it is just the expectation over the PDF

$$V\left(X\right) = \mathbb{E}_{f_X}\left\{f_X\left(x\right)\right\}. \tag{3.71}$$

For estimation purposes, we can just compute the IP and then measure the uncertainty, since the second-order Rényi entropy is a monotonic function of the IP. As it turns out, this particular case (from all the Rényi entropies) copes particularly well with the Parzen-Rosenblatt window estimate. To see that, let us take the estimator from (3.51) and plug-in to the IP expression. Consider the samples $\{x(i)\}_{i=0,...,L-1}$ drawn from the random variable $X \in \mathcal{X}$ with $\mathcal{X} \subseteq \mathbb{R}$. The plug-in window estimate of the IP is then as follows:

$$\hat{V}\left(X\right) = \int_{\mathbb{R}} \hat{f}_X^2\left(x\right) \mathrm{d}x = \int_{\mathbb{R}} \frac{1}{L} \sum_{i=0}^{L-1} g_h\left(x - x(i)\right) \frac{1}{L} \sum_{j=0}^{L-1} g_h\left(x - x(j)\right) \mathrm{d}x \tag{3.72a}$$

$$= \frac{1}{L^2} \sum_{i,j=0}^{L-1} \int_{\mathbb{R}} g_h\left(x - x(i)\right) g_h\left(x - x(j)\right) \mathrm{d}x. \tag{3.72b}$$

Assuming that the Gaussian window is used, we then have

$$\hat{V}\left(X\right) = \frac{1}{L^2} \sum_{i,j=0}^{L-1} k_{\sqrt{2}h}\left(x(i), x(j)\right), \tag{3.73}$$

where

$$k_\sigma\left(x(i), x(j)\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x(i) - x(j)\right)^2}{2\sigma^2}\right). \tag{3.74}$$

While this result can be directly obtained by observing that it corresponds to a single point of the convolution between two Gaussian functions, which is also Gaussian, the full derivation of (3.73) is detailed in Appendix 7.2.1. As can be seen, this procedure completely avoids the need of estimating the PDF and measuring the integral. It allows then for a direct estimation solely based on pairwise differences of the data. As the dissimilarity increases, i.e. the data is more spread out, the IP decreases and the second-order Rényi entropy (thus the uncertainty measure) increases, and vice versa.

Furthermore, note that the notation of the associated function has changed from g to k. This is due to k_σ not being a window function, but a kernel. In this regard, it can be easily seen that $k_\sigma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which is also a symmetric and positive definite function. In fact, (3.74) is just a normalized (to unit area) version of the Gaussian kernel from (3.4). Due to the kernel *trick*, we can say that (3.73) is intrinsically computing some norm in an infinite-dimensional Hilbert space by evaluating the relationship among input data through a kernel function[2]. Due to the exponential form of the window g_h, the constructed kernel is specifically shift-invariant, which entails further kernel properties, such as a one to one relationship with a PDF. The definition and repercussions of these types of kernels will be particularly addressed in the next chapter.

The main difference between the kernel methods presented in Section 3.2 and (3.73) is that in the latter the norm in the infinite-dimensional space has been computed explicitly. Both approaches fulfil the kernel signal processing tag by coming from opposite ways. In the classical approach, an inner product in an unknown feature space is interchanged with a kernel. In the approach presented here, the kernel is attained after the inner product has been calculated. This two-way relationship allows to locate this special case of entropy estimation within the RKHS framework. In fact, in [Xu+08] is is shown that this framework can be generalized to the set of square-integrable PDFs with a given L^2 norm.

To further see this relation, let us recover the reproducing property

$$k_\sigma\left(x(i), x(j)\right) = \left\langle \phi\left(x(i)\right), \phi\left(x(j)\right)\right\rangle_{\mathcal{H}}. \tag{3.75}$$

Then, we may express

$$\hat{V}\left(X\right) = \frac{1}{L^2} \sum_{i,j=0}^{L-1} \left\langle \phi\left(x(i)\right), \phi\left(x(j)\right)\right\rangle_{\mathcal{H}} = \left\langle \frac{1}{L} \sum_{i=0}^{L-1} \phi\left(x(i)\right), \frac{1}{L} \sum_{j=0}^{L-1} \phi\left(x(j)\right) \right\rangle_{\mathcal{H}} = \left\|\hat{\mu}_{\phi X}\right\|_2^2, \tag{3.76}$$

where $\hat{\mu}_{\phi X}$ is the mean of the mapped data in the feature space. Therefore, the estimation of entropy is closely related to a measure of the first-order statistics in an infinite-dimensional feature space.

Due to the symmetry of the kernel function, as well as that $k_{\sqrt{2}h}\left(x(i), x(j)\right)$ is a constant data-independent value for $i = j$, we can further reduce the summation of samples such that

$$\hat{U}\left(X\right) = \frac{2}{L(L-1)} \sum_{0 \leq i < j \leq L-1} k_{\sqrt{2}h}\left(x(i), x(j)\right). \tag{3.77}$$

[2] Generally speaking, other window functions could be used instead, but the evaluation of the integral is not as convenient as with Gaussian functions.

Its relation with (3.73) is then given by

$$\hat{V}\left(X\right) = \frac{L-1}{L}\hat{U}\left(X\right) + \frac{1}{L}\left(4\pi h^2\right)^{-1/2}. \tag{3.78}$$

In (3.77), while the $L(L-1)$ terms indicate the number of elements that correspond to $i < j$, the 2 in the numerator just replicates these elements such that it includes $i > j$. For $L \to \infty$, both estimators are asymptotically the same measure, since $(L-1)/L \to 1$ and the data-independent second term tends to 0. However, we are more interested in $\hat{U}\left(X\right)$ as it has a distinguishable expression. Concretely, $\hat{U}\left(X\right)$ is actually a U-statistic, as it is defined in Definition 10, with $g\left(x(i), x(i)\right) = k_{\sqrt{2}h}\left(x(i), x(j)\right)$ and $M = 2$. Therefore, $\hat{U}\left(X\right)$ is an unbiased estimate of $\mathbb{E}_{f_X}\{k_{\sqrt{2}h}\left(x(i), x(j)\right)\}$. However, it is noteworthy that the expected outcome of the estimator, $\mathbb{E}_{f_X}\{\hat{U}\left(X\right)\}$, is *not* the true IP. This is to be expected since, as we have started with the Parzen-Rosenblatt estimator, we have added some Gaussian contamination from the beginning. In [Prí10] the bias of the estimator is computed [3] and it results in

$$\text{Bias}\left\{\hat{U}\left(X\right)\right\} \approx h^2 \mathbb{E}_{f_X}\left\{\frac{\partial^2 f_X\left(x\right)}{\partial x^2}\right\}. \tag{3.79}$$

Similarly as with the Parzen-Rosenblatt estimator, if we let $h \to 0$ at the same time as $L \to \infty$, then $\hat{V}\left(X\right)$ and $\hat{U}\left(X\right)$ are also asymptotically unbiased estimates of $V\left(X\right)$. Even so, for known $f_X\left(x\right)$, it is still possible to compute the true unbiased estimate $\mathbb{E}_{f_X}\{k_{\sqrt{2}h}\left(x(i), x(j)\right)\}$. This knowledge can be particularly useful for applications where the estimation of entropy is not the main focus, but to measure some parameter that may depend on it. For example, let $X \in \mathcal{X}$ be a univariate random variable with PDF $f_X\left(x, \gamma\right)$, where γ is some unknown parameter that controls its shape, location, etc. Then, the known expected value of the IP estimator becomes

$$\mathbb{E}_{f_X}\left\{k_{\sqrt{2}h}\left(x(i), x(j)\right)\right\} = \int_{\mathcal{X}}\int_{\mathcal{X}} k_{\sqrt{2}h}\left(x, x'\right) f_X\left(x, \gamma\right) f_X\left(x', \gamma\right) \mathrm{d}x\mathrm{d}x' = g\left(h, \gamma\right). \tag{3.80}$$

If the previous integral can be solved, then g is known and only depends on the kernel bandwidth and on the parameter to be estimated. If g has inverse function, then one can recover the parameter γ through just an entropy estimate. We will see some particular applications of this rationale in Chapter 5.

While the asymptotic case of $h \to 0$ provides the rationale for unbiasedness, we can do an equivalent asymptotic enquiry for $h \to \infty$. In particular, for an increasing kernel bandwidth, the estimator unveils an inherent dependence with the sample variance estimator. To see this relation, let us take $k_{\sqrt{2}h}\left(x(i), x(j)\right)$ from (3.77) and make a change of variable $\delta^2 = \left(x(i) - x(j)\right)^2/4h^2$ such that

$$k_{\sqrt{2}h}\left(x(i), x(j)\right) = \frac{1}{\sqrt{4\pi h^2}}\exp\left(\delta\right) = \frac{\delta}{\sqrt{\pi\left(x(i) - x(j)\right)^2}}\exp\left(-\delta^2\right). \tag{3.81}$$

By assuming that the kernel bandwidth is very large, hence δ goes to zero, we can approximate $k_{\sqrt{2}h}\left(x(i), x(j)\right)$ by its Taylor expansion at the point $\delta = 0$, i.e. its Maclaurin series. Then we have

$$k_{\sqrt{2}h}\Big|_{\delta=0} = 0, \qquad \frac{\partial}{\partial\delta}k_{\sqrt{2}h}\Big|_{\delta=0} = \frac{1}{\sqrt{\pi\left(x(i) - x(j)\right)^2}}, \tag{3.82a}$$

$$\frac{\partial^2}{\partial\delta^2}k_{\sqrt{2}h}\Big|_{\delta=0} = 0, \qquad \frac{\partial^3}{\partial\delta^3}k_{\sqrt{2}h}\Big|_{\delta=0} = \frac{-6}{\sqrt{\pi\left(x(i) - x(j)\right)^2}}, \tag{3.82b}$$

[3] In [Prí10] the bias is actually measured for $\hat{V}\left(X\right)$. However, since $\hat{V}\left(X\right)$ is asymptotically $\hat{U}\left(X\right)$ for $L \to \infty$, then we can say that the bias is approximately the same.

which yields the following second-order Taylor expansion:

$$k_{\sqrt{2}h}\left(x(i),x(j)\right) \approx \frac{2}{L(L-1)} \sum_{0\leq i<j\leq L-1} \left(\frac{1}{\sqrt{4\pi h^2}} - \frac{\left(x(i)-x(j)\right)^2}{8h^3\sqrt{\pi}} \right) \tag{3.83a}$$

$$= \frac{1}{\sqrt{4\pi h^2}} - \frac{\left(x(i)-x(j)\right)^2}{8h^3\sqrt{\pi}}. \tag{3.83b}$$

By plugging-in this approximation into the IP estimator in (3.77), we then have

$$\hat{U}\left(X\right) \approx \frac{2}{L(L-1)} \sum_{0\leq i<j\leq L-1} \left(\frac{1}{\sqrt{4\pi h^2}} - \frac{\left(x(i)-x(j)\right)^2}{8h^3\sqrt{\pi}} \right) \tag{3.84a}$$

$$= \frac{1}{\sqrt{4\pi h^2}} - \frac{1}{L(L-1)4h^3\sqrt{\pi}} \sum_{0\leq i<j\leq L-1} \left(x(i)-x(j)\right)^2. \tag{3.84b}$$

From the U-statistics standpoint, the sum is then directly the sample variance, as in Example 10.2. Consequently, for very large kernel bandwidth, the IP estimator is a biased and scaled estimator of the sample variance, but a second-order statistic nonetheless. It is worth noting that the first term corresponds to the IP of a Gaussian variable [4] with variance h^2. This observation shows that, if the variance of the contamination is sufficiently high to conceal the underlying true PDF (from the Parzen-Rosenblatt perspective), the resulting IP is governed by the IP of this contamination. While an expected result, here it is appropriately described. A similar analysis can be encountered in [Prí10], albeit for a generic kernel function with unknown derivatives.

The generalization to multivariate random variables can be obtained directly from the univariate case, if the proper window function is used. Let $\mathbf{x}(i) = [x_1(i), ..., x_N(i)]^T$ be L i.i.d. samples from the set $\mathcal{X} \subseteq \mathbb{R}^N$ with $i = 0, ..., L-1$. By taking (3.57) as the density estimate with an isotropic Gaussian window (3.59) and variance h_G^2 from (3.61), then the multivariate IP estimator just becomes

$$\hat{V}\left(X\right) = \frac{1}{L^2} \sum_{i,j=0}^{L-1} k_{\sqrt{2}h}\left(\mathbf{x}(i),\mathbf{x}(j)\right). \tag{3.85}$$

Similarly, due to the properties of kernels, one can define the U-statistic for multivariate random variables as

$$\hat{U}\left(X\right) = \frac{2}{L(L-1)} \sum_{0\leq i<j\leq L-1} k_{\sqrt{2}h}\left(\mathbf{x}(i),\mathbf{x}(j)\right). \tag{3.86}$$

To end with this section, we will further simplify the IP estimates by taking advantage of kernel matrices (Definition 7). First, in (3.85) it can be easily seen that it corresponds to the summation of all the terms of the kernel matrix $\mathbf{K} \in \mathbb{R}^{L\times L}$ with entries $[\mathbf{K}]_{i,j} = k_{2h^2\mathbf{I}}\left(\mathbf{x}(i),\mathbf{x}(j)\right)$. Then, we can express the estimator as

$$\hat{V}\left(X\right) = \frac{1}{L^2}\mathbf{1}^T\mathbf{K}\mathbf{1}. \tag{3.87}$$

The U-statistic version can be easily obtained by again noticing the symmetry of the Gram kernel matrix. As we only need to compute the cases for which $i < j$, or vice versa, we can express (3.86) as follows:

$$\hat{U}\left(X\right) = \frac{2}{L(L-1)}\mathbf{1}^T\left(\mathbf{K}\odot\mathbf{L}\right)\mathbf{1}, \tag{3.88}$$

[4] The proof is omitted here because it will be derived later on. Specifically, it can be obtained through the Appendix 7.3.5 by looking at the argument of the natural logarithm in the case of a single Gaussian component.

where $\mathbf{L}$ is a $L \times L$ lower triangular matrix with ones in its entries below the diagonal, and the diagonal elements equal to zero:

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \cdots & 1 & 0 \end{bmatrix}. \tag{3.89}$$

This estimate of the second-order Rényi entropy is actually related to the KPCA. The fact that we work with the same kernel matrix allows us to make connections that go beyond the qualitative link. Let $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H$ be the eigendecomposition of the kernel matrix. The estimator in (3.87) then becomes

$$\hat{V}(X) = \frac{1}{L^2}\mathbf{1}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H\mathbf{1} = \frac{1}{L^2}\sum_{i=0}^{L-1}\lambda_i\left(\mathbf{u}_i^H\mathbf{1}\right)^2, \tag{3.90}$$

where $\mathbf{u}_i$ are the eigenvectors and λ_i the eigenvalues. Consequently, the estimator of second-order Rényi entropy is composed of projections onto the directions obtained by the KPCA [Jen09]. Furthermore, it can be shown that by performing KPCA, one is minimizing the IP, thus maximizing the second-order Rényi entropy [PXP06]. All these notions and connections will become apparent in the next chapter.

3.3.3.1 A remark on the information potential bias

A final remark on the bias induced on the IP estimate is in order. In view of (3.54), the resulting estimator

$$\hat{U}_{h^2}(X) = \frac{2}{L(L-1)}\sum_{0\leq i<j\leq L-1} k_{\sqrt{2}h}(x(i), x(j)) \tag{3.91}$$

is, in fact, estimating

$$\mathbb{E}_{f_X}\left\{\hat{U}_{h^2}(X)\right\} = \int_{\mathbb{R}}\left(\left(g_h * f_X\right)(x)\right)^2 \mathrm{d}x. \tag{3.92}$$

Note that the subindex term, which corresponds to the kernel variance h^2, has been added to highlight the relation between the contamination and the estimator, for reasons that will become apparent. The interpretation [5] is that the measured IP has an average value that corresponds to the IP of the original random variable at the output of an additive Gaussian noise channel with noise variance h. A natural question then arises on how sensitive is the IP of a random variable in front of a Gaussian perturbation. As a matter of fact, the approximation in (3.79) suggests that the second derivative (the rate of change) of the original PDF has a prominent role in the amount of bias, and the obtained expression resembles a negative Fisher-like expression, albeit evaluated on the PDF itself instead of on the logarithm of the likelihood. To see that clearly, let us express the normalized bias (in terms of h^2) in (3.79) such that

$$\frac{\mathbb{E}_{f_X}\left\{\hat{U}_{h^2}(X)\right\} - U(X)}{h^2} \approx \mathbb{E}_{f_X}\left\{\frac{\partial^2}{\partial x^2}f_X(x)\right\}. \tag{3.93}$$

The left side of the equation is, in fact, the derivative with respect to h^2 for $h^2 \to 0$. Therefore, we can write

$$\lim_{h^2\to 0}\frac{\mathbb{E}_{f_X}\left\{\hat{U}_{h^2}(X)\right\} - V(X)}{h^2} = \frac{\partial}{\partial h^2}\mathbb{E}_{f_X}\left\{\hat{U}_{h^2}(X)\right\}\bigg|_{h^2=0} \approx \mathbb{E}_{f_X}\left\{\frac{\partial^2}{\partial x^2}f_X(x)\right\}. \tag{3.94}$$

[5] All these notions will be specified later on, in Chapter 4.

It is worth noting that the previous expression is always negative, given that a contamination process decreases the value of IP (increases the second-order Rényi entropy).

Following the rationale, it is well-known that Shannon entropy sensitivity to the variance of an additive Gaussian perturbation is directed related to the Fisher information through the well-known de Bruijn identity [Sta59]. These observations motivate the study of whether a similar relationship exists for other entropic measures. This is the work performed in our publication [RC19], where both the classical Fisher and a Fisher-like measures are provided under a unified view based on the generalization of Price's theorem. While the development of this theoretical line is outside the scope of this dissertation, the main results resonate into multiple facets of this section. Therefore, we will outline some of the most important aspects in [RC19] to gain intuition of the relationship between the Gaussian perturbation and the IP estimate.

Consider the random variable X with finite second-order moment, $Z \sim \mathcal{N}(0,1)$, and g is any memoryless nonlinear transformation. The additive Gaussian noise model is then defined as $Y = X + \sqrt{\eta}Z$ with PDF $f_{Y|\eta}(y)$ on the set $\mathcal{Y}$, where $\eta > 0$ is the *variance*[6] of the contamination. A generalization of Price's theorem [RC19] is given by

$$\frac{\partial}{\partial \eta}\mathbb{E}_{f_{Y|\eta}}\left\{g\left(X + \sqrt{\eta}Z\right)\right\} = \frac{1}{2}\mathbb{E}_{f_{Y|\eta}}\left\{\frac{\partial^2}{\partial y^2}\, g\left(y\right)\Big|_{y=X+\sqrt{\eta}Z}\right\}. \tag{3.95}$$

The result is that the sensitivity of the system output $g(Y)$ in front of the Gaussian perturbation can be computed in terms of the second derivative of the output itself. The relevance of this version of Price's theorem is that it can actually be related to an entropic measure, or, more generally, to a measure of information, which then further establishes a connection with de Bruijn identity. For that, consider the IP of Y such that

$$V\left(Y\right) = V\left(X + \sqrt{\eta}Z\right) = \int_{\mathcal{Y}} f^2_{Y|\eta}\left(y\right) \mathrm{d}y. \tag{3.96}$$

Using the chain rule, we can view the derivative of the contaminated IP with respect to η from the perspective of the generalized Price's theorem such that

$$\frac{\partial}{\partial \eta}V\left(X + \sqrt{\eta}Z\right) = 2\int_{\mathcal{Y}} f_{Y|\eta}\left(y\right)\frac{\partial}{\partial v}f_{Y|\eta}\left(y\right) \mathrm{d}y. \tag{3.97}$$

Following similar steps than those in [RC19], the resulting memoryless transformation is defined as

$$g\left(y\right) = -2f_{Y|\eta}\left(y\right). \tag{3.98}$$

Finally, in view of the previous expressions, we can derive the final Bruijn-like identity for the IP:

$$\frac{\partial}{\partial \eta}V\left(X + \sqrt{\eta}Z\right) = -J_{IP}\left(X + \sqrt{\eta}Z\right), \tag{3.99}$$

where $J_{IP}(Y)$ is a Fisher-like information associated to the PDF of Y with respect to a location parameter such that

$$J_{IP}\left(Y\right) = -\mathbb{E}_{f_{Y|\eta}}\left\{\frac{\partial^2}{\partial y^2}f_{Y|\eta}\left(y\right)\right\}. \tag{3.100}$$

The negative sign in (3.99) is required to obtain an expression that is in accordance with (3.94). Similarly to (3.94), the previous expressions show that the sensitivity of the IP to a Gaussian perturbation, or Gaussian smoothing, depends on the shape of the resulting PDF. The more disrupted distributions by the Gaussian contamination are then those with strong curvatures in the shape of

[6] Note that h in (3.91) refers to the standard deviation, or kernel bandwidth, while here we define η as the variance. This is done in order to comply with Price's theorem [Pri58], which will be used in subsequent derivations. However, we can ultimately recover the kernel bandwidth perspective by the change of variable $h = \sqrt{\eta}$.

the PDF. For example, “sharp” distributions, or with fine details, are more susceptible to contamination. This relationship is then translated to the estimation of the IP, with similar results. Given that estimating the IP through (3.91) is equivalent to estimating the IP of the additive Gaussian channel $Y = X + \sqrt{\eta}Z$, the bias induced by this implicit contamination is directly influenced by the shape of the PDF itself. In comparison with (3.94), the equation in (3.99) is expressed in terms of the contaminated PDF $f_{Y|\eta}\left(y\right)$, and not $f_X\left(x\right)$. Nonetheless, the equivalence is ensued if we let $\eta \to 0$, given that $f_{Y|\eta}\left(y\right)\big|_{\eta=0} = f_X\left(x\right)$, similarly to how we let $h^2 \to 0$. It is also worth mentioning that this problem is also studied by Valero-Toranzo et al. in [VZB17] for the case of the ϕ-entropies, where they propose a generalization of de Bruijn identity for multivariate random variables and with any kind of noise, and not only Gaussian perturbations.

3.4 General remarks

This chapter has discussed some kernel-based measures in the literature that are connected to information-theoretical descriptors. In particular, we realize that, by estimating correlation and covariance in a given feature space, those measures that are expressed as second-order moments of some functional arise naturally. From the HGR, and interchanging multiple points of view on how to properly deal with the kernel matrices, we have ended up with the SMI, as can be seen in Figure 3.1. Hence, the rationale provided in Chapter 2 is strengthened.

However, kernel methods are still susceptible to some undesired properties. The first is the so-called “curse of dimensionality”, addressed in this chapter, which makes any infinite-dimensional-based mapping prone to sparsity problems, a trademark of kernel methods. To avoid computing inverses is, in summary, to not estimate a proper dependence measure. The required regularization is then performed blindly since kernel methods do not require to directly operate in the feature space. The second issue is their computational complexity. As the quantity of available data increases, kernel methods tend to be computationally prohibitive [SS00; BJ02].

Apart from that, the estimation of the second-order Rényi entropy has been addressed from the perspective of a plug-in estimate. This entropy surrogate estimator is a particular case that directly leads to kernels. Unlike the kernel methods built from the ground up, this estimator has a known feature map, which provides some advantages with respect to other kernel-based derivations. While regularization is still conducted, which is embedded in the kernel bandwidth parameter, no inverses are required and the repercussion of this regularization is manageable. Nonetheless, the computational complexity issue is still present.

In fact, there has been an active pursuit of techniques that tackle the problem of regularization, but especially so the problem of computational complexity. While the first is usually performed by some kind of penalization to the weight vector in feature space (see [SSB+02], Chapter 4 and references therein), which ends up reassembling the Tikhonov regularization as seen in (3.27) and (3.29), several attempts have been performed to cope with the latter. Most of them are focused on substituting the kernel matrix with low-rank approximations. For example, by means of the incomplete Cholesky decomposition [SC04; SP09], or by using the Nyström method to approximate the kernel matrix from a subspace spanned by a subset of its columns [WS00; KMT12]. However, these approaches still require computing the kernel matrix, apart from performing the chosen low-rank matrix approximation method. Another approach is studied by Rahimi et al. in [RR07] by mapping the data into a random “low-”dimensional feature space, where the desired operations can be computed to approximate the infinite-dimensional feature space. Recently, following this last method, Li et al. have proposed to cope with the computational complexity issue by defining an explicit feature map of reduced dimension [LP19b; LP20]. Their approach is to construct an *equivalent* kernel with reproducing kernel properties. Therefore, their interest relies on the approximation of the inner product performed by the kernels.

From here, we move to the next chapter with the objective of rightfully addressing the two challenges with kernel methods. Generally speaking, we will first focus on the dimensionality of the mapping, not only to reduce the computational complexity, but also to discuss which is the required dimension of the finite-dimensional feature space with the purpose of estimating information. Moreover, by limiting the feature space the question of the regularization will arise naturally, combining both issues into a single argument.

Chapter 4

Leveraging second-order statistics to measure information

This chapter tackles the estimation of entropy and dependence in a feature space with a higher dimension than the data space by means of second-order statistics. Similarly to the approximation methods pursuing a finite-dimensional space, the one proposed here is also explicit and of finite dimension. This finite dimension will be achieved by sampling a function space, of inherent "infinite" dimension (provided that it is sufficiently dense). In contrast with [RR07] and [LP20], the proposed approach is neither an approximation of the original infinite-dimensional feature space nor its inner product. In fact, our proposal is akin to returning to the primal model of the SVM, where instead of casting the problem to an infinite-dimensional space, we directly stay at the finite-dimensional representation. It is therefore a change of the paradigm, which will be characterized by the outer products in the mapped space, instead of reproducing or approximating inner products. The difference will be apparent because, instead of pairwise data differences, we will encounter pairwise differences in the sampled function space. As a result, the problem gains in interpretability, which further allows for a regularization strategy that benefits from well-known concepts in the field of spectral estimation.

This chapter is structured as follows. Section 4.1 begins by focusing on estimating information for discrete sources. The objective is twofold: to show the fundamental link between the proposed surrogates and second-order statistics, and to unveil the implications of casting the problem to a finite-dimensional feature space, which translates into just a smaller dimension than the source cardinality in the case of discrete sources. Then, Section 4.2 moves the problem into analog sources, first by analyzing the function space to be sampled, and then by assessing the sampling and regularization strategies. Lastly, Section 4.3 particularizes the second-order Rényi entropy and SMI estimators with the tools developed in the previous sections, and their performance is illustrated by computer simulations.

4.1 Second-order statistics in the simplex feature space

The following case studies will be based on information measures that can be expressed as the second-order moment of some relation between distribution functions. These particular cases are, as previously pointed out, the second-order Rényi entropy and the SMI. Here, we will first focus our attention on discrete sources and on the mapping to the feature space, whose dimension will be based on the cardinality of the source. Since our objective is the estimation of measures of information for continuous random variables, we are particularly interested in the case in which the dimension of the mapping is lower than the dimension of the original space, presuming that the required mapping to do so in the analog case is given by a function of implicit infinite dimension. The derivation for discrete random variables will then serve as the bridge to the case of analog sources. In short, this section studies the implications of the mapping and determines an adequate pre-conditioning of the data in pursuit of developing a tool that is capable of estimating information based on second-order statistics.

4.1.1 Collision entropy

Consider that X is a discrete random variable with alphabet $\mathcal{X} = \{x_n\}_{n=1,2,...,N}$ and PMF $p_X(x)$. Let us define the probability column vector as $\mathbf{p} \in \mathbb{R}_+^N$, whose elements are $p_n = \Pr\{X = x_n\} = p_X(x_n)$. Then the second-order Rényi entropy, or collision entropy in the case of discrete sources, can be expressed as follows:

$$H_2(X) = -\ln \sum_{n=1}^{N} p_n^2 = -\ln \|\mathbf{p}\|_2^2 . \tag{4.1}$$

In order to prepare for the forthcoming mapping, let us express $H_2(X)$ in a matrix form as

$$H_2(X) = -\ln \|\mathbf{p}\|_2^2 = -\ln \operatorname{tr}\left(\mathbf{P}\mathbf{P}^T\right) = -\ln \|\mathbf{P}\|_{\mathrm{F}}^2 , \tag{4.2}$$

where $\mathbf{P} = \operatorname{diag}(\mathbf{p}) \in \mathbb{R}_+^{N\times N}$ is a diagonal matrix such that $[\mathbf{P}]_{n,n} = p_n$. The advantage of this formulation is given by the properties of the Frobenius norm. Clearly, since the Frobenius norm is invariant under unitary transformations, it allows for a formulation of the collision entropy by mapping the data to a concrete feature space, provided the mapping is given by a unitary matrix. To see that more clearly, let us move into the estimation of the collision entropy.

Consider a sequence of L i.i.d. observations $x(i) \in \mathcal{X}$ for $i = 0, ..., L-1$. An estimate of $\mathbf{p}$, can be obtained as follows:

$$\hat{\mathbf{p}} = \frac{1}{L}\mathbf{D}\mathbf{1}_L, \tag{4.3}$$

where $\mathbf{D} \in \mathbb{R}^{N\times L}$ is the data matrix with elements

$$[\mathbf{D}]_{n,i} = \mathbb{1}_{\{x(i)=x_n\}}. \tag{4.4}$$

Here, $\mathbb{1}_{\{x(i)=x_n\}}$ is the indicator function and $\mathbf{1}_L$ is a L-dimensional column vector that contains all ones. The data matrix $\mathbf{D}$ is the result of a one-to-one mapping from the observations $x(i)$ to the canonical basis of dimension equal to the set cardinality $|\mathcal{X}|$. That is, column i of matrix $\mathbf{D}$ contains a one in the n-th row and $N-1$ zeros. Consequently, we can write

$$\tilde{\hat{H}}_2(X) = -\ln \left\| \frac{1}{L}\mathbf{D}\mathbf{D}^T \right\|_{\mathrm{F}}^2 , \tag{4.5}$$

where the $\tilde{}$ notation indicates a nonmapped version of the estimator. Given that $\mathbf{D}$ is referred to as the data matrix, then the mass function estimate required in the computation of the collision entropy can be seen as just an autocorrelation matrix

$$\hat{\mathbf{P}} = \frac{1}{L}\mathbf{D}\mathbf{D}^T, \tag{4.6}$$

which leads to the following estimator:

$$\tilde{\hat{H}}_2(X) = -\ln \|\hat{\mathbf{P}}\|_{\mathrm{F}}^2. \tag{4.7}$$

It is worth noting that, for the discrete random variables case, (4.6) just refers to an empirical estimate of the PMF that simply reckons the number of times that an event x_n occurs divided by the total number of observations.

Let us proceed by mapping the data to the feature space, and to measure the uncertainty after the mapping. First, we will consider that the dimensionality of the mapping is the same as the cardinality of the source, which encompasses the desired properties and results. Then, the mapping will be limited, whose impact on the estimate of the collision entropy will be discussed. The following lemma establishes the implications of the first case:

Lemma 4.1. *Let $\mathbf{X} \in \mathbb{C}^{N \times L}$ be the data matrix obtained by $\mathbf{X} = \mathbf{FD}$, where $\mathbf{F} \in \mathbb{C}^{N \times N}$ is a unitary matrix. The estimated collision entropy is then*

$$\hat{H}_2(X) = -\ln \|\hat{\mathbf{R}}\|_F^2 = -\ln \left\| \frac{1}{L}\mathbf{X}\mathbf{X}^H \right\|_F^2 = \hat{\hat{H}}_2(X), \tag{4.8}$$

where $\hat{\mathbf{R}} = \mathbf{X}\mathbf{X}^H/L$ is the autocorrelation matrix of the mapped data.

Proof. The equality can be easily proven by

$$\begin{aligned} \hat{H}_2(X) &= -\ln \left\| \frac{1}{L}\mathbf{X}\mathbf{X}^H \right\|_\mathrm{F}^2 = -\ln \operatorname{tr}\left(\frac{1}{L^2}\mathbf{X}\mathbf{X}^H\mathbf{X}\mathbf{X}^H\right) && \text{(4.9a)} \\ &= -\ln \operatorname{tr}\left(\frac{1}{L^2}\mathbf{F}\mathbf{D}\mathbf{D}^T\mathbf{F}^H\mathbf{F}\mathbf{D}\mathbf{D}^T\mathbf{F}^H\right) = -\ln \operatorname{tr}\left(\mathbf{F}^H\mathbf{F}\hat{\mathbf{P}}\hat{\mathbf{P}}\right) && \text{(4.9b)} \\ &= -\ln \operatorname{tr}\left(\hat{\mathbf{P}}\hat{\mathbf{P}}\right) = -\ln \|\hat{\mathbf{P}}\|_\mathrm{F}^2 = \hat{\hat{H}}_2(X), && \text{(4.9c)} \end{aligned}$$

since $\mathbf{F}^H\mathbf{F} = \mathbf{I}_N$, $\hat{\mathbf{P}}^T = \hat{\mathbf{P}}$, and due to the circularity of the trace. ■

The full-rank case just proves the direct but useful consequence of orthonormal matrices being isometric, thus preserving any given norm. We refer to matrix $\mathbf{F}$ as the mapping matrix since its role is to map the events of the source X onto its own columns. Consequently, the mapping matrix behaves as the *code-book* that contains all possible column vectors $[\mathbf{F}]_{:,n}$. The new mapped data matrix is then used to construct the correlation matrix, whose norm determines the estimate of uncertainty of the random variable X. It is also noteworthy that the data is considered to be real and the mapping is defined in the complex field. Nonetheless, the whole methodology is not restricted to consider otherwise, such as to determine a mapping to the space of real numbers. In our case, the complex mapping matrices are used in preparation of the mapping into the space of the characteristic function.

Thereupon, we analyze how a mapping of lower dimensionality than the cardinality of the source compromises the collision entropy estimator. Let $\mathbf{X} \in \mathbb{C}^{N' \times L}$ be the data matrix obtained by $\mathbf{X} = \mathbf{FD}$ with $N' \leq N$, where $\mathbf{F} \in \mathbb{C}^{N' \times N}$ is a semi-unitary matrix with $\mathbf{F}\mathbf{F}^H = \mathbf{I}_{N'}$ for $N' < N$, and a unitary matrix for $N' = N$. The estimated collision entropy is then

$$\hat{H}_2(X) = -\ln \|\hat{\mathbf{R}}\|_\mathrm{F}^2 = -\ln \left\| \frac{1}{L}\mathbf{X}\mathbf{X}^H \right\|_\mathrm{F}^2, \tag{4.10}$$

where $\hat{\mathbf{R}} = \mathbf{X}\mathbf{X}^H/L$ is the autocorrelation matrix of the mapped data.

Proposition 4.1. *Let $\hat{\hat{H}}_2(X)$ be the collision entropy estimator given in* (4.7)*, and $\hat{H}_2(X)$ the estimator given in* (4.10)*. For $N' \leq N$, the following inequality holds*

$$\hat{H}_2(X) \geq \hat{\hat{H}}_2(X), \tag{4.11}$$

where the equality is met for $N' = N$.

Proof. See Appendix 7.3.1. ■

A sufficient condition for the equality in (4.11) is that $\mathbf{F} = \mathbf{I}_N$, i.e. to map the data onto the orthonormal canonical basis, or that $\mathbf{F}$ is a full-rank unitary matrix, such as the discrete Fourier transform matrix as portrayed in (2.18). The relevance of sufficient condition is, however, encountered in analog sources rather than with discrete sources. In the first case, they are inherently of infinite dimensionality, thus an infinite-dimensional feature map is required, directly appealing to kernel methods and their capabilities of implicitly mapping into infinite-dimensional spaces.

For discrete sources specifically, the inequality is rather observational than a real constraint, but it shows what to expect when moving to analog sources. On a side note, one can see the mapping to a lower dimension than the cardinality of the source related to the problem of *compressed sensing* [Don06; CW08], where the interest relies on finding the adequate orthonormal basis, or code-book as previously mentioned, that better describes the original signal without loosing too much information about the observations themselves. For the case of estimating entropy, this translates to the choice of $\mathbf{F}$, and consequently also N', that allows for the best estimate while mapping to the smallest dimension as possible. Nevertheless, while in compressed sensing the goal is to reconstruct the original signal, here we are only interested the mapping itself, and how it determines the final second-order Rényi entropy estimate. The discussion regarding the dimensionality reduction and its consequences will be properly addressed in Section 4.2.

4.1.1.1 Relation to principal component analysis and kernel methods

Before moving to the estimate of a measure of information, it is of relevance to establish connections to previously addressed methods. To be concrete, the problem of estimating uncertainty is linked with the problem of PCA, and particularly to its kernelized version KPCA. We have previously seen this link in the form of (3.90). To see this link from (4.7), and for clarity of exposition, let us express the estimate of the IP with

$$\hat{\hat{V}}\,(X) = \|\hat{\mathbf{P}}\|_{\mathrm{F}}^2. \tag{4.12}$$

By operating with correlation matrices, this measure is related to the PCA problem from (2.32), albeit here the entire spectrum of the correlation matrix is computed instead of only the largest singular value. This difference is akin to that between the COCO and the HSIC, or between the KCCA and the NOCCO, as shown in Figure 3.1. However, while in (2.32) a covariance matrix is required (or zero-mean data), in (4.12) the estimation is performed through the correlation matrix.

On the other hand, the estimate of the IP by mapping the data onto a feature space is given by

$$\hat{V}\,(X) = \|\hat{\mathbf{R}}\|_{\mathrm{F}}^2. \tag{4.13}$$

Due to the mapping, this approach is more akin to the KPCA from (3.17), given that we are measuring some norm operator of a matrix conformed by the correlation mapped data in the feature space. Nevertheless, the main difference between operating with correlation or covariance is still present, rendering the relevance of the link less useful. While it is still possible that the expectation of the mapped data in the feature space is zero, hence closing the gap between both methods, that is not generally the case. Fortunately, this is only the case for the problem of estimating the IP, and we will see that the equivalent formulation for estimating information does allow for building relationships with its equivalent problem, namely CCA.

Still, the structure of both formulations is sufficiently rich to provide more insights on the problem of estimating the IP. To be concrete, from (4.10), let us express the estimator as follows:

$$\begin{aligned}\hat{V}\,(X) = \|\hat{\mathbf{R}}\|_{\mathrm{F}}^2 = \left\|\frac{1}{L}\mathbf{X}\mathbf{X}^H\right\|_{\mathrm{F}}^2 &= \mathrm{tr}\left(\frac{1}{L^2}\mathbf{X}\mathbf{X}^H\mathbf{X}\mathbf{X}^H\right) &\text{(4.14a)}\\ &= \mathrm{tr}\left(\frac{1}{L^2}\mathbf{X}^H\mathbf{X}\mathbf{X}^H\mathbf{X}\right) = \mathrm{tr}\left(\mathbf{K}_x\mathbf{K}_x\right) = \|\mathbf{K}_x\|_{\mathrm{F}}^2, &\text{(4.14b)}\end{aligned}$$

where $\mathbf{K}_x = \mathbf{X}^H\mathbf{X}/L \in \mathbb{R}^{L\times L}$. Clearly, the previous formulation presents a duality of the problem of estimating the IP thanks to the properties of the Frobenius norm. On the one hand, the expression derived from Lemma 4.1 is constructed by outer products of the mapped data matrices. The problem is then focused on the dimensionality of the space, rather than in the number of data observations. Due to its similarities with the SVM dual representations, we refer to this approach as the primal model. On the other hand, (4.14) is drawn from inner products of the mapped data, which ends up constructing a kernel-like Gram matrix. This time, the matrix dimension is dictated by the data size L. This represents the dual problem, which also restores to kernels as with the SVM dual model.

This rationale exemplifies the difference between kernel methods and the proposed approach, and resonates with the main objectives of this thesis. By controlling the dimensionality of the problem in terms of the feature space dimension, we are not only moving the involved matrices dimension from $L \times L$ to $N \times N$, thus reducing the associated computational complexity for $N < L$, but also the feature space is known and manageable. Furthermore, we can formulate the problem with the reduced dimension N'. The problem then becomes even more computationally convenient at the cost of adding a certain bias in the estimation. However, it is worth noting that the kernel estimate of the IP (as in (3.87)) is also biased in terms of the kernel bandwidth, which is how kernel methods address the regularization of an infinite-dimensional and probably sparse feature space. In conclusion, we are moving the problem of determining a pertinent kernel function that operates to an unvisited space to the determination of the new problem parameter N'.

4.1.2 Squared-loss mutual information

The case of measuring the SMI for discrete sources follows closely the one of the collision entropy. Nevertheless, its relation to a classic second-order processing technique is more intricate, and it shall be properly addressed. In the sequel, we first establish an estimate of the SMI by mapping the data through the corresponding mapping matrices. Afterwards, we will unveil the full link between the SMI and CCA.

Consider the discrete random variables X and Y with alphabets $\mathcal{X} = \{x_n\}_{n=1,2,\ldots,N}$ and $\mathcal{Y} = \{y_m\}_{m=1,2,\ldots,M}$, whose PMFs are $p_X(x)$ and $p_Y(y)$. The joint PMF is defined as $p_{XY}(x,y)$. The probability column vectors are now $\mathbf{p} \in \mathbb{R}_+^N$ and $\mathbf{q} \in \mathbb{R}_+^M$, and their elements are $p_n = \Pr\{X = x_n\} = p_X(x_n)$ and $q_m = \Pr\{Y = y_m\} = p_Y(y_m)$, respectively. Let us define the marginal probability matrices as $\mathbf{P} = \operatorname{diag}(\mathbf{p}) \in \mathbb{R}_+^{N\times N}$ and $\mathbf{Q} = \operatorname{diag}(\mathbf{q}) \in \mathbb{R}_+^{M\times M}$, which are diagonal matrices with elements $[\mathbf{P}]_{n,n} = p_n$ and $[\mathbf{Q}]_{m,m} = q_m$, and the joint probability matrix $\mathbf{J} \in \mathbb{R}_+^{N\times M}$ whose elements are $[\mathbf{J}]_{n,m} = \Pr\{X = x_n, Y = y_m\} = p_{XY}(x_n, y_m)$. The SMI for discrete random variables, as can be deduced from (2.74), then corresponds to

$$I_s(X;Y) = \sum_{n=1}^{N}\sum_{m=1}^{M} \frac{\left(p_{XY}(x_n,y_m) - p_X(x_n)\,p_Y(y_m)\right)^2}{p_X(x_n)\,p_Y(y_m)} \tag{4.15a}$$

$$= \sum_{n=1}^{N}\sum_{m=1}^{M} [\mathbf{C}]_{n,m}^2 = \operatorname{tr}(\mathbf{C}^T\mathbf{C}) = \|\mathbf{C}\|_{\mathrm{F}}^2, \tag{4.15b}$$

where

$$\mathbf{C} = \mathbf{P}^{-1/2}\left(\mathbf{J} - \mathbf{p}\mathbf{q}^T\right)\mathbf{Q}^{-1/2}. \tag{4.16}$$

We refer to matrix $\mathbf{C}$ as the *information coherence* matrix due to its link with the *coherence* matrix from (2.40) used in CCA. This matrix form is also encountered in linear information coupling problems [HSZ15], whose objective is to simplify information theory problems by linearly approximating local features. To see this relation, let us express the information coherence matrix as:

$$\mathbf{C} = \left(\mathbf{B} - \mathbf{q}^{1/2}\mathbf{p}^{T/2}\right)^T, \tag{4.17}$$

with

$$\mathbf{B} = \mathbf{Q}^{-1/2}\mathbf{J}^T\mathbf{P}^{-1/2}, \tag{4.18}$$

where $\mathbf{q}^{1/2}$ and $\mathbf{p}^{T/2} = \left(\mathbf{p}^{1/2}\right)^T$ are element-wise powers of vectors. Matrix $\mathbf{B} \in \mathbb{R}_+^{M\times N}$ is referred to as the Divergence Transition Matrix (DTM) in [HSZ15]. Particularly, we can express the DTM as

$$\mathbf{B} = \mathbf{Q}^{-1/2}\mathbf{M}^T\mathbf{P}^{1/2}, \tag{4.19}$$

where $\mathbf{M} \in \mathbb{R}_{+}^{M \times N}$ is the channel transition matrix with elements

$$[\mathbf{M}]_{m,n} = \Pr\left\{Y = y_m | X = x_n\right\}, \tag{4.20}$$

and therefore

$$\mathbf{J} = \mathbf{P}^T \mathbf{M}^T = \mathbf{P}\mathbf{M}^T. \tag{4.21}$$

This matrix is used for studying the local geometry of the χ^2-divergence (2.66), whose SVD becomes a descriptor of dependencies between distributions that are close to each other [HZ12]. Moreover, the information coherence matrix $\mathbf{C}$ is the same as the Canonical Dependence Matrix (CDM) studied in the framework of the modal decomposition of the joint probability mass function matrix [Mak19; Hua+19]. Both the DTM and the CDM accomplish a similar task: they decompose information in a series of features, embedded in their SVD, which is then used to analyze phenomena among random variables. In this sense, the following steps are guided by this decomposition with the perspective of estimating information.

We first begin by analysing the properties of the DTM and the CDM, and then to use these properties to discuss its relationship with the SMI. In particular, a fundamental property of the DTM is that its largest singular value is $\sigma_1(\mathbf{B}) = 1$ [Mak19]. The corresponding right and left singular vectors of the largest singular value are $\mathbf{p}^{1/2}$ and $\mathbf{q}^{1/2}$, respectively. Since the largest singular value is always 1, we focus our attention to the second largest σ_{smax}, whose singular vectors are $\mathbf{u}_{\text{smax}}$ and $\mathbf{v}_{\text{smax}}$. Therefore, the SVD of $\mathbf{B}$ can be expressed as follows:

$$\mathbf{B} = \mathbf{p}^{1/2}\mathbf{q}^{T/2} + \sigma_{\text{max}}(\mathbf{B})\,\mathbf{u}_{\text{max}}\mathbf{v}_{\text{max}} + \sum_{i=3}^{\min\{N,M\}} \sigma_i(\mathbf{B})\,\mathbf{u}_i\mathbf{v}_i. \tag{4.22}$$

Given (4.17), it is straightforward to see that the SVD of $\mathbf{C}$ corresponds to

$$\begin{aligned}
\mathbf{C} &= \left(\mathbf{B} - \mathbf{p}^{-1/2}\mathbf{q}^{T/2}\right)^T && (4.23a)\\
&= \left(\mathbf{p}^{1/2}\mathbf{q}^{T/2} + \sigma_{\text{smax}}(\mathbf{B})\,\mathbf{u}_{\text{smax}}\mathbf{v}_{\text{smax}} + \sum_{i=3}^{\min\{N,M\}} \sigma_i(\mathbf{B})\,\mathbf{u}_i\mathbf{v}_i^T - \mathbf{q}^{1/2}\mathbf{p}^{T/2}\right)^T && (4.23b)\\
&= \left(\sigma_{\text{smax}}(\mathbf{B})\,\mathbf{u}_{\text{smax}}\mathbf{v}_{\text{smax}} + \sum_{i=3}^{\min\{N,M\}} \sigma_i(\mathbf{B})\,\mathbf{u}_i\mathbf{v}_i^T\right)^T && (4.23c)\\
&= \left(\sigma_{\text{smax}}(\mathbf{B})\,\mathbf{u}_{\text{smax}}\mathbf{v}_{\text{smax}} + \sum_{i=2}^{\min\{N,M\}-1} \sigma_i(\mathbf{C})\,\mathbf{u}_{i+1}\mathbf{v}_{i+1}^T\right)^T. && (4.23d)
\end{aligned}$$

Consequently, it is the largest singular value of $\mathbf{C}$ the one that corresponds to the second largest singular value of $\mathbf{B}$. Furthermore, the minimum singular value of $\mathbf{C}$ corresponds to 0 as a consequence of (4.23). In essence, the spectrum of $\mathbf{C}$ coincides with the one of $\mathbf{B}$ up to the $\min\{N, M\} - 1$ value, and the remaining singular value is either 0 or 1, respectively. Note that, while the expression inside of the parenthesis in (4.23) is the SVD of $\mathbf{C}^T$. Nevertheless, as we are interested only in the singular values, the assessment of the corresponding singular vectors of $\mathbf{C}$ is not required, and (4.23) is equally valid for studying $\sigma_i(\mathbf{C})$.

In view of the previous characterization of $\mathbf{C}$, we can now establish the following proposition:

Proposition 4.2. *Consider $I_s(X;Y)$ the SMI defined in* (4.15). *It is bounded as follows:*

$$0 \leq I_s(X;Y) \leq \min\{N, M\} - 1. \tag{4.24}$$

Proof. Let $\{\sigma_i(\mathbf{C})\}_{i=1,\dots,\min\{N,M\}}$ be the singular values of $\mathbf{C}$. We know from (4.23) that the largest singular value corresponds to the second largest singular value of $\mathbf{B}$, which is bounded by one such that $\sigma_{\text{smax}}(\mathbf{B}) \leq 1$, and the minimum singular value is 0. The bound is then directly obtained due to the squared Frobenius norm of (4.15), which corresponds to the summation of all the squared singular values of $\mathbf{C}$ (or, equivalently, the summation of all the singular values of $\mathbf{C}^T\mathbf{C}$). Since $0 \leq \sigma_i(\mathbf{C}) \leq 1$, we obtain the stated bounds of the SMI. ■

The reason behind the bound of $\min\{N,M\}-1$ is that the vectors $\mathbf{p}$ and $\mathbf{q}$ are probability simplexes, which is the subset of the unit simplex whose vectors satisfy

$$\mathbf{x} \geq 0, \qquad \mathbf{1}^T\mathbf{x} = 1. \tag{4.25}$$

Clearly, the probability vectors satisfy this condition:

$$\mathbf{1}_N^T\mathbf{p} = 1, \qquad \mathbf{1}_M^T\mathbf{q} = 1. \tag{4.26}$$

Since the probability simplex has one dimension less than the unit simplex[1] due to the unit-sum constraint, this property is being reflected in bounding the SMI. In particular, the contribution of one of the elements is lost, i.e. a singular value becomes zero, and only $N-1$ (or $M-1$) become relevant. We refer to this property as the simplex condition, and it will play a major role in the forthcoming subsection.

The relation between the decomposition of $\mathbf{C}$ and the χ^2-divergence (between any given distribution) is indeed well-known in the literature. In [Hir35], Hirschfeld determined the relationship between the mean-square contingency (as addressed in Subsection 2.3.2) and the SVD of the information coherence matrix. In [Pin+17] this decomposition of dependence between X and Y is referred to as the *principal inertia components*, which are studied for privacy applications, and then expanded in [Mak19] for the study of strong data processing inequalities for certain classes of f-divergences. Moreover, they determine that the second largest singular value $\sigma_{\text{smax}}(\mathbf{B})$ corresponds to the HGR coefficient, which will be discussed after introducing the mapping matrices for estimating the SMI in the next subsection.

Next, we proceed with the empirical estimate of the SMI. Consider L i.i.d. observations $\{x(i), y(i)\} \in \mathcal{X} \times \mathcal{Y}$ for $i = 0, ..., L-1$. The empirical estimates of the marginal probability column vectors and the joint probability matrix are now

$$\hat{\mathbf{p}} = \frac{1}{L}\mathbf{D}_x\mathbf{1}_L, \qquad \hat{\mathbf{q}} = \frac{1}{L}\mathbf{D}_y\mathbf{1}_L, \qquad \hat{\mathbf{J}} = \frac{1}{L}\mathbf{D}_x\mathbf{D}_y^T, \tag{4.27}$$

where the data matrices $\mathbf{D}_x \in \mathbb{R}^{N\times L}$ and $\mathbf{D}_y \in \mathbb{R}^{M\times L}$ are defined as in (4.4) with

$$[\mathbf{D}_x]_{n,i} = \mathbb{1}_{\{x(i)=x_n\}}, \qquad [\mathbf{D}_y]_{m,i} = \mathbb{1}_{\{y(i)=y_m\}}. \tag{4.28}$$

Equivalently, the diagonal matrices containing the probability vectors are

$$\hat{\mathbf{P}} = \frac{1}{L}\mathbf{D}_x\mathbf{D}_x^T, \qquad \hat{\mathbf{Q}} = \frac{1}{L}\mathbf{D}_y\mathbf{D}_y^T. \tag{4.29}$$

An estimate of the SMI can then be constructed as follows:

$$\hat{\hat{I}}_{\text{s}}(X;Y) = \left\| \hat{\mathbf{P}}^{-1/2}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\hat{\mathbf{Q}}^{-1/2} \right\|_{\text{F}}^2. \tag{4.30}$$

For simplicity, we can also take advantage of the projection matrix $\mathbf{P}_{\mathbf{1}}^{\perp} = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}_N^T/L$ to express the centered joint probability matrix in the following way:

$$\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T = \frac{1}{L}\mathbf{D}_x\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{D}_y^T. \tag{4.31}$$

[1] The unit simplex also includes the zero vector and satisfies $\mathbf{1}^T\mathbf{x} \leq 1$ instead of the equality in (4.25).

As a result, the required matrices to estimate the SMI are just two sample autocorrelation matrices and a sample cross-covariance matrix.

Similarly to the mapping matrices introduced in the estimation of the collision entropy in Proposition 4.1, we can also take advantage of the Frobenius norm properties for the estimation of the SMI. Let $\mathbf{X} \in \mathbb{C}^{N\times L}$ and $\mathbf{Y} \in \mathbb{C}^{M\times L}$ be the data matrices obtained by $\mathbf{X} = \mathbf{F}\mathbf{D}_x$ and $\mathbf{Y} = \mathbf{G}\mathbf{D}_y$, respectively, where $\mathbf{F} \in \mathbb{C}^{N\times N}$ and $\mathbf{G} \in \mathbb{C}^{M\times M}$ are unitary matrices. An estimate of the SMI is given by the squared Frobenius norm of the information coherence matrix, that is:

$$\hat{I}_{\mathrm{s}}\left(X;Y\right) = \|\hat{\mathbf{C}}\|_{\mathrm{F}}^2, \tag{4.32}$$

where

$$\hat{\mathbf{C}} = \hat{\mathbf{R}}_x^{-1/2}\hat{\mathbf{C}}_{xy}\hat{\mathbf{R}}_y^{-1/2}, \tag{4.33}$$

being

$$\hat{\mathbf{R}}_x = \frac{1}{L}\mathbf{X}\mathbf{X}^H, \qquad \hat{\mathbf{R}}_y = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H, \qquad \hat{\mathbf{C}}_{xy} = \frac{1}{L}\mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H \tag{4.34}$$

the sample autocorrelation matrices and the sample cross-covariance matrix, respectively. The following lemma introduces the mapping of the data matrices $\mathbf{D}_x$ and $\mathbf{D}_y$, and provides a preliminary link to CCA:

Lemma 4.2 (Preliminary link SMI-CCA)**.** *Let $\hat{\hat{I}}_s\left(X;Y\right)$ be the SMI estimator given in* (4.30)*, and $\hat{I}_{\mathrm{s}}\left(X;Y\right)$ the estimator given in* (4.32)*. The following equality holds:*

$$\hat{I}_{\mathrm{s}}\left(X;Y\right) = \hat{\hat{I}}_s\left(X;Y\right) \tag{4.35}$$

Proof. Lemma 4.2 is a direct consequence of the SMI (Frobenius norm) being invariant to nonsingular transformations. Nonetheless, the proof is fully shown in Appendix 7.3.2. ■

The previous lemma states a first link between the SMI and second-order statistics. The information coherence matrix $\hat{\mathbf{C}}$ is similar to the one required to perform the CCA, as shown in (2.40). Nevertheless, it is not a coherence matrix as it is depicted in CCA, i.e. two autocovariance matrices are needed for a full comparison instead of two autocorrelation matrices. Furthermore, the previous lemma only concerns mappings with dimensionality equal to the cardinality of the sources.

4.1.2.1 Relation to canonical correlation analysis

In order to address all previous considerations, we shall first address the case of reduced dimensionality, and full link with CCA will be shown afterwards. Let $\mathbf{F} \in \mathbb{C}^{N'\times N}$ and $\mathbf{G} \in \mathbb{C}^{M'\times M}$ be semi-unitary mapping matrices with $\mathbf{F}\mathbf{F}^H = \mathbf{I}_{N'}$ and $\mathbf{G}\mathbf{G}^H = \mathbf{I}_{M'}$, and let $\mathbf{X} = \mathbf{F}\mathbf{D}_x \in \mathbb{C}^{N'\times L}$ and $\mathbf{Y} = \mathbf{G}\mathbf{D}_y \in \mathbb{C}^{M'\times L}$ be the data matrices for $N' \le N$ and $M' \le M$. Consider the small-size sample coherence matrix as

$$\hat{\mathbf{K}}_{N',M'} = \hat{\mathbf{C}}_x^{-1/2}\hat{\mathbf{C}}_{xy}\hat{\mathbf{C}}_y^{-1/2}, \tag{4.36}$$

where $\hat{\mathbf{C}}_x = \mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{X}^H/L$, $\hat{\mathbf{C}}_y = \mathbf{Y}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H/L$, and $\hat{\mathbf{C}}_{xy} = \mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H/L$ are the sample autocovariance and cross-covariance matrices, respectively. The corresponding SMI estimator is given by

$$\hat{I}_{\mathrm{s}}'\left(X;Y\right) = \|\hat{\mathbf{K}}_{N',M'}\|_{\mathrm{F}}^2. \tag{4.37}$$

The following theorem establishes the relation between the previous expression and the full-size SMI estimator:

Theorem 4.1 (Full link SMI-CCA)**.** *Let $\hat{I}_{\mathrm{s}}\left(X;Y\right)$ be the SMI estimator given in* (4.32)*, and $\hat{I}_{\mathrm{s}}'\left(X;Y\right)$ the estimator given in* (4.37)*. For $N' = N-1$ and $M' = M-1$, the following equality holds:*

$$\hat{I}_{\mathrm{s}}'\left(X;Y\right) = \|\hat{\mathbf{K}}_{N-1,M-1}\|_{\mathrm{F}}^2 = \hat{I}_{\mathrm{s}}\left(X;Y\right). \tag{4.38}$$

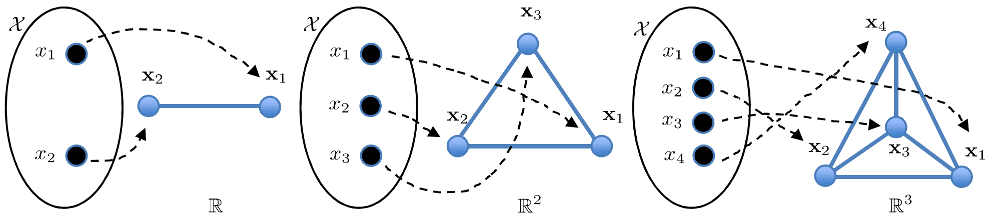


Figure 4.1: Illustration of the mapping $\mathcal{X} \to \mathbb{R}^{|\mathcal{X}|-1}$ into the $(|\mathcal{X}|-1)$-simplex.

Proof. See Appendix 7.3.3. ■

The implication of Theorem 4.1 is that we can estimate the SMI by either measuring the Frobenius norm of $\hat{\mathbf{K}}_{N-1,M-1}$ or $\hat{\mathbf{C}}$. In other words, the sample autocovariance and autocorrelation matrices are interchangeable if the dimensionality of the mapping is equal to the cardinality of the sources minus one, thus $N' = N - 1$ and $M' = M - 1$, or directly $N' = N$ and $M' = M$. As a result, the information coherence matrix $\hat{\mathbf{C}}$ is equivalent to the sample coherence matrix required in CCA, and we conclude that the estimate of the SMI can be expressed as the summation of the squared canonical correlations, which correspond to the singular values of $\hat{\mathbf{C}}$:

$$\hat{I}_{\mathrm{s}}\left(X;Y\right) = \sum_{i=1}^{\min\{N,M\}-1} \sigma_i^2\left(\hat{\mathbf{C}}\right). \tag{4.39}$$

This result can also be seen as a consequence of the simplex condition addressed in Proposition 4.2. Given that the autocovariance matrices are conformed by the probability vectors, the contribution of one singular value is lost, and only $N' = N - 1$ and/or $M' = M - 1$ are required. A sufficient condition for the equality in Theorem 4.1 to be true is that the columns of $\mathbf{F}$ and $\mathbf{G}$ are given by the $(N-1)$-simplex and the $(M-1)$-simplex, respectively. As a result, autocorrelation matrices (and specifically the information coherence matrix $\mathbf{C}$) are equally valid as autocovariance matrices for the purpose of estimating information. It is also worth noting that, given the resulting estimator is measured as the squared Frobenius norm of a coherence matrix, the SMI estimate is related to the local test for correlated Gaussian vectors from [Ram+13]. However, (4.39) applies to any kind of data mapped on a specific feature space.

Figure 4.1 illustrates the mapping into the simplex and the intuition behind Theorem 4.1: binary data (i.e. a discrete source with two possible outcomes) can be mapped to 1-dimensional points in the set

$$\{-1, 1\}, \tag{4.40}$$

ternary data (i.e. a discrete source with three possible outcomes) can be mapped to 2-dimensional points in the set

$$\left\{[1,0], \left[-0.5, \sqrt{3}/2\right], \left[-0.5, -\sqrt{3}/2\right]\right\}, \tag{4.41}$$

and so on. Note that these vectors are not probability simplexes, as described in (4.25), but they conform semi-unitary matrices, as required in Theorem 4.1.

In short, Theorem 4.1 allows to estimate an information measure through second-order statistics by mapping the events of the sources onto $\mathbf{F}$ and $\mathbf{G}$. For discrete sources, the SMI benefits from invariance under linear invertible transformations. Consequently, the *code-books* used for estimating the SMI are irrelevant, as long as the columns of $\mathbf{F}$ and $\mathbf{G}$ are linearly independent. The minimum dimensions of the spaces spanned after the mapping are required to be equal to the cardinality minus one. Otherwise, if only one (or either) of the dimension of the mapping is smaller than $N-1$ or $M-1$, then the contribution of some canonical correlations is lost. This property is also another consequence of the CCA standpoint, where any dimensionality reduction applied to a pair of vectors will result in a bounded sum of singular vectors. The following corollary gathers this result:

Corollary 4.1.1. *Given Theorem 4.2, and the SMI estimator* $\hat{I}'_{\mathrm{s}}(X;Y)$ *defined in* (4.37), *then we have:*

$$\hat{I}'_{\mathrm{s}}(X;Y) = \sum_{i=1}^{\min\{N',M'\}} \sigma_i^2\left(\hat{\mathbf{K}}_{N',M'}\right) \leq \sum_{i=1}^{\min\{N,M\}-1} \sigma_i^2\left(\hat{\mathbf{C}}\right) = \hat{I}_{\mathrm{s}}(X;Y), \tag{4.42}$$

and

$$\hat{I}'_{\mathrm{s}}(X;Y) \leq \min\left\{N', M'\right\} \leq \min\left\{N-1, M-1\right\} \tag{4.43}$$

for $N' \leq N$ *and* $M' \leq M$.

The resulting bounds contrast with Proposition 4.2, which are now given by N' and M', and not by the cardinality of the sources minus one. If we have either $N' \leq N-1$ or $M' \leq M-1$, then the Moore-Penrose inverse is usually required to cope with the rank-deficient matrices [Pez+04]. Moreover, Theorem 4.1 also implicitly states that a higher dimensionality than required ($N' > N-1$ and/or $M' > M-1$) also yields to a low-rank structure on the autocovariance matrices.

4.1.2.2 Relation to the Hirschfeld-Gebelein-Rényi coefficient

To finish this segment, we will show how the SMI, and its measure as determined in Theorem 4.1, can be linked to the estimate of the HGR coefficient from Subsection 2.3.4.1. While a similar observation can be also encountered in [Hua+19; Xu+22], our interest relies on putting the HGR coefficient estimation within the context of this thesis. Next, a particular example (mapping the events of the sources onto the reals) of the HGR is provided to examine the link between both measures.

Consider again the discrete sources $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathbf{u} \in \mathbb{R}^N$ and $\mathbf{v} \in \mathbb{R}^M$ be the vectors containing the values on which the events of the sources X and Y are mapped. That is, $[\mathbf{u}]_n = f(x_n)$ and $[\mathbf{v}]_m = g(y_m)$ for $n = 1, ..., N$ and $m = 1, ..., M$, respectively. Consider then a sequence of L i.i.d. pairs $\{x(i), y(i)\}$, for which we obtain the data matrices $\mathbf{D}_x \in \mathbb{R}^{N\times L}$ and $\mathbf{D}_y \in \mathbb{R}^{M\times L}$, and then the L-th length mapped samples

$$\mathbf{x} = \mathbf{u}^T\mathbf{D}_x, \qquad \mathbf{y} = \mathbf{v}^T\mathbf{D}_y. \tag{4.44}$$

An estimate of the HGR is then expressed as follows:

$$\hat{\rho}_{\mathrm{HGR}}(X;Y) = \max_{\mathbf{u},\mathbf{v}} \frac{\left(\mathbf{x}-\mathbf{1}^T\mathbf{x}\right)^T\left(\mathbf{y}-\mathbf{1}^T\mathbf{y}\right)/L}{\sqrt{\left(\mathbf{x}-\mathbf{1}^T\mathbf{x}\right)^T\left(\mathbf{x}-\mathbf{1}^T\mathbf{x}\right)/L}\sqrt{\left(\mathbf{y}-\mathbf{1}^T\mathbf{x}\right)^T\left(\mathbf{y}-\mathbf{1}^T\mathbf{y}\right)/L}} \tag{4.45a}$$

$$= \max_{\mathbf{u},\mathbf{v}} \frac{\left(\mathbf{D}_x^T\mathbf{u}-\mathbf{1}^T\mathbf{D}_x^T\mathbf{u}\right)^T\left(\mathbf{D}_y^T\mathbf{v}-\mathbf{1}^T\mathbf{D}_y^T\mathbf{v}\right)}{\sqrt{\left(\mathbf{D}_x^T\mathbf{u}-\mathbf{1}^T\mathbf{D}_x^T\mathbf{u}\right)^T\left(\mathbf{D}_x^T\mathbf{u}-\mathbf{1}^T\mathbf{D}_x^T\mathbf{u}\right)}\sqrt{\left(\mathbf{D}_y^T\mathbf{v}-\mathbf{1}^T\mathbf{D}_y^T\mathbf{v}\right)^T\left(\mathbf{D}_y^T\mathbf{v}-\mathbf{1}^T\mathbf{D}_y^T\mathbf{v}\right)}} \tag{4.45b}$$

$$= \max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^T\left(\mathbf{D}_x-\hat{\mathbf{p}}\right)\left(\mathbf{D}_y-\hat{\mathbf{q}}\right)\mathbf{v}}{\sqrt{\mathbf{u}^T\left(\mathbf{D}_x-\hat{\mathbf{p}}\right)\left(\mathbf{D}_x-\hat{\mathbf{p}}\right)\mathbf{u}}\sqrt{\mathbf{v}^T\left(\mathbf{D}_y-\hat{\mathbf{q}}\right)\left(\mathbf{D}_y-\hat{\mathbf{q}}\right)\mathbf{v}}} \tag{4.45c}$$

$$= \max_{\mathbf{u},\mathbf{v}} \frac{\mathbf{u}^T\hat{\mathbf{C}}_{xy}\mathbf{v}}{\sqrt{\mathbf{u}^T\hat{\mathbf{C}}_x\mathbf{u}}\sqrt{\mathbf{v}^T\hat{\mathbf{C}}_y\mathbf{v}}}. \tag{4.45d}$$

Clearly, this problem is the same as solving the CCA, which is given by the maximum singular value of the coherence matrix $\hat{\mathbf{C}}$, hence also implying $0 \leq \hat{\rho}_{\mathrm{HGR}}(X;Y) \leq 1$. As a result, the

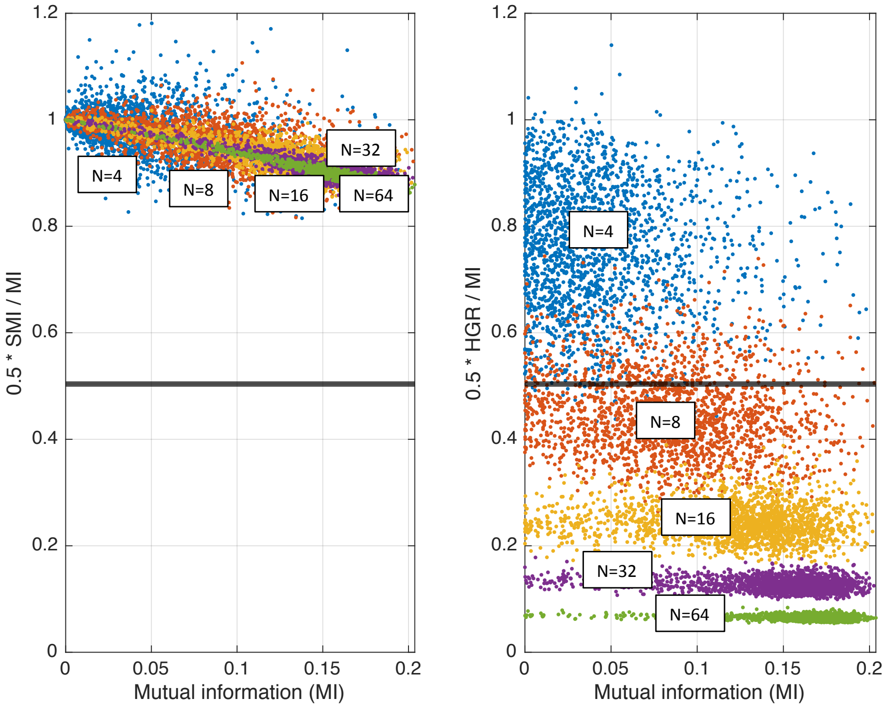


Figure 4.2: Comparison between the SMI and the HGR coefficient in terms of their half of the ratio of the MI for random discrete channels with random input distributions, and different alphabet sizes with $N = M$.

HGR coefficient does only provide the largest canonical correlation. In contrast, the SMI is given by the sum of all the squared canonical correlations, which are potentially nonzero. The SMI then can be seen as a more thorough measure, one that not only tells the best single mapping to the reals, as the HGR coefficient does, but one that also looks to other mappings to canonical coordinates of the coherence. The SMI is then expected to be more sensitive to complex hidden relationships between the observed data.

Figure 4.2 illustrates these ideas by comparing the two measures of information and their relation with the MI. The ratio of half the SMI/HGR coefficient is inspired from the local relationship between the SMI and the MI from (2.76). In particular, we are interested in seeing the dispersion of both measures at the small dependence regime. In this sense, the SMI not only exhibits a noticeably more consistent behaviour for different alphabet sizes, but also less dispersion when close to independence. Although this concrete case can be seen as an extreme case of measuring the HGR coefficient by mapping the sources to reals instead of mapping them onto linearly independent vectors, it still exemplifies the loss of information with respect to the MI. In conclusion, since both information measures are tied and estimated by performing CCA in the feature space, in terms of estimating information the SMI provides a potentially better representation of the dependence between two random variables.

4.2 Second-order statistics in the function space

Once the case for discrete sources has been addressed, we proceed to move into the case of analog sources. In this section, the core idea of the mapping for continuous random variables is established. Particularly, the mapping is inspired by the CF from Section 2.1.1 due to its property of translating independence to uncorrelation, and vice versa. Thanks to the fact that the CF is the inverse Fourier transform of a PDF, we will be able to naturally link the proposed mapping with power spectral density estimation theory. What follows is an analysis of a mapping to the space of functions and its relation with kernel methods. We are particularly interested in the relationship between the derivation of kernels provided in Chapter 3 and the CF space. However, we will also take advantage of the properties of kernels to enhance the intended mapping.

4.2.1 The characteristic function space

We start by considering a mapping from reals to the CF space $\mathcal{Z}$. Let X be a continuous random variable defined in the set $\mathcal{X} \subseteq \mathbb{R}$ and PDF $f_X(x)$. Consider L i.i.d. observations $\{x(i)\}$ for $i = 0, ..., L-1$. A tentative mapping $\tilde{z} : \mathcal{X} \to \mathcal{Z}$ is then:

$$\tilde{z}_{x(i)}(\omega) = e^{j\omega x(i)}, \qquad \forall \omega \in \mathbb{R}. \tag{4.46}$$

The main motivation of this mapping comes from the quality of the CF of linking nonlinear dependence with linear dependence (correlation). Given that our objective is to measure the SMI with just the first and second-order statistics in the feature space, the CF proves to be an excellent candidate for a feature space. However, we are still not in position to determine whether (4.46) is a valid mapping or not. In fact, we require that the scalar product in this feature space is well-defined. This question is relevant since it allows for bridging between the proposed mapping and kernel methods. For that, we need the following expression to be finite:

$$\left\langle \tilde{z}_{x(i)}(\omega), \tilde{z}_{x(j)}(\omega) \right\rangle = \int_{\mathbb{R}} \tilde{z}_{x(i)}(\omega)\, \tilde{z}^*_{x(j)}(\omega)\, \mathrm{d}\omega = \int_{\mathbb{R}} e^{j\omega(x(i)-x(j))} \mathrm{d}\omega, \tag{4.47}$$

where $j = 0, ..., L-1$. As can be seen, the scalar product is not finite, thus the mapped observations in the feature space are not of finite norm. This property poses a problem, hence a modification is needed in order to ensure a well-defined scalar product. Particularly, let us consider the following mapping:

$$z_{x(i)} = e^{j\omega x(i)} G(\omega) \qquad \forall \omega \in \mathbb{R}, \tag{4.48}$$

where $G(\omega)$ acts as a window function with finite-norm constraint such that

$$\int_{\mathbb{R}} |G(\omega)|^2 \mathrm{d}\omega < \infty. \tag{4.49}$$

This mapping can be seen as a pure frequency shift of the deterministic function $G(\omega)$, whose shifting is determined by the magnitude of $x(i)$. As a result, we have now a well-defined scalar product

$$\begin{aligned} \left\langle z_{x(i)}(\omega), z_{x(j)}(\omega) \right\rangle &= \int_{\mathbb{R}} \tilde{z}_{x(i)}(\omega)\, G(\omega)\, \tilde{z}^*_{x(j)}(\omega)\, G^*(\omega)\, \mathrm{d}\omega & \text{(4.50a)} \\ &= \int_{\mathbb{R}} e^{j\omega(x(i)-x(j))} |G(\omega)|^2 \mathrm{d}\omega, & \text{(4.50b)} \end{aligned}$$

which can be shown to be finite given the Cauchy-Schwarz inequality:

$$\begin{aligned} \left| \left\langle z_{x(i)}(\omega), z_{x(j)}(\omega) \right\rangle \right| &\leq \sqrt{\int_{\mathbb{R}} \left| \tilde{z}_{x(i)}(\omega)\, G(\omega) \right|^2 \mathrm{d}\omega} \sqrt{\int_{\mathbb{R}} \left| \tilde{z}_{x(j)}(\omega)\, G(\omega) \right|^2 \mathrm{d}\omega} & \text{(4.51a)} \\ &= \sqrt{\int_{\mathbb{R}} \left| e^{j\omega x(i)} \right|^2 |G(\omega)|^2 \mathrm{d}\omega} \sqrt{\int_{\mathbb{R}} \left| e^{j\omega x(j)} \right|^2 |G(\omega)|^2 \mathrm{d}\omega} & \text{(4.51b)} \\ &= \sqrt{\int_{\mathbb{R}} |G(\omega)|^2 \mathrm{d}\omega} \sqrt{\int_{\mathbb{R}} |G(\omega)|^2 \mathrm{d}\omega} = \|G(\omega)\|_2^2 . & \text{(4.51c)} \end{aligned}$$

As a result from (4.50), the scalar product of the *windowed* mapping based on the CF is just the squared modulus of the Fourier transform of $G(\omega)$, namely $g(\alpha) = \mathcal{F}\{G(\omega)\}$. It also means that, following Definition 4, we are now in front of a Hilbert space, and the scalar product can be seen as a kernel, albeit explicitly defined. Particularly, we have

$$\left\langle z_{x(i)}(\omega), z_{x(j)}(\omega)\right\rangle = \langle \phi(x(i)), \phi(x(j))\rangle = k(x(i), x(j)), \tag{4.52}$$

where the kernel function k is given by

$$k(x(i), x(j)) = \int_{\mathbb{R}} e^{j\omega(x(i)-x(j))} |G(\omega)|^2 \,\mathrm{d}\omega, \tag{4.53}$$

which is just the inverse Fourier transform of $|G(\omega)|^2$ evaluated at the pairwise difference $x(i) - x(j)$. As a result, we can express the kernel function as the autocorrelation of $g(\alpha)$ such that

$$k(x(i), x(j)) = \int_{\mathbb{R}} g((x(i) - x(j)) + \beta)\, g^*(\beta)\,\mathrm{d}\beta. \tag{4.54}$$

Since $k(x(i), x(j))$ comes from an autocorrelation function, its derivation entails further implications on k. On the one hand, it is a shift-invariant kernel, which can be directly written as $k(x(i), x(j)) = k(x(i) - x(j))$ (with some abuse of notation). We have actually observed this property in the IP estimator from (3.73), where it originated from the Gaussian window and its expression as an exponential function. Here, the reason is equivalent, given that (4.53) is a direct consequence of the feature map, which comes from the separability of the CF, i.e. $e^{j\omega(x(i)-x(j))} = e^{j\omega x(i)}e^{-j\omega x(j)}$. On the other hand, the resulting kernel is also bounded by the squared L_2 norm of $G(\omega)$, i.e. $|k(x(i), x(j))| \leq \|G(\omega)\|_2^2$, which is an additional consequence of “controlling” the feature space.

In the literature, these type of kernels are called autocorrelation kernels [Roj+18, Chapter 4]. While autocorrelation kernels have found their use in image signal processing [PT01; Hor04; ZZJ04] as a way to avoid an explicit measure of a costly autocorrelation, they have also been used in communications problems [Fig+12; Fig+14]. Nonetheless, the autocorrelation point of view comes from (4.53), which spans a more general kind of kernels: the shift-invariant kernel. These kernels actually come from Bochner’s theorem [Rud90, Sec. 1.4], which states that a continuous function k on $\mathbb{R}$ is positive definite if and only if it is the Fourier transform of a non-negative measure $p(\omega)$ such that:

$$k_p(x(i), x(j)) = \int_{\mathbb{R}} e^{j\omega(x(i)-x(j))} p(\omega)\,\mathrm{d}\omega. \tag{4.55}$$

Under this setting, if k_p is properly scaled, then $p(\omega)$ is a density function, which implies

$$k_p(x(i), x(j)) = \mathbb{E}_p\left\{e^{j\omega(x(i)-x(j))}\right\}. \tag{4.56}$$

This relationship with Bochner’s theorem is what drives the approximation of the kernel function with a combination of N features [RR07; LP20] such that

$$k_p(x(i), x(j)) = \mathbb{E}_p\left\{\tilde{z}_{x(i)}(\omega)\, \tilde{z}^*_{x(j)}(\omega)\right\} \approx \frac{1}{N}\sum_{n=0}^{N-1} \tilde{z}_{x(i)}(\omega_n)\, \tilde{z}^*_{x(j)}(\omega), \tag{4.57}$$

where ω_n are samples of ω referred to as nodes. Therefore, by imposing that $p(\omega)$ is a PDF, the feature space is explicitly known and related to the CF, which can then be used to approximate any shift-invariant kernel. This final expression is relevant since, as our objective is to implicitly reduce the dimension of the mapping, it provides a point of comparison with forthcoming derivations of the proposed method, which will be addressed in Subsection 4.2.1.3.

In contrast to (4.56), (4.53) does not impose any constraint on $G(\omega)$. From the tentative mapping to the shift-invariant autocorrelation kernel, we just require a square-integrable window function. The result is that the well-defined scalar product is what ensures a proper nonnegative kernel function, and not the opposite. While still endorsed by the theory of RKHS, the derivation here gains in intuition, and provides a fresh look into kernel functions.

Nonetheless, there is still some advantages by imposing additional constraints to $G(\omega)$. In particular, since we want to deal with CFs, we may impose that $G(\omega)$ is also the CF of a given random variable. Therefore it is $g(\alpha)$ that will become a PDF (unit area and nonnegative function), following Theorem 2.2, and not $|G(\omega)|^2$. This approach subverts the usual kernels derived from Bochner's theorem, given that we are primarily interested in the feature space, rather than in the kernel measure itself. Interestingly, by doing so, the resulting kernel k is an autocorrelation between two PDFs. This property coincides with the estimate of the IP based on the Parzen-Rosenblatt window estimate in (3.72), which also leads to the shift-invariant kernel (3.74) and originates from the interest on the feature space itself. Again, while in (3.72) the kernel is defined by the autocorrelation itself, here the kernel is derived from the scalar product in the function space, hence operating in the inverse path. Later, we will leverage the notion of $G(\omega)$ being a CF to define a mapping based on the *contamination* of random variables.

4.2.1.1 First and second-order statistics in the characteristic space

Once the mapping is defined, and following the case for discrete sources, we are now interested in studying the first and second-order statistics in the characteristic space. Since the feature map is based on the CF, we will see that the expectation and the correlation among the mapped data in this space can be evaluated in a closed form. To get a better insight, let us begin from (4.46). The expectation over the mapped data is

$$\mathbb{E}_{f_X}\left\{\tilde{z}_x(\omega)\right\} = \int_{\mathbb{R}} e^{j\omega x} f_X(x)\,\mathrm{d}x = \varphi_X(\omega), \tag{4.58}$$

which is directly the CF of X. The sample mean estimator is then

$$\hat{\varphi}_X(\omega) = \frac{1}{L}\sum_{i=0}^{L-1} e^{j\omega x(i)}. \tag{4.59}$$

Clearly, this directly translates to the empirical estimator of the CF, which is known to be consistent for a wide class of probability distributions [FM77]. By introducing the window function, the expected value becomes

$$\mathbb{E}_{f_X}\left\{z_x(\omega)\right\} = \int_{\mathbb{R}} G(\omega)\, e^{j\omega x} f_X(x)\,\mathrm{d}x = G(\omega)\,\varphi_X(\omega) = \xi(\omega), \tag{4.60}$$

and the sample mean is

$$\hat{\xi}(\omega) = G(\omega)\,\frac{1}{L}\sum_{i=0}^{L-1} e^{j\omega x(i)} = G(\omega)\,\hat{\varphi}_X(\omega). \tag{4.61}$$

Given that $G(\omega)$ is data invariant, the estimator is also consistent for a wide class of probability distributions.

Next, we proceed to consider the second-order statistics in the windowed CF space. For that, consider two random variables X and Y defined in the sets $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{Y} \subseteq \mathbb{R}$, whose PDFs are $f_X(x)$ and $f_Y(y)$, respectively, and the joint PDF is $f_{X,Y}(x,y)$. First, we consider the autocovariance of the mapped data:

$$\begin{aligned}
&\operatorname{Cov}\left\{z_x\left(\omega_1\right), z_x\left(\omega_2\right)\right\} \\
&\quad = \mathbb{E}_{f_X}\left\{z_x\left(\omega_1\right) z_x^*\left(\omega_2\right)\right\} - \mathbb{E}_{f_X}\left\{z_x\left(\omega_1\right)\right\}\mathbb{E}_{f_X}\left\{z_x^*\left(\omega_2\right)\right\} && \text{(4.62a)}\\
&\quad = \mathbb{E}_{f_X}\left\{e^{j\omega_1 x}G\left(\omega_1\right)e^{-j\omega_2 x}G^*\left(\omega_2\right)\right\} - \mathbb{E}_{f_X}\left\{e^{j\omega_1 x}G\left(\omega_1\right)\right\}\mathbb{E}_{f_X}\left\{e^{-j\omega_2 x}G^*\left(\omega_2\right)\right\} && \text{(4.62b)}\\
&\quad = G\left(\omega_1\right)G^*\left(\omega_2\right)\left(\int_{\mathbb{R}} e^{j\left(\omega_1-\omega_2\right)x} f_X\left(x\right)\mathrm{d}x - \int_{\mathbb{R}} e^{j\omega_1 x} f_X\left(x\right)\mathrm{d}x\int_{\mathbb{R}} e^{-j\omega_2 x} f_X\left(x\right)\mathrm{d}x\right) && \text{(4.62c)}\\
&\quad = G\left(\omega_1\right)G^*\left(\omega_2\right)\left(\varphi_X\left(\omega_1-\omega_2\right) - \varphi_X\left(\omega_1\right)\varphi_X\left(-\omega_2\right)\right). && \text{(4.62d)}
\end{aligned}$$

As can be seen, the autocovariance becomes a function of the univariate CF, and consequently to the first-order statistics. This is a particular property of the proposed mapping, which will entail further advantages when operating with the finite-dimensional feature space. On the other hand, by proceeding analogously as in (4.62), the cross-covariance between the mapped data is the following:

$$\begin{aligned}
&\operatorname{Cov}\left\{z_x\left(\omega_1\right), z_y\left(\omega_2\right)\right\} \\
&\quad = \mathbb{E}_{f_{X,Y}}\left\{z_x\left(\omega_1\right) z_y^*\left(\omega_2\right)\right\} - \mathbb{E}_{f_X}\left\{z_x\left(\omega_1\right)\right\}\mathbb{E}_{f_Y}\left\{z_y^*\left(\omega_2\right)\right\} && \text{(4.63a)}\\
&\quad = \mathbb{E}_{f_{X,Y}}\left\{e^{j\omega_1 x}G\left(\omega_1\right)e^{-j\omega_2 y}G^*\left(\omega_2\right)\right\} - \mathbb{E}_{f_X}\left\{e^{j\omega_1 x}G\left(\omega_1\right)\right\}\mathbb{E}_{f_Y}\left\{e^{-j\omega_2 y}G^*\left(\omega_2\right)\right\} && \text{(4.63b)}\\
&\quad = G\left(\omega_1\right)G^*\left(\omega_2\right)\left(\int_{\mathbb{R}^2} e^{j\omega_1 x - j\omega_2 y} f_{X,Y}\left(x,y\right)\mathrm{d}x\mathrm{d}y - \int_{\mathbb{R}} e^{j\omega_1 x} f_X\left(x\right)\mathrm{d}x\int_{\mathbb{R}} e^{-j\omega_2 y} f_Y\left(y\right)\mathrm{d}y\right) && \text{(4.63c)}\\
&\quad = G\left(\omega_1\right)G^*\left(\omega_2\right)\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right) - \varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right). && \text{(4.63d)}
\end{aligned}$$

By considering a set of L i.i.d. observations of the original sources $\left\{x\left(i\right), y\left(i\right)\right\}$ for $i = 0, ..., L-1$, then the empirical estimate of (4.63) [2] can be directly computed as follows:

$$\begin{aligned}
&\hat{\operatorname{Cov}}\left\{z_x\left(\omega_1\right), z_y\left(\omega_2\right)\right\} \\
&\quad = G\left(\omega_1\right)G^*\left(\omega_2\right)\left(\frac{1}{L}\sum_{i=0}^{L-1} e^{j\omega_1 x(i)}e^{-j\omega_2 y(i)} - \frac{1}{L}\sum_{i=0}^{L-1} e^{j\omega_1 x(i)}\frac{1}{L}\sum_{i=0}^{L-1} e^{-j\omega_2 y(i)}\right). && \text{(4.64)}
\end{aligned}$$

As it turns out, the cross-covariance just becomes a windowed version of the joint CF with reverse sign in one of its arguments. Following (4.61), the previous expression is also a consistent estimate of the covariance of the mapped data, provided that $G\left(\omega\right)$ is known. Thanks to the property of the CF of translating uncorrelation to independence, as addressed in Section 2.1.1, the correlation between the mapped data can be seen as a detector of dependence. However, due to the nature of functions, the uncorrelation has to be checked for all possible values of ω_1 and ω_2, i.e. $\forall\left(\omega_1,\omega_2\right) \in \mathbb{R}^2$.

In order to limit the feature space with the purpose of estimating an information measure, we proceed to further determine the implications, properties and behaviour of the first and second-order statistics in the windowed feature space. In the sequel, we analyze what the previous expressions entail while studying further constraints to the window function to comply with the desired finite-dimensional mapping.

On complete second-order statistics When dealing with complex random variables, the complete information is not given by the *proper* second-order statistics, but their *improper* depiction. The improper perspective can be achieved by gathering the usual proper representation, for example $\mathbb{E}_{f_{X,Y}}\left\{z_x\left(\omega_1\right) z_y^*\left(\omega_2\right)\right\}$, with the extended one, $\mathbb{E}_{f_{X,Y}}\left\{z_x\left(\omega_1\right) z_y\left(\omega_2\right)\right\}$. The complementary

[2] The empirical derivation of (4.62) is omitted because it is equivalent to performing (4.61) due to the first-order statistic point of view.

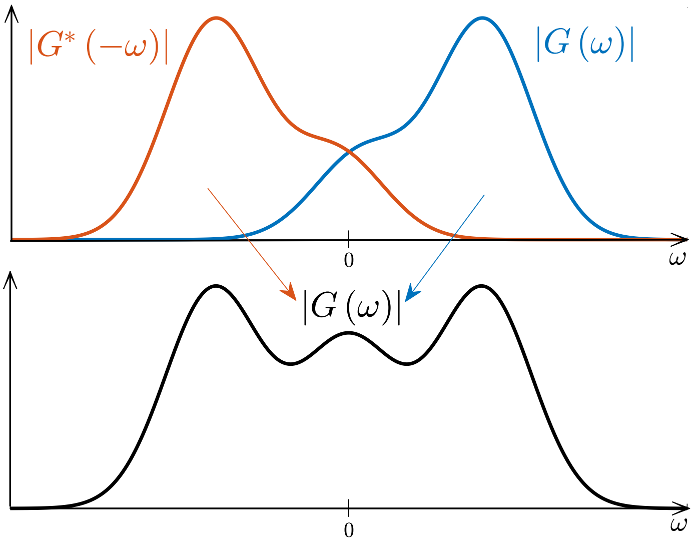


Figure 4.3: Illustrative rationale for a Hermitian window function.

information, referred to as the complete characterization of second-order statistics, leads to the consideration of the augmented covariance (or augmented matrix in the case of random vectors). For a complete overview of improper statistical processing, the reader is referred to the works of Schreier et al. [SS10]. This is relevant in our problem given that the characteristic function space is defined in the complex set. Therefore, given (4.62) and (4.63), the complete characterization of the proposed complex feature space is, in principle, required.

In order to comply with the complete characterization, we would need to compute the proper covariance and add the improper representation. Nevertheless, this may pose a problem when limiting the feature space, given that, when moving from functions to matrices, the augmented covariance matrix (improper) has double the size of the nonaugmented one (proper). In order to avoid this derivation, and, at the same time, to not loose the information given by the complete representation of the second-order statistics, the window function will be constrained to be Hermitian with $G^*(\omega) = G(-\omega)$.

Figure 4.3 exemplifies the previous argument. If the window is Hermitian from the start, then one “just” needs to explore $\forall(\omega_1, \omega_2) \in \mathbb{R}^2$ on the (only) proper statistics to include the complete characterization of the second-order statistics. It is also worth noting that the Hermitian constraint is, in fact, in agreement with $\varphi_{X,Y}$, φ_X and φ_Y, since the CF is always a Hermitian function. Therefore, by imposing this additional constraint, the complete representation of second-order statistics is achieved both for the window function and CFs.

Another consequence of imposing a Hermitian window is that its Fourier transform $g(\alpha)$ is real-valued. Therefore, from (4.54), the kernel k is a real-valued positive definite function, which is in agreement with Definition 5. Thanks to considering the complete second-order statistics and imposing some additional restrictions on $G(\omega)$, the resulting derivation entails some relevant properties that guarantee k to be a proper reproducing kernel for $x \in \mathbb{R}$. Or, in the opposite side, kernels must be real-valued and positive definite so that the complete characterization is performed. The resulting constraint is the same as derived from the mathematical kernel background, albeit expressed in the terms of this dissertation, with a more intuitive point of view.

On the contamination of the CF As mentioned before, a straightforward approach to the design of Hermitian window functions is to also assume that $G(\omega)$ is a CF. This constraint also implies that $|G(\omega)| \leq G(0) = 1$, hence $g(\alpha)$ is a PDF and the kernel is properly scaled, similarly to (4.56). Not only that, but it also entails the characterization of first and second-order statistics under a contamination point of view. Here, we consider this particular case to study its implications.

First, we consider the first-order statistics from (4.60). Generally speaking, $\xi\left(\omega\right)$ is just a windowed version of the CF of X. However, if $G\left(\omega\right)$ is a CF, we can ensure that $\xi\left(\omega\right)$ is also a CF. Given that the product of CFs implies the convolution of PDFs, (4.61) can be seen as directly the first-order statistic of some virtual source X' with density function $f_{X'}\left(x\right)$. To better see this rationale, let us consider i.i.d. observations of X' such that

$$x'\left(i\right)=x\left(i\right)+v_x\left(i\right) \tag{4.65}$$

where $v_x\left(i\right)$ are independent realizations of a random variable with PDF $g\left(x\right)$, and also independent from $x\left(i\right)$. We refer to this approach as a contamination of the original random variable due to its resemblance with an additive noise channel. The PDF of the virtual source is just the convolution of the original PDF with g, that is

$$f_{X'}\left(x\right)=\left(f_X * g\right)\left(x\right), \tag{4.66}$$

and its CF becomes the product of CFs such that

$$\varphi_{X'}\left(\omega\right)=\xi\left(\omega\right)=\varphi_X\left(\omega\right)G\left(\omega\right). \tag{4.67}$$

Consequently, to window the feature space and measure the first-order statistic is equivalent to directly operate with the virtual source X' mapped in the original, windowless, feature space. Therefore, we can write

$$\mathbb{E}_{f_{X'}}\left\{\tilde{z}_{x'}\left(\omega\right)\right\}=\mathbb{E}_{f_X}\left\{z_x\left(\omega\right)\right\}, \tag{4.68}$$

recalling that $\tilde{z}$ denotes the nonregularized feature space (4.46), but X' implies that the regularization is performed from the contamination point of view.

Henceforth, we will directly deal with the virtual sources instead, given that the contamination in (4.65) will be required for the adequate evaluation of the second-order statistics, and operations further on, from the point of view of a necessary regularization. It is thanks to $G\left(\omega\right)$ being a CF that we can benefit from the idea of convolutions between PDFs, and it will be further addressed in Subsection 4.2.1.2.

Next, we proceed to consider the autocovariance and cross-covariance of the virtual sources, following (4.62) and (4.63). On the one hand, the autocovariance of the contaminated mapped sources yields:

$$\begin{aligned}
&\operatorname{Cov}\left\{\tilde{z}_{x'}\left(\omega_1\right),\tilde{z}_{x'}\left(\omega_2\right)\right\}\\
&\quad=\mathbb{E}_{f_{X'}}\left\{\tilde{z}_{x'}\left(\omega_1\right)\tilde{z}^*_{x'}\left(\omega_2\right)\right\}-\mathbb{E}_{f_{X'}}\left\{\tilde{z}_{x'}\left(\omega_1\right)\right\}\mathbb{E}_{f_{X'}}\left\{\tilde{z}^*_{x'}\left(\omega_2\right)\right\} && \text{(4.69a)}\\
&\quad=\mathbb{E}_{f_{X'}}\left\{e^{j\omega_1 x'}e^{-j\omega_2 x'}\right\}-\mathbb{E}_{f_X}\left\{e^{j\omega_1 x'}\right\}\mathbb{E}_{f_X}\left\{e^{-j\omega_2 x'}\right\} && \text{(4.69b)}\\
&\quad=\int_{\mathbb{R}}e^{j\left(\omega_1-\omega_2\right)x'}f_{X'}\left(x\right)\mathrm{d}x-\left(\int_{\mathbb{R}}e^{j\omega_1 x'}f_{X'}\left(x\right)\mathrm{d}x\right)\left(\int_{\mathbb{R}}e^{-j\omega_2 x'}f_{X'}\left(x\right)\mathrm{d}x\right) && \text{(4.69c)}\\
&\quad=\varphi_{X'}\left(\omega_1-\omega_2\right)-\varphi_{X'}\left(\omega_1\right)\varphi_{X'}\left(-\omega_2\right) && \text{(4.69d)}\\
&\quad=\varphi_X\left(\omega_1-\omega_2\right)G\left(\omega_1-\omega_2\right)-G\left(\omega_1\right)G\left(-\omega_2\right)\varphi_X\left(\omega_1\right)\varphi_X\left(-\omega_2\right). && \text{(4.69e)}
\end{aligned}$$

In contrast to (4.62), the window that governs over $\varphi_X\left(\omega_1-\omega_2\right)$ is now also $G\left(\omega_1-\omega_2\right)$, and not $G\left(\omega_1\right)G^*\left(\omega_2\right)$. This variation is a consequence of the contamination point of view, which imposes some constraints to the window function, now considered a CF. Thanks to this expression, we will be able to simplify the computation of the autocovariance function when limiting the feature space. On the other hand, the cross-covariance is as follows:

$$\begin{aligned}
&\operatorname{Cov}\left\{\tilde{z}_{x'}\left(\omega_1\right),\tilde{z}_{y'}\left(\omega_2\right)\right\}\\
&\quad=\mathbb{E}_{f_{X',Y'}}\left\{\tilde{z}_{x'}\left(\omega_1\right)\tilde{z}^*_{y'}\left(\omega_2\right)\right\}-\mathbb{E}_{f_{X'}}\left\{\tilde{z}_{x'}\left(\omega_1\right)\right\}\mathbb{E}_{f_{Y'}}\left\{\tilde{z}^*_{y'}\left(\omega_2\right)\right\} && \text{(4.70a)}\\
&\quad=\mathbb{E}_{f_{X`,Y'}}\left\{e^{j\omega_1 x'}e^{-j\omega_2 y'}\right\}-\mathbb{E}_{f_{X'}}\left\{e^{j\omega_1 x'}\right\}\mathbb{E}_{f_{Y'}}\left\{e^{-j\omega_2 y'}\right\} && \text{(4.70b)}
\end{aligned}$$

$$= \int_{\mathbb{R}^2} e^{j\omega_1 x' - j\omega_2 y'} f_{X',Y'}(x,y)\,\mathrm{d}x\mathrm{d}y - \int_{\mathbb{R}} e^{j\omega_1 x} f_{X'}(x)\,\mathrm{d}x \int_{\mathbb{R}} e^{-j\omega_2 y} f_{Y'}(y)\,\mathrm{d}y$$

$$= \varphi_{X',Y'}(\omega_1, -\omega_2) - \varphi_{X'}(\omega_1)\,\varphi_{Y'}(-\omega_2) \tag{4.70c}$$

$$= G(\omega_1, -\omega_2)\,\varphi_{X,Y}(\omega_1, -\omega_2) - G(\omega_1)\,\varphi_X(\omega_1)\,G(-\omega_2)\,\varphi_Y(-\omega_2)\,, \tag{4.70d}$$

where $G(\omega_1, \omega_2)$ is the joint CF (the bivariate case in (2.5)) of the chosen distribution for regularization purposes. Similarly to (4.69), we will use this particular derivation of the cross-correlation when measuring information after the limitation of the feature space.

On the window effective finite support As mentioned before, an uncountable number of (ω_1, ω_2) pairs is required to measure dependence in this feature space due to the nature of the characteristic function space. This approach is studied, for instance, in the distance covariance method from Subsection 2.3.4.3. Specifically, (2.86) checks the magnitude of correlation $\forall\,(\omega_1, \omega_2) \in \mathbb{R}^2$ by performing an integral. Furthermore, the distance covariance is also regularized by a window function to ensure the convergence of the integral, consolidating the idea that the proposed feature space needs regularization. With the purpose of limiting the number of pairs required to asses uncorrelation, we explore here further considerations of the window function to achieve it.

The key point is to assume that $G(\omega)$ has an effective finite support with a given bandwidth, namely ν. Since $G(\omega)$ is deterministic and ν is determined beforehand, its shape and effective finite support give some clues to deciding the largest values of ω_1 and ω_2 that need to be tested. For example, we can benefit from the Chebyshev's inequality to determine an approximate decay of the contaminated CF. Moreover, the addition of the window function can be seen as a *tapering* of the CF that helps with improving the estimation of the CF itself. This property exhibits an insightful duality with the classical spectral estimation problem [Kay88]. In particular, Blackman-Tukey spectral estimation "tapers" the estimated autocorrelation function to trade-off bias and variance of the spectral estimate. In fact, we will see that, by determining the shape of $G(\omega)$ with a given bandwidth, the trade-off of the taper function is replicated in terms of the window bandwidth. All these notions will be leveraged for deciding the required bandwidth for the estimation of information measures in Section 4.3.

Nevertheless, the finite effective support of $G(\omega)$ may not be a sufficient condition for a proper estimation of information. In particular, we are also interested in requiring that $g(\alpha)$ also has a finite effective support. This is in preparation for an eventual limitation of the function space based on sampling ω_1 and ω_2 (for example, as in (4.57)). Consequently, following the Nyquist-Shannon sampling theorem, it is desirable to reduce the potential impact of the replicas from the contaminated PDF $f_{X'}(x)$ by considering that the resulting PDF also has an effective finite support[3]. Given that we are working in the frequency domain, a possible way to manage the decay of the Fourier transform is to consider $G(\omega)$ to be a smooth function. In this regard, it is well-known that the smoothness of a function, i.e. a function that is differentiable everywhere, is tied to the decay of its Fourier transform (see, for instance, [SS11] Section 2.2). In particular, a function that is n-th differentiable has a Fourier transform that decays with $O\left(|\omega|^{-n}\right)$ as $|\omega| \to \infty$. It is therefore desirable that $G(\omega)$ is smooth. The result is no surprising since smoothness was also assumed when addressing kernel methods in Chapter 3. In particular, the smoothness condition also has been assumed in Subsection 3.2.2 (which addressed the KCCA) in order to properly estimate the HGR coefficient.

By gathering all previous observations, an appropriate window function is then a smooth CF that provides the minimum spreading over both $g(\alpha)$ and $G(\omega)$. A window function that fulfills all the previous conditions is, for instance, the Gaussian window. By imposing $G(\omega)$ to be the CF of a Gaussian random variable, then the contamination from (4.65) is an Additive White Gaussian Noise (AWGN) process. The Gaussian window is also known to be optimal in terms of minimizing

[3] The full implications of the sampling will be detailed in Section 4.3. Right now, the sampling standpoint is only provided to identify an adequate window function.

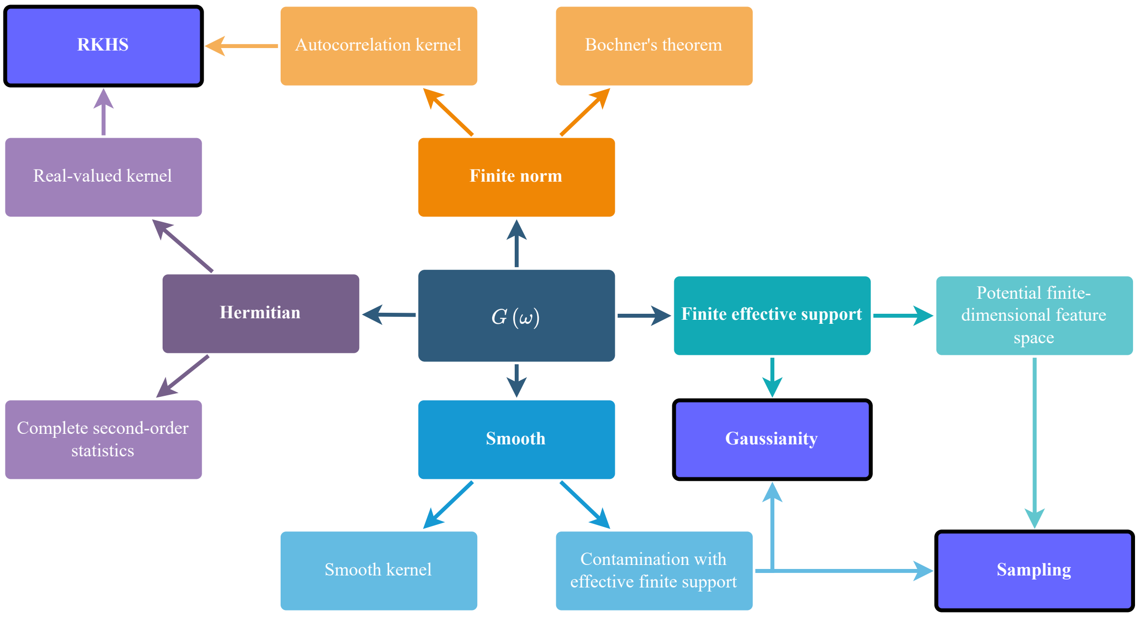


Figure 4.4: Diagram of the window function properties and their implications.

the spreading in both time and frequency domains in the context of spectral estimation [Jan91], and in time-frequency analysis methods [JB95], thus complying with the effective finite supports requirement. This property of the Gaussian window is also studied in radar detection problems, which characterizes the spread of the so-called ambiguity function, where it is also desirable to be as sharp as possible in both time and frequency domains [Lie90].

Unified window map In order to gather all previous particularizations and implications of the window function $G(\omega)$, Figure 4.4 shows an orientation map of the constraints imposed and their respective outcomes. We have begun this subsection by imposing a finite scalar product to the feature space by means of $G(\omega)$. Thanks to the addition of the window function, we can relate the proposed feature map with Bochner's theorem, which also yields the autocorrelation kernels. Furthermore, if a kernel comes from an autocorrelation function, then we can ensure that it fulfills the reproducing property, hence it spans a RKHS. Next, by considering the complete representation of the second-order statistics, we have imposed that $G(\omega)$ is Hermitian so that it considers both proper and improper statistics. Consequently, the resulting kernel is real-valued, thus also complying with the reproducing property of kernels. In another regard, we require the window function and its Fourier transform, at the same time, to have an effective finite support. On the one hand, we have assumed that the decay of the window function in the frequency domain is governed by a certain kernel bandwidth. This has been done in preparation for a forthcoming limitation of the feature space based on sampling ω. On the other hand, a smooth window function also ensures the decaying in the transformed domain (and also that the kernel is smooth), thus minimizing the potential aliasing introduced by the sampling. By uniting the previous statements, we have concluded that the Gaussian function provides the desired behaviour in both domains. Nonetheless, there are further advantages to choosing a Gaussian function, which will be addressed in the following subsection.

Relation to second-order statistics in the data space Lastly, before advancing to the next subsection, it is also of interest to characterize how the second-order statistics in the CF space are related to the second-order statistics in the data space. In particular, consider again that $G(\omega)$ has a window bandwidth ν. If we let ν to be very close to 0, then we can approximate (4.69) and (4.70) by means of the Taylor expansion in its origin, as it is addressed in Subsection 2.1.1.1. As a study case, let us take (4.70) to analyze the resulting approximation [4]. Following Appendix 7.3.4, the approximation for small values of ν yields the following:

$$\text{Cov}\left\{\tilde{z}_{x'}\left(\omega_1\right), \tilde{z}_{y'}\left(\omega_2\right)\right\} \approx \omega_1\omega_2\left(\mathbb{E}_{f_{X,Y}}\left\{XY\right\} + \mathbb{E}_{f_X}\left\{X\right\}\mathbb{E}_{f_Y}\left\{Y\right\}\right). \tag{4.71}$$

The resulting cross-covariance is then expressed as a function of the first and second-order statistics of the original random variables. However, it is not completely the covariance, given that the marginal variances are added instead of subtracted. Also note that a very small value of ν corresponds to a large value of the bandwidth of $g\left(x\right)$, since it is the Fourier pair of $G\left(\omega\right)$. Therefore, as the variance of the contamination process increases, the windowed mapping restores to second-order statistics. This behaviour resembles, again, the asymptotic IP estimator provided in (3.84), where an increase of the kernel bandwidth allowed us to link the kernel estimate with the sample variance. In this case, however, it is performed from the CF point of view.

4.2.1.2 Gaussian regularization

We are now in conditions of detailing the particular feature space given by imposing a Gaussian window function. To be concrete, consider $G\left(\omega\right)$ to be the CF of a zero-mean Gaussian random variable such that

$$G\left(\omega\right) = e^{-\sigma^2\omega^2/2} \tag{4.72}$$

for nonzero variance σ^2. Consequently, its Fourier pair (with the appropriate $1/2\pi$ normalization to make the transformation unitary) is

$$g\left(\alpha\right) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\alpha^2}{2\sigma^2}} = f_V\left(\alpha\right), \tag{4.73}$$

which results in the PDF of the zero-mean Gaussian random variable V, i.e. $\mathcal{N}\left(0, \sigma^2\right)$. The feature map is then directly as follows:

$$z_{x(i)} = e^{j\omega x(i)}e^{-\sigma^2\omega^2/2} = e^{j\omega x(i)-\sigma^2\omega^2/2} \qquad \forall\omega\in\mathbb{R}, \tag{4.74}$$

which is the CF of a Gaussian random variable with mean $x\left(i\right)$ and variance σ^2, i.e. $\mathcal{N}\left(x\left(i\right), \sigma^2\right)$. Then, the scalar product in the feature space can now be computed (as in Appendix 7.2.1) with

$$k\left(\alpha\right) = \int_{\mathbb{R}} g\left(\alpha+\beta\right)g^*\left(\beta\right)\mathrm{d}\beta = \frac{1}{\sqrt{4\pi\sigma^2}}e^{-\frac{\alpha^2}{4\sigma^2}}. \tag{4.75}$$

The resulting kernel function is equivalent to the one from (3.74) for $\alpha = x\left(i\right) - x\left(j\right)$, which just confirms the expected result of obtaining a Gaussian kernel by imposing a Gaussian shape on the window in the characteristic space.

Following the perspective of virtual sources from (4.65), the CF becomes

$$\varphi_{X'}\left(\omega\right) = \varphi_X\left(\omega\right)\varphi_V\left(\omega\right), \qquad \varphi_{Y'}\left(\omega\right) = \varphi_Y\left(\omega\right)\varphi_V\left(\omega\right), \tag{4.76}$$

where $\varphi_V\left(\omega\right) = G\left(\omega\right)$. The implications of the Gaussian virtual contamination are particularly meaningful in the form of the first-order statistics in the characteristic space. For instance, it is worth noting that the Fourier transform of the sample estimate is equivalent to estimating the PDF of X through the Parzen-Rosenblatt method shown in (3.51):

$$\mathcal{F}\left\{\mathbb{E}_{f_X}\left\{z_{x(i)}\right\}\right\} = \mathcal{F}\left\{\frac{1}{L}\sum_{i=0}^{L-1}e^{j\omega x(i)-\sigma^2\omega^2/2}\right\} = \frac{1}{L}\sum_{i=0}^{L-1}\mathcal{F}\left\{e^{j\omega x(i)-\sigma^2\omega^2/2}\right\} \tag{4.77a}$$

[4] A similar procedure can be performed with (4.69). However, for clarity, only one of the cases will be shown.

$$= \frac{1}{L}\sum_{i=0}^{L-1}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-x(i))^2}{2\sigma^2}}. \tag{4.77b}$$

This means that (4.74) is intrinsically estimating the sample CF of X by the summation of Gaussian random data in the original space, whose expectation is the convolution between the original PDF and the window function as in (3.54). However, here we have a direct control of the shape of the Gaussian function in the feature space in terms of σ. Particularly, we can directly determine the bandwidth of the window function that operates in the inherently infinite-dimensional feature space in terms of the bandwidth h required for estimating the PDF with the Parzen-Rosenblatt method (3.51). This behavior allows to determine an specific trade-off between the required (ω_1, ω_2) pairs and the sensitivity of the CF in front of $\varphi_V(\omega)$. Clearly, we want to limit the CF but also to let the bandwidth to be sufficiently large to contain all the relevant areas of the CF. The question is then how to determine the window bandwidth without tampering too much the CF for a given set of observations. This topic will be addressed later, on Section 4.3.

Lastly, another advantage of this particular case is that smoothing with Gaussian convolutions is known to act as a natural regularization of the problem of estimating measures of information. Goldfeld et al. [Gol+19; Gol+20] proposed to employ the contamination with a Gaussian variable to improve the convergence rate of estimation with respect to the sample size, even achieving the parametric rate of convergence with nonparametric estimators. Particularly, in [Gol+20] the case of the χ^2-divergence is addressed, which specifically ties into the estimation of SMI. The consequence of smoothing is reflected on an additional bias of the estimators. For the estimation of the second-order Rényi entropy, the smoothed density function becomes “more” Gaussian, or sub-Gaussian, resulting in a positive bias. For the estimation of the SMI, the contributing bias is negative due to the general data processing inequality for f-divergences (see, for instance, [PPW17; Col19]). Both of these behaviors will be verified on Section 4.3 via computer simulations.

For all the reasons above, we will fix $G(\omega) = \varphi_V(\omega)$ to be the Gaussian CF and it will be used as a basis for windowing the feature map on the characteristic space.

4.2.1.3 From the function space to the finite-dimensional feature space

Before moving into the finite-dimensional feature space formulation, we will provide some insight on its properties, advantages, disadvantages and challenges. So far, we have seen that the feature space requires some sort of regularization, whose implementation resonates in multiple facets of the mapping. On the one hand, by windowing the feature space to allow for a well-defined scalar product, we are intrinsically “kernelizing” the problem. This kernel is uniquely determined by the window function, and it is directly related to the kernel methods addressed in Chapter 3. The imposition of a windowed CF space also helps on gaining interpretability of the feature map, whose study has lead to Figure 4.4. On the other hand, the regularization of the feature space is equivalent to contaminating the original random variable, which, by imposing a bandwidth to the window function, permits to control the decay of the CF. Furthermore, the window function $G(\omega)$ helps in the universality standpoint. Thanks to imposing an effective finite support on the characteristic space, any underlying CF of the data is admitted, and the resulting estimate will only be concerned by a bias penalty. With the objective of limiting the feature space, we have concluded that the Gaussian window not only meets all the required constraints, but also has a strong relationship with the concept of smoothing convolutions for the estimation of information measures. Under this assumption, the first-order statistics is also equivalent to the Parzen-Rosenblatt window estimate. However, this course is intrinsically tied to kernel methods. Since the path took by the kernel methods is well-studied, we are more interested in operating with the first and second-order statistics in this particular space, opening the possibility of applying classical signal processing techniques. In essence, we want to establish the dual problem sketched in (4.14): outer products in front of inner products, kernel methods in front of second-order statistics.

We are now interested on limiting the dimension of the feature space to perform any kind of processing based on first and second-order statistics efficiently. As we move to analog sources, a mapping to functions is required, in principle. This can be theoretically done by rearranging a combination of the covariances studied in Subsection 4.2.1.1 so that the second-order Rényi entropy or the SMI are obtained. In essence, this would mean to translate the expressions from (4.8) and (4.32) to an integral form. However, the evaluation of this integral is still required to be solved numerically. The immediate solution is then to sample the function space, so it has a limited dimension. By doing so, we are inherently introducing bias to the estimate, a direct consequence of Proposition 4.1 and Theorem 4.1, but we recover the matrix and Frobenius norm structure. Under this setting, a relevant question becomes how is the required sampling of ω.

This is a similar problem as the one posed in (4.57), where the selected features ω_n, called nodes, have to be determined appropriately for a good approximation of the inner product. Rahimi et al. [RR07] proposed to randomize the selected features by generating samples from a uniform distribution. This approach, however, does not account for the shape of the underlying distribution, and it is let to converge in probability as the number of features increases. In this sense, the approach proposed in this section is that the Gaussian shape of $G(\omega)$ is known, which can be leveraged to improve the required mapping. Dao et al. [DDR17] proposed, in return, to approximate the integral given in (4.55) by means of a Gaussian quadrature. The approximate is then cast as follows

$$k_p\left(x\left(i\right), x\left(j\right)\right) = \int_{\mathbb{R}} e^{j\omega\left(x(i)-x(j)\right)} p\left(\omega\right) \mathrm{d}\omega \approx \sum_{n=1}^{N} \alpha_n e^{j\omega_n\left(x(i)-x(j)\right)}, \tag{4.78}$$

where the challenge is then to determine α_n and ω_n, based on the distribution $p(\omega)$, that better approximate the inner product with the minimum value of N. In particular, given (4.74), we are interested in the Gauss-Hermite quadrature ([AS64], 25.4.46), which is generally used for approximating integrals with infinite interval of the form

$$\int_{\mathbb{R}} e^{-\omega^2} f\left(\omega\right) \mathrm{d}\omega \approx \sum_{n=1}^{N} \alpha_n f\left(\omega_n\right). \tag{4.79}$$

From here, we are only interested on the nodes ω_n. The reason is that our objective is not to approximate (4.53), that is the inner product kernel approach, but to translate the problem of sampling ω_n into the required one in Subsection 4.2.1.1. The values of ω_n for the Gauss-Hermite quadrature are given by the roots of the Hermite polynomial

$$H_N\left(\omega\right) = \left(-1\right)^N e^{\omega^2} \frac{\partial^N}{\partial\omega^N} e^{-\omega^2}. \tag{4.80}$$

As it turns out, the roots of the previous expression are approximately equally spaced (see, for instance, [GM48]). Figure 4.5 illustrates this outcome by comparing the mean distance between nodes given by the roots of (4.80) which is proportional to $1/\sqrt{N}$. This result denotes that, by imposing a uniform sampling on ω, we are approximately complying with the Gauss-Hermite quadrature. Moreover, a uniform sampling will be advantageous since then the involved matrices for estimating information measures will be Toeplitz. Consequently, we can benefit from Szegö's theorem to determine asymptotic behaviours of these matrices. These concepts will be clarified as the finite-dimensional mapping is specified.

4.3 Estimating information with second-order statistics

In view of the observations and phenomena addressed in the previous section, we now proceed to propose a finite-dimensional space. We are particularly interested on three key aspects. First, to analyze how to translate a function space to a finite one, which will be the first question to be

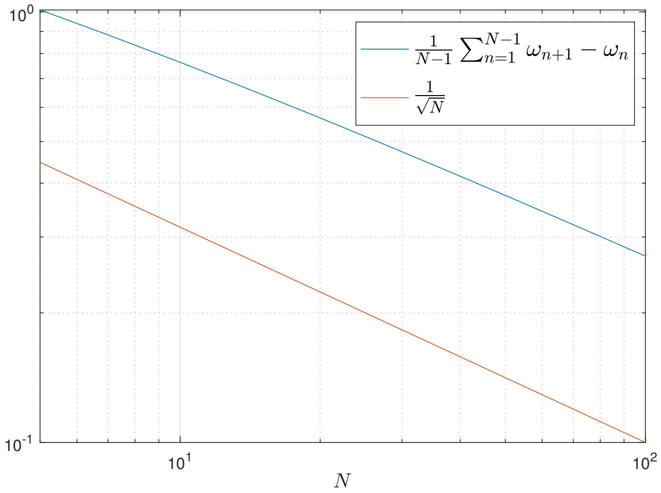


Figure 4.5: Mean value of the distance between nodes for an increasing value of N.

addressed. Secondly, since we want to deal with analog sources, any limitation of the dimension of the space entails a bounded estimation that depends on the feature space dimension itself. This covers the results observed in Proposition 4.1, in the case of the second-order Rényi entropy, and in Theorem 4.1, in the case of the SMI. And finally, the dimension of the feature space becomes a meaningful parameter that has to be chosen appropriately. This question will become undoubtedly tied to the choice of window $G(\omega)$ and associated bandwidth. In the sequel, the finite-dimensional mapping is proposed and discussed, the second-order Rényi entropy and SMI estimators are derived, and some numerical results are provided.

4.3.1 An explicit finite-dimensional feature space

Consider a sequence of L i.i.d. pairs $\{x(i), y(i)\}_{i=0,1,\ldots L-1}$ drawn from the random variables X and Y defined in the sets $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{Y} \subseteq \mathbb{R}$, respectively. Let us begin by considering the mapping from (4.48), and by proposing a sampling of the windowed CF such that the resulting complex-valued vector is the following:

$$\tilde{\mathbf{z}}_{x'(i)} = \frac{1}{\sqrt[4]{N}} e^{jx'(i)\alpha\mathbf{n}} \quad \in \mathbb{C}^N, \tag{4.81a}$$

$$\tilde{\mathbf{z}}_{y'(i)} = \frac{1}{\sqrt[4]{N}} e^{jy'(i)\alpha\mathbf{m}} \quad \in \mathbb{C}^M, \tag{4.81b}$$

where $\mathbf{n} \in \mathbb{N}^N$ and $\mathbf{m} \in \mathbb{N}^M$ with

$$\mathbf{n} = \begin{bmatrix} -K & \cdots & 0 & \cdots & K \end{bmatrix}^T, \qquad \mathbf{m} = \begin{bmatrix} -P & \cdots & 0 & \cdots & P \end{bmatrix}^T, \tag{4.82}$$

hence $N = 2K+1$ and $M = 2P+1$. Here $e^{jx'(i)\alpha\mathbf{n}}$ and $e^{jy'(i)\alpha\mathbf{m}}$ denote an element-wise exponential vector. This mapping represents a uniform sampling of the ω domain with a given sampling period α. Without loss of generality, we will consider the case $M = N$ for simplicity, and therefore $\mathbf{m} = \mathbf{n}$. As a result, we determine $z : \mathbb{R} \to \mathbb{C}^{2N+1}$ as the general function that maps the data onto a high-dimensional (but finite) space. The case $M \neq N$ can be considered if the true CFs φ_X or φ_Y require either a narrower or a broader sampling limit.

The proposed uniform sampling goes as follows. Given that analog sources can be considered of inherent "infinite-dimensionality", a finite-dimensional mapping infringes, by default, the orthonormality requirement imposed for discrete sources (see, for instance, Proposition 4.1). The new parameters that control the gap between the function and the finite-dimensional mappings are N and α. In particular, if we let $N \to \infty$ and $\alpha \to 0$ simultaneously such that $N\alpha \to \infty$, we would be then mapping the sources onto asymptotically orthonormal vectors. For example, $\alpha = N^{-1/2}$ fulfils the condition for $N \to \infty$. If such condition is achieved, then the posteriors estimates that take advantage of this mapping would be also asymptotically unbiased (Proposition 4.1 and Theorem 4.1). However, we are interested on limiting N such that the computational complexity of the problem does not grow exponentially, and particularly for it to be lower than the one of kernel methods, i.e. $N \ll L$. For that, we will determine the values of N and α from the data itself, trying to limit as much as possible the dimensionality while providing a sufficiently accurate estimation. Also note that the factor of $1/\sqrt[4]{N}$ is added to obtain asymptotically (with respect to N) orthonormal vectors, which will be crucial for estimation purposes.

Using the mapping defined in (4.81) we can then construct the data matrices $\mathbf{X} \in \mathbb{C}^{N\times L}$ and $\mathbf{Y} \in \mathbb{C}^{N\times L}$ such that

$$\mathbf{X} = \left[\begin{array}{ccc} \tilde{\mathbf{z}}_{x'(0)} & \cdots & \tilde{\mathbf{z}}_{x'(L-1)} \end{array}\right], \qquad \mathbf{Y} = \left[\begin{array}{ccc} \tilde{\mathbf{z}}_{y'(0)} & \cdots & \tilde{\mathbf{z}}_{y'(L-1)} \end{array}\right]. \tag{4.83}$$

These matrices conform the mapped data matrices from Lemma 4.2. Consequently, it is straightforward to define the sample correlation and covariance matrices required for the estimation. The autocorrelation matrices[5] are

$$\hat{\mathbf{R}}_{x'} = \mathbb{E}_{f_{X'}}\left\{\tilde{\mathbf{z}}_{x'(i)}\tilde{\mathbf{z}}_{x'(i)}^H\right\} = \frac{1}{L}\sum_{i=0}^{L-1}\tilde{\mathbf{z}}_{x'(i)}\tilde{\mathbf{z}}_{x'(i)}^H = \frac{1}{L}\mathbf{X}\mathbf{X}^H, \tag{4.84a}$$

$$\hat{\mathbf{R}}_{y'} = \mathbb{E}_{f_{Y'}}\left\{\tilde{\mathbf{z}}_{y'(i)}\tilde{\mathbf{z}}_{y'(i)}^H\right\} = \frac{1}{L}\sum_{i=0}^{L-1}\tilde{\mathbf{z}}_{y'(i)}\tilde{\mathbf{z}}_{y'(i)}^H = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H, \tag{4.84b}$$

and the cross-covariance matrix is

$$\hat{\mathbf{C}}_{x'y'} = \mathbb{E}_{f_{X',Y'}}\left\{\tilde{\mathbf{z}}_{x'(i)}\tilde{\mathbf{z}}_{y'(i)}^H\right\} - \mathbb{E}_{f_X}\left\{\tilde{\mathbf{z}}_{x'(i)}\right\}\mathbb{E}_{f_Y}\left\{\tilde{\mathbf{z}}_{y'(i)}\right\}^H \tag{4.85a}$$

$$= \frac{1}{L}\sum_{i=0}^{L-1}\tilde{\mathbf{z}}_{x'(i)}\tilde{\mathbf{z}}_{y'(i)}^H - \left(\frac{1}{L}\sum_{i=0}^{L-1}\tilde{\mathbf{z}}_{x'(i)}\right)\left(\frac{1}{L}\sum_{i=0}^{L-1}\tilde{\mathbf{z}}_{y'(i)}\right)^H \tag{4.85b}$$

$$= \frac{1}{L}\mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H. \tag{4.85c}$$

Note that, following the concept of Gaussian convolutions, the sample autocorrelation and cross-covariance matrices refer to the contaminated sources X' and Y'. We can then specify the regularized feature space by following the same rule as in (4.69) and (4.70).

One the one hand, from (4.69), the elements of the autocorrelation matrices can be written as:

$$\left[\hat{\mathbf{R}}_{x'}\right]_{n,n'} = \frac{1}{\sqrt{N}}\left(\frac{1}{L}\sum_{i=0}^{L-1}e^{jx(i)\alpha(n-n')}\right)\varphi_V\left(\alpha\left(n-n'\right)\right) \tag{4.86a}$$

$$\left[\hat{\mathbf{R}}_{y'}\right]_{n,n'} = \frac{1}{\sqrt{N}}\left(\frac{1}{L}\sum_{i=0}^{L-1}e^{jy(i)\alpha(n-n')}\right)\varphi_V\left(\alpha\left(n-n'\right)\right) \tag{4.86b}$$

for

$$\varphi_V\left(\alpha\left(n-n'\right)\right) = e^{-\sigma^2\alpha^2(n-n')^2/2}. \tag{4.87}$$

Thanks to this particular expression of the autocorrelation matrix, which only depends on the first-order statistics, we observe that (4.86) is a Toeplitz matrix for $n, n' = 0, ..., 2K$. For example,

[5] Only the autocorrelation matrices and the cross-covariance matrix will be shown, as these are the only matrices required for estimating both the second-order Rényi entropy and the SMI.

the autocorrelation matrix $\hat{\mathbf{R}}_{x'}$ can be expressed as the matrix whose diagonals are constant and correspond to the sample mean of the regularized feature space:

$$\hat{\mathbf{R}}_{x'} = \begin{bmatrix} 1 & \left[\hat{\mathbf{R}}_{x'}\right]_{0,1} & \left[\hat{\mathbf{R}}_{x'}\right]_{0,2} & \cdots & \left[\hat{\mathbf{R}}_{x'}\right]_{0,2K} \\ \left[\hat{\mathbf{R}}_{x'}\right]_{1,0} & 1 & \left[\hat{\mathbf{R}}_{x'}\right]_{0,1} & \ddots & \vdots \\ \left[\hat{\mathbf{R}}_{x'}\right]_{2,0} & \left[\hat{\mathbf{R}}_{x'}\right]_{1,0} & \ddots & \ddots & \left[\hat{\mathbf{R}}_{x'}\right]_{0,2} \\ \vdots & \ddots & \ddots & 1 & \left[\hat{\mathbf{R}}_{x'}\right]_{0,1} \\ \left[\hat{\mathbf{R}}_{x'}\right]_{2K,0} & \cdots & \left[\hat{\mathbf{R}}_{x'}\right]_{2,0} & \left[\hat{\mathbf{R}}_{x'}\right]_{1,0} & 1 \end{bmatrix}. \tag{4.88}$$

As a result, we can construct both autocorrelation matrices as follows:

$$\hat{\mathbf{R}}_{x'} = \frac{1}{\sqrt{N}}\text{Toe}\left(\hat{\mathbf{p}}_{\text{a}}\right), \qquad \hat{\mathbf{R}}_{y'} = \frac{1}{\sqrt{N}}\text{Toe}\left(\hat{\mathbf{q}}_{\text{a}}\right), \tag{4.89}$$

where $\hat{\mathbf{p}}_{\text{a}}$ and $\hat{\mathbf{q}}_{\text{a}}$ are the *extended* weighed first-order statistics

$$\hat{\mathbf{p}}_{\text{a}} = \left(\frac{1}{L}\sum_{i=0}^{L-1} e^{jx(i)\alpha\mathbf{n}_{\text{a}}}\right) \odot \mathbf{w}_{\text{a}}, \qquad \hat{\mathbf{q}}_{\text{a}} = \left(\frac{1}{L}\sum_{i=0}^{L-1} e^{jy(i)\alpha\mathbf{n}_{\text{a}}}\right) \odot \mathbf{w}_{\text{a}} \tag{4.90}$$

for $\mathbf{n}_{\text{a}} = [0, 1, \cdots, 2K]^T = [0, 1, \cdots, N-1]^T$ and the asymmetric window vectors

$$\mathbf{w}_{\text{a}} = \varphi_V\left(\alpha\mathbf{n}_{\text{a}}\right) = e^{-\sigma^2\alpha^2\mathbf{n}_{\text{a}}^2/2}. \tag{4.91}$$

Note that (4.89) is expressed following Definition 2 in order to simplify and operate with Hermitian-Toeplitz matrices.

On the other hand, from (4.70), the elements of the cross-covariance matrices can be expressed as follows:

$$\begin{aligned}\left[\hat{\mathbf{C}}_{x'y'}\right]_{n,n'} = & \frac{1}{\sqrt{N}}\left(\frac{1}{L}\sum_{i=0}^{L-1} e^{jx(i)n}e^{-jy(i)n'}\right)\varphi_V\left(\alpha n\right)\varphi_V\left(-\alpha n'\right) \\ & - \frac{1}{\sqrt[4]{N}}\left(\frac{1}{L}\sum_{k=0}^{L-1} e^{jx(k)n}\right)\frac{1}{\sqrt[4]{N}}\left(\frac{1}{L}\sum_{k=0}^{L-1} e^{-jy(k)n'}\right)\varphi_V\left(\alpha n\right)\varphi_V\left(-\alpha n'\right). \end{aligned} \tag{4.92}$$

Unfortunately, this structure does not yield a Toeplitz matrix, and no further simplifications can be done. We can, however, express the full matrix as follows:

$$\hat{\mathbf{C}}_{x'y'} = \frac{1}{\sqrt{N}}\left(\frac{1}{L}\sum_{i=0}^{L-1} e^{jx(i)\mathbf{n}}e^{-jy(i)\mathbf{n}^T}\right) \odot \left(\mathbf{w}\mathbf{w}^T\right) - \frac{\alpha}{\sqrt{N}}\hat{\mathbf{p}}\hat{\mathbf{q}}^H, \tag{4.93}$$

where the weighted first order statistics are

$$\hat{\mathbf{p}} = \left(\frac{1}{L}\sum_{i=0}^{L-1} e^{jx(i)\mathbf{n}}\right) \odot \mathbf{w}, \qquad \hat{\mathbf{q}} = \left(\frac{1}{L}\sum_{i=0}^{L-1} e^{jy(i)\mathbf{n}}\right) \odot \mathbf{w} \tag{4.94}$$

and the symmetric window vector

$$\mathbf{w} = \varphi_V\left(\alpha\mathbf{n}\right) = e^{-\sigma^2\alpha^2\mathbf{n}^2/2}. \tag{4.95}$$

Finally, with the definition of the required matrices for estimation of information measures, we are now in terms of determining a new class of estimators based on the second-order statistics in the finite-dimensional windowed characteristic space.

4.3.2 Entropy estimation via empirical characteristic function

Consider the sample correlation matrix $\hat{\mathbf{R}}_{x'}$ from (4.84), whose elements are defined in (4.86). From (4.10), an estimate of the second-order Rényi entropy just becomes

$$\hat{h}_2(X) = -\ln\left(\frac{\alpha}{2\pi}\left\|\hat{\mathbf{R}}_{x'}\right\|_{\mathrm{F}}^2\right). \tag{4.96}$$

The normalization factor $\alpha/2\pi$ is required since we are sampling ω rather than $2\pi f$, which is typical when defining the Fourier transform. This relation will become clear as we analyze the implications of the mapping and its asymptotic behaviour. To simplify further equations, we will just consider the estimate of the IP, namely $\hat{V}(X)$. By gathering all previous definitions, the estimate is as follows:

$$\hat{V}(X) = \frac{\alpha}{2\pi}\left\|\hat{\mathbf{R}}_{x'}\right\|_{\mathrm{F}}^2 = \frac{\alpha}{2\pi}\sum_{k,k'=-K}^{K}\left|\left[\hat{\mathbf{R}}_{x'}\right]_{k,k'}\right|^2 \tag{4.97a}$$

$$= \frac{\alpha}{2\pi}\sum_{k,k'=-K}^{K}\left|\frac{1}{\sqrt{N}}\left(\frac{1}{L}\sum_{i=0}^{L-1}e^{jx(i)\alpha(k-k')}\right)e^{-\sigma^2\alpha^2(k-k')^2/2}\right|^2 \tag{4.97b}$$

$$= \frac{\alpha}{2\pi}\sum_{k,k'=-K}^{K}\frac{1}{N}\left|\left(\frac{1}{L}\sum_{i=0}^{L-1}e^{jx(i)\alpha(k-k')}\right)\right|^2\left|e^{-\sigma^2\alpha^2(k-k')^2/2}\right|^2. \tag{4.97c}$$

Additionally, we can further simplify the estimation by exploiting the Toeplitz structure of $\hat{\mathbf{R}}_{x'}$. Thanks to the correlation matrix only relying on first-order statistics, it may be more efficient to just operate with the vector of sample means $\hat{\mathbf{p}}_{\mathrm{a}}$ given in (4.90). This observation can be easily deduced from the matrix structure in (4.88), where only the first column/row is required, and the summation of all the squared-modulus elements is equivalent to sum all $[\hat{\mathbf{p}}_{\mathrm{a}}]_n$ for $n = 0, ..., 2K$ and multiplied by a unilateral triangular window. In particular, we can write the IP estimator as

$$\hat{V}(X) = \frac{\alpha}{2\pi}\left(\sum_{n=0}^{2K}\left|\frac{1}{\sqrt{N}}[\hat{\mathbf{p}}_{\mathrm{a}}]_n\right|^2 2\,[\mathbf{v}_{\mathrm{a}}]_n - \frac{1}{N}N\right) \tag{4.98a}$$

$$= \frac{\alpha}{2\pi}\left(\sum_{n=0}^{N-1}\frac{1}{N}\left|\frac{1}{L}\sum_{i=0}^{L-1}e^{jx(i)\alpha n}\right|^2 e^{-\sigma^2\alpha^2n^2}2(N-n) - 1\right), \tag{4.98b}$$

where

$$\mathbf{v}_{\mathrm{a}} = [2K+1, 2K, 2K-1, \cdots, 2, 1]^T = [N, N-1\cdots, 1]^T, \tag{4.99}$$

and $\mathbf{w}_{\mathrm{a}}$ as defined in (4.91). Note that N elements of value $1/N$ are subtracted since the main diagonal is only computed once, while all the m-th diagonals, for $m = 1, ..., 2K$ are computed twice, hence the factor of 2 in front of $\mathbf{v}_{\mathrm{a}}$. This can be done since (4.98) sums squared-modulus elements of a Hermitian matrix. At last, the computationally efficient second-order Rényi entropy estimator is the following:

$$\hat{h}_2(X) = -\ln\left(\hat{V}(X)\right) = -\ln\left(\frac{\alpha}{2\pi}\mathbf{1}_N^T\left(\frac{2}{N}\left|\hat{\mathbf{p}}_{\mathrm{a}}\right|^2\odot\mathbf{v}_{\mathrm{a}} - 1\right)\right). \tag{4.100}$$

4.3.2.1 Regularization and choice of parameters

Given the second-order Rényi estimator, the only left question is how to choose N, α and σ^2. These parameters determine all the relevant aspects regarding the uniform sampling. Our objective is to establish a rule so that, provided that the effective support of the CF is known, a single parameter determines the required sampling hyper-parameters. To do so, we will again take advantage of the known shape of $\varphi_V(\omega)$ and follow a similar regularization as the one considered in [Gol+20]. However, while in [Gol+20] a smoothing variance is added to improve the convergence rate of the estimator, here a smoothing variance is added to regularize the problem by limiting the required samples. Nevertheless, we will see the first is also true, and a strong regularization/smoothing increases the convergence rate of the estimator at the cost of an added bias.

We will begin by fixing the variance of the contamination process σ^2, where we will leverage the known decay of the Gaussian window according to the approximated bandwidth of the CF of the data. As it is addressed in Subsection 4.2.1.1, we assume that the Gaussian window has an effective finite support, outside of which $\varphi_V(\omega)$ is approximately zero. We define this support as the ω_{max} that yields

$$|\varphi_V(\omega)| < \varepsilon, \qquad \forall\, |\omega| > \omega_{\text{max}}, \tag{4.101}$$

being $\varepsilon > 0$ an arbitrary small number, and

$$\omega_{\text{max}} = k\sigma^{-1}, \tag{4.102}$$

with $k > 0$. Note that the ω_{max} is expressed in terms of the standard deviation σ. Consequently, due to the separability of CFs, we can determine that

$$|\varphi_{X'}(\omega)| = |\varphi_X(\omega)\,\varphi_V(\omega)| \leq \varepsilon, \qquad \forall\, |\omega| > \omega_{\text{max}} \tag{4.103}$$

given the global bound $|\varphi_X(\omega)| \leq 1$. Therefore, we just need to choose a sufficiently low value of ε given $\omega_{\text{max}} = k\sigma^{-1}$. For that, we will generally use the Chebyshev's inequality, where the interval is decided in accordance with a sufficiently high percentage of the population. However, since the normal distribution is well-known, we can particularize the inequality to be tighter. Specifically, in [Das00] it is shown that, for normal distributions with mean μ and nonzero standard deviation ν, the inequality then becomes

$$\Pr\{|X-\mu| \geq k\nu\} \leq \frac{1}{3k^2}. \tag{4.104}$$

For the case of the Gaussian CF, we have the equivalent $k\sigma^{-1}$ instead of ν, but k remains unchanged. Therefore, we will usually use $k = 2.5$ as a general rule, since this provides approximately 98.67% of the population, and a further increase of k would negatively sway the number of sampling points. While the choice of k may seem to follow an arbitrary rule, it is based on the controlled decay of a Gaussian shape, providing a certain degree of robustness in front of the real (unknown) effective support.

The next step is to determine N, which is tied to both the observed lags in the CF and the dimension of the mapping. Previously, we have fixed in (4.81) that we perform a uniform sampling in ω with sampling period α. Since the CF and the PDF are Fourier pairs, the sampling in ω implies a periodicity of the PDF equal to $2\pi/\alpha$. Specifically, the implicit density of X' becomes

$$f_{X'}(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (f_X * f_V)\left(x - k\frac{2\pi}{\alpha}\right) & \frac{-\pi}{\alpha} \leq x \leq \frac{\pi}{\alpha} \\ 0 & \text{otherwise} \end{cases}. \tag{4.105}$$

Thus the smaller is α, the more separated are the replicas of the $(f_X * f_V)(x)$ and the smaller is the aliasing. Since we want to avoid the overlapping between the replicas as much as possible, then the sampling period can be determined as the inverse of the expected dynamic range of the PDF of the sources. In particular, we will let

$$\alpha = \frac{1}{q\sigma_x}, \tag{4.106}$$

where σ_x denotes the standard deviation of X, and $q > 0$. In contrast with the choice of k, wherein the Gaussian CF has a well-behaved shape, q needs to be high enough to contain most of the PDF for a wide class of probability distributions. Hence, we will now strictly follow the Chebyshev's inequality [PP02], that is $\Pr\{|X-\mu_x| \geq q\sigma_x\} \leq q^{-2}$. In view of this, we propose to use $q=6$, as it guarantees that approximately 97.22% of the population is not overlapping. Note that, while lower values may become critical in terms of overlapping replicas in (4.105), higher values mainly contribute to an increase of N. Therefore, while q is chosen as a somewhat heuristic approach between overlapping and sampling size, and similar with k, it is reinforced by the Chebyshev's inequality.

Next, we will combine N, α and σ into a single parameter by letting $\omega_{\text{max}} = \alpha K$, i.e. the last sample from (4.81), and combining the result with (4.102). The following relation is then obtained:

$$\omega_{\text{max}} = \frac{k}{\sigma} = \alpha K = \frac{K}{q\sigma_x}, \tag{4.107}$$

and consequently we can determine $N = 2K+1$ such that

$$N = 2\left\lceil kq\frac{\sigma_x}{\sigma}\right\rceil + 1. \tag{4.108}$$

As a result, we have tied the choice of N and α to only the smoothing variance σ^2. Note that while the choices of k and q are driven by two different aspects (regularization and sampling, respectively), both equally influence the value of N. In fact, an increase in q can be mitigated by decreasing k accordingly, in terms of N. On another note, the stronger the smoothing of the CF, less sampling points are required. This rationale is in agreement with the observations made in Subsection 4.2.1.2 regarding Gaussian convolutions, where a "strong" regularization adds a higher bias but improves the rate of convergence of the estimator. In terms of the proposed estimator and consequent choice of parameters, the regularization narrows the required observed support of the CF, not only improving the rate in which the estimator approximates the true (contaminated) value, but also by decreasing the computational complexity of the overall process. The final interpretation of (4.108) is that we are moving the problem to a finite parametrization of the PDF, which originally belongs to the nonparametric class. As the implicit number of parameters becomes finite, the problem of estimating the second-order Rényi entropy becomes consistent.

Finally, the only parameter left to be chosen is σ^2. The role of σ^2 is to determine smoothing of the underlying PDF and to regularize the required sampling points. However, σ^2 also implicitly governs the convergence rate of the estimator by adding a trade-off between bias and variance. Therefore, we will take advantage of the duality with the taper function in the context of spectral density estimation, where the relation between data size and perturbation variance is determined after minimizing the MSE with respect to the taper bandwidth [HA17]. Generally speaking, the required contamination is reduced as the number of observations L increases, and vice versa. This is a similar observation to that made by Silverman's rule of thumb in (3.55), which is also portrayed in link between the empirical estimate of the CF and the Parzen-Rosenblatt method in (4.77). Following these arguments, the choice of σ^2 goes as follows. Consider the bias and variance of the IP estimator for large L such that

$$\text{Bias}\left\{\hat{V}(X)\right\} = -O\left(\sigma^2\right) + O\left(\sigma^{-1}L^{-1}\right), \qquad \text{Var}\left\{\hat{V}(X)\right\} = O\left(\sigma^{-1}L^{-1}\right). \tag{4.109}$$

Note that the term $-O\left(\sigma^2\right)$, for $O\left(\sigma^2\right) \geq 0$ is the result of smoothing the original PDF with a normal distribution, as addressed in Subsection 4.2.1.2. If the second-order Rényi entropy were analyzed here, the signs of the contributed bias would be reversed, but the case of the IP is exhibited to avoid further approximations with the natural logarithm. However, note that as the general rule for determining the bias-variance trade-off is implemented, it will also serve as a valid rule for approximating the desired smoothing for both the second-order Rényi entropy and the SMI in the next subsection. The key point is that both bias and variance are let to be decreased with $O\left(\sigma^{-1}L^{-1}\right)$ so that

$$\lim_{L\to\infty} \sigma^2 = \lim_{L\to\infty} \sigma^{-1}L^{-1} = 0. \tag{4.110}$$

As a result, the desired limit $\lim_{L\to\infty} \text{MSE}\left\{\hat{V}\left(X\right)\right\} = 0$ is attained, where

$$\text{MSE}\left\{\hat{V}\left(X\right)\right\} = \text{Bias}^2\left\{\hat{V}\left(X\right)\right\} + \text{Var}\left\{\hat{V}\left(X\right)\right\}. \tag{4.111}$$

To achieve this limit, the variance σ^2 is required to be a monotonically decreasing function of L such that $\sigma^{-1}L^{-1}$ is also monotonically decreasing. For this, let us adopt a power law $\sigma = O(L^{-\gamma})$, for $0 < \gamma < 1$, that guarantees the desired convergence with

$$\text{MSE}\left\{\hat{V}\left(X\right)\right\} = O\left(L^{-\min\{4\gamma, 1-\gamma\}}\right). \tag{4.112}$$

Then, the value of γ can finally be chosen by the following MinMax rule:

$$\gamma = \arg\max_{\gamma}\left\{\min\left\{4\gamma, 1-\gamma\right\}\right\} = \frac{1}{5}. \tag{4.113}$$

Notice that the resulting power law of the standard deviation σ is the same as the Silverman's rule of thumb, that is $\sigma = O(L^{-1/5})$. Consequently, we will set the value of σ^2 as the following:

$$\sigma^2 = \frac{p}{L^{2/5}}, \tag{4.114}$$

where p is the new relative free parameter of the estimator. However, the choice of p is less restrictive as a general choice of σ^2, since the contamination is now always governed by the desired power law. The selection of p can then be done by choosing a sufficiently small value, which will be confirmed with computer simulations.

4.3.2.2 Szegö's limit theorem

Before delving into the numerical results, the proposed estimator will be analyzed by means of Szegö's theorem. For clarity, let us define the Hermitian-Toeplitz matrix $\mathbf{P}_{x'} = \text{Toe}\left(\mathbf{p}_{\text{a}}\right) \in \mathbb{C}^{N\times N}$, where $\mathbf{p}_{\text{a}}$ is the sampled CF whose sample estimate can be found in (4.90). Note that, by construction, the difference between $\mathbf{P}_{x'}$ and $\mathbf{R}_{x'}$ is the normalization factor $1/\sqrt{N}$, i.e. it is composed by the elements

$$\left[\mathbf{P}_{x'}\right]_{n,n'} = \varphi_X\left(\alpha\left(n-n'\right)\right)\varphi_V\left(\alpha\left(n-n'\right)\right) \tag{4.115}$$

for $n, n' = 0, ..., N-1$, and $\mathbf{R}_{x'} = \mathbf{P}_{x'}/\sqrt{N}$. The elements of $\mathbf{P}_{x'}$ are then samples of the CF of the contaminated random variable X'. From (2.1) we have

$$\left[\mathbf{p}_{\text{a}}\right]_n = \int_{-\pi/\alpha}^{\pi/\alpha} f_{X'}\left(x\right)e^{j\alpha nx}\mathrm{d}x, \tag{4.116}$$

where the integration limits are added due to the sampling of the CF, as described in (4.105). In view of this relationship, the involvement of Theorem 2.4 is straightforward. In this case the sign of the exponential is reversed given that $\left[\mathbf{p}_{\text{a}}\right]_n^* = \left[\mathbf{p}_{\text{a}}\right]_{-n}$. Moreover, note that in (4.116) the normalization factor (adjusted to α) that makes the Fourier transform unitary is missing, as it is also not present in (2.1). The factor $2\pi/\alpha$ is required, however, in the inverse transform, which is manifested by the inverted factor in the integral. In light of Theorem 2.5 and the previous considerations, for large feature dimension N we can write:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\lambda_n^2\left(\mathbf{P}_{x'}\right)\frac{\alpha}{2\pi} = \int_{-\pi/\alpha}^{\pi/\alpha} f_{X'}^2\left(x\right)\mathrm{d}x = \int_{-\pi/\alpha}^{\pi/\alpha}\left(f_X * f_V\right)^2\left(x\right)\mathrm{d}x. \tag{4.117}$$

As a result, the integral is just the IP of the contaminated variable X', namely $V\left(X'\right)$, which is the measure that we want to estimate. In comparison with the original expression in (2.20), here the CF is uniformly sampled with sampling period α, hence the integration limits are changed to encompass the region for which we assume, for a sufficiently small α, that no replica is present. Given $\mathbf{P}_x$ (or $\mathbf{R}_x$) is a Hermitian matrix, the eigenvalues are real and nonnegative, and equal to the singular values. Hence, we can write the particular form of Szegö's theorem in terms of the Frobenius norm:

$$\lim_{N\to\infty} \frac{1}{N} \left\|\mathbf{P}_{x'}\right\|_{\mathrm{F}}^2 \frac{\alpha}{2\pi} = \int_{-\pi/\alpha}^{\pi/\alpha} \left(f_X * f_V\right)^2(x)\,\mathrm{d}x, \tag{4.118}$$

where we finally have

$$\frac{1}{N} \left\|\mathbf{P}_{x'}\right\|_{\mathrm{F}}^2 \frac{\alpha}{2\pi} = \left\|\mathbf{R}_{x'}\right\|_{\mathrm{F}}^2 \frac{\alpha}{2\pi}. \tag{4.119}$$

In light of this observation, we can deduce that for $\alpha \to 0$ and $N \to \infty$ at the same time, which corresponds to the same constraint made in Subsection 4.3.1, the estimator becomes unbiased with respect to the contaminated IP. In particular, we can assert that

$$\lim_{N\to\infty,\alpha\to 0} \mathbb{E}_{f_X}\left\{\hat{V}\left(X\right)\right\} = V\left(X'\right). \tag{4.120}$$

Note that, given (4.108), to decrease α is intrinsically tied to an increase of N, ensuring that the asymptotic limits are attained simultaneously. In conclusion, by estimating the elements of $\mathbf{R}_{x'}$ as the sample mean of the CF, its squared Frobenius norm tends to the true IP of X'. The resulting estimator is not only consistent in terms of L [FM77], but also in terms of N.

The application of Szegö's theorem for estimating an entropy measure has also been studied in [Ram+09]. In contrast with the proposed approach, the authors of [Ram+09] explore the reverse path by exploiting the analogy between a PDF and a PSD, and the estimation of the differential entropy and the KL divergence arises from Szegö's theorem. Instead, the estimator in (4.100) is conceived from the feature map perspective and analyzed through Szegö's theorem lenses. Nevertheless, the process to regularize the estimator is also very different. In [Ram+09] the data is normalized so that the PDF is inside of the interval $[-1/2, 1/2]$, and the CF is sampled with sampling period 1, hence the integral limits in (4.116) are directly inside of $[-1/2, 1/2]$ and no normalization is required. Afterwards, the dimension N is also chosen in view of classical spectral estimating techniques, but it is done by modelling the PDF as an autoregressive model [Kay98; BV00] and with a posterior minimum description length criterion to determine the required lags of the CF. In the proposed approach, the normalization and regularization is performed through α and N. Thanks to the Gaussian convolutions, the overall regularization problem is tied with a physical sense of contaminating random variables, and the required samples of the CF are governed by the degree of contamination. As a result, the estimation process is endowed with a more direct and manageable approach, which allows formulations as the ones made to control the curve of learning in (4.114).

4.3.2.3 Numerical results

In this subsection, the performance of the proposed second-order Rényi entropy estimator is analyzed by means of Monte Carlo simulations. Generally speaking, we will evaluate the estimator by modelling the data as GMMs. Particularly, we will consider L i.i.d. observations $x\left(i\right)$, with $i = 0, ..., L-1$, from the following distribution

$$X \sim \sum_{m=0}^{M-1} p_m \mathcal{N}\left(\mu_m, \sigma_m^2\right). \tag{4.121}$$

One can also write the probability weight, mean and variance vectors as

$$\mathbf{p} = \left[p_0, ..., p_{M-1}\right]^T, \quad \boldsymbol{\mu} = \left[\mu_0, ..., \mu_{M-1}\right]^T, \quad \boldsymbol{\sigma}_X^2 = \left[\sigma_0^2, ..., \sigma_{M-1}^2\right]^T, \tag{4.122}$$

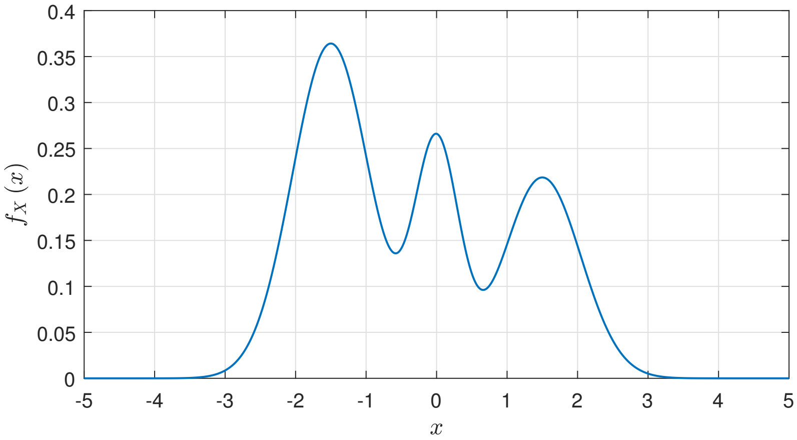


Figure 4.6: PDF of the GMM with $M = 3$, $\mathbf{p} = [0.5, 0.2, 0.3]^T$, $\boldsymbol{\sigma}^2 = [0.3, 0.1, 0.3]^T$ and $\boldsymbol{\mu} = [-1.5, 0, 1.5]^T$.

respectively, with $\sum_{m=0}^{M-1} p_m = 1$. The reason for modelling the data this way is that it allows for an extra flexibility on generating different kind of models, while $h_2(X)$ is always known in closed form. Particularly, by following Appendix 7.3.5, the second-order Rényi entropy of a GMM is

$$h_2(X) = -\ln(V(X)) = -\ln\left(\sum_{m'=0}^{M-1}\sum_{m=0}^{M-1} \frac{p_{m'}p_m}{\sqrt{2\pi\left(\sigma_{m'}^2+\sigma_m^2\right)}} \exp\left(\frac{-\left(\mu_m-\mu_{m'}\right)^2}{2\left(\sigma_{m'}^2+\sigma_m^2\right)}\right)\right). \tag{4.123}$$

Furthermore, for any smoothing variance σ^2, the second-order Rényi entropy of the contaminated variable, i.e. $h_2(X')$, can be immediately obtained by letting the new variance vector be $\boldsymbol{\sigma}_{X'}^2 = \left[\sigma_0^2+\sigma^2, \sigma_1^2+\sigma^2, ..., \sigma_{M-1}^2+\sigma^2\right]^T$, which yields

$$h_2(X') = -\ln\left(\sum_{m'=0}^{M-1}\sum_{m=0}^{M-1} \frac{p_{m'}p_m}{\sqrt{2\pi\left(2\sigma^2+\sigma_{m'}^2+\sigma_m^2\right)}} \exp\left(\frac{-\left(\mu_m-\mu_{m'}\right)^2}{2\left(2\sigma^2+\sigma_{m'}^2+\sigma_m^2\right)}\right)\right). \tag{4.124}$$

Unless otherwise specified, the choice of parameters will follow the rationale from Subsection 4.3.2, where $k = 2.5$, $q = 6$ and therefore

$$N = 2\left\lceil 15\frac{\hat{\sigma}_x}{\sigma}\right\rceil + 1, \tag{4.125}$$

with the sample variance being $\hat{\sigma}_x = \frac{1}{L-1}\sum_{i=0}^{L-1}\left(x(i) - \hat{\mu}_x\right)^2$, and $\hat{\mu}_x = \frac{1}{L}\sum_{i=0}^{L-1} x(i)$. The smoothing variance (or contamination) σ^2 will be determined following (4.114), that is $\sigma^2 = pL^{-2/5}$, where p will be indicated per simulation.

To begin with, the repercussion of choosing p is analyzed. The distribution used to this end is the GMM illustrated in Figure 4.6. These parameters are chosen so that the distribution has a distinctive non-Gaussian shape while maintaining a relative simple model. The resulting distribution has three clear distinguishable clusters that provide a particular vulnerability in front of the added contamination. When the smoothing effect of the Gaussian regularization is strong, the nuance of the original distribution is lost. For this reason, the proposed distribution is useful for studying if the regularization is well-behaved, i.e. if it maintains the information embedded in the underlying distribution, or if the smoothing effect is overly strong so that it disrupts it.

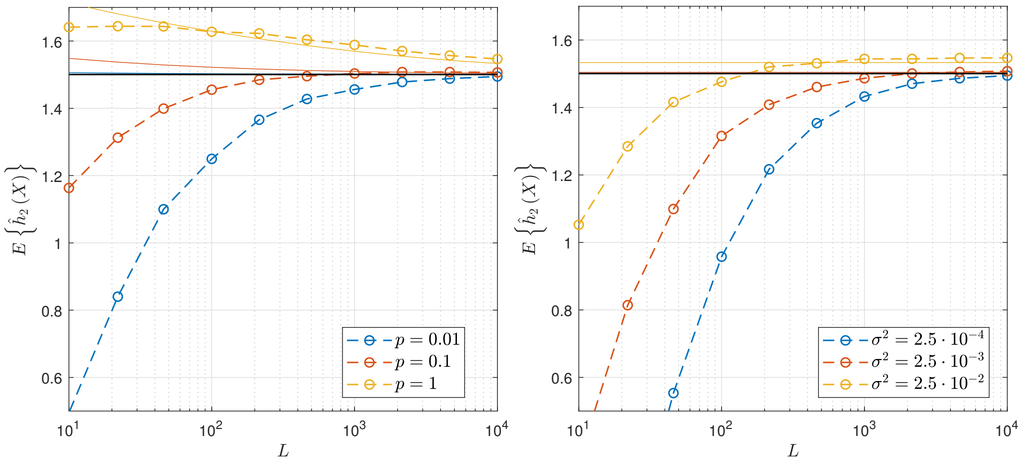


Figure 4.7: Expected value of $\hat{h}_2(X)$ with different values of p (left) and with different values of σ^2 (right). The solid black line represents the true $h_2(X)$ value, while the color-coded solid lines represent the contaminated entropies $h_2(X')$ for each value of p or σ^2.

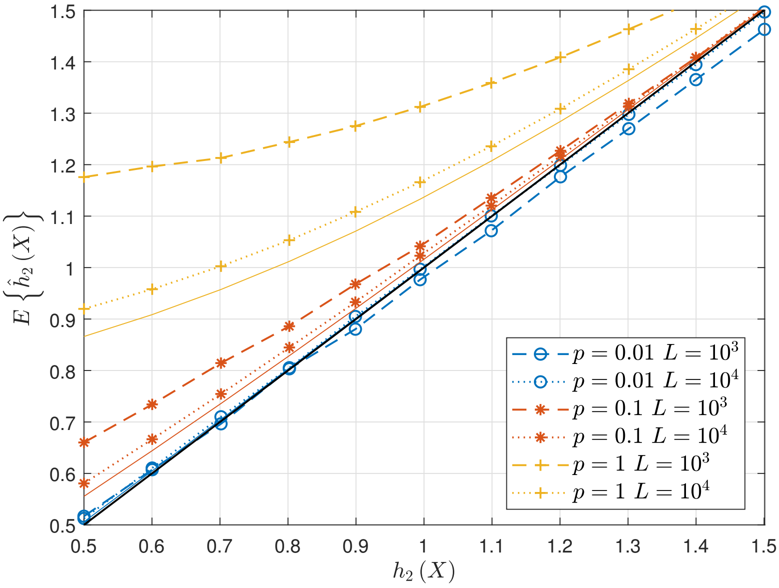


Figure 4.8: Expected value of the second-order Rényi entropy estimator with different values of p and L. The color-coded solid lines represent the contaminated entropies $h_2(X')$ for each value of p with fixed $L = 10^4$.

Figure 4.7 exhibits the effectiveness of letting σ^2 to be decided by the rule in (4.114), where the distribution is the one from the previous figure. Consequently, the second-order Rényi entropy of the contaminated random variable $h_2(X')$ changes as L grows, approximating the true second-order Rényi entropy $h_2(X)$. On the one hand, in the left figure, it is shown that high values of p the estimator tends faster to $h_2(X')$, but not necessarily to $h_2(X)$, than with lower p values. As a drawback, the added positive bias is noticeable for large values of p. Nevertheless, given (4.114), the smoothing variance is reduced as L increases for any given choice of p, attaining an asymptotically unbiased estimator. On the other hand, in the right figure, it is shown what happens if the smoothing variance σ^2 is fixed [6]. While the expected value and bias results the same around

[6] The choice of σ^2 in this case is based on the "objective" σ^2 from the left figure for $L = 10^4$, that is $\sigma^2 = p10^{-8/5}$, and p equal to its corresponding value in the left figure.

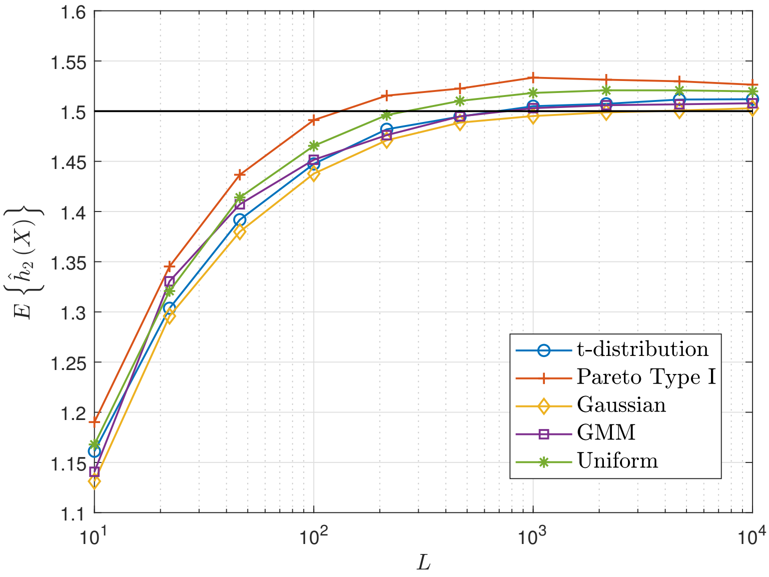


Figure 4.9: Expected value of the second-order Rényi entropy estimator for different distributions for $h_2(X) = 1.5$.

the targeted L, which in this case corresponds to the rightmost part of the curve, the convergence rate is not properly managed, resulting in an increased bias in the overall estimate. Furthermore, since σ^2 is not a function of L, then the estimator is not asymptotically unbiased. In conclusion, by adopting the power law described in (4.114), we are capable of both improving the convergence rate and the introduced bias. While the choice of p remains undecided, ultimately left to be determined by a rule of thumb, its consequences are less strict than those of choosing a fixed value of σ^2.

Figure 4.8 iterates on the display of the behaviour of the estimator for different values of the parameter p and data size L. In this figure, we examine different values of $h_2(X)$ and how these may influence the choice of parameters. To obtain these different values, we have taken advantage of the GMM from Figure 4.6 by just changing σ_0^2. It can be seen that lower values of $h_2(X)$ become more susceptible to higher contamination values. This is as expected since the underlying PDF becomes narrower in some portions of the overall dynamic range, and the convolution with a smoothing distribution is more noticeable. On the contrary, while higher values of $h_2(X)$ are more robust to the contamination, for p sufficiently low the bias's sign is ultimately reversed. This observation reveals that, for a sufficiently low p, the estimator becomes asymptotically close from below, and not from above. The consequence is that lower levels of $h_2(X)$ become reliant on the dimension, requiring a lower p, while higher levels become dependent on the data size, requiring higher L.

Next, we test the convergence rate with fixed p for different distributions. Figure 4.9 shows this comparison for the Gaussian distribution, the GMM from Figure 4.6, the Student's t-distribution, the Pareto uniform distributions. The reason behind each one of these distributions goes as follows: The Gaussian distribution serves as the basis for comparison, whose second-order Rényi entropy is $h_2(X) = 0.5 \ln\left(4\pi\sigma_G^2\right)$ where σ_G^2 denotes the variance, jointly with the GMM distribution from Figure 4.6. The Student's t-distribution, whose second-order Rényi entropy is known, represents a distribution with a long tail, which is, in principle, a favorable distribution as it endures better the induced contamination, and whose effective support of the CF is more limited. However, it requires a higher sampling rate in order to avoid aliasing from the long tails of the PDF. The Pareto Type I and the uniform distributions serve a similar function by presenting a discontinuity in the PDF. While the first is a unilateral exponential distribution, with a discontinuity in $x = 0$, the latter is just defined inside of the support $x \in [a, b]$. As a result, both distributions are

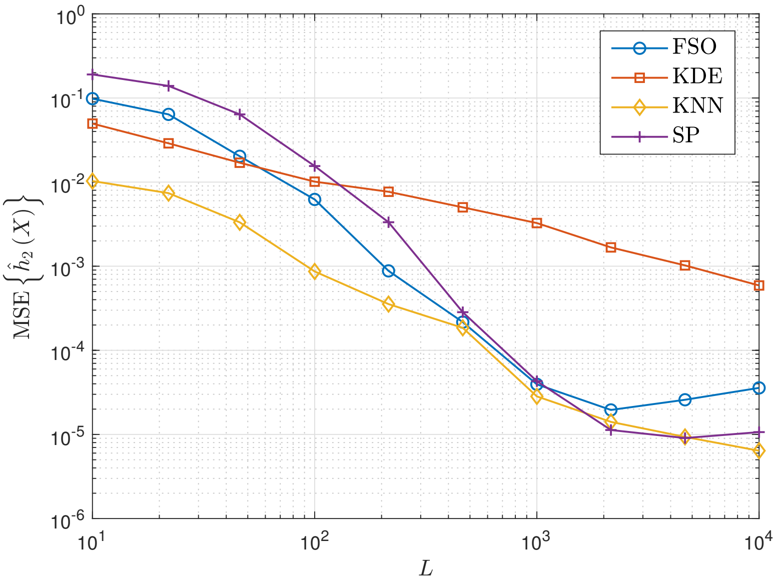


Figure 4.10: MSE of the estimated second-order Rényi entropy as a function of data size L for $h_2(X) = 1.5$ and $p = 0.1$.

susceptible to the smoothing variance, and it is expected that the introduced bias is greater. The second-order Rényi entropy of the Pareto distribution can be found in [NZ03], while for the uniform distribution is $h_2(X) = \ln(b - a)$. All distributions have been calibrated so that $h_2(X) = 1.5$. The phenomenon of the data-dependent regime from Figure 4.8 can also be seen in Figure 4.9.

For this second-order Rényi entropy value, the estimator becomes asymptotically close from below, introducing a negative bias. However, the estimate tends to the contaminated entropy $h_2(X')$, and not to the true value (as happens, for instance, in (3.92) for the case of the IP). Ultimately, this is translated to a reversed bias sign, which is visibly manifested for higher values of L. As expected, the bias introduced by the regularization is especially distinguishable for the uniform and Pareto distributions. While the Gaussian distribution is the most robust in front of the contamination, either Student's t-distribution or the GMM come close, and the end result is almost not perceivable. Even so, apart from the implication of σ^2, all five distribution share a similar learning curve for an increasing value of L, showing the qualities of the proposed second-order Rényi entropy estimator.

Finally, the proposed estimator based on the Frobenius norm of the mapped correlation matrix, denoted Frobenius Second-Order statistics (FSO) in the figures, is compared with other second-order Rényi entropy estimators in the literature. Specifically, the other shown techniques are the following: the KDE estimator defined in (3.77), where the kernel bandwidth is decided as in [Che15], the estimator based on the k-Nearest Neighbors (KNN) method [LPS08; LLC09], with $k = 3$, and the one based on Sample Spacings (SP) of the data [Wac+05; HLE05], which is a direct expansion of the Vasicek's entropy estimation [Vas76], and with internal parameter $m = \left\lfloor\sqrt{L}\right\rfloor$. We use the information theoretical estimators toolbox [Sza14] for the estimation of the Rényi entropy with both KNN and SP methods. Note that, in the case of estimating the second-order Rényi entropy, the classical approach of plugging-in the estimated PDF through the Parzen-Rosenblatt window method into the functional of interest is equivalent to performing (3.77) [AH84].

Figure 4.10 illustrates the MSE of the aforementioned estimators for the GMM from Figure 4.6, defined as

$$\text{MSE}\left\{\hat{h}_2(X)\right\} = \mathbb{E}_{f_X}\left\{\left(\hat{h}_2(X) - h_2(X)\right)^2\right\}. \tag{4.126}$$

Clearly, the proposed estimator competes with the KNN and SP estimators, while the KDE one struggles to reach a similar performance. It can be observed, however, how the bias sign is reversed when increasing L, ultimately deteriorating the MSE. This is a consequence of the regime change illustrated in Figure 4.8. This outcome can be avoided by decreasing p so it behaves better at higher L, but the cost is then a worse performance for low values of L. Note that the contrary is also true, in order to improve the performance of the estimator at the low L regime, a stronger smoothing effect can be applied by increasing p, at the cost of a worse performance at high L. These phenomena can be implicitly observed in Figure 4.7.

4.3.3 Information estimation via empirical characteristic function

Once the second-order Rényi entropy estimator has been developed and analyzed, the equivalent for estimating the SMI is straightforward. From (4.39) we have

$$\hat{I}_{\mathrm{s}}\left(X;Y\right) = \|\hat{\mathbf{C}}\|_{\mathrm{F}}^{2} = \left\|\hat{\mathbf{R}}_{x'}^{-1/2}\hat{\mathbf{C}}_{x'y'}\hat{\mathbf{R}}_{y'}^{-1/2}\right\|_{\mathrm{F}}^{2}, \tag{4.127}$$

where the sample autocorrelation and cross-covariance matrices are defined in (4.89) and in (4.93), respectively. It is worth noting that, unlike with $h_2\left(X\right)$, the SMI estimate does not require any additional normalization factor, given that it just cancels out by the autocorrelation matrices. Nevertheless, similar to $h_2\left(X\right)$, we are moving the problem of estimating information measures to the problem of estimating sample correlation and covariance matrices, which are known to be a consistent estimate for i.i.d. data.

The choice of parameters is performed equally to the second-order Rényi entropy case. Specifically, we let

$$\alpha = \frac{1}{q\sigma_{\mathrm{max}}} \tag{4.128}$$

where $\sigma_{\mathrm{max}} = \max\left\{\sigma_x, \sigma_y\right\}$. In other words, the choice of sampling period is determined by the most restrictive distribution in terms of overlapping replicas, which is the one that requires a denser sampling in the ω domain. The maximum value of the effective support is maintained with $\omega_{\mathrm{max}} = k\sigma^{-1}$, since it is determined by the contamination rather than the distribution. It is noteworthy that this assumption can be done since the joint CF is also bounded, similar to (4.103). Specifically, assume that we have

$$|\varphi_{X'}(\omega_1)| < \varepsilon, \quad |\varphi_{Y'}(\omega_2)| < \varepsilon, \quad \left|\varphi_{X',Y'}(\omega_1,\omega_2)\right| < \varepsilon^2, \quad \forall\omega_1,\omega_2 > \omega_{\mathrm{max}}. \tag{4.129}$$

The last bound is derived by following the separability of the CF for independent random variables

$$\varphi_{X',Y'}(\omega_1,\omega_2) = \varphi_{X,Y}(\omega_1,\omega_2)\varphi_V(\omega_1)\varphi_V(\omega_2), \tag{4.130}$$

given that $|\varphi_{X,Y}(\omega_1,\omega_2)| \leq 1$ and

$$|\varphi_V(\omega_1)| < \varepsilon, \qquad |\varphi_V(\omega_2)| < \varepsilon, \qquad \forall\omega_1,\omega_2 > \omega_{\mathrm{max}}. \tag{4.131}$$

Lastly, the choice of N is the same as in (4.108), and the smoothing variance as in (4.114).

It is also worth noting that, thanks to the Gaussian convolutions perspective, the inversion of $\hat{\mathbf{R}}_{x'}$ and $\hat{\mathbf{R}}_{y'}$ can be done without issues. Contrary to the diagonal loading that is usually required to invert the kernel matrices, for instance in (3.27), the proposed approach succeeds in regularizing the feature space by providing a physical interpretation of the overall effect on the final estimate.

4.3.3.1 Large feature space dimension

While the information coherence matrix $\hat{\mathbf{C}}$ that is required for estimating the SMI is not Toeplitz, the autocorrelation matrices $\hat{\mathbf{R}}_{x'}$ and $\hat{\mathbf{R}}_{y'}$ indeed are. Like with the second-order Rényi entropy estimation, we can further exploit their asymptotic behaviour. In this case, we are particularly interested in facilitating the matrix inversions.

Recall from Theorem 2.5 and Corollary 2.5.1 that Toeplitz matrices behave asymptotically like circulant matrices, which are diagonalizable by the unitary Fourier matrix. A weak condition for this asymptotic behaviour is that the columns of the Toeplitz matrix are square integrable for $N \to \infty$ [Gra+06]. While this condition may not be guaranteed for any $\hat{\mathbf{R}}_x$, since it depends on the unknown CF of the data, thanks to the addition of the Gaussian windowing this condition is ensured for $\hat{\mathbf{R}}_{x'}$. Particularly, for $\sigma^2 > 0$, and given that

$$\left| \frac{1}{L} \sum_{i=0}^{L-1} e^{jx(i)\alpha n} \right| \leq 1, \qquad \left| \frac{1}{L} \sum_{i=0}^{L-1} e^{jy(i)\alpha n} \right| \leq 1, \tag{4.132}$$

then the sample vectors $\hat{\mathbf{p}}_{\mathrm{a}}$ and $\hat{\mathbf{q}}_{\mathrm{a}}$ from (4.90) are square-integrable such that they have finite L^2-norm:

$$\lim_{N \to \infty} \|\hat{\mathbf{p}}_{\mathrm{a}}\|_2^2 < \infty, \qquad \lim_{N \to \infty} \|\hat{\mathbf{p}}_{\mathrm{a}}\|_2^2 < \infty. \tag{4.133}$$

The following lemma sets the required theoretical framework:

Lemma 4.3. *Let $t_n \in \mathbb{C}$ be a Hermitian sequence, that is $t_n = t^*_{-n}$, with $t_0 = 1$ and*

$$\lim_{N \to \infty} \sum_{n=0}^{N-1} |t_n|^2 < \infty. \tag{4.134}$$

Let us define vector $\mathbf{t} \in \mathbb{C}^N$ and associated Hermitian-Toeplitz matrix $\mathbf{T} \in \mathbb{C}^{N \times N}$ as $[\mathbf{t}]_n = t_n$ and $\mathbf{T} = \mathrm{Toe}\{\mathbf{t}\}$, respectively. Let $\mathbf{W} \in \mathbb{C}^{N \times N}$ be the unitary discrete Fourier transform matrix from (2.18). *Then*

$$\lim_{N \to \infty} \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} \sum_{n'=0}^{N-1} \left| [\mathbf{W}\mathbf{T}\mathbf{W}^H]_{n,n'} - \left[\mathrm{diag}\left(\frac{2}{\sqrt{N}} \mathrm{Re}\left\{ \mathbf{W} \left(\mathbf{t} \odot \mathbf{v}_{\mathrm{a}} \right) \right\} - \mathbf{1}_N \right) \right]_{n,n'} \right|^2} = 0 \tag{4.135}$$

for $n = 0, 1, \ldots, N-1$, where $\mathbf{v}_{\mathrm{a}}$ is a unilateral triangular window

$$\mathbf{v}_{\mathrm{a}} = [N, N-1 \cdots, 1]^T. \tag{4.136}$$

Proof. See Appendix 7.3.6. ■

Next, we define the computationally efficient SMI estimator, and show that is asymptotically $\hat{I}_{\mathrm{s}}(X;Y)$ for $N \to \infty$. Consider the sample cross-covariance matrix $\hat{\mathbf{C}}_{x'y'} \in \mathbb{C}^{N \times N}$ from (4.93), $\hat{\mathbf{p}}_{\mathrm{a}} \in \mathbb{C}^N$ and $\hat{\mathbf{q}}_{\mathrm{a}} \in \mathbb{C}^N$ the sample weighted first-order statistics from (4.90), $\mathbf{W} \in \mathbb{C}^{N \times N}$ the unitary Fourier matrix as in (2.18) and $\mathbf{v}$a $\in \mathbb{R}^N$ a unilateral triangular window with elements $[\mathbf{v}_{\mathrm{a}}]_n = N - n$ for $n = 0, ..., N-1$, as in (4.136). Then, the reduced computational complexity estimator of the SMI is as follows:

$$\hat{I}_{\mathrm{as}}(X;Y) = \left\| \hat{\mathbf{P}}'^{-1/2} \mathbf{W} \hat{\mathbf{C}}_{x'y'} \mathbf{W}^H \hat{\mathbf{Q}}'^{-1/2} \right\|_F^2, \tag{4.137}$$

where $\hat{\mathbf{P}}' = \frac{1}{\sqrt{N}} \mathrm{diag}\{\hat{\mathbf{p}}'\}$, $\hat{\mathbf{Q}}' = \frac{1}{\sqrt{N}} \mathrm{diag}\{\hat{\mathbf{q}}'\}$, and

$$\hat{\mathbf{p}}' = 2\mathrm{Re}\left(\mathbf{W}^H \left(\hat{\mathbf{p}}_{\mathrm{a}} \odot \mathbf{v}_{\mathrm{a}} \right) \right) - \mathbf{1}_N, \qquad \hat{\mathbf{q}}' = 2\mathrm{Re}\left(\mathbf{W}^H \left(\hat{\mathbf{q}}_{\mathrm{a}} \odot \mathbf{v}_{\mathrm{a}} \right) \right) - \mathbf{1}_N. \tag{4.138}$$

Theorem 4.2. *Consider the SMI estimator $\hat{I}_{\mathrm{s}}(X;Y)$ from* (4.127). *Let*

$$\hat{I}_{\mathrm{s}}(X;Y) = \left\| \left(\mathbf{W} \hat{\mathbf{R}}_{x'} \mathbf{W}^H \right)^{-1/2} \mathbf{W} \hat{\mathbf{C}}_{x'y'} \mathbf{W}^H \left(\mathbf{W} \hat{\mathbf{R}}_{y'} \mathbf{W}^H \right)^{-1/2} \right\|_F^2 = \|\hat{\mathbf{A}}\|_F^2, \tag{4.139}$$

and

$$\hat{I}_{\mathrm{as}}(X;Y) = \left\| \hat{\mathbf{P}}'^{-1/2} \mathbf{W} \hat{\mathbf{C}}_{x'y'} \mathbf{W}^H \hat{\mathbf{Q}}'^{-1/2} \right\|_F^2 = \|\hat{\mathbf{B}}\|_F^2 \tag{4.140}$$

from (4.137). *For an increasing feature space dimension N, then*

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \left\| \hat{\mathbf{A}} - \hat{\mathbf{B}} \right\|_F = 0, \tag{4.141}$$

and consequently

$$\lim_{N\to\infty} \hat{I}_s (X;Y) = \lim_{N\to\infty} \hat{I}_{as} (X;Y). \tag{4.142}$$

Proof. See Appendix 7.3.7. ■

The implication of Theorem 4.2 is that we can approximate the Fourier transform of the autocorrelation matrices as diagonal matrices, whose inverse is then an element-wise operation, hence reducing the computational complexity typically associated with the inversion of matrices of high dimension. As the dimension increases, the individual elements of the matrices involved in the estimation of the SMI tend to have the same value, and both estimators yield virtually the same SMI. The cost of this approximation is that a high value of N is required to cope with the limit behaviour. Nevertheless, we will see through computer simulations that this high-dimensional regime is actually quite fast to achieve. The main advantage of this approximation is that, as the N required increases, e.g. the required smoothing variance σ^2 is low, the approximate estimator becomes more enticing, and the overall computational complexity does not increase exponentially.

4.3.3.2 Numerical results

Lastly, the performance of the proposed estimators and the impact of their free parameters are evaluated by means of Monte Carlo simulations. Similarly to the second-order Rényi entropy case, we will exploit the GMMs in order to model the data and to easily generate different scenarios. Unless otherwise stated, the following GMM will be used:

$$(X,Y) \sim \frac{1}{4}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{4}\mathcal{N}(-\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{4}\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \frac{1}{4}\mathcal{N}(-\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \tag{4.143}$$

where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -\sqrt{\lambda\rho} \\ \sqrt{\lambda\rho} \end{bmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} \sqrt{\lambda\rho} \\ \sqrt{\lambda\rho} \end{bmatrix}, \tag{4.144a}$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1-\lambda\rho & (1-\lambda)\rho \\ (1-\lambda)\rho & 1-\lambda\rho \end{bmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1-\lambda\rho & -(1-\lambda)\rho \\ -(1-\lambda)\rho & 1-\lambda\rho \end{bmatrix}, \tag{4.144b}$$

with $\rho \in [0,1)$, and $\lambda \in [0,1]$ so that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive semi-definite matrices. The marginal random variables are then distributed as follows:

$$X \sim \frac{1}{2}\mathcal{N}\left(\sqrt{\lambda\rho}, 1-\lambda\rho\right) + \frac{1}{2}\mathcal{N}\left(-\sqrt{\lambda\rho}, 1-\lambda\rho\right), \tag{4.145a}$$

$$Y \sim \frac{1}{2}\mathcal{N}\left(\sqrt{\lambda\rho}, 1-\lambda\rho\right) + \frac{1}{2}\mathcal{N}\left(-\sqrt{\lambda\rho}, 1-\lambda\rho\right). \tag{4.145b}$$

The parameter ρ determines the degree of dependence between both random variables, performing a similar role as a correlation coefficient, where 0 yields to independent random variables, hence $I_s(X;Y) = 0$. Conversely, λ determines different scenarios with varying degree of difficulty, and it shall be used to evaluate the performance of the estimator in front of different distributions. Following this model, L i.i.d. data samples can be generated by

$$x(i) = h_x(i)\left(-\sqrt{\lambda} + z(i)\sqrt{1-\lambda}\right)\sqrt{\rho} + w_x(i)\sqrt{1-\rho} \tag{4.146a}$$

$$y(i) = h_y(i)\left(\sqrt{\lambda} + z(i)\sqrt{1-\lambda}\right)\sqrt{\rho} + w_y(i)\sqrt{1-\rho}, \tag{4.146b}$$

for $i = 0, ..., L-1$, where $z, w_x, w_y \sim \mathcal{N}(0,1)$ are i.i.d. random variables and h_x, h_y are discrete random variables that can take the values in $\{-1, 1\}$ with equal probability.

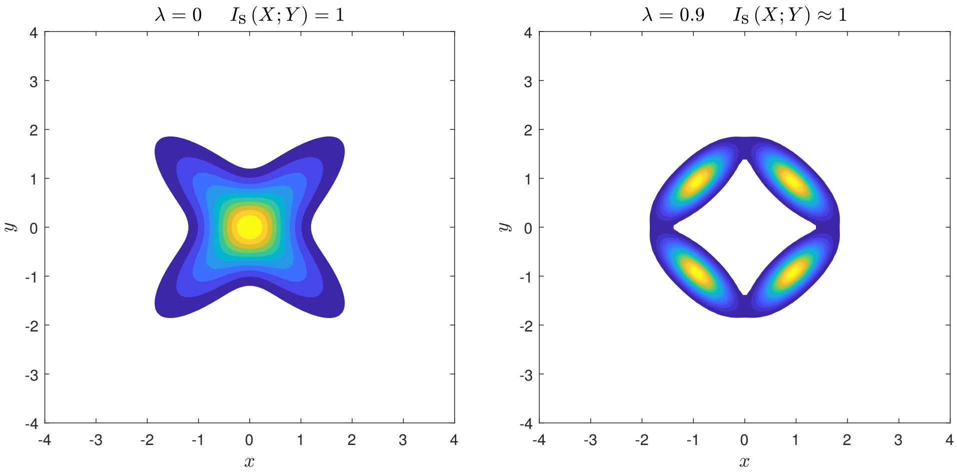


Figure 4.11: Contour plots of the proposed GMM model for two λ. The SMI has been calibrated to be approximately the same.

The usefulness of this model is that the correlation between X and Y is always null, i.e. $\mathbb{E}_{f_{X,Y}}\{XY\} = 0$, thanks to the decorrelating random variables h_x and h_y. The reason behind this constraint is to demonstrate that the dependence measure can be estimated by measuring correlation in the feature space, thus forcing the estimator to discover dependence from originally uncorrelated data. Moreover, for simplicity, the model has been calibrated so that $\mathbb{E}_{f_X}\{X^2\} = \mathbb{E}_{f_Y}\{Y^2\} = 1$. Note, however, that this has been done so that the $\sigma_{\max}$ from (4.128) is always around the same value for comparison purposes.

From (4.143) two distributions with different λ value will be used: $\lambda = 0$ and $\lambda = 0.9$. The joint distributions for both these cases can be seen in Figure 4.11. The advantage of $\lambda = 0$ is that the marginal distributions are then just normal with $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$. Consequently, the SMI can be computed for any given smoothing variance σ^2, which is the following:

$$I_{\mathrm{s}}(X,Y)_{\lambda=0} = \frac{\rho^4}{(1+\sigma^2)^4 - \rho^4}. \tag{4.147}$$

This property allows to characterize the true SMI values for any given Gaussian regularization. This model has also been previously used to test independence detectors in [Res+11; dR19]. This is also the model that has been used in Figure 2.1 under the name of GMM. On the other hand, for $\lambda = 0.9$ the SMI needs to be measured by a genie-aided estimator based on the empirical average of the SMI under the knowledge of the marginal and joint distributions [WKV09]. A similar distribution to this one has also been used to test an independence detector based on kernel methods, as can be seen in [Gre+07; GG10]. However, while in these works independence is generated by rotating the nonoverlapping ($\lambda = 1$) Gaussian components, here it is embedded in the parameter ρ.

We begin the analysis by performing a similar test to that from Figure 4.8 in the case of $h_2(X)$ estimation. Figure 4.12 shows the expected value of the estimated SMI values as a function of small ($I_{\mathrm{s}}(X;Y) \in [0, 0.1]$, left figures) and moderate ($I_{\mathrm{s}}(X;Y) \in [0,1]$, right figures) SMI values with different combinations of p and L. Similar to the case of estimating the second-order Rényi entropy with the same tool as here, we can identify two different regimes. On the one hand, the small dependence regime is the data-limited regime, where a large value of data size L is required. On the other hand, the strong dependence is the dimension-limited regime, requiring a smaller value of p to reduce the estimation bias. Nonetheless, it is worth noting that the regimes have pivoted with respect to the $h_2(X)$ case. On the one hand, here the bias is inherently negative, following the general data processing inequality for f-divergences [Col19], which yields $I_{\mathrm{s}}(X;Y) \geq I_{\mathrm{s}}(X';Y')$. On the other hand, the estimator becomes asymptotically close from *above* to the contaminated SMI for $L \to \infty$. Furthermore, the case of estimating the SMI entails some further advantages

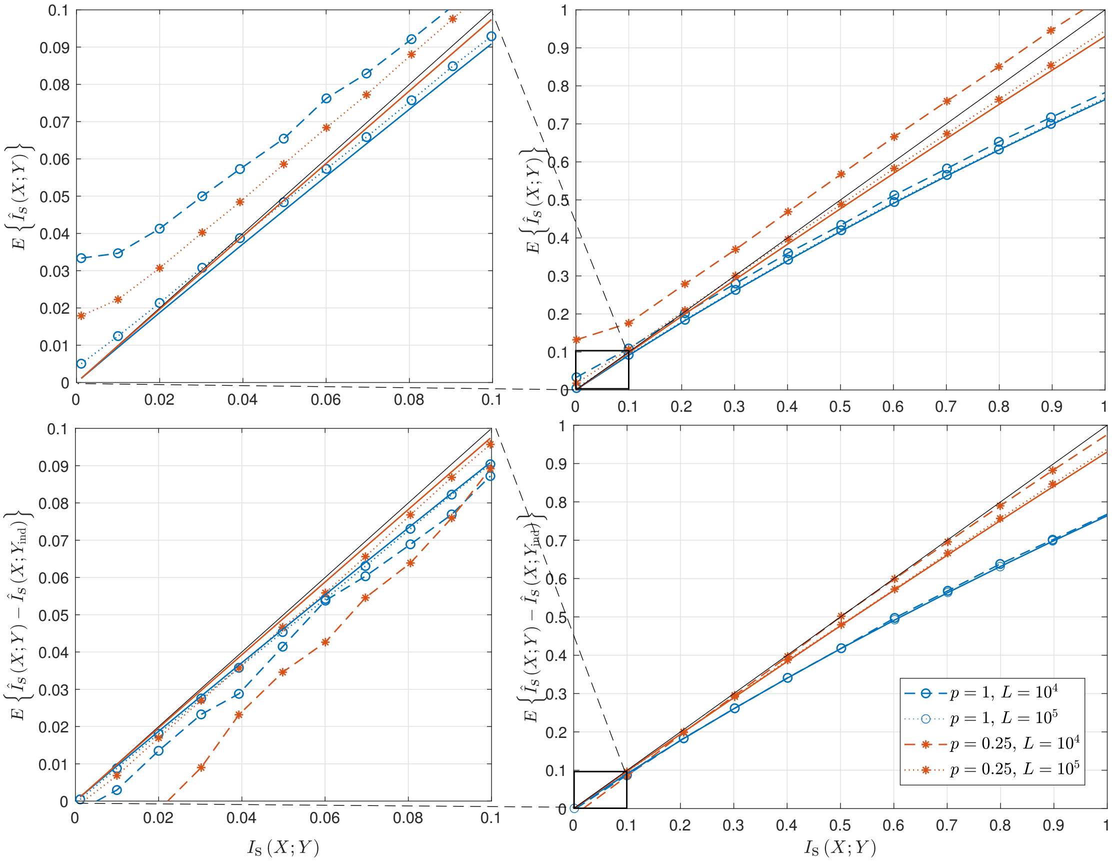


Figure 4.12: Expected value of the SMI estimator (up) and the reduced-bias estimator (down) as a function of the true SMI value for the GMM model with $\lambda = 0.9$. The solid black line denotes $I_s(X;Y)$, while the solid color-based lines denote $I_s(X';Y')$ for their respective value of p.

that are not possible when estimating $h_2(X)$. In particular, following (4.127), and consequently (4.39), some of the singular values of $\hat{\mathbf{C}}$ yield an arbitrarily small value, which correspond to the theoretically null squared canonical correlations. The estimator then aggregates these residuals, which originates a ground level and a bigger estimator error. This can be seen in Figure 4.12, and particularly for low L values, given that the estimation error of these squared canonical correlations increases. However, in the case of SMI estimation we can compensate this behaviour thanks to a reduced bias estimator $\hat{I}_s(X;Y) - \hat{I}_s(X;Y_{\text{ind}})$, portrayed in the bottom figures, where Y_{ind} is another independent random variable with the same distribution than Y. This data can be obtained by circularly shifting the data sequence associated to Y by $j \in \mathbb{N}_*$ positions, where $j \in [1, L-1]$. The overall result is an improvement over the original estimator, reducing the bias in the small data regime regardless of the kind of data statistics.

Next, we test the convergence rate of the method with different distributions. Figure 4.13 shows the expected value of the estimator for a varying value L. The distributions shown are the same as in Figure 2.1, and are the corresponding bidimensional distributions from Figure 4.9. Similarly to the case of entropy estimation, the distributions with discontinuities are of special difficulty for the proposed estimator. This time, however, the SMI estimation with the Student's t-distribution is not as well-behaved as in the second-order Rényi entropy case. Another relevant observation is that, within the GMM model proposed in (4.143), the case $\lambda = 0$ presents the slower convergence rate, while this rate is improved by increasing λ. Furthermore, for $\lambda = 0$ the rate of learning is even faster than the Gaussian case, manifesting the negative bias given by the Gaussian regularization for relatively low values of L.

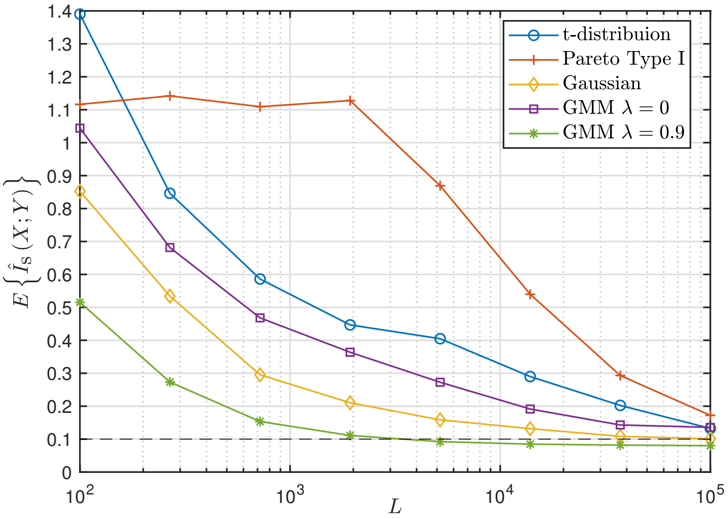


Figure 4.13: Expected value of the SMI estimator for different distributions, with $I_{\text{s}}(X;Y) = 0.1$ and $p = 2.5$.

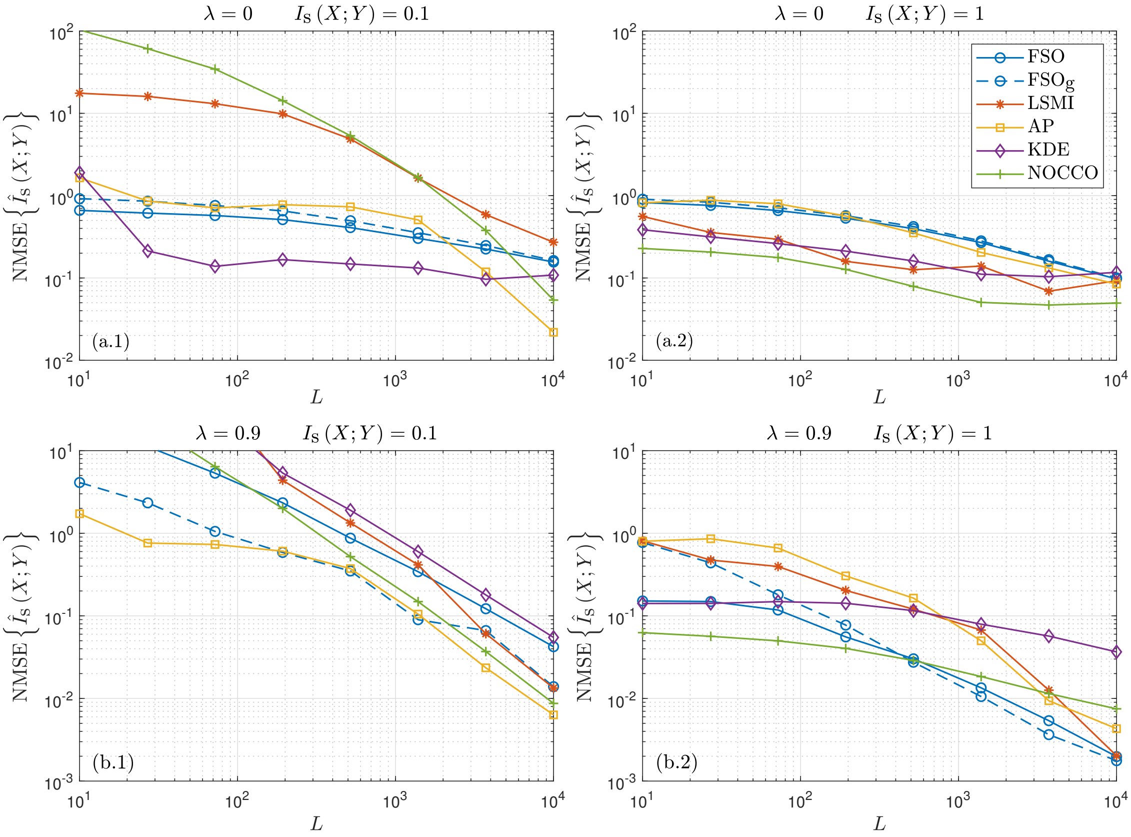


Figure 4.14: NMSE of the estimated SMI as a function of the data size L. Parameters of choice for the FSO estimator: (a.1) $p = 5$, (a.2) $p = 2.5$, (b.1) $p = 0.5$, (b.2) $p = 0.25$.

Figure 4.14 depicts the Normalized Mean Squared Error (NMSE) of the proposed estimator, denoted as FSO in the figure, along with the reduced bias version $\hat{I}_{\text{s}}(X;Y) - \hat{I}_{\text{s}}(X;Y_{\text{ind}})$, denoted as FSO_g in the figure, where $\cdot_\text{g}$ stands for ground. The MSE is computed as follows:

$$\text{NMSE}\left\{\hat{I}_\text{s}\left(X;Y\right)\right\} = \frac{\mathbb{E}_{f_{X,Y}}\left\{\left(\hat{I}_\text{s}\left(X;Y\right) - I_\text{s}\left(X;Y\right)\right)^2\right\}}{I_\text{s}\left(X;Y\right)^2}, \tag{4.148}$$

and it is used to fairly compare different values of SMI, given that it is now relevant to study the small information regime in particular due to its links with the MI. For the sake of comparison, four other SMI estimators are shown: the Least-Squares Mutual Information (LSMI) estimator [Suz+09], the estimator based on the Adaptive Partitioning (AP) of the observation space [DV99], the NOCCO [Fuk+07], addressed in Subsection 3.2.5, and the one that employs the KDE [MRL95]. Both the AP and the KDE based have been adapted to specifically estimate the SMI, since they are plug-in estimators. The LSMI estimator is an explicit estimator of the LSMI, and no further adjustment is required. The kernel bandwidth for the KDE is again chosen following [CK19], and the hyper-parameter of the LSMI is chosen through the cross-validation procedure depicted in [Suz+09]. Generally speaking, the proposed method and choice of parameters are shown to be effective and comparable with the other estimators from the literature. It is only for $\lambda = 0$ and small dependence that the FSO struggles a bit, jointly with the LSMI, which corresponds to the known previously analyzed case of slow convergence in Figure 4.13. It can also be seen that the FSO_g is particularly effective for $\lambda = 0.9$. This is a result of this model requiring a higher mapping dimension N, but with the same amount of SMI, hence contributing to a higher ground value. For $\lambda = 0$, fewer samples of the CF are required, and the ground level can be ultimately neglected. Regarding the choice of p, it is observed that "easier" scenarios require very small values of p, while more difficult scenarios require higher values of p, thus a stronger regularization.

The advantage of the FSO is that, contrary to other methods, its computational complexity scales with the dimension of the mapping N, rather than with the sample size L. This is especially true for the KDE and LSMI, whose computational complexity is $O\left(L^2\right)$ [Nos+21], and for the NOCCO, which is $O\left(L^3\right)$ if no low-rank approximation is computed [LP20]. Furthermore, the LSMI estimator requires the cross-validation procedure, which increases the overall computational complexity well above the rest of the methods. The computational complexity of the AP estimator depends on the partitioning algorithm [SN12], although it still scales with L. In the case of the proposed estimator, the complexity stays at $O\left(N^2L\right)$, since it is required to multiply $N \times L$ matrices. The advantage becomes appealing for $N \ll L$. For reference, the largest N required in Figure 4.14 is $N = 601$, which is only attained for $L = 10^4$. Although it varies from model to model, choosing the estimation method is then a trade-off between accuracy and complexity.

To conclude, Figure 4.15 shows the performance of the asymptotic frequency-domain estimator $\hat{I}_\text{as}\left(X;Y\right)$ depicted in Theorem 4.2. First, since we are testing the bias with respect to the true SMI value, note that the bias does not tend to 0 but to the difference between the contaminated SMI and $I_\text{s}\left(X;Y\right) = 1$. That is, except for the case $p = 0$, which denotes that no regularization is being performed and the bias of the estimator diverges as N increases. This behaviour ratifies the need of regularizing the proposed estimator, and particularly for the case of estimating the SMI. Besides that, for an increasing value of N the asymptotic estimator converges to the regular estimator, as stated in (4.141). Another phenomenon that can be observed is that the stronger the regularization is, which entails a large value of p, the faster the convergence becomes, in the sense of the difference between the frequency-domain estimator and the original performance. This can be seen as a consequence of the increased convergence rate given a large smoothing variance, which not only governs the learning rate of $\hat{I}_\text{s}\left(X;Y\right)$, but also how rapidly $\hat{I}_\text{as}\left(X;Y\right)$ tends to the original estimator. The reason is that, given a strong decay of the CF, matrices $\hat{\mathbf{R}}_{x'}$ and $\hat{\mathbf{R}}_{y'}$ become virtually banded (as in Theorem 2.5), and the asymptotic behaviour of Szegö's theorem is rapidly achieved.

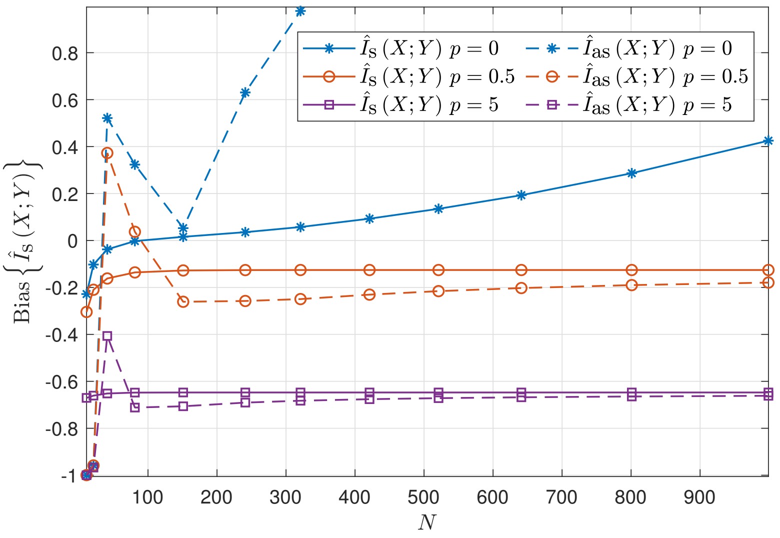


Figure 4.15: Bias of the estimated SMI as a function of the mapping dimension N for $I_{\text{s}}(X;Y) = 1$ and $L = 10^5$. The data is modelled following the GMM with $\lambda = 0.9$.

4.4 Concluding remarks

In this chapter a framework for estimating information measures based on measuring correlation and covariance in a given feature space has been developed. This tool maps the data onto a feature space of higher dimension to perform the required measurements of second-order statistics, avoiding the classic approach of plug-in methods that require a prior estimate of the density functions. Instead of relying on the kernel “trick”, as it is usual in kernel methods, our approach thrives on explicitly determining the feature map. Thanks to this, we can perform the estimation of information measures, provided that the desired surrogate can be expressed as the Frobenius norm of a quadratic measure of uncertainty/dependence.

While the required dimension in the discrete case is known and limited by the source cardinality, in the case of analog sources the feature space is limited by sampling the feature space. To reduce the implicit dimension entails an inherent error in the estimate, which becomes asymptotically unbiased in terms of the dimension N. Consequently, the estimate poses a classical trade-off between accuracy and complexity, but reformulated in terms of a simple sampling. The paradigm then becomes to determine the required sampling points, tied in the required dimension of the mapping itself.

To bestow the tool with the required capability of estimating the second-order Rényi entropy and the SMI under any kind of data or underlying distribution, the approach engages with different classical techniques from other fields, such as information theory, signal processing, and mathematical analysis. In front the imposed regularization for any high-dimensional mapping, we have proposed an approach based on a Gaussian regularization. It does not only improve the convergence rate of the estimator, but it also allows the inverse of the autocorrelation matrices in the feature space. Given that the feature space is based on the CF, the repercussions of the windowing in this space overlap with the ideas of classical spectral estimation problems. Finally, the required matrix operations have been linked with Szegö’s theorem, providing both an asymptotic theoretic guarantee (in the case of the second-order Rényi entropy) and a computation complexity reduction (in the case of the SMI). The choice of parameters is embedded in the parallelisms traced with the regularization approach, and, although there are some additional hyper-parameters, these have been unified in a single approach. Nonetheless, the choice of this single parameter is still vulnerable in front of different values of second-order Rényi entropy or SMI. The simulations support the good behaviour of the estimator and reveal that its performance can be on pair with well-known methods in the literature.

Following this chapter, we move to the other side of the balance. Instead of leveraging second-order statistics to estimate information measures, we next proceed to leverage information measures to solve problems usually performed by second-order statistics. The perspective is then shifted, but for the sake of unifying both paths. Specifically, we will focus on methods based on estimating an entropic measure, specifically the IP, and how the properties of these measures enhance the original measure.

Chapter 5

Entropy-based applications

This chapter deals with specific problems that benefit from an information-theoretic point of view. In particular, three different parameter estimation problems are analyzed by performing an entropy estimate first, and afterward the measure of the parameter of interest. While the previous chapter has focused on utilizing second-order statistics to estimate a measure of information, this chapter strives to substitute the second-order statics approach in concrete applications. The use of an entropic measure is driven by its property of being a more sensitive measure of the underlying PDF than the first and second-order moments. The estimator used in this chapter is the one addressed in Chapter 3. The reasons to choose this estimator are its inherent relation with second-order statistics, as it is studied in Subsection 3.3.3, as well as its depiction as a U-statistic in (3.86). The first reason allows us to generalize the second-order statistics by an entropy-based approach by tuning the kernel variance of the entropy estimator. For high values of kernel variance, the estimator behaves as the sample variance, which can be beneficial when the data is Gaussian or quasi-Gaussian. However, the entropy estimator gains in awareness of the complexity of the data for low values of kernel variance, resulting in a more robust estimate when the assumed distribution does not correspond with the real one. The second reason, regarding the U-statistics, pursues a specific modification of the entropy estimator that helps with the robustness of one of the entropy-based applications addressed here. Thanks to the computation of the absolute value of pair-wise differences, one can develop strategies that take advantage of this particular expression. As this section is being unravelled, all of these properties, advantages, and results will become clear.

The structure of this chapter is straightforward, where each section addresses a different application. First, Section 5.1 establishes the entropic measure utilized throughout this chapter, and analyzes its expected value and variance under GMMs. Section 5.2 studies the problem of estimating the determinant of the covariance matrix. The objective here is to gain robustness in front the of classical Gaussian assumption. Section 5.3 carries on the rationale of the previous section, but this time it is extended to the estimation of the coherence between two random sequences. Lastly, 5.4 changes the perspective and tackles a different problem, which is the estimation of the SNR in a digital communication channel.

5.1 Preliminaries

Before we delve into the core contents of this chapter, we will define again the IP estimator from (3.86), but particularized for complex-valued random variables. The objective is to determine the bias and variance for a given signal model based on a GMM, and to employ this knowledge to improve the entropy-based methods afterwards.

Consider the random variable $X \in \mathcal{X}$ with $\mathcal{X} \subseteq \mathbb{C}^N$, and a set of L i.i.d. observations $\mathbf{x}(i) = [x_1(i), ..., x_N(i)]^T$ for $i = 0, ..., L-1$. The IP estimator is then defined as

$$\hat{U}(X) = \frac{2}{L(L-1)} \sum_{0 \leq i < j \leq L-1} k_{\mathbf{V}}(\mathbf{x}(i), \mathbf{x}(j)), \tag{5.1}$$

where

$$k_{\mathbf{V}}(\mathbf{x}(i), \mathbf{x}(j)) = \frac{1}{\pi^N |\mathbf{V}|} \exp\Big(-(\mathbf{x}(i) - \mathbf{x}(j))^H \mathbf{V}^{-1} (\mathbf{x}(i) - \mathbf{x}(j))\Big). \tag{5.2}$$

Note that, in contrast with (3.86), here we let the kernel variance (the square of the kernel bandwidth) matrix $\mathbf{V} \in \mathbb{R}^{N \times N}$ to be any given positive semi-definite matrix, which represents a general expression of the covariance matrix of a multivariate Gaussian distribution. This generalization is used as the basis for characterizing the IP estimator, whose structure will be specified per application. Nevertheless, in most cases a diagonal matrix is used, where a particular kernel bandwidth is associated to each one of the dimensions of the original random variable.

We use this particular estimator for two reasons. On the one hand, it is an unbiased estimate of $\mathbb{E}_{f_X}\{k_{\mathbf{V}}(\mathbf{x}(i), \mathbf{x}(j))\}$, given that (5.1) is a U-statistic. Recall that, from Definition 10, a U-statistic is an unbiased estimate of a given parameter based on the average of a symmetric function whose argument is composed of combinations of i.i.d. observations. Consequently, if this expectation can be computed, we can guarantee that the expected value of the estimator is known, jointly with the bias to the true value of IP. On the other hand, some data models work particularly well with the Gaussian function in (5.2), and we can determine other relevant properties of the IP estimator, such as its variance. Specifically, by employing a general assumption that the data is modelled as a GMM the problem becomes manageable. This assumption can be seen as an extension to the well-known and widely used Gaussian assumption. As a matter of fact, we shall see how this assumption can be used, in tandem with the entropy-based processing, to enhance the solution to problems typically undertaken with just the Gaussian assumption. In the sequel, we proceed to determine the statistical model of the data, its corresponding true IP value, and the bias and variance of the estimator.

Let X be a random variable defined on the set $\mathcal{X}$ with PDF $f_X(\mathbf{x})$ that is distributed as a GMM:

$$X \sim \sum_{m=0}^{M-1} p_m \mathcal{CN}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m). \tag{5.3}$$

Recall that the IP is defined as

$$V(X) = \int_{\mathcal{X}} f_X^2(\mathbf{x})\, d\mathbf{x}. \tag{5.4}$$

Hence, the corresponding IP of the multivariate complex GMM (see Appendix 7.4.1) is the following:

$$V(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \\ \sum_{m,m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^N |\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'}|} \exp\Big(-(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'})^H (\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'})^{-1} (\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'})\Big). \tag{5.5}$$

Next we determine the expected value of the IP estimator based on the following proposition and corollary:

Proposition 5.1. *Let Z be a N-dimensional complex random variable defined on the set $\mathcal{Z}$ and modelled as a GMM such that $Z \sim \sum_{q=0}^{Q-1} p_q \mathcal{CN}(\mathbf{a}_q, \mathbf{W}_q)$ with PDF $f_Z(\mathbf{z})$, where $\mathbf{a}_q \in \mathbb{R}^N$ is the mean vector of the q-th component, and $\mathbf{W}_q \in \mathbb{R}^{N \times N}$ its covariance matrix. Then*

$$\mathbb{E}_{f_Z}\{k_{\mathbf{V}}(\mathbf{z})\} = \int_{\mathcal{Z}} k_{\mathbf{V}}(\mathbf{z}) f_Z(\mathbf{z})\, d\mathbf{z} = \sum_{q=0}^{Q-1} \frac{p_q}{\pi^N |\mathbf{W}_r + \mathbf{V}|} \exp\Big(-\mathbf{a}_q^H (\mathbf{W}_q + \mathbf{V})^{-1} \mathbf{a}_q\Big), \tag{5.6}$$

where $k_{\mathbf{V}}(\mathbf{z})$ is defined as in (5.2)*.*

Proof. See Appendix 7.4.2. ■

Corollary 5.1.1. *Given Proposition 5.1, the expected value of the IP estimator from* (5.1) *is as follows:*

$$\mathbb{E}_{f_X}\left\{\hat{U}\left(X\right)\right\} = \frac{2}{L(L-1)} \sum_{0\leq i<j\leq L-1} \mathbb{E}_{f_X}\left\{k_{\mathbf{V}}\left(\mathbf{x}(i),\mathbf{x}(j)\right)\right\} = \mathbb{E}_{f_X}\left\{k_{\mathbf{V}}\left(\mathbf{x}(i),\mathbf{x}(j)\right)\right\} \tag{5.7a}$$

$$= \sum_{m,m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^N \left|\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'} + \mathbf{V}\right|} \exp\left(-\left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\right)^H \left(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'} + \mathbf{V}\right)^{-1} \left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\right)\right) \tag{5.7b}$$

$$= V\left(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m + \mathbf{V}/2\right) \tag{5.7c}$$

Proof. From the argument of (5.1) we can determine that $\mathbf{z} = \mathbf{x}(i) - \mathbf{x}(j)$, which yields $Q = M^2$, $\mathbf{a}_q = \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}$ and $\mathbf{W}_q = \boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'}$. ■

As a consequence of employing (5.1), the expected value of the estimator just becomes a contaminated version of the true IP, determined by the kernel variance matrix $\mathbf{V}$. In contrast with (5.1), the resulting expected value does take the collision between two equal Gaussian components into consideration. In other words, while in (5.1) we discard the data-independent additive constant provided by the samples with $i = j$, the resulting IP recovers the original exponential summation of pairwise Gaussian components differences. Given this result, the bias of the estimator is directly computed as

$$\text{Bias}\left\{\hat{U}\left(X\right)\right\} = V\left(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right) - V\left(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m + \mathbf{V}/2\right). \tag{5.8}$$

Clearly, the bias then tends to zero as the elements of $\mathbf{V}$ also tend to zero. This result is a particularization of the analysis in (3.79), but applied to multivariate random variables.

Next, we perform a similar analysis to determine the variance of the estimator. To do so, we will take advantage of the following lemma [1]:

Lemma 5.1. *Let Z_1 and Z_2 be D-dimensional complex random variables modelled as a GMM such that $Z_1 \sim \sum_{r=0}^{R-1} p_r \mathcal{CN}\left(\mathbf{a}_r, \mathbf{W}\right)$ and $Z_2 \sim \sum_{r'=0}^{R'-1} p_{r'} \mathcal{CN}\left(\mathbf{a}_{r'}, \mathbf{W}\right)$, where $\mathbb{E}_{f_{Z_1,Z_2}}\left\{\mathbf{z}_1\mathbf{z}_2\right\} = \gamma\mathbf{W}$, $\mathbf{a}_r \in \mathbb{R}^D$ and $\mathbf{a}'_r \in \mathbb{R}^D$ are the mean vectors of the r-th and r'-th components, respectively, and $\mathbf{W} \in \mathbb{R}^{D\times D}$ is a shared covariance matrix. Then*

$$\mathbb{E}_{f_{Z_1,Z_2}}\left\{k_{\mathbf{V_1}}\left(\mathbf{z}_1\right) k_{\mathbf{V}_2}\left(\mathbf{z}_2\right)\right\} = \int_{\mathcal{Z}} k_{\mathbf{V_1}}\left(\mathbf{z}_1\right) k_{\mathbf{V}_2}\left(\mathbf{z}_2\right) f_{Z_1,Z_2}\left(\mathbf{z}_1,\mathbf{z}_2\right) \mathrm{d}\mathbf{z}_1 \mathrm{d}\mathbf{z}_2 \tag{5.9a}$$

$$= \frac{1}{\pi^{2D}\left|\mathbf{U}\right|} \sum_{r=0}^{R-1} \sum_{r'=0}^{R'-1} \exp\left(-\begin{bmatrix} \mathbf{a}_r \\ \mathbf{a}_{r'} \end{bmatrix}^H \mathbf{U}^{-1} \begin{bmatrix} \mathbf{a}_r \\ \mathbf{a}_{r'} \end{bmatrix}\right), \tag{5.9b}$$

where

$$\mathbf{U} = \begin{bmatrix} \mathbf{W} + \mathbf{V}_1 & \gamma\mathbf{W} \\ \gamma\mathbf{W} & \mathbf{W} + \mathbf{V}_2 \end{bmatrix}. \tag{5.10}$$

Proof. Let $\mathbf{z} = \left[\mathbf{z}_1\ \mathbf{z}_2\right]^T$. Then, (5.9) can be obtained through Proposition 5.1 by fixing

$$\mathbf{W}_q = \begin{bmatrix} \mathbf{W} & \gamma\mathbf{W} \\ \gamma\mathbf{W} & \mathbf{W} \end{bmatrix}, \qquad \mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix}, \qquad \mathbf{a}_q = \begin{bmatrix} \mathbf{a}_r \\ \mathbf{a}_{r'} \end{bmatrix}, \tag{5.11}$$

$N = 2D$, and $Q = RR'$. ■

[1] For simplicity, the covariance matrices of each Gaussian component are defined now equally, thus we have dropped the subindex associated with each of them.

Using Proposition 5.1 and Lemma 5.1, we can now finally state the following proposition:

Proposition 5.2. *Let $\hat{U}(X)$ be the IP estimator given in* (5.1)*, where X is a N-dimensional complex random variable with $X \sim \sum_{m=0}^{M-1} p_m \mathcal{CN}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}_m \in \mathbb{R}^N$ the mean vector, and $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ the covariance matrix. The "cross-covariance" of the IP estimator given two kernel variance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ is the following*

$$\mathrm{Cov}\left\{\hat{U}_{\mathbf{V}_1}(X)\hat{U}_{\mathbf{V}_2}(X)\right\} = \mathbb{E}_{f_X}\left\{\hat{U}_{\mathbf{V}_1}(X)\hat{U}_{\mathbf{V}_2}(X)\right\} - \mathbb{E}_{f_X}\left\{\hat{U}_{\mathbf{V}_1}(X)\right\}\mathbb{E}_{f_X}\left\{\hat{U}_{\mathbf{V}_2}(X)\right\} \tag{5.12a}$$

$$= \frac{4(L-2)(a-c)+2(b-c)}{L(L-1)}, \tag{5.12b}$$

where

$$a = \frac{1}{\pi^{2N}|\mathbf{U}_a|}\sum_{m,m',n=0}^{M-1} p_m p_{m'} p_n \exp\left(-\begin{bmatrix}\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\\ \boldsymbol{\mu}_n-\boldsymbol{\mu}_{m'}\end{bmatrix}^H \mathbf{U}_a^{-1}\begin{bmatrix}\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\\ \boldsymbol{\mu}_n-\boldsymbol{\mu}_{m'}\end{bmatrix}\right), \tag{5.13}$$

$$b = \frac{1}{\pi^{2N}|\mathbf{U}_b|}\sum_{m,m'=0}^{M-1} p_m p_{m'} \exp\left(-\begin{bmatrix}\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\\ \boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\end{bmatrix}^H \mathbf{U}_b^{-1}\begin{bmatrix}\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\\ \boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\end{bmatrix}\right), \tag{5.14}$$

and

$$c = \mathbb{E}_{f_X}\left\{\hat{U}_{\mathbf{V}_1}(X)\right\}\mathbb{E}_{f_X}\left\{\hat{U}_{\mathbf{V}_2}(X)\right\} \tag{5.15a}$$

$$= V(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}+\mathbf{V}_1/2)\,V(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}+\mathbf{V}_2/2)\,. \tag{5.15b}$$

as in Corollary 5.1.1, for

$$\mathbf{U}_a = \begin{bmatrix} 2\boldsymbol{\Sigma}+\mathbf{V}_1 & \boldsymbol{\Sigma}\\ \boldsymbol{\Sigma} & 2\boldsymbol{\Sigma}+\mathbf{V}_2\end{bmatrix}, \qquad \mathbf{U}_b = \begin{bmatrix} 2\boldsymbol{\Sigma}+\mathbf{V}_1 & 2\boldsymbol{\Sigma}\\ 2\boldsymbol{\Sigma} & 2\boldsymbol{\Sigma}+\mathbf{V}_2\end{bmatrix}. \tag{5.16}$$

Proof. See Appendix 7.4.3 ■

From this analysis is it clear that $\hat{U}(X)$ is a consistent estimator for any finite value of a, b and c. In particular, the variance of the estimator decreases inversely proportional to L as $L \to \infty$. The rationale is then similar to the one in Subsection 3.3.3. As L increases, both variance and bias tend to zero, given that the required kernel variance also tends to $\mathbf{0}$. It is noteworthy that a similar analysis is provided by Príncipe in [Prí10], albeit for univariate random variables and for the estimator in (3.73). The term b is ignored in [Prí10] as it becomes asymptotically negligible. However, as we advance in the analysis of bias and variance of the particular estimators for different applications, we will see that we are interested in the case in which the kernel variance matrix becomes close to $\mathbf{0}$ with a fixed L. In these cases, b is prone to dominate the variance of the estimators, and it should be computed likewise as a and c.

With this, we are now in terms to move into the entropy-based applications based on the U-statistic estimator in (5.1).

5.2 Covariance determinant estimation

The first entropy-based application to be addressed is the variance estimation of a random scalar sequence. The classical approach to this problem is to make a prior assumption that the data is Gaussian, referred to as nominal conditions. Under this setting, it is well-known that the maximum likelihood estimator of the covariance matrix is the sample covariance matrix (see, for instance, (2.27)). However, the Gaussian assumption does not always hold in practice. In many applications,

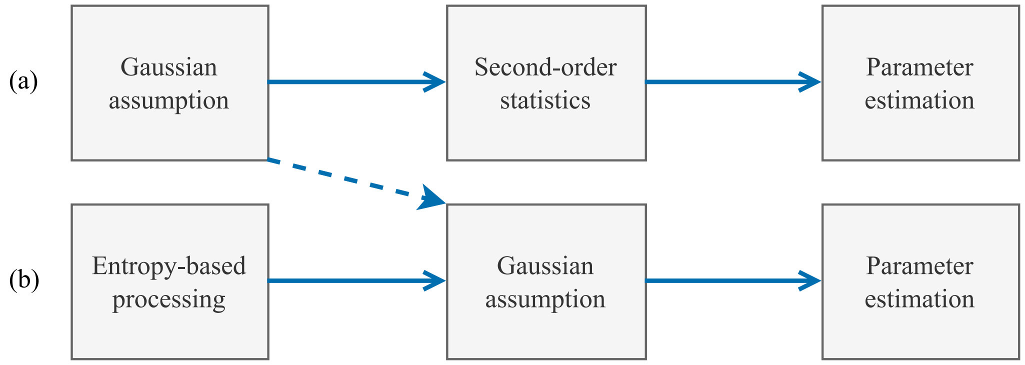


Figure 5.1: Main rationale of the robust entropy-based approach (b) in front of the classical Gaussian assumption (a).

the underlying distributions tend to be leptokurtic, i.e. the distribution has a longer tail than the mesokurtik (zero kurtosis regardless of the parameters) Gaussian distribution. A typical occurrence of long tails is given by faulty observations, considered outliers that are detached from the expected range of operation. In these cases, the Gaussian assumption may become detrimental to the overall processing (see [PSQ13] and references therein).

We refer to robust signal processing to these methods that are conscious of the faulty Gaussian assumption and develop an alternative that is resistant against the effects of deviations. In particular, a common problem is to develop a robust estimation of the mean and covariance matrix from a random sequence [SBP15], which arises in multiple areas [Zou+12]. A common practice is to directly assume that the underlying model has a heavy-tailed distribution due to the outlier contamination [Hub64], or some elliptical distribution [Tyl87; KT88].

Here, we will address the estimate of just the *determinant* of the covariance matrix in a robust manner. In some applications, only the determinant is required, such as in multichannel signal detection [LV01; Ram+11]. We first focus on the determinant because it naturally links with the IP of a Gaussian distribution. In particular, we can see from (5.5) that, provided the data is distributed as a zero-mean Gaussian variable (corresponding to a GMM with just one component), the IP is proportional to the determinant of the covariance matrix. As a consequence, the retrieval of the determinant of the covariance matrix from an entropy estimate becomes straightforward.

An entropic measure may be more favorable to estimate the determinant of the covariance matrix than the classic sample approach. In particular, a favorable property of entropy (or the IP in this case) is that only depends on the probability of the events, and not on their magnitude. This property may be particularly advantageous in front of outliers, which is an occurrence with low probability, but with great impact for the sample estimate. This property contrasts with the use of the Huber loss function [Hub64], typically used in robust statistics, which penalizes the low probability events so their magnitude is not as impactful in the overall estimate. An entropy-based approach acquires robustness in a natural manner, thanks to the properties of the entropic measure itself rather than by a trade-off between accuracy and closeness between the assumed and the real model. At the end, the proposed approach here is an information-theoretic method to the estimation of variance, which substitutes a classical second-order statistics approach (to measure sample covariance matrices) to one based on measures of information.

Nevertheless, the proposed approach still considers a Gaussian assumption in its formulation. The expected IP is assumed to be from a normal distribution for a proper isolation of the determinant of the covariance matrix. However, this assumption is not made from the point of view of the underlying distribution, but from the perspective of the IP. In particular, we subvert the Gaussian assumption not by omitting it, but by a displacement of its typical order. Figure 5.1 illustrates this philosophy by comparing the classical approach versus the one proposed here. The Gaussian assumption is then taken in a second step, once the nonparametric entropy measure has been performed. This allows us to relax the penalty of infringing the nominal condition, gaining in both robustness and simplicity with respect to other formulations that consider more complex underlying distributions.

In the sequel, the estimator is derived, the impact of the free parameters of the IP estimator are unveiled, and an analysis of the performance is provided.

5.2.1 Problem formulation

Let us assume that the samples are drawn from a N-dimensional complex Gaussian distribution $X \sim \mathcal{CN}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with

$$f_X\left(\mathbf{x}\right) = \frac{1}{\pi^N \left|\boldsymbol{\Sigma}\right|} \exp\Big(- \left(\mathbf{x} - \boldsymbol{\mu}\right)^H \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\Big). \tag{5.17}$$

In nominal conditions, the IP can directly be obtained from (5.5), which results in

$$V\left(X\right) = \frac{1}{\pi^N \left|2\boldsymbol{\Sigma}\right|}. \tag{5.18}$$

Clearly, since the determinant of the covariance matrix is inversely proportional to the IP, our focus will be on first estimating the IP and then measuring the determinant. Another relevant property is that, for a sole Gaussian component, the IP is invariant in front of the mean value of the distribution, a shared property with other entropic measures.

For convenience, we will use the Modified Information Potential (MIP), which will further simplify the consequent expressions and the bias and variance analysis. Consider L observations of the random variable X denoted as $\mathbf{x}(i) \in \mathbb{C}^N$ for $i = 0, ..., L-1$. The MIP estimator is then defined as follows:

$$\hat{W}\left(X\right) = \frac{2\pi^N \left|\mathbf{V}\right|}{L(L-1)} \sum_{0 \leq i < j \leq L-1} k_{\mathbf{V}}\left(\mathbf{x}(i), \mathbf{x}(j)\right), \tag{5.19}$$

where $k_{\mathbf{V}}\left(\mathbf{x}(i), \mathbf{x}(j)\right)$ is given in (5.2). It can be seen that the only difference between $\hat{U}\left(X\right)$ and the MIP estimator $\hat{W}\left(X\right)$ is the constant factor $1/\pi^N \left|\mathbf{V}\right|$, i.e.

$$\hat{W}\left(X\right) = \pi^N \left|\mathbf{V}\right| \hat{U}\left(X\right). \tag{5.20}$$

Therefore, there is no significant change in the previous analyses of the base estimator apart from a constant value that depends on the determinant of the kernel variance matrix. The expected value can be directly obtained from Corollary 5.1.1 (for $M = 1$), and becomes the following:

$$\mathbb{E}_{f_X}\left\{\hat{W}\left(X\right)\right\} = \frac{\left|\mathbf{V}\right|}{\left|2\boldsymbol{\Sigma} + \mathbf{V}\right|}, \tag{5.21}$$

which is just $\mathbb{E}_{f_X}\left\{\hat{W}\left(X\right)\right\} = \pi^N \left|\mathbf{V}\right| \mathbb{E}_{f_X}\left\{\hat{U}\left(X\right)\right\}$. Consequently, if we want to retrieve the determinant of the covariance matrix, the contamination added by the kernel variance matrix $\mathbf{V}$ must be considered. On the other hand, the variance of the MIP can be obtained from Proposition 5.2 by letting $M = 1$ and $\mathbf{V}_1 = \mathbf{V}_2$, which results in

$$\text{Var}\left\{\hat{W}\left(X\right)\right\} = \mathbb{E}_{f_X}\left\{\hat{W}^2\left(X\right)\right\} - \mathbb{E}_{f_X}\left\{\hat{W}\left(X\right)\right\}^2 = \frac{4\left(L-2\right)\left(a-c\right) + 2\left(b-c\right)}{L\left(L-1\right)}, \tag{5.22}$$

where

$$a = \frac{\left|\mathbf{V}\right|^2}{\begin{vmatrix} 2\boldsymbol{\Sigma} + \mathbf{V} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & 2\boldsymbol{\Sigma} + \mathbf{V} \end{vmatrix}} = \frac{\left|\mathbf{V}\right|^2}{\left|\boldsymbol{\Sigma} + \mathbf{V}\right| \left|3\boldsymbol{\Sigma} + \mathbf{V}\right|} \tag{5.23a}$$

$$b = \frac{|\mathbf{V}|^2}{\begin{vmatrix} 2\boldsymbol{\Sigma}+\mathbf{V} & 2\boldsymbol{\Sigma} \\ 2\boldsymbol{\Sigma} & 2\boldsymbol{\Sigma}+\mathbf{V} \end{vmatrix}} = \frac{|\mathbf{V}|^2}{|\mathbf{V}|\,|4\boldsymbol{\Sigma}+\mathbf{V}|} \tag{5.23b}$$

$$c = \frac{|\mathbf{V}|^2}{|2\boldsymbol{\Sigma}+\mathbf{V}|^2}. \tag{5.23c}$$

It is worth noting that the resulting variance tends to zero as $\mathbf{V}$ goes to zero, as can be seen from a, b and c. This result is different from the classical IP estimate, whose variance increases as the kernel variance decreases, and further requiring $L \to \infty$ for its variance to go to zero. However, the MSE of both estimators tend to zero as L increases, where $\hat{W}(X)$ and $\hat{U}(X)$ converge to the real value of MIP and IP, respectively. This property can be retrieved from the KDE estimator in Subsection 3.3.1, given that both IP and MIP estimators are constructed from the Parzen-Rosenblatt method.

in both cases the estimator is consistent, since for fixed a, b and c the estimator converges in probability to the real value $W(X)$ in the case of the MIP, or $U(X)$ in the case of the IP, with respect to L.

These observations reinforce the idea that the MIP may be more appropriate for a parameter estimation. However, note that the resulting expression could not be used as a pure IP estimator, since the expected value also goes to zero as $\mathbf{V}$ decreases, and it is only used here for the purpose of determining the determinant of the covariance matrix. The consequences of this choice and the analysis of the parameters will be provided in the following subsections.

With the knowledge of the expected value and the variance, the problem then moves to constructing an estimator of the form

$$\hat{D}_{IP} = \xi_{\mathbf{V}}\left(\hat{W}(X)\right), \tag{5.24}$$

where $\xi_{\mathbf{V}} : \mathbb{R} \to \mathbb{R}$ is a monotonic decreasing function of $\hat{W}(X)$, whose bias and variance depend on the parameter $\mathbf{V}$. Note that $\hat{D}_{IP}$ corresponds to the bottom rationale from Figure 5.1. On the contrary, the upper rationale would be the equivalent of estimating the sample covariance matrix with

$$\hat{\boldsymbol{\Sigma}}_S = \frac{1}{L}\mathbf{X}\mathbf{P}\mathbf{X}^H, \tag{5.25}$$

where $\mathbf{X} = [\mathbf{x}(0), ..., \mathbf{x}(L-1)]$, $\mathbf{P}_{\mathbf{1}}^{\perp} = \mathbf{I}_N - \mathbf{1}\mathbf{1}^T/L$, and then by

$$\hat{D}_{\mathrm{S}} = \left|\hat{\boldsymbol{\Sigma}}_S\right|. \tag{5.26}$$

These expressions illustrate the main differences between both approaches, where the nominal condition is assumed separately in different stages. The entropy-based rationale is again illustrated in Figure 5.2, where it has now been updated (from Figure 5.1) to specify the problem of the estimation of the determinant of the covariance matrix.

5.2.2 Performance analysis

For the purpose of properly characterizing the estimator, and for clarity, we focus here on the univariate case $N = 1$, where $\mathbf{V}$ and $\boldsymbol{\Sigma}$ are now v^2 and σ^2, respectively. To be concrete, the kernel function employed to estimate the MIP in (5.19) is the following:

$$h_v\left(x(i), x(j)\right) = \exp\left(-\frac{|x(i)-x(j)|^2}{v^2}\right). \tag{5.27}$$

This particular case transforms the estimation of the determinant of the covariance matrix into the estimation of variance σ^2. In this case, by isolating (5.21) we have that

$$\hat{D}_{\mathrm{IP}} = \hat{\sigma}^2 = \xi_v\left(\hat{W}(X)\right) = \frac{v^2}{2}\left(\frac{1}{\hat{W}(X)} - 1\right), \tag{5.28}$$

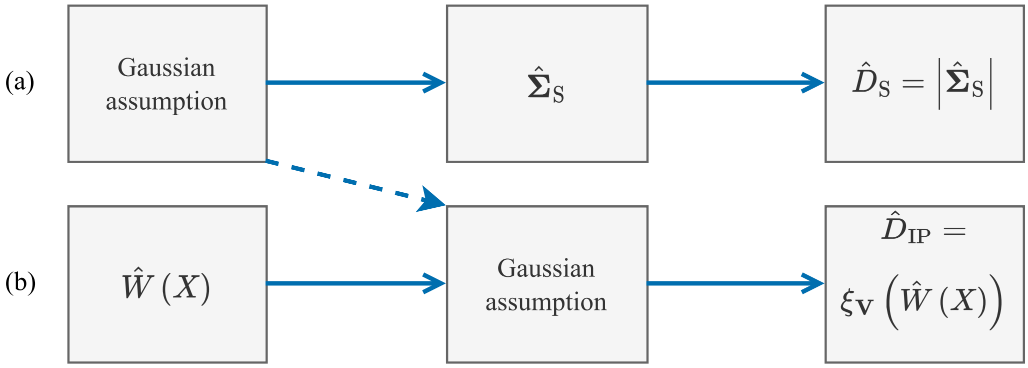

(a)
Gaussian assumption
$\hat{\mathbf{\Sigma}}_{\mathrm{S}}$
$\hat{D}_{\mathrm{S}} = \left|\hat{\mathbf{\Sigma}}_{\mathrm{S}}\right|$
(b)
$\hat{W}(X)$
Gaussian assumption
$\hat{D}_{\mathrm{IP}} = \xi_{\mathbf{V}}\left(\hat{W}(X)\right)$

Figure 5.2: Entropy-based approach for the problem of estimating the determinant of the covariance matrix (b) in front of the sample covariance estimator approach (a).

which is, ultimately, an entropy-based estimator of the variance. Henceforth, we also define

$$\mathbb{E}_{f_X}\left\{\hat{W}(X)\right\} = W(X) = \frac{v^2}{2\sigma^2 + v^2}. \tag{5.29}$$

Note that this assessment can be done following the unbiased property of the U-statistic in (5.21).

Next, we analyze the bias and variance of this new estimator with the purpose of determining the role of the main free parameter, the kernel bandwidth v, as well as determining the efficiency of the estimator. For the bias, we just use Jensen's inequality to state that

$$\mathbb{E}_{f_X}\left\{\hat{\sigma}^2\right\} \geq \frac{v^2}{2}\left(\frac{1}{\mathbb{E}_{f_X}\left\{\hat{W}(X)\right\}} - 1\right) = \frac{v^2}{2}\left(\frac{2\sigma^2 + v^2}{v^2} - 1\right) = \sigma^2, \tag{5.30}$$

given that $1/x$ is a concave function for $x > 0$, and $\hat{W}(X) > 0$ for $v > 0$. As a consequence, we have

$$\mathrm{Bias}\left\{\hat{\sigma}^2\right\} = \mathbb{E}_{f_X}\left\{\hat{\sigma}^2\right\} - \sigma^2 \geq 0. \tag{5.31}$$

On the one hand, the previous expression just confirms that the bias is always nonnegative. Therefore, for any given value of v, the estimate tends to σ^2 from above. This property contrasts with the second-order Rényi entropy estimator from Chapter 4, which had two different regimes depending on the number of available data and the added contamination. On the other hand, the bias is well-behaved, given that the estimator $\hat{W}(X)$ is a consistent estimate of $W(X)$, thus the estimator $\hat{\sigma}^2$ is also well-behaved.

In order to determine the variance of the estimator, a small error analysis is performed, following a similar approach as López-Valcarce et al. in [LM07]. Consider a Taylor expansion of $\xi_v\left(\hat{W}(X)\right)$ up to the first order around the point $\hat{W}(X) = W(X)$ is given by

$$\hat{\sigma}^2 \approx \sigma^2 + \left(\hat{W}(X) - W(X)\right)\left.\frac{\partial \xi_v(z)}{\partial z}\right|_{z=W(X)}, \tag{5.32}$$

where the derivative yields the following result:

$$\left.\frac{\partial \xi_v(z)}{\partial z}\right|_{z=W(X)} = \left.\frac{\partial}{\partial z}\left(\frac{v^2}{2}\left(\frac{1}{z} - 1\right)\right)\right|_{z=W(X)} = -\frac{v^2}{2U_M(X)^2} = -\frac{\left(2\sigma^2 + v^2\right)^2}{2v^2}. \tag{5.33}$$

Consequently, the estimator variance can be approximated by

$$\mathrm{Var}\left\{\hat{\sigma}^2\right\} \approx \mathrm{Var}\left\{\hat{W}(X)\right\}\left(\left.\frac{\partial \xi_v(z)}{\partial z}\right|_{z=W(X)}\right)^2 = \mathrm{Var}\left\{\hat{W}(X)\right\}\frac{\left(2\sigma^2 + v^2\right)^4}{4v^4}, \tag{5.34}$$

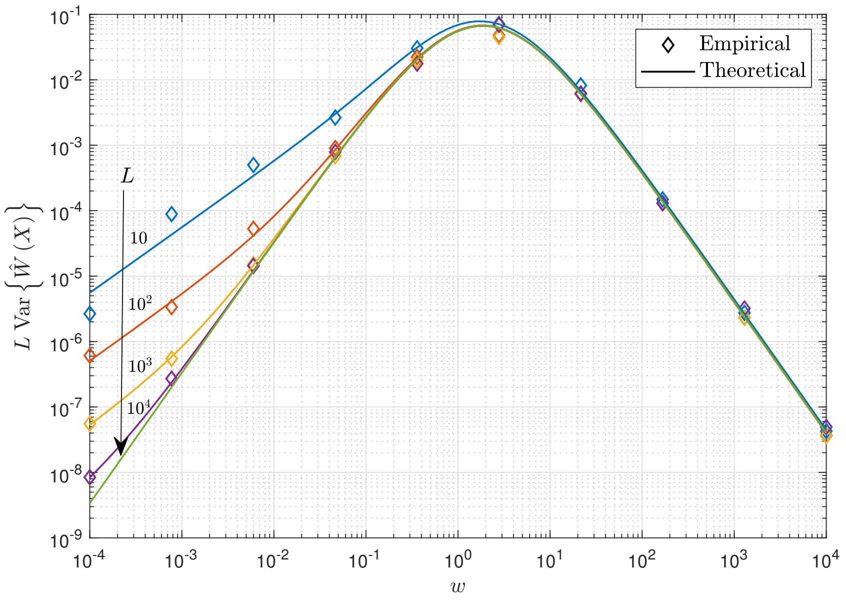


Figure 5.3: Normalized variance of the MIP estimator as a function of the relative kernel variance for different values of data size.

where Var $\left\{\hat{W}(X)\right\}$ is given in (5.22). We define here the relative kernel variance

$$w = \frac{v^2}{\sigma^2} \tag{5.35}$$

in order to better analyze the variance of the estimator. Notice that it is a common practice in kernel methods to let the kernel bandwidth/variance depend on the true standard deviation/variance of the data, e.g. (3.55). It is then within reason to evaluate the impact of v^2 in a relative manner, which also generalizes for any given σ^2. The relative variance is then written as follows:

$$\widetilde{\text{Var}}\left\{\hat{\sigma}^2\right\} = \frac{\text{Var}\left\{\hat{\sigma}^2\right\}}{\sigma^4} \approx \text{Var}\left\{\hat{W}(X)\right\} \frac{(2+w)^4}{4w^2}. \tag{5.36}$$

The variance of the MIP estimator is given in (5.22), which can be written in terms of the relative variance as follows:

$$a = \frac{w^2}{(w+1)(w+3)}, \qquad b = \frac{w^2}{w(w+4)}, \qquad c = \frac{w^2}{(w+2)^2}. \tag{5.37}$$

Note that, unlike the normalization in (5.36), the previous expressions are not normalized, but only a change of variable has been applied, which yields

$$\begin{aligned}\text{Var}\left\{\hat{W}(X)\right\} =& \frac{4(L-2)}{L(L-1)}\left(\frac{w^2}{(w+1)(w+3)} - \frac{w^2}{(w+2)^2}\right) \\ &+ \frac{2}{L(L-1)}\left(\frac{w^2}{w(w+4)} - \frac{w^2}{(w+2)^2}\right).\end{aligned} \tag{5.38}$$

From this result, one can observe that the relative variance of the estimator goes to infinity as $w \to 0$, independently from the fact that $a, b, c \to 0$. The reason for this is that b goes to zero as $O(w)$ instead of $O\left(w^2\right)$ as happens with a and c, i.e. b decreases slowly than a and c, and this is why b was not neglected in (5.12) nor in (5.22) as it is done in [Príl10, Sec. 2.5]. Figure 5.3 illustrates this effect by showing the normalized variance (with respect to L) of the MIP estimator both empirically and theoretically, given in (5.38). While for low and moderate values of w the normalized variance depends on the value of L, high values of w mitigate this dependency and tend to the same value. These discrepancies at low w values are mostly due to b, which increases the relative variance as L decreases. Nonetheless, the penalty term provided by b becomes negligible as L increases.

In terms of the variance estimate $\hat{\sigma}^2$, it should be noted that its relative estimators variance $\widetilde{\text{Var}}\left\{\hat{\sigma}^2\right\}$ does not go to zero as w decreases, but to infinity. This property is recovered from the nonmodified IP estimator, which only goes to zero as $L \to \infty$. This property can be seen as a consequence of estimating a parameter of the data, instead of just the MIP. As a result, the choice of w is again a rather complex choice, provided that the final application is to estimate true descriptor of the data and not some modified intermediate value.

Lastly, we would like to point out that the strategy for evaluating the relative variance here in (5.34) differs slightly from the one used in our published paper [CRV17], which is based on a small perturbation analysis. The reason behind this choice is to unify all three applications addressed in this chapter, given that the analysis of the variance of the SNR is cumbersome with the small perturbation approach. Although in [CRV17] the approximation is performed up to the second order, which ultimately conveys an additional term in (5.36), they are the same result up to the first order. Nevertheless, both approaches will render the same results in the hereunder asymptotic analysis.

5.2.2.1 Asymptotic performance

We are now interested in determining the behaviour of the estimator in the case of large data size L. For $w > 0$, we can gather from (5.22) that

$$\lim_{L\to\infty} L\text{Var}\left\{\hat{W}\left(X\right)\right\} = 4\left(a-c\right). \tag{5.39}$$

Therefore, by joining (5.36) and (5.38) we have

$$\lim_{L\to\infty} L\widetilde{\text{Var}}\left\{\hat{\sigma}^2\right\} = 4\left(a-c\right)\frac{\left(2+w\right)^4}{4w^2} = \left(\frac{w^2}{\left(w+1\right)\left(w+3\right)} - \frac{w^2}{\left(w+2\right)^2}\right)\frac{\left(2+w\right)^4}{4w^2} \tag{5.40a}$$

$$= \frac{\left(w+2\right)^2}{\left(w+1\right)\left(w+3\right)}. \tag{5.40b}$$

The limits regarding the relative kernel variance are given by $w \to 0$ and $w \to \infty$, yielding $4/3$ and 1, respectively[2]. We can therefore write the following

$$1 \leq \lim_{L\to\infty} L\widetilde{\text{Var}}\left\{\hat{\sigma}^2\right\} \leq \frac{4}{3}. \tag{5.41}$$

From this result, it can be seen that the estimator is asymptotically efficient as $w \to \infty$, since it reaches the Cramér-Rao Bound (CRB) of the sample variance estimator, namely $\hat{\sigma}_S^2$, which is σ^4/L for circular complex data (and $2\sigma^4/L$ for real-valued data, see [Kay93]). This is not a surprising outcome since it is known from the large kernel bandwidth analysis in Subsection 3.3.3 that the IP estimator (and equivalently the MIP) becomes a function of the sample variance estimator. On the other hand, the variance increases as w decreases, but never more than $4/3$, which represents the maximum asymptotic penalty. In fact, the small kernel bandwidth values are the ones interesting for the purpose of gaining robustness in front of outliers, as we will see in Subsection 5.2.3.

5.2.2.2 Threshold effect

Before advancing to the robust side of the proposed method, we will determine the required condition for the previous asymptotic analysis to be valid. In particular, it is assumed that L is large

[2] For $w \to \infty$ it is also required that (5.40) is monotonically decreasing. Since its derivative $-2(w+2)/((w+1)^2(w+3)^2)$ is negative for $w > 0$, then it is strictly decreasing.

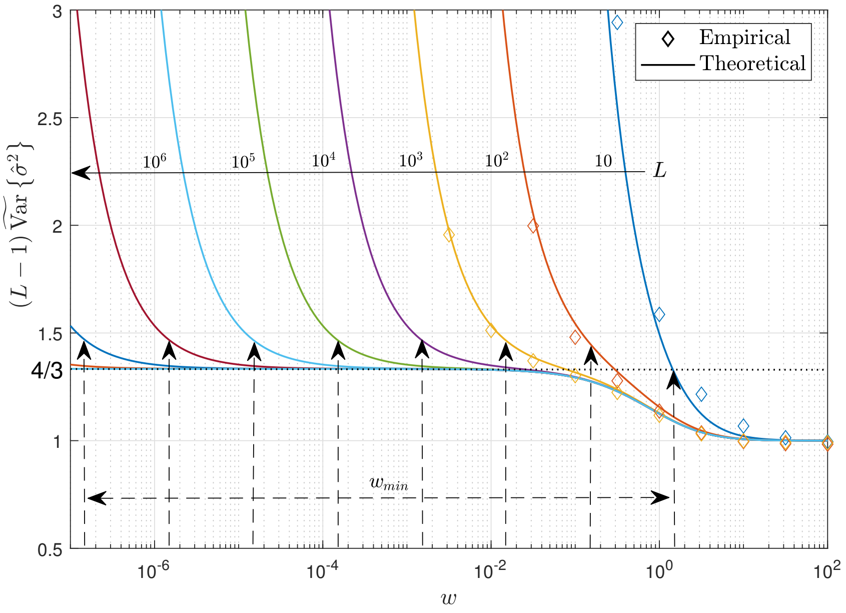


Figure 5.4: Variance amplification with respect the CRB as a function of the relative kernel variance for complex-valued data with different L.

enough such that the impact of b in Var $\left\{\hat{\sigma}^2\right\}$ from (5.22) is negligible. To be concrete, the following is assumed for $\alpha \gg 1$:

$$\frac{2}{L\,(L-1)}\,(b-c) < \frac{1}{\alpha}\frac{4\,(L-2)}{L\,(L-1)}\,(a-c)\,, \tag{5.42}$$

which translates to

$$(b-c) < \frac{1}{\alpha}2\,(L-2)\,(a-c)\,. \tag{5.43}$$

Using (5.37) we can state the following relation between the data size L, the relative kernel variance w and α:

$$L > 2\left(\alpha\frac{(w+1)\,(w+3)}{w\,(w+4)}+1\right). \tag{5.44}$$

It can be seen that for small values of w, higher values of L are required so that the estimator reaches the asymptotic behaviour. If this inequality is not fulfilled, the variance of the estimator will not correspond to the given approximation with a much higher value than desired.

This condition also serves as a lead hint to the determination of w as a function of L. For this purpose, consider that we have a very small relative kernel variance w so that (5.44) can be simplified to

$$w > \frac{3\alpha}{2L}. \tag{5.45}$$

If we fix $\alpha = 10$, for example, we obtain an approximate minimum relative kernel variance such that

$$w_{\text{min}} \approx \frac{15}{L}. \tag{5.46}$$

Although the choice of kernel variance is crucial in the estimation of the MIP, we can at least guarantee that the variance of the estimator will not be more amplified than approximately $4/3$ with respect the CRB if (5.46) is fulfilled, following (5.41).

To illustrate this threshold effect, Figure 5.4 shows the variance of the estimator under nominal conditions as a function of w and for different values of L. On the one hand, it can be seen that the normalized variance tends to the factor of $4/3$ with respect to the CRB for moderately small

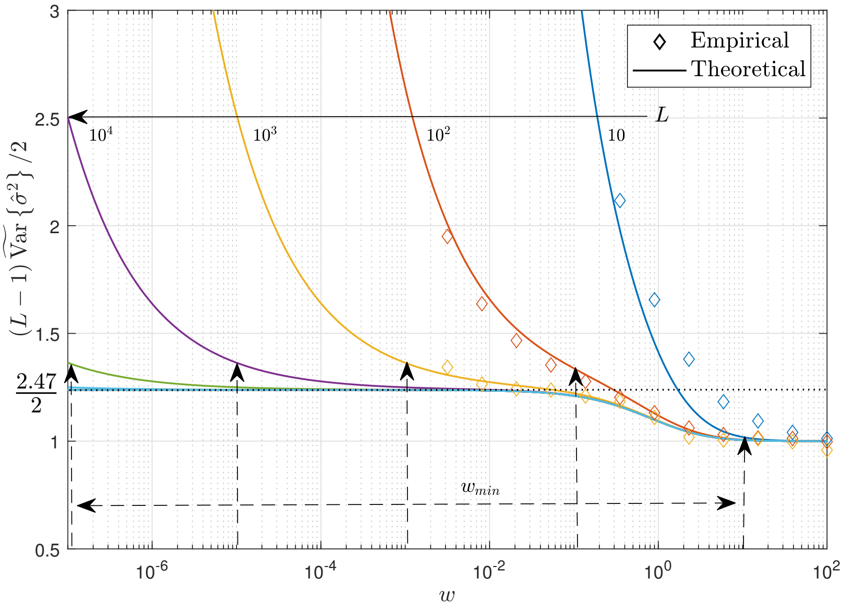


Figure 5.5: Variance amplification with respect the CRB for real-valued variables.

w and increasing values of L from (5.41). Consequently, the larger is L the smaller can be the relative kernel variance before the variance of the estimator is overly amplified. This also implies that the larger is L the less critical becomes to choose an adequate value of w in order to have sufficiently high accuracy under nominal conditions. Figure 5.4 also shows the required minimum w for achieving the asymptotic regime for each value of L, following (5.46) and indicated by the upward arrows. The normalized variances of the estimator corresponding to these w_{min} values are quite close the $4/3$ factor, proving that the rule for determining a floor value of w for each L is behaving as expected.

Real-valued case It is also worth noting that, in the case of real-valued data, the resulting asymptotic relative variance of the estimator results in (the computation is quickly sketched in Appendix 7.4.4):

$$\lim_{L\to\infty} L\widetilde{\mathrm{Var}}\left\{\hat{\sigma}^2\right\} = 4\left(a-c\right)\frac{\left(w+2\right)^3}{w} = 4\left(w+2\right)^2\left(\frac{w+2}{\sqrt{\left(w+1\right)\left(w+3\right)}}-1\right), \tag{5.47}$$

whose limits are

$$2 \leq \lim_{L\to\infty} L\widetilde{\mathrm{Var}}\left\{\hat{\sigma}^2\right\} \leq 2.4752. \tag{5.48}$$

The minimum kernel variance is then determined by

$$w > \left(\frac{\alpha}{L}\right)^2 \frac{3}{4}\left(\frac{1}{\left(2-\sqrt{3}\right)^2}\right). \tag{5.49}$$

As it turns out, the w required for the asymptotic behaviour is then inversely proportional to L^2. The minimum kernel variance is now roughly $w_{\text{min}} \approx 1044/L^2$ for $\alpha = 10$. Although the resulting w_{min} becomes small faster than the complex-valued case with respect to L, the constant factor is one order of magnitude above the complex-valued case in (5.46). As a result, for very small values of L the kernel variance required in the complex-valued case may be lower than the real-valued case, but this tendency is quickly pivoted and w_{min} decreases much faster in the real-valued case (for $L > 70$, approximately). This result implies that the performance of the estimator is more robust in front of the choice of w, since the possible values of w that achieve the asymptotic threshold are less limiting.

Figure 5.5 also shows the normalized variance of the estimator, following Figure 5.4, but for the real-valued case. The floor value from (5.48) is also shown. Clearly, the margin for choosing w is much broader than in the complex-valued case, given that for the same value of L the variance requires a much smaller relative kernel variance before it is amplified. This property can be observed in the distances between the asymptote of different values of L, which are now greater than in Figure 5.4. The upward arrows now point to the corresponding relative variance by following the minimum relative kernel variance from (5.49) for $\alpha = 10$. It is also worth noting that the penalty term is divided by 2 for the sake of comparison, given that the CRB in the real-valued case is two times greater than in the complex-valued case. Consequently, the resulting estimator is also efficient for $w \to \infty$.

5.2.2.3 Kernel variance determination

Following the previous subsections, the strategy for determining the kernel variance is provided hereunder. Until now, the analysis of the estimator's behaviour has featured the relative kernel variance w. The reason is that the kernel variance needs to operate in a range around the true variance σ^2. Precisely, (5.41) illustrates this requirement by showing that very *relative* small values have a penalty to the variance and very *relative* high values tend to the sample variance case. However, given that σ^2 is precisely the parameter to be estimated, we propose to determine the kernel variance v^2 from the available data in a iterative manner. Moreover, the relative kernel variance w has to meet (5.46) for any value of L and data variance in order to attain the desired asymptotic variance results. Consequently, $v^2 = 15\hat{\sigma}_S^2/L$ will be used as a tentative value for determining the kernel variance, and it shall be used to initiate the iterative method.

Algorithm 1 Iterative procedure for determining the kernel variance.

```
$\hat{\sigma}^2[0] = \hat{\sigma}_{\mathrm{S}}^2;\ \Delta = 1;\ q = 1;\ 0 < \delta \ll 1;\ k = 1$
while $\Delta > \delta$
    $v^2[k] = 15\hat{\sigma}^2[k]/L$
    $\hat{W}[k] = \frac{2}{L(L-1)} \sum_{0 \le i < j \le L-1} h_{v[k]}(x(i) - x(j))$
    $\hat{\sigma}^2[k] = \frac{v^2[k]}{2}\left(\frac{1}{\hat{W}[k]} - 1\right)$
    $\Delta = \left|\hat{\sigma}^2[k] - \hat{\sigma}^2[k-1]\right| / \hat{\sigma}^2[k]$
    $k \leftarrow k + 1$
end
```

Algorithm 1 specifies the proposed iterative method in terms of the sample variance estimator $\hat{\sigma}_{\mathrm{S}}^2$. The method measures the MIP through (5.19) in order to retrieve a tentative value of σ^2 from the expected value of the MIP in (5.28). The operation is repeated until no significant change is observed in the estimate of the variance. At every step, the algorithm uses a conservative value of the kernel variance as a function of the estimated variance and the available number of samples, following (5.46). As a result, the final iteration of the algorithm directly provides an estimation of the variance, as intended.

5.2.3 Robustness to outliers

Finally, the sensitivity of the estimator in front of outliers is quantified. To do so, we will use a contaminated additive model given by

$$x_\varepsilon(i) = x(i) + z(i)\, y(i)\,, \tag{5.50}$$

where X_ε is the new random variable whose PDF is defined as $f_{X_\varepsilon}(x)$. Here, $z(i)$ can be either one or zero, whose probabilities are $\Pr\{z(i)=1\}=\varepsilon$ and $\Pr\{z(i)=0\}=1-\varepsilon$, respectively, and assuming that ε is an arbitrarily small value. This is a common model utilized for studying distributions that are close to Gaussian models, but with some error probability to departure from normality [Mar+19, Sec. 2.2]. However, while the new random variable Y is usually considered to be a Gaussian process independent of X, thus constructing a GMM, here we will assume that it is a discrete random variable such that

$$\Pr\{Y=y_m\}=p_Y(y_m)=p_m, \tag{5.51}$$

with $m=1,...,M$ and PMF $p_Y(y)$. This model allows for a better analysis in terms of the proposed estimator, which will result in deterministic and well-defined bounds of the contaminated IP value. With (5.50), we assume that there is a small probability ε that an outlier $y(i)$ is disturbing the original observation. This is particularly harmful for the sample variance estimator, whose expected value under this model becomes the following (see Appendix 7.4.5 for details on the computation):

$$\mathbb{E}_{f_X}\{\hat{\sigma}_{\mathrm{S}}^2\}=\frac{1}{2}\mathbb{E}_{f_{X_\varepsilon}}\left\{|x_\varepsilon(i)-x_\varepsilon(j)|^2\right\}=\sigma^2+\varepsilon\left(\sigma_y^2+\mu_y^2(1-\varepsilon)\right), \tag{5.52}$$

where μ_y and σ_y^2 are the mean and variance of the random variable Y, respectively. As a result, the variance estimation is biased with a term that is proportional to the probability of an outlier, as well as to the mean and variance of the contamination process. While ε can be expected to be a fairly low value, the fact that it also depends μ_y and σ_y^2 is not a desirable property, since both mean and variance can be very large in comparison to σ^2.

Once the sample variance case is outlined, we are now interested in analyzing the impact of the contamination model on the entropy-based estimator. Since the proposed estimator is based on the estimated IP [3], we will examine how the outliers and their probabilities deteriorate the IP of the random variable X_ε. For this, recall that the IP is expressed as

$$V(X_\varepsilon)=\int_{\mathcal{X}} f_{X_\varepsilon}^2(x)\,\mathrm{d}x. \tag{5.53}$$

The PDF of the contaminated data is expressed as follows:

$$f_{X_\varepsilon}(x)=(1-\varepsilon)f_X(x)+\varepsilon\sum_{m=1}^{M}p_m f_X(x-y_m). \tag{5.54}$$

This density distribution corresponds to a weighted sum of shifted replicas of the original PDF $f_X(x)$. Alternatively, we can write

$$f_{X_\varepsilon}(x)=\sum_{m=0}^{M}\tilde{p}_m f_X(x-y_m), \tag{5.55}$$

where

$$\tilde{p}_m=\begin{cases}1-\varepsilon & \text{for} \quad m=0\\ \varepsilon p_m & \text{for} \quad 1\le m\le M\end{cases} \tag{5.56}$$

and $y_0=0$. First, let us examine an upper bound of the IP through the Cauchy-Schwarz inequality:

$$f_{X_\varepsilon}(x)=\sum_{m=0}^{M}\sqrt{\tilde{p}_m}\sqrt{\tilde{p}_m}f_X(x-y_m)\le\sqrt{\sum_{m=0}^{M}\left(\sqrt{\tilde{p}_m}\right)^2\sum_{m'=0}^{M}\left(\sqrt{\tilde{p}_{m'}}f_X(x-y_{m'})\right)^2}. \tag{5.57}$$

[3] Or equivalently, the MIP. However, the analysis of robustness in front of outliers will be executed in terms of the IP, since it simplifies the computation and shows the exact same result.

We have that

$$\sum_{m=0}^{M}\left(\sqrt{\tilde{p}_m}\right)^2=\sum_{m=0}^{M}\tilde{p}_m=1-\varepsilon+\varepsilon\sum_{m=1}^{M}p_m=1-\varepsilon+\varepsilon=1, \tag{5.58}$$

which yields

$$f_{X_\varepsilon}(x)\leq\sqrt{\sum_{m=0}^{M}\tilde{p}_m f_X^2(x-y_m)}. \tag{5.59}$$

The IP of the contaminated random variable is then bounded by

$$V(X_\varepsilon)\leq\int_{\mathcal{X}}\sum_{m=0}^{M}\tilde{p}_m f_X^2(x-y_m)\,\mathrm{d}x=\sum_{m=0}^{M}\tilde{p}_m\int_{\mathcal{X}}f_X^2(x-y_m)\,\mathrm{d}x \tag{5.60a}$$

$$=\sum_{m=0}^{M}\tilde{p}_m V(X)=V(X), \tag{5.60b}$$

where in the last equality we consider that the support of X is all the complex numbers.

On the other hand, let us now examine a lower bound. By plugging-in (5.55) into (5.53) we can express

$$V(X_\varepsilon)=\sum_{m=0}^{M}\sum_{m'=0}^{M}\tilde{p}_m\tilde{p}_{m'}\zeta(y_m-y_{m'}), \tag{5.61}$$

where

$$\zeta(z)=\int_{\mathcal{X}}f_X(x+z)\,f_X(x)\,\mathrm{d}x\leq\zeta(0)=V(X). \tag{5.62}$$

Notice that $\zeta(z)$ is expressed as an autocorrelation, and that is why it is bounded by its value at the origin, which coincides with $\int_{\mathcal{X}}f_X^2(x)\,\mathrm{d}x$ and thus the IP. Consequently, we can write

$$V(X_\varepsilon)\geq\sum_{m=0}^{M}\tilde{p}_m^2\zeta(0)=V(X)\sum_{m=0}^{M}\tilde{p}_m^2. \tag{5.63}$$

The resulting sum of squared $\tilde{p}_m$ can be seen as the argument of the logarithm in the collision entropy expression (2.47), and therefore will be denoted as the collision probability $\mathrm{Col}(Z;Y)$, where

$$\mathrm{Col}(Z;Y)=(1-\varepsilon)^2+\varepsilon^2\sum_{m=1}^{M}p_m^2\leq 1. \tag{5.64}$$

In particular, this expression captures the probability that two different realizations of Y will take the same value, wrapped by the contamination rate. By gathering the previous bounds, we can state the following inequalities:

$$V(X)\,\mathrm{Col}(Z;Y)\leq V(X_\varepsilon)\leq V(X). \tag{5.65}$$

The outcome of contaminating the random variable X is then assimilated by a shrinking of the IP. From (5.65) it can be seen that the deterioration of the IP is governed by the collision probability, whose relevance is that it *only* depends on ε and p_m, rather than μ_y and σ_y^2. Therefore, the values that Y can take do not have any impact whatsoever for the entropy-based estimator, unlike with the sample variance estimator. This is the key observation that justifies to use an entropy-based approach to the problem of estimating the variance with the purpose of achieving robustness.

Once we have analyzed the bounds of the contaminated IP, we proceed to give further insights on the implications of these bounds in terms of the estimation of σ^2. The upper-bound in (5.65) corresponds to the noncontaminated case, and therefore it is implicitly addressed in previous subsections. The relevant case now is the lower-bound in (5.65), which yields the worst-case scenario, i.e. it is the farthest from the true IP value $V(X)$. We then proceed by considering the MIP, since it is the entropic measure used for estimating σ^2. Given the expected value of the MIP in (5.29), its worst-case expected estimate is the following:

$$\mathbb{E}_{f_{X_\varepsilon}}\left\{\hat{W}(X_\varepsilon)\right\} = \text{Col}(Z;Y)\,\mathbb{E}_{f_X}\left\{\hat{W}(X)\right\} = \frac{v^2\text{Col}(Z;Y)}{2\sigma^2+v^2}. \tag{5.66}$$

Therefore, by plugging the previous expression into (5.28), the bound of the expected variance estimate becomes

$$\mathbb{E}_{f_{X_\varepsilon}}\left\{\hat{\sigma}^2\right\} \geq \frac{v^2}{2}\left(\frac{1}{\mathbb{E}_{f_X}\left\{\hat{W}(X_\varepsilon)\right\}}-1\right) = \frac{v^2}{2}\left(\frac{2\sigma^2+v^2}{v^2\text{Col}(Z;Y)}-1\right) \tag{5.67a}$$

$$= \frac{\sigma^2}{\text{Col}(Z;Y)} + \frac{v^2}{2}\left(\frac{1}{\text{Col}(Z;Y)}-1\right). \tag{5.67b}$$

Since we want the variance estimate to be as close as possible to σ^2, it is clear from (5.67) that we are interested on small kernel variances. As L grows to infinity, the kernel variance goes to zero naturally[4], providing an asymptotic floor value of

$$\lim_{L\to\infty}\mathbb{E}_{f_{X_\varepsilon}}\left\{\hat{\sigma}^2\right\} = \frac{\sigma^2}{\text{Col}(Z;Y)}. \tag{5.68}$$

Again, it is worth noting that this is only attained by considering the worst case in (5.65) (the lower-bound), which is influenced by the variance of the contamination σ_y^2. Nevertheless, as σ_y^2 increases, (5.68) is quickly achieved and then it is saturated, preventing the bias to further increase, unlike with the sample variance. This result will become apparent when showing the computer simulations.

It is also worth analyzing the minimum penalty in front of the "best" contamination case. On the one hand, for a fixed value of ε in (5.68), the minimum penalty occurs when $\text{Col}(Z;Y) = (1-\varepsilon)^2+\varepsilon^2$ (which requires that p_m is 1 only once, and the remaining terms are 0). Then the asymptotic expectation of the variance estimator becomes

$$\lim_{L\to\infty}\mathbb{E}_{f_{X_\varepsilon}}\left\{\hat{\sigma}^2\right\} = \frac{\sigma^2}{(1-\varepsilon)^2+\varepsilon^2}. \tag{5.69}$$

On the other hand, from (5.52), the expectation of the sample variance becomes

$$\mathbb{E}_{f_X}\left\{\hat{\sigma}_\text{S}^2\right\} = \sigma^2 + \mu_y^2\varepsilon(1-\varepsilon). \tag{5.70}$$

Therefore, even in the most favorable case, the sample variance estimator is still dependent on the squared magnitude of the single outlier, unlike the entropy-based approach.

Figure 5.6 exhibits the robustness of the proposed entropy-based method in front of the sample variance. The normalized bias, namely $\text{NBias}\left\{\hat{\sigma}^2\right\} = \left(\mathbb{E}_{f_X}\left\{\hat{\sigma}^2\right\}-\sigma^2\right)/\sigma^2$, is shown in terms of the normalized contamination variance for two different contamination rates ε and two values of L. The dotted lines indicate the asymptotic floor values assuming (5.68), which results in

$$\lim_{L\to\infty}\text{NBias}\left\{\hat{\sigma}^2\right\} = \frac{\frac{\sigma^2}{\text{Col}(Z;Y)}-\sigma^2}{\sigma^2} = \frac{1}{\text{Col}(Z;Y)}-1. \tag{5.71}$$

It can be seen that, while the sample variance increases with σ_y^2, as in (5.52), the bias of the entropy-based approach exhibits a ceiling effect, showing that ε becomes relevant instead of σ_y^2.

[4] It is a shared desirable property of the chosen IP estimator, as addressed in Subsection 3.3.3. Generally speaking, it follows from the KDE estimator, hence it can be seen as analogous to Silverman's rule given in (3.55), which also decreases the kernel variance as the data size increases.

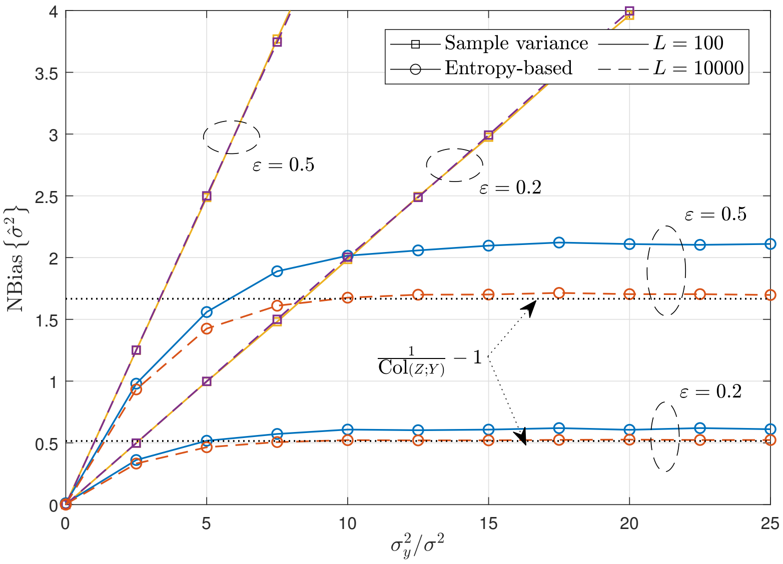


Figure 5.6: Normalized bias of the variance estimators as a function of the relation between the variance of the contaminating random variable and the true variances.

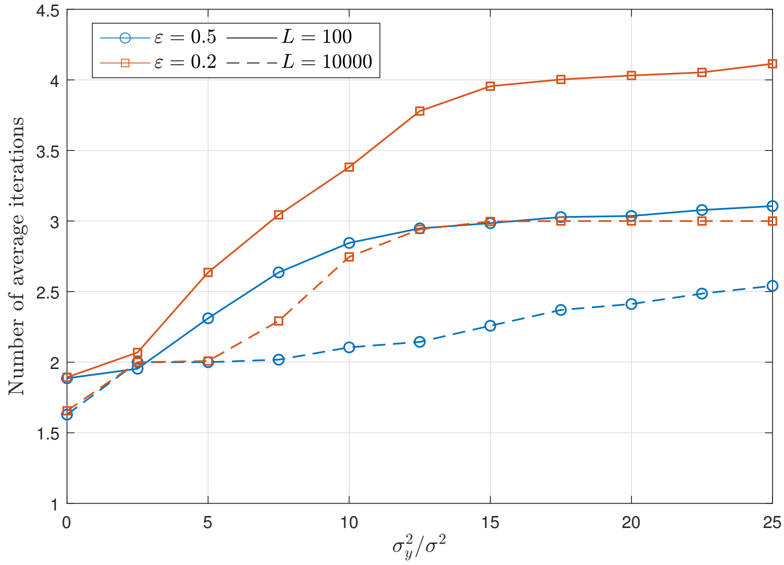


Figure 5.7: Average iterations for each measurement of the kernel variance in Algorithm 1 as a function of the relative contamination variance.

In order to demonstrate the effectiveness of Algorithm 1, Figure 5.7 shows the average number of iterations required for the robust measure of variance in relation to Figure 5.6. On the one hand, for a small contamination rate, i.e. close to nominal conditions, the algorithm proves to converge quickly and only two or three iterations are required. On the other hand, a higher ε requires, in average, more iterations. After some value of σ_y^2/σ^2 the number of average iterations still increases, albeit slowly. However, generally speaking, these are still very few iterations, proving Algorithm 1 to be an effective tool for determining v^2.

5.2.4 Conclusion

In this section, a method for estimating the determinant of the covariance matrix, without explicitly estimating that matrix, has been studied. It has been shown that performing the Gaussian assumption in a second stage, after an entropy-based estimate, is a valid strategy for gaining robustness in front of, precisely, the lack of nominal conditions. The key point is that entropy is sensitive to the probability of the outliers, rather than to their value. Consequently, robustness is acquired by introducing an information measure to a problem that is classically performed under the second-order statistics perspective. Furthermore, the proposed entropy estimate has an intrinsic relation with second-order statistics in form of an asymptotic trend of the kernel variance. Therefore, the entropy-based method for estimating the determinant of the covariance matrix can be seen as a generalization of the second-order statistics approach, where, in the *worst* case, are comparable. We will precisely see that this is a recurrent topic when using the IP estimator from Section 3.3. Next, we continue to study this entropy-based proposition to widen its applications with a more general case than the one addressed until now.

5.3 Coherence estimation

Once the estimation of the determinant of the covariance matrix has been addressed from the point of view of an entropy estimate, we proceed to broaden this concept by addressing an entropy-based *coherence* estimation. We refer to coherence as the statistic that provides a measure of similarity between two different signals. In its most general form, this statistic translates to the Generalized Coherence (GC) [GC88], which is commonly encountered as a nonparametric detector of a common signal on two noisy channels [CGS95; GPC06; KAS14].

To get more insights on this problem, we begin by defining the GC. Consider an i.i.d. complex sequence of the form

$$\mathbf{x}(i) = \begin{bmatrix} \mathbf{x}_1(i) \\ \mathbf{x}_2(i) \end{bmatrix} \tag{5.72}$$

for $\mathbf{x}_1(i) \in \mathbb{C}^{N_1}$ and $\mathbf{x}_2(i) \in \mathbb{C}^{N_2}$, whose covariance matrix is a block-composite matrix such that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{1,2} \\ \boldsymbol{\Sigma}_{2,1} & \boldsymbol{\Sigma}_2 \end{bmatrix}, \tag{5.73}$$

where

$$\boldsymbol{\Sigma}_1 = \mathbb{E}_{f_{X_1}}\left\{\left(\mathbf{x}_1 - \mathbb{E}_{f_{X_1}}\{\mathbf{x}_1\}\right)\left(\mathbf{x}_1 - \mathbb{E}_{f_{X_1}}\{\mathbf{x}_1\}\right)^H\right\} \tag{5.74a}$$

$$\boldsymbol{\Sigma}_2 = \mathbb{E}_{f_{X_2}}\left\{\left(\mathbf{x}_2 - \mathbb{E}_{f_{X_2}}\{\mathbf{x}_2\}\right)\left(\mathbf{x}_2 - \mathbb{E}_{f_{X_2}}\{\mathbf{x}_2\}\right)^H\right\} \tag{5.74b}$$

$$\boldsymbol{\Sigma}_{1,2} = \boldsymbol{\Sigma}_{2,1}^H = \mathbb{E}_{f_X}\left\{\left(\mathbf{x}_1 - \mathbb{E}_{f_{X_1}}\{\mathbf{x}_1\}\right)\left(\mathbf{x}_2 - \mathbb{E}_{f_{X_2}}\{\mathbf{x}_2\}\right)^H\right\} \tag{5.74c}$$

are the autocovariance and cross-covariance matrices, respectively. The GC is then defined as

$$\gamma = 1 - \frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_1|\,|\boldsymbol{\Sigma}_2|}, \tag{5.75}$$

where $0 \leq \gamma \leq 1$. The second term in the right-hand side of (5.75) corresponds to the Hadamard ratio [Ram+10], which is just the determinant of the covariance matrix over the product of the determinants of the marginal covariance matrices. In this setting, the Hadamard ratio becomes the generalized likelihood ratio test for distinguishing a block diagonal matrix (Gaussian data) from an arbitrary Hermitian matrix [Ram+13; Hua+14]. Since it is a measure that only depends on determinants of covariance matrices, the GC can be seen as an extension to the previous subsection, hence indicating a good predisposition to be estimated based on an entropy-based approach. In fact, this observation has also been exploited for measuring the time delay among spatially separated sensors based on an entropy estimate [BHC07], albeit with Shannon's entropy.

For completeness, the GC can also be seen under the perspective of the CCA. To see that, let us rearrange the GC as follows:

$$\gamma = 1 - \frac{|\mathbf{\Sigma}_2| \left|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_{1,2}\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_{2,1}\right|}{|\mathbf{\Sigma}_1| \, |\mathbf{\Sigma}_2|} = 1 - \left|\mathbf{\Sigma}_1^{-1}\left(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_{1,2}\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_{2,1}\right)\right| = 1 - |\mathbf{I} - \mathbf{C}| \, , \quad (5.76)$$

where $\mathbf{C} = \mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_{1,2}\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_{2,1}$ is the squared coherence matrix, as it is addressed in Subsection 2.2.2. The GC can be then expressed as a function of the eigenvalues of $\mathbf{C}$ such that

$$\gamma = 1 - \prod_{n=1}^{\min\{N_1,N_2\}} \left(1 - \lambda_n\left(\mathbf{C}\right)\right). \quad (5.77)$$

Consequently, to estimate the GC is equivalent to perform the CCA of the sample coherence matrix and to retrieve γ afterwards. Under normal distributions, this link between the GC and CCA is translated to the MI. In particular, the MI between X_1 and X_2 then becomes

$$I\left(X_1; X_2\right) = -\ln\left(\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}_1| \, |\boldsymbol{\Sigma}_2|}\right) = -\ln\left(1 - \gamma\right) \quad (5.78)$$

This set of relationships helps us to locate the GC within the framework of the thesis, and justifies our interest in estimating this statistic through an entropic measure.

5.3.1 Problem formulation

Similar to the case of the covariance determinant estimation, we focus here on the particular case of $N_1 = N_2 = 1$ for simplicity, which corresponds to the bivariate case in Section 5.1 with $N = 2$. The covariance matrix is then defined as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sqrt{\sigma_1^2\sigma_2^2} \\ \rho\sqrt{\sigma_1^2\sigma_2^2} & \sigma_2^2 \end{bmatrix} \quad (5.79)$$

for $\rho \in [0, 1)$. As a result, the parameter to be estimated is the following:

$$\gamma = 1 - \frac{\sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2} = \rho^2, \quad (5.80)$$

where ρ is just the coherence factor or Pearson coefficient. In this case, γ is commonly referred to as the Magnitude-Squared Coherence (MSC) [CGS95]. To justify this particular case, we would like to point out that the MSC is a well-known approach to test whether there is a common signal in different noisy channels or not, whose statistical behaviour under the null hypothesis is well-known for Gaussian noise [Nut81; GC87].

The intention is then to take advantage of the entropy-based estimator from sections 5.1 and 5.2 for the purpose of estimating the MSC. Given that we now have to deal with the bivariate case, the kernel variance is now defined as the matrix $\mathbf{W} \in \mathbb{R}^{2\times 2}$. For general purpose, we will assume that $\mathbf{W}$ is a diagonal matrix with elements $[\mathbf{W}]_{n,n} = v_n^2$ and $n = 1, 2$. Note that this is a common assumption for estimating a multivariate PDF through the Parzen-Rosenblatt window estimate, and therefore a common assumption for estimating the IP (see, for instance, (3.58) and (3.85)). For simplicity and cohesion with the previous section, the relative kernel variance will be defined as follows:

$$w = \frac{v_1^2}{\sigma_1^2} = \frac{v_2^2}{\sigma_2^2}. \quad (5.81)$$

If the marginal variances are known, which is uncommon, then only the marginal kernel variances v_1^2 and v_1^2 need to be estimated. For example, by running Algorithm 1 once for each kernel variance. Otherwise, the variance of each complex sequence also needs to be estimated, for example with

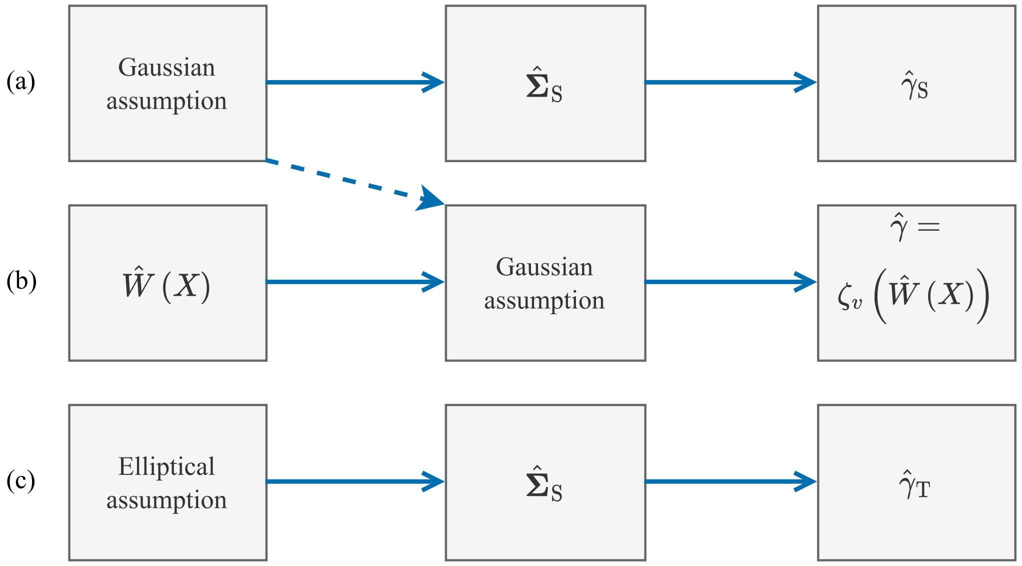

(a)
Gaussian assumption
$\hat{\mathbf{\Sigma}}_\mathrm{S}$
$\hat{\gamma}_\mathrm{S}$
(b)
$\hat{W}(X)$
Gaussian assumption
$\hat{\gamma} = \zeta_v\left(\hat{W}(X)\right)$
(c)
Elliptical assumption
$\hat{\mathbf{\Sigma}}_\mathrm{S}$
$\hat{\gamma}_\mathrm{T}$

Figure 5.8: Entropy-based approach for the problem of estimating the MSC (b) in front of the sample estimator (a), and Tyler's approach (c).

the univariate robust estimator addressed in Section 5.2 and then used to normalize the data. By performing such pre-processing, the estimate of the MSC, namely $\hat{\gamma}$, then only requires a single relative kernel variance w.

We begin by expressing the expected value of the estimation of the MIP. From (5.21) we have

$$\mathbb{E}_{f_X}\left\{\hat{W}(X)\right\} = \frac{|\mathbf{V}|}{|2\mathbf{\Sigma}+\mathbf{V}|} = \frac{v_1^2 v_2^2}{\left(2\sigma_1^2+v_1^2\right)\left(2\sigma_2^2+v_2^2\right) - 4\gamma\sigma_1^2\sigma_1^2}, \tag{5.82}$$

which, by expressing it in terms of the relative kernel variance, yields

$$\mathbb{E}_{f_X}\left\{\hat{W}(X)\right\} = \frac{w^2}{(w+2)^2 - 4\gamma}. \tag{5.83}$$

Consequently, the estimate of the MSC is expressed as follows:

$$\hat{\gamma} = \zeta_w\left(\hat{W}(X)\right) = \frac{w^2}{4}\left(1 - \frac{1}{\hat{W}(X)}\right) + w + 1. \tag{5.84}$$

The previous expression provides the baseline on which the robust estimator will be constructed, performing a similar function to $\hat{\sigma}^2 = \xi_v\left(\hat{W}(X)\right)$ from the previous section in (5.28), but now the kernel variance is normalized from the start.

A comparison of different approaches for estimating the MSC is illustrated in Figure 5.8, following the general rationale from Figure 5.1. The sample and entropy-based approaches are equivalent to the estimation of the determinant of the covariance matrix, but now the entropy-based approach, which corresponds to the middle row (b), is based on the specific function $\hat{\gamma} = \zeta_w\left(\hat{W}(X)\right)$. Tyler's approach [Tyl87] is also shown in the lower row (c), which assumes an elliptical distribution with long tails to gain robustness to potential outliers, given that it will be used to contrast the entropy-based approach with another robust method.

In the sequel, a bias and variance analysis will be performed following the steps of the variance estimator. Again, we are particularly interested on the interplay between the data size L and w under nominal conditions as a means of learning about the problem of estimating the MSC in a robust manner.

5.3.2 Performance analysis

In terms of the bias of the estimator, by the convexity of $-1/x$ for $x > 0$ and Jensen's inequality, we can now state

$$\mathbb{E}_{f_X}\left\{\hat{\gamma}\right\} \leq \frac{w^2}{4}\left(1 - \frac{1}{\mathbb{E}_{f_X}\left\{\hat{W}\left(X\right)\right\}}\right) + w + 1 = \frac{w^2}{4}\left(1 - \frac{\left(w+2\right)^2 - 4\gamma}{w^2}\right) + w + 1 = \gamma, \tag{5.85}$$

which resolves into the following bias:

$$\operatorname{Bias}\left\{\hat{\gamma}\right\} = \mathbb{E}_{f_X}\left\{\hat{\gamma}\right\} - \gamma \leq 0. \tag{5.86}$$

In contradistinction to the variance estimator, the MSC estimate has a strict negative bias. Nevertheless, given that $\mathbb{E}_{f_X}\left\{\hat{W}\left(X\right)\right\}$ converges to the real MIP for a fixed value of w, then $\hat{\gamma}$ turns out to be asymptotically unbiased.

For determining the variance of the estimator, the same small error approximation as in (5.32) will be used. From the function $\zeta\left(\hat{W}\left(X\right)\right)$ in (5.84) and the equivalent Taylor expansion from (5.32), we have now

$$\left.\frac{\partial \zeta_w\left(z\right)}{\partial z}\right|_{z=W(X)} = \left.\frac{w^2}{4}\frac{1}{z^2}\right|_{z=W(X)} = \frac{\left(\left(w+2\right)^2 - 4\gamma\right)^2}{4w^2}. \tag{5.87}$$

The *relative* variance is then approximated as

$$\widetilde{\operatorname{Var}}\left\{\hat{\gamma}\right\} \approx \operatorname{Var}\left\{\hat{W}\left(X\right)\right\}\left(\left.\frac{\partial \zeta_w\left(z\right)}{\partial z}\right|_{z=W(X)}\right)^2 = \operatorname{Var}\left\{\hat{W}\left(X\right)\right\}\left(\frac{\left(w+2\right)^2 - 4\gamma}{2w}\right)^4. \tag{5.88}$$

The variance of the MIP estimator is composed as in (5.22), namely

$$\operatorname{Var}\left\{\hat{W}\left(X\right)\right\} = \frac{4\left(L-2\right)\left(a-c\right) + 2\left(b-c\right)}{L\left(L-1\right)}, \tag{5.89}$$

where now we have

$$a = \frac{w^4}{\left(\left(w+1\right)^2 - \gamma\right)\left(\left(w+3\right)^2 - 9\gamma\right)}, \qquad b = \frac{w^4}{\left(\left(w+4\right)^2 - 16\gamma\right)w^2}, \tag{5.90a}$$

$$c = \frac{w^4}{\left(\left(w+2\right)^2 - 4\gamma\right)^2}. \tag{5.90b}$$

While a, b, and c may seem similar to the case of variance estimation, their dependence on w entails different asymptotic tendencies. In this case, the variance increases without limit as both $w \to 0$ and $w \to \infty$ for a fixed value of L, whereas the entropy-based variance estimator only tends to infinity for $w \to 0$. All the same, the relevant relative kernel variances for both variance and MSC estimators to gain robustness are those close to 0. As a result, we will only provide the analysis for small values of w in the subsequent asymptotic analysis.

In view of (5.88), let us now examine the asymptotic performance of the MSC estimator. First, from (5.40), the equivalent expression for the MSC is as follows:

$$\lim_{L\to\infty} L\widetilde{\operatorname{Var}}\left\{\hat{\gamma}\right\} = 4\left(a-c\right)\left(\frac{\left(w+2\right)^2 - 4\gamma}{2w}\right)^4 \tag{5.91a}$$

$$= \left(\frac{\left((w+2)^2 - 4\gamma\right)^2 - \left((w+1)^2 - \gamma\right)\left((w+3)^2 - 9\gamma\right)}{\left((w+1)^2 - \gamma\right)\left((w+3)^2 - 9\gamma\right)} \right) \left(\frac{(w+2)^2 - 4\gamma}{2} \right)^2 . \tag{5.91b}$$

In this case, it is direct to see that the variance reaches a minimum for $w = 0$, instead of an asymptotic trend as in (5.41). In particular, using (5.91) the following is obtained:

$$\lim_{L\to\infty} L\widetilde{\mathrm{Var}}\left\{\hat{\sigma}^2\right\}\Big|_{w=0} = \left(\frac{(4-4\gamma)^2 - (1-\gamma)(9-9\gamma)}{(1-\gamma)(9-9\gamma)} \right) \left(\frac{4-4\gamma}{2} \right)^2 = \frac{28}{9}(1-\gamma)^2 . \tag{5.92}$$

Clearly, this value becomes maximum for $\gamma = 0$ and tends to zero for $\gamma \to 1$.

We can also determine the new required w for attaining the asymptotic behaviour. From Subsection 5.2.2.2 and following the same rule as in (5.43), we now have

$$L > \frac{\alpha}{2}\left(\frac{b-c}{a-c}\right) + 2 \tag{5.93a}$$

$$= \frac{\alpha}{2} \frac{\left(\left((w+2)^2 - 4\gamma\right)^2 - \left((w+4)^2 - 16\gamma\right) w^2 \right) \left((w+1)^2 - \gamma\right)\left((w+3)^2 - 9\gamma\right)}{\left(\left((w+2)^2 - 4\gamma\right)^2 - \left((w+1)^2 - \gamma\right)\left((w+3)^2 - 9\gamma\right) \right) \left((w+4)^2 - 16\gamma\right) w^2} + 2, \tag{5.93b}$$

which, assuming that w is very small, translates into the minimum w required such that

$$w^2 > \frac{\alpha}{2L} \frac{(4-4\gamma)^2 (1-\gamma)(9-9\gamma)}{\left((4-4\gamma)^2 - (1-\gamma)(9-9\gamma)\right)(16-16\gamma)} \tag{5.94a}$$

$$= \frac{\alpha}{2L} \frac{16(1-\gamma)^2 (1-\gamma) 9 (1-\gamma)}{7(1-\gamma)^2 16 (1-\gamma)} \tag{5.94b}$$

$$= \frac{9\alpha}{14L}(1-\gamma) . \tag{5.94c}$$

For $\alpha = 10$, we finally obtain the approximate minimum relative kernel variance

$$w_{\text{min}} \approx \sqrt{\frac{45}{7L}(1-\gamma)} \leq \sqrt{\frac{45}{7L}} . \tag{5.95}$$

Similarly to (5.46), this approximate w_{min} is the one used as a tentative value for determining the kernel variance in Algorithm 1 for the case of estimating the MSC. Since γ is the parameter to be estimated, the rational choice is to be as conservative as possible and assume the “worst-case”, which is given by assuming $\gamma = 0$, resulting in the upper bound given in (5.95). In terms of the dependency with the data size L, w_{min} is this time inversely proportional to the square root of L, which translates into a slower dependence than in (5.46). As a consequence, the choice of kernel variance for the MSC estimate is more restrictive. Nevertheless, considering again the real case, the dependency is then inversely proportional to L, similar to the difference between (5.45) and (5.49).

All the effects described above are illustrated in Figure 5.9. As mentioned, the variance of the MSC estimator tends to infinity for $w \to \infty$, regardless of the data size L, and to the value described in (5.92) for $w \to 0$ at the same time as $L \to \infty$. The minimum relative kernel variance from (5.95) is also shown. As expected, the higher is γ, the lower is the minimum w. Moreover, it can be seen that the relative variance indicated by w_{min} also corresponds, approximately, to its minimum value for a given L, suggesting that (5.95) is a fitting rule for deciding the kernel variance.

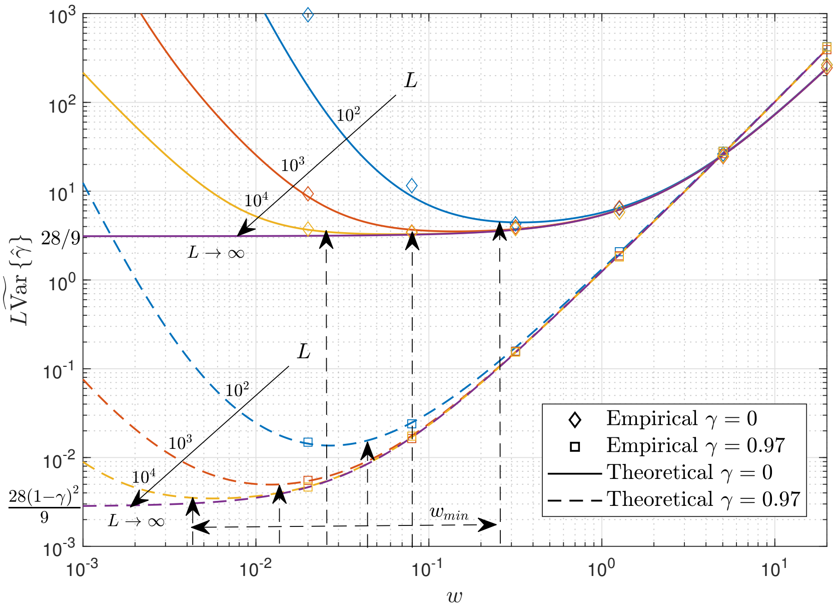


Figure 5.9: Variance of the MSC estimator as a function of the relative kernel variance.

5.3.3 Robustness to outliers

Next, we focus on the robustness of the entropy-based MSC estimate. This time, we consider a bivariate model of replacement outliers [Zou+12; Mar+19], where the outliers do not modify the marginal variances. The observed data under this model is defined as

$$\mathbf{x}_\varepsilon(i) = (1 - z(i))\,\mathbf{x}(i) + z(i)\,\mathbf{y}(i) \tag{5.96}$$

for $i = 0, ..., L-1$ and $z(i)$ defined as in (5.50). The random variables X and Y will be considered to be complex Gaussian variables with $X \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_x)$ and $Y \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y)$, independent to each other, where

$$\boldsymbol{\Sigma}_x = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_y = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}. \tag{5.97}$$

The model in (5.96) can actually be seen as a GMM of two components with weights $1-\varepsilon$ and ε, whose covariance matrix is

$$\begin{aligned}\boldsymbol{\Sigma} &= (1-\varepsilon)\left(\boldsymbol{\Sigma}_x + \boldsymbol{\mu}\boldsymbol{\mu}^T\right) + \varepsilon\left(\boldsymbol{\Sigma}_y + \boldsymbol{\mu}\boldsymbol{\mu}^T\right) - \left((1-\varepsilon)\,\boldsymbol{\mu} + \varepsilon\boldsymbol{\mu}\right)\left((1-\varepsilon)\,\boldsymbol{\mu} + \varepsilon\boldsymbol{\mu}\right)^T && (5.98a)\\ &= (1-\varepsilon)\,\boldsymbol{\Sigma}_x + \varepsilon\boldsymbol{\Sigma}_y + (1-\varepsilon)\left(\boldsymbol{\mu}\boldsymbol{\mu}^T\right) + \varepsilon\left(\boldsymbol{\mu}\boldsymbol{\mu}^T\right) - \boldsymbol{\mu}\boldsymbol{\mu}^T && (5.98b)\\ &= (1-\varepsilon)\,\boldsymbol{\Sigma}_x + \varepsilon\boldsymbol{\Sigma}_y. && (5.98c)\end{aligned}$$

Consequently, the MSC of X_ε is directly given by

$$\begin{aligned}\gamma_{\varepsilon,\mathrm{s}} &= 1 - \left|(1-\varepsilon)\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} + \varepsilon\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\right| = 1 - \left|\begin{bmatrix} 1 & \rho\,(1-2\varepsilon) \\ \rho\,(1-2\varepsilon) & 1 \end{bmatrix}\right| && (5.99a)\\ &= 1 - 1 + |\rho|^2\,(1-2\varepsilon)^2 = \gamma\,(1-2\varepsilon)^2, && (5.99b)\end{aligned}$$

where the subscript $\cdot_\varepsilon$ refers to the contaminated signal (5.96), and the subscript $\cdot\mathrm{s}$ denotes the sample approach to the MSC, similar to the one given in (5.52). Given that $(1-2\varepsilon)^2 \leq 1$ for $\varepsilon \in [0,1]$, then $\gamma_{\varepsilon,\mathrm{s}}$ shrinks for any value of ε. In addition, for $\varepsilon = 0.5$ the resulting MSC becomes null, resulting in a particular difficult case for the sample estimate [5]. In contrast, the entropy-based

[5] Nevertheless, it is not expected to reach such values, taking into account that the probability of an outlier ε is expected to be rather low.

MSC estimate is sensitive to the PDF of the contaminated data, which is a GMM given by

$$\begin{aligned} f_{X_\varepsilon}(\mathbf{x}) =& \frac{(1-\varepsilon)}{\pi^2(1-\gamma)} \exp\Big(-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_x^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big) \\ &+ \frac{\varepsilon}{\pi^2(1-\gamma)} \exp\Big(-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_y^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big). \end{aligned} \tag{5.100}$$

From here, we can determine the resulting IP of the contaminated random variable such that

$$V(X_\varepsilon) = \int_{\mathcal{X}} f_{X_\varepsilon}^2(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{5.101a}$$

$$\begin{aligned} =& \int_{\mathcal{X}} \frac{(1-\varepsilon)^2}{\pi^4(1-\gamma)^2} \exp\Big(-2(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_x^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\mathrm{d}\mathbf{x} \\ &+ \int_{\mathcal{X}} \frac{\varepsilon^2}{\pi^4(1-\gamma)^2} \exp\Big(-2(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_y^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\mathrm{d}\mathbf{x} \\ &+ 2\int_{\mathcal{X}} \frac{(1-\varepsilon)\varepsilon}{\pi^4(1-\gamma)^2} \exp\Big(-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_x^{-1}(\mathbf{x}-\boldsymbol{\mu}) - (\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_y^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\mathrm{d}\mathbf{x}. \end{aligned} \tag{5.101b}$$

The first two integrals correspond to scaled versions the IPs of the random variables X and Y, namely $(1-\varepsilon)^2 V(X)$ and $\varepsilon^2 V(Y)$, which can be solved through as in (**??**) and correspond to

$$(1-\varepsilon)^2 V(X) = \frac{(1-\varepsilon)^2}{\pi^2 4(1-\gamma)}, \qquad \varepsilon^2 V(Y) = \frac{\varepsilon^2}{\pi^2 4(1-\gamma)}. \tag{5.102}$$

The third integral is solved following similar steps to those in Appendix 7.4.1, which yields

$$A = \int_{\mathcal{X}} \frac{2(1-\varepsilon)\varepsilon}{\pi^4(1-\gamma)^2} \exp\Big(-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_x^{-1}(\mathbf{x}-\boldsymbol{\mu}) - (\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}_y^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)\mathrm{d}\mathbf{x} \tag{5.103a}$$

$$= \frac{2(1-\varepsilon)\varepsilon}{\pi^4(1-\gamma)^2} \int_{\mathcal{X}} \exp\Big(-(\mathbf{x}-2\boldsymbol{\mu})^H \left(\boldsymbol{\Sigma}_x^{-1}+\boldsymbol{\Sigma}_y^{-1}\right)(\mathbf{x}-2\boldsymbol{\mu})\Big)\mathrm{d}\mathbf{x} \tag{5.103b}$$

$$= \frac{2(1-\varepsilon)\varepsilon}{\pi^4(1-\gamma)^2} \int_{\mathcal{X}} \exp\left(-\frac{2\|\mathbf{x}-2\boldsymbol{\mu}\|^2}{1-\gamma}\right)\mathrm{d}\mathbf{x} = \frac{2(1-\varepsilon)\varepsilon}{\pi^4(1-\gamma)^2}\frac{\pi^2(1-\gamma)^2}{4} = \frac{(1-\varepsilon)\varepsilon}{2\pi^2}. \tag{5.103c}$$

By rearranging and substituting, we finally obtain

$$V(X_\varepsilon) = \frac{(1-\varepsilon)^2+\varepsilon^2}{4\pi^2(1-\gamma)} + \frac{(1-\varepsilon)\varepsilon}{2\pi^2} = \frac{(1-\varepsilon)^2+\varepsilon^2+2(1-\gamma)(1-\varepsilon)\varepsilon}{4\pi^2(1-\gamma)} \tag{5.104a}$$

$$= V(X)\left(1-2\gamma\varepsilon(1-\varepsilon)\right), \tag{5.104b}$$

where $V(X) = \left(4\pi^2(1-\gamma)\right)^{-1}$ is the IP of the noncontaminated random variable. The consequence of the contamination is that the IP is decreased as a function of γ and ε, and the inferred MSC (assuming that $w \to 0$) becomes

$$\gamma_\varepsilon = 1 - \frac{1-\gamma}{1-2\gamma\varepsilon(1-\varepsilon)}. \tag{5.105}$$

Once the entropy-based estimator has been addressed, we are now interested in comparing the resulting MSC estimate with the one given by the sample estimate. In particular, from (5.99), the bias of the sample MSC is

$$\mathrm{Bias}\{\hat{\gamma}_{\mathrm{s}}\} = \gamma(1-2\varepsilon)^2 - \gamma = 4\gamma\varepsilon(\varepsilon-1). \tag{5.106}$$

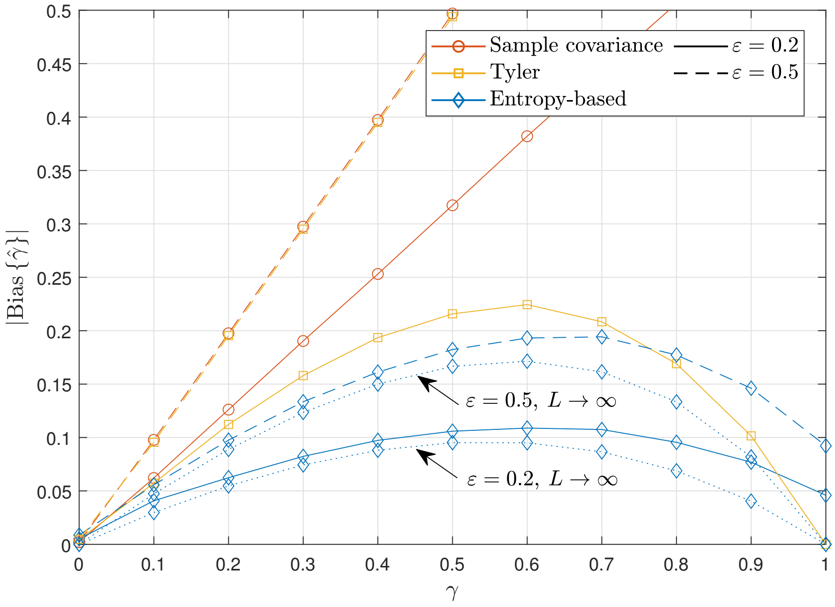


Figure 5.10: Absolute value of the bias of the MSC estimators as a function of the true value of MSC for $L = 500$.

On the other hand, the asymptotic bias of the entropy-based estimator is

$$\lim_{w\to\infty} \mathrm{Bias}\left\{\hat{\gamma}\right\} = 1 - \frac{1-\gamma}{1-2\gamma\varepsilon\left(1-\varepsilon\right)} - \gamma = \frac{2\gamma\varepsilon\left(1-\varepsilon\right)\left(\gamma-1\right)}{1-2\gamma\varepsilon\left(1-\varepsilon\right)}. \tag{5.107}$$

By testing which is bigger or smaller between (5.106) and (5.107), the resulting inequality [6] is

$$\left|\lim_{w\to\infty} \mathrm{Bias}\left\{\hat{\gamma}\right\}\right| \leq \left|\mathrm{Bias}\left\{\hat{\gamma}_{\mathrm{s}}\right\}\right|, \tag{5.108}$$

where the equality is met for $\varepsilon = 0$ and equal to γ, meaning that both approaches are unbiased without contamination. However, the inequality is strict for $\varepsilon > 0$. As a result, the theoretical entropy-based MSC estimate proves to be more robust than the sample MSC estimate, given that the shrinkage is always lower. Although the required condition for (5.105) is that $w \to 0$, which translates into $L \to \infty$, generally speaking the kernel variance will be a relatively low value, following the rule in (5.95), and maintaining the robustness of the entropy-based method.

Figure 5.10 shows the absolute value of the bias (given that is it negative) of the entropy-based MSC estimator in comparison with the sample covariance and the Tyler [Tyl87] approach (with 10 iterations). While the sample covariance method is nonrobust, Tyler's method is a robust approach that is shown to compare with the proposed, also robust, method. Two different contamination rates are shown, $\varepsilon = 0.2$ and the particularly difficult case of $\varepsilon = 0.5$. In both of these, the sample covariance grows linearly with γ with a slope proportional to ε, as expected. Tyler's approach shows its robustness for moderate value of ϵ, albeit it does not hold up for $\varepsilon = 0.5$. On the contrary, the entropy-based method achieves robustness in all scenarios. Moreover, it can be seen that with only $L = 500$ it is considerably close to the asymptotic value given in (5.105), and only for higher values of γ a higher value of L is required. It is only in these last cases where Tyler's method outperforms the proposed entropy-based approach. However, as pointed out in [SBP15], Tyler's requires a prior estimation of the mean, leading to severe problems in practice depending on the nature of the outlier process, while the entropy-based method is invariant to it.

[6] The proof is omitted since it just requires to simplify the absolute value of the previous expressions by isolating γ and testing whether it is bigger or lower than the resulting inequality, which is always true for $\gamma \in [0, 1)$. While it may be cumbersome, it is not complex, and the empirical results will shown the inequality to be true.

5.3.4 Conclusion

After addressing the case of just the determinant of the covariance matrix, the case of estimating the coherence, and in particular the MSC, has been studied. Similarly to the previous section, the entropy-based approach achieves a certain degree of robustness in front of outliers by relying on an information measure. The Gaussian assumption is done again after an estimate of the IP (or its surrogate), providing certain advantages over the usual preliminary Gaussian assumption. The result is that the entropy-based is less harmed by the probability of outliers than other methods, and consolidates the information-theoretic perspective in the problem of estimating determinants of covariance matrices.

Next, we move into a different problem, albeit similar in many aspects. Instead of estimating a single parameter of a random variable, the objective is to estimate the relation between two of them: the signal power and the noise level of a noisy channel. The signal model will now be characterized as a GMM, which allow us to seize the complete derivations from Section 5.1. The desired parameters are then retrieved from the expected IP and the second-order Rényi entropy of a complex GMM. The objective is to determine not only the intravariance (equivalent to the variance for a univariate random variable), but also the intervariance between different components of the GMM. In contrast with the second-order statistics approach, which is classically used for this problem and provides good results when the intravariance is large, entropy provides a more throughout characterization of the signal and becomes sensitive to both inter and intravariance.

5.4 Signal-to-noise ratio estimation

To end this chapter, the estimation of the SNR in a digital communications system will be addressed. An accurate assessment of the link quality of a channel is a fundamental task for adaptive coding and modulation, optimum signal detection, or turbo decoding, among others (see [SW98; PB00] and references therein). The constant growth of the complexity of communication systems requires both increasing accuracy and faster, in-service, algorithms. The main interest relies on determining this link quality, embedded in the SNR, with as little information as possible. Moreover, it is also desirable to design algorithms that are invariant/robust in front of as many parameters as possible, such as the transmitted symbols, the carrier phase, carrier frequency or modulation, among others.

To comply with these specifications, the most common approach is a Nondata-Aided (NDA) algorithm, which are designed to tackle the estimation of the SNR without the knowledge of the transmitted data. Conversely, Data-Aided (DA) methods require pilot information and have the advantage of a known closed-form maximum likelihood estimator. Nevertheless, although NDA methods provide a worse CRB at low SNR values than the DA methods, it proves to be too little of an advantage in front of the capability of NDA methods of measuring the SNR in-service and without prior information [PB00; Gap08; BMA14].

Within the NDA category, there are many different estimator classes. From an entropy perspective, and for reasons that will become clear in short, we are particularly interested in the classification between Envelope-Based (EVB) and coherent methods. Coherent methods compute the SNR by preserving the in-phase and quadrature components, while EVB methods directly compute the magnitude, i.e. the envelope, of the received signal. Another main category is encountered between baud-rate sampled versus oversampled methods [Sev+07; RVV10], and further distinctions are usually portrayed by the characteristics of the channel, either by considering multiple-input multiple-output systems [MDW16; Wau+07] or different kind of channel distributions [HKS15; JH13], among many others.

The advantage of EVB methods is that are invariant in front of phase and frequency errors. However, due to the suppression of the phase information of the signal, their CRB is two times bigger than the CRB of coherent methods at high SNR [GLM09]. Moreover, EVB methods also tend to perform poorly for multilevel constellations, particularly in the medium to high SNR range. The most notable and widely used EVB approach is based on measuring the even-order moments of the signal and to solve a linear system of equations from these estimates. In its most basic form, namely the M_2M_4 method [ME94], it measures the SNR through the second and fourth-order statistics. However, since the M_2M_4 only leverages low-order statistics, it is prone to being only sensitive to the noise level (intravariance), and looses perception of the complexity of the different GMM clusters (intervariance). As a result, the M_2M_4 provides good result for low SNR values and constant modulus constellations, but worsens its performance otherwise. To cope with this problem, the use of higher-order moment have been proposed. These are the moments-based estimators that encompasses up to the eight-order statistics [ÁLM10] (hence using the eight-, sixth-, fourth- and second-order moments) and up to the sixth-order statistics [LM07] (with the sixth-, fourth- and second-order moments). Nonetheless, constellations with three or more levels remain of particularly difficulty for the method of moments family. This approach has also been explored for different channel configurations, such as with Multiple-Input Multiple-Output (MIMO) systems [MDW16] or under a Nakagami fading channel [HKS15]. More recent EVB methods utilize the Kolmogorov-Smirnov test between the empirical cumulative distribution function and a set of stored cumulative distribution functions [Fu+15; WSW20]. Their estimation is improved for a wide range of SNR values with respect to the method of moments, but it is required to store a predetermined set of cumulative functions, and the performance is still lacking at medium and high SNR values.

On the other hand, NDA coherent methods are narrowed down to the Expectation-Maximization (EM) algorithm [Das08], a commonly used algorithm that provides an iterative solution to the maximum likelihood parameter estimation. The EM approach attains a variance of the estimator very close to the CRB, but at the cost of a high computational complexity. Moreover, it requires a perfect synchronization of the channel, or at least that the phase and frequency errors are estimated and corrected [GLM10]. However, by performing such joint estimation, its computational complexity is then further increased, opposing the desired characteristic of an in-service algorithm. These conditions preclude the EM for applications in which the standard conditions are not guaranteed, or where its computational intensity is detrimental.

The purpose of this section is to study an SNR estimator from an entropy measure, following the trail of the entropy-based MSC and variance estimators. The use of entropy provides more awareness to the intricacies of the GMM, akin to the use of higher-order moments in the case of EVB methods. However, entropy generalizes the idea and inherently includes all higher-order moments, both even and odd. Moreover, we will observe a similar situation of gaining robustness than in previous sections, albeit it comes from a different direction. From the SNR estimation point of view, the central feature is that entropy is invariant in front of rotations, meaning that an estimate of an entropy measure remains unaffected in front of carrier phase errors. Moreover, thanks to the estimator provided in Section 5.1, it is also possible to build an estimator that is robust in front of carrier frequency errors following the properties of incomplete U-statistics (Subsection 3.3.2). The resulting estimator is then considered a hybrid between EVB and coherent methods, but it has the capability of achieving the CRB of coherent methods. As a result, we call this entropy-based approach a *semi*-EVB method.

5.4.1 Signal model

The entropy-based approach to the estimation of the SNR addressed here considers a complex additive white Gaussian noise channel where the quality of the signal is estimated at the output of the matched filter at the optimal sampling instants. The method is generalized for any linearly modulated signal, although we will assume that the constellation and symbol probabilities are known. The signal follows the same model contemplated in [GLM10]. Considering that the signal power S' and noise power N' are constant over a block of L symbols, the symbol-rate samples at the receiver are the following:

$$x'\left(i\right) = \sqrt{S'}s\left(i\right)e^{j\left(2\pi\triangle_f i+\phi\right)} + \sqrt{N'}z\left(i\right) \tag{5.109}$$

for $i = 0, ..., L-1$. Here $z(i)$ are i.i.d. samples of a complex zero-mean Gaussian process with unit variance such that $Z \sim \mathcal{CN}(0,1)$. The transmitted symbols are represented by $s(i)$ and are independently drawn from a constellation $\mathcal{C}$, which is composed by M complex symbols $s_m = r_m e^{j\theta_m}$ whose probabilities are $\Pr\{s(k) = s_m\} = p_m$ for $m = 0, ..., M-1$. We will also assume that the constellation symbols have zero mean and unit variance such that

$$\sum_{m=0}^{M-1} p_m s_m = 0, \qquad \sum_{m=0}^{M-1} p_m |s_m|^2 = 1. \tag{5.110}$$

The parameter ϕ corresponds to the phase error, which is constant during all the observation window, and $\triangle_f$ denotes the frequency error, which is normalized to the symbol rate and is considered sufficiently small so that the intersymbol interference can be neglected. The problem of estimating the SNR is then cast as the estimation of $\rho = S'/N'$ given L samples from (5.109).

Along this section, we refer to standard conditions then both carrier phase and frequency offsets are null, i.e. $\triangle_f = \phi = 0$. This is a common assumption in coherent NDA SNR estimators. Otherwise, the carrier phase has to be synchronized beforehand. Conversely, EVB methods are invariant to phase and frequency errors, and the measure of SNR can be done before any synchronization stage. In the case of the entropy-based estimator, we will consider both scenarios for a proper evaluation of the semi-EVB property.

In addition, we will also consider that the data has been normalized with

$$x(i) = \frac{x'(i)}{\sqrt{\hat{M}_2}} = \frac{x'(i)}{\sqrt{\frac{1}{L}\sum_{j=0}^{L-1} |x'(j)|^2}}, \tag{5.111}$$

where $\hat{M}_2$ is the estimate of the second-order moment whose expectation is

$$\mathbb{E}_{f_X}\left\{\hat{M}_2\right\} = M_2 = S' + N'. \tag{5.112}$$

The SNR is then $\rho = S/N$, where $S = S'/\hat{M}_2$ and $N = N'/\hat{M}_2$. Since $\hat{M}_2$ is a consistent estimate of the envelope [ÁLM10], we ensure that the normalized signal and noise powers S and N are roughly between 0 and 1. Although this normalization is not strictly necessary for the entropy-based estimator, it guarantees some consistency for determining the kernel variance. In particular, we know from the previous sections that it is desirable to determine the kernel variance in terms of the variance of the data itself. Thanks to the normalization, the whole search is limited, which will ultimately help with the choice of parameters of the estimator.

With all this into consideration, we can now define the second-order Rényi entropy of the received data. Following the signal model in (5.109) with the standard condition assumption and the normalization from (5.111), X is distributed as follows:

$$X \sim \sum_{m=0}^{M-1} p_m \mathcal{CN}\left(\sqrt{S} s_m, N\right). \tag{5.113}$$

Hence, given Appendix 7.4.1 for univariate random variables, the second-order Rényi entropy of the noisy constellation $\mathcal{C}$ is directly given by

$$h_2(X) = -\ln(V(X)) = -\ln\left(\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1} \frac{p_m p_{m'}}{2\pi N} \exp\left(-\frac{S|s_m - s_{m'}|^2}{2N}\right)\right). \tag{5.114}$$

As it can be seen, $h_2(X)$ only depends on S and N, considering that s_m and p_m are known. On top of that, the invariance in front of phase errors is gained thanks to the computation of the squared

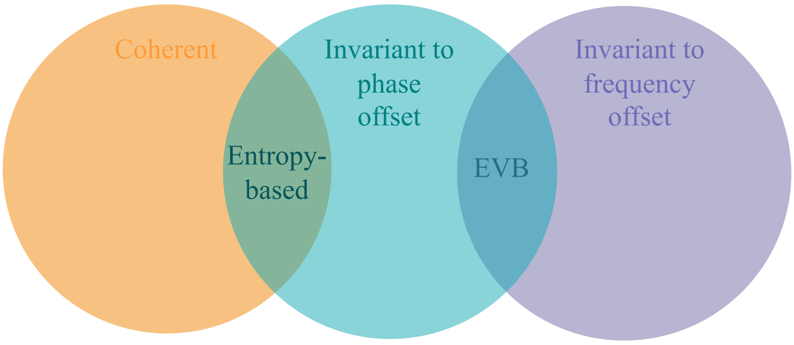


Figure 5.11: Diagram of the entropy-based approach in terms of EVB and coherent methods.

absolute value of symbols pairwise differences. Concretely, for $\triangle_f = 0$ and $\phi \neq 0$, the argument of the exponential becomes

$$\frac{-S\left|s_m e^{j\phi} - s_{m'} e^{j\phi}\right|^2}{2N} = \frac{-S\left|e^{j\phi}\left(s_m - s_{m'}\right)\right|^2}{2N} = \frac{-S\left|s_m - s_{m'}\right|^2}{2N}. \tag{5.115}$$

As a result, we can ensure that

$$\left.h_2\left(X\right)\right|_{\phi\neq 0} = \left.h_2\left(X\right)\right|_{\phi=0}. \tag{5.116}$$

Given this particular signal model, this property can be seen as the equivalent of the most well-known mean invariance property of any given entropic measure. In this case, thanks to the use of the second-order Rényi entropy, we gain the property of rotation invariance. However, carrier frequency errors do still influence the entropy measure. Since gaining robustness in this last case is more complex, we will address it in Subsection 5.4.4. For now, the data will be assumed that just have an arbitrary carrier phase error, which is the same as assuming the standard conditions for the entropy-based approach.

In view of the previous properties, Figure 5.11 illustrates the allocation of the entropy-based approach as a function of their invariance in front of carrier phase and/or frequency errors. While coherent methods are not, generally speaking, invariant to carrier offsets, the entropy-based approach *is* a coherent method, and it is invariant to constant rotations. That is why we refer to the proposed method here as a semi-EVB approach.

5.4.2 Problem formulation

Next, we proceed to evaluate the entropy estimator from (5.1) in the context of this section with the purpose of determining the SNR estimation. Given L i.i.d. observations from (5.109), the second-order Rényi entropy is given by

$$\hat{h}_{2,v}\left(X\right) = -\ln\left(\frac{2}{L\left(L-1\right)}\sum\sum_{0\leq i<j\leq L-1} k_v\left(x\left(i\right) - x\left(j\right)\right)\right) = -\ln\left(\hat{U}_v\left(X\right)\right) \tag{5.117}$$

It is worth noting that, this time, the kernel bandwidth v is included in the subscript of both second-order Rényi entropy and IP estimator. The subscript is included in preparation for an eventual requirement of various different kernel bandwidths, so every entropy estimate with different parameters is properly differentiable. To be concrete, we define the second-order Rényi entropy of X contaminated by a variance v as follows:

$$h_{2,v}\left(X\right) = -\ln\left(V_v\left(X\right)\right) = -\ln\left(\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1}\frac{p_m p_{m'}}{\pi\left(2N+v^2\right)}\exp\left(-\frac{S\left|s_m - s_{m'}\right|^2}{2N+v^2}\right)\right), \tag{5.118}$$

where $V_v(X)$ is also the contaminated IP. As a result, the implicit distribution of the contaminated X is now

$$X' \sim \sum_{m=0}^{M-1} p_m \mathcal{CN}\left(\sqrt{S}s_m, N + v/2\right). \tag{5.119}$$

Focusing now only on the IP, the expected value of the estimator is

$$\mathbb{E}_{f_X}\left\{\hat{U}_v(X)\right\} = \sum_{m=0}^{M-1}\sum_{m'=0}^{M-1} \frac{p_m p_{m'}}{\pi\left(2N + v^2\right)} \exp\left(-\frac{S\left|s_m - s_{m'}\right|^2}{2N + v^2}\right) = V_v(X), \tag{5.120}$$

which can be obtained directly from Corollary 5.1.1 for univariate GMM with $\mu_m = \sqrt{S}s_m$ and $\sigma_m^2 = N$. As can be seen, the expected value of the IP estimator is in agreement with the contaminated IP in (5.118). Another relevant observation is that the amount of contamination given by the kernel variance v^2 directly sways the noise power N in an additive manner. Although both parameters, S and N, are to be estimated by an entropic measure, v^2 will be chosen in terms of the noise power. This makes sense because, by contaminating all the GMM components equally, the affected variance is the intravariance, and not the intervariance, which translates to the additive behaviour of v^2 in front of N.

We are now in terms of determining the expression of the SNR estimator. Given (5.120), we will employ the estimate of the second-order Rényi entropy to retrieve S and N in a similar fashion to the method of moments. In particular, the moments-based estimators are defined by two (or more) equations based on the measurement of the even statistical moments (odd moments are not considered, since it would infringe the EVB approach with, apparently, no gain in performance). In the case of the entropy-based approach, the equivalent equations are given by different kernel variances. This way, we can construct a nonlinear system with any desired number of equations.

Given a set of Q kernel bandwidths $\{v_1, v_2, ..., v_Q\}$, the SNR estimator can be constructed by fitting the model of the true second-order Rényi entropy h_{2,v_q} to the estimated one by minimizing with respect to the desired parameters S and N. The estimation is then cast as the following Nonlinear Least-Squares (NLS) problem:

$$\left\{\hat{S}, \hat{N}\right\} = \arg\min_{S,N} \sum_{q=1}^{Q} \left(h_{2,v_q}(X) - \hat{h}_{2,v_q}(X)\right)^2, \tag{5.121}$$

and then by computing the SNR with $\hat{\rho} = \hat{S}/\hat{N}$. This optimization problem can be solved with the Levenberg–Marquardt algorithm [NW06], which behaves both as a gradient-descent method and as the Gauss-Newton method depending on the distance of the parameters to their optimal value. As a result, this algorithm provides both fast and good local convergence. Due to the normalization in (5.111), the signal and noise powers are assessed only between 0 and 1, hence facilitating the minimization problem.

Regarding the initialization of parameters, a moments-based method is generally used. For instance, an initial guess from the sixth-order statistics estimator [LM07] provides a good enough first approach with a very low computational cost. Although the error increases for multilevel constellations at high SNR values, it is usually enough for the entropy-based estimator to find the values that minimize (5.121). Otherwise, the bias is propagated into the entropy-based estimator, and the global minimum may not be achieved.

It is worth noting that the second-order Rényi entropy is used instead of the IP for the purpose of estimating the SNR, unlike the variance and MSC estimators. The main reason is given by the concavity of $h_{2,v_q}(X)$ with respect to S and N. To see that, consider (5.118) as follows:

$$h_{2,v_q}(X) = \ln\left(\pi\left(2N + v_q^2\right)\right) - \ln\left(\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1} p_m p_{m'} \exp\left(-\frac{S\left|s_m - s_{m'}\right|^2}{2N + v_q^2}\right)\right). \tag{5.122}$$

On the one hand, we have that $\ln\left(\pi\left(2N+v_q^2\right)\right)$ is strictly concave. On the other hand, it is known that the logarithm of the sum of exponential functions is convex [BV04], hence its reverse sign is concave. Given that the sum of concave functions is another concave function, hence $h_{2,v_q}(X)$ is concave. The quadratic adjustment of the logarithm is then used since it does not alter the performance of the fitting. Nonetheless, the concavity of $(h_{2,v_q}(X)-\hat{h}_{2,v_q}(X))^2$ in (5.121) cannot be guaranteed since the difference between the second-order Rényi entropy and its estimate (the argument of the previous expression) is not strictly nonpositive. This observation may pose a problem in terms of the convergence of the NLS. However, we will see in the computer simulations that the optimization process generally behaves as expected, with the exception for very high SNR values, which is the most difficult regime for the chosen initial estimate $\hat{M}_6$. Nonetheless, the convergence can be controlled by choosing an adequate kernel bandwidth, as we will see in forthcoming subsections.

5.4.3 Performance analysis

Once the base estimator is established, we are then interested on unraveling the performance of the SNR estimator. First, we analyze the behaviour of the second-order Rényi entropy at different values of true SNR in the case of a GMM. Then, the variance of the estimator is analyzed, which will be used to determine the strategy for choosing the adequate value of kernel variances.

5.4.3.1 Asymptotic behaviour

Consider the expression of the contaminated second-order Rényi entropy $h_{2,v_q}(X)$ in (5.122). Given that the signal model is now based on a GMM, rather than a single Gaussian, we are now interested on addressing the behaviour of this entropic measure when the true value of SNR increases or decreases indefinitely.

First, let us examine the low SNR case. For that, and for clarity, we will express the SNR as a function of a single parameter, which can be either S or N. This can be done thanks to the normalization in (5.111), with which we can express

$$S=\frac{S'}{N'+S'},\qquad N=\frac{N'}{N'+S'},\tag{5.123}$$

where $S+N=1$ and

$$\rho=\frac{1-N}{N}=\frac{S}{1-S}.\tag{5.124}$$

Then, it is clear that $\rho\to 0$ is equivalent to $N\to 1$ and $S\to 0$ at the same time. With this in mind, the asymptotic second-order Rényi entropy at low SNR yields

$$\lim_{\rho\to 0} h_{2,v_q}(X)=\ln\left(\pi\left(2N+v_q^2\right)\right)-\ln\left(\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1}p_m p_{m'}\right)=-\ln\left(\frac{1}{\pi\left(2N+v_q^2\right)}\right).\tag{5.125}$$

As a result, the second-order Rényi entropy of a GMM at low SNR becomes the one of a contaminated Gaussian variable (as be seen, for example, from the IP in (5.18) for univariate random variables). This is an expected behaviour since, as the signal power decreases and the signal noise increases, the components of the GMM have a strong overlap and the distribution becomes Gaussian. In this low SNR regime, it is possible to retrieve the SNR just by isolating N from an estimate of (5.125). Therefore, only a single kernel variance and entropy estimation ($Q=1$) is required.

Under the context of parameter estimation with inherent Gaussian distributions, it is well-known that second-order becomes a sufficient statistic for estimating the parameters of interest [VV05]. However, for the estimation of SNR, two parameters are assessed, S and N. The second-order moment is indeed a sufficient statistic for estimating N at low SNR. This can be seen by recalling that $M_2=S+N$, which becomes $M_2\approx N$ as $\rho\to 0$. This is a similar outcome than

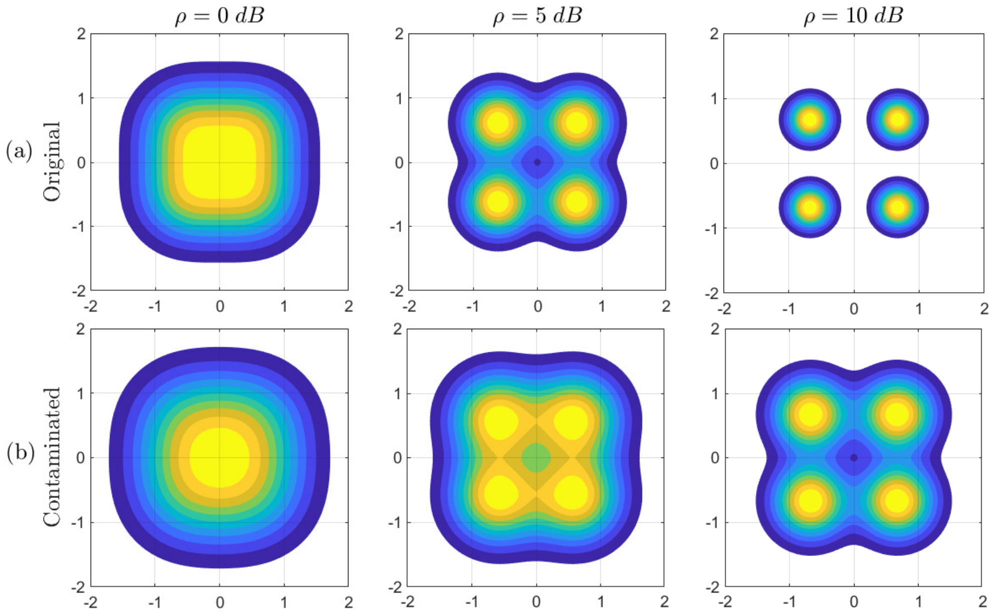


Figure 5.12: Contour plot of the complex GMM distribution in the complex plane for different values of SNR for $p_m = 1/M$. Both original (a) and contaminated (b) distributions are shown.

in (5.125). However, the SNR requires the estimation of both signal and noise *powers*, requiring two equations for two unknowns. Therefore, it is the combination of the two lower (even) moments that become a sufficient statistic. This is exemplified by the M_2M_4 estimator, which yields very good results for low values of SNR, particularly for single level constellations. Moreover, we know from (3.84) in Subsection 3.3.3 that the U-statistics entropy estimator becomes a scaled version of the sample variance. Consequently, very large values of v_q^2 become beneficial for the estimation of SNR, albeit only at the low SNR regime. In fact, the entropy-based estimator approximates the M_2M_4 estimator for very large values of v_q^2. This behaviour will be later confirmed through both theoretical estimator variance analysis and computer simulations.

To illustrate the second-order Rényi entropy for different values of SNR, Figure 5.12 shows the low SNR scenario with the quasi-Gaussian distribution for the Quadrature Phase-Shift Keying (QPSK) constellation, along with other scenarios that will be addressed shortly after. The upper row shows the true PDF as it is expressed in (5.113), while the lower row shows the contaminated PDF in (5.119), where $v_q^2 = 10^{-1}$ is chosen for illustration purposes. The low SNR regime can be appreciated for $\rho = 0$ dB, which is enough for a good approximation of (5.125). Furthermore, it can be seen that the contaminated PDF becomes even more Gaussian, provided that the contamination is strong enough, thus providing a better approximate of (5.125).

On the other hand, for $\rho \to \infty$ the components of the GMM become clearly distinguishable. Consequently, a very small kernel variance is preferred in order to preserve the separation among different clusters. Otherwise, high values of v_q^2 would conceal the rich structure provided by the corresponding GMM, and the second-order Rényi entropy would derive again to the Gaussian case. This outcome can be observed in Figure 5.12 for $\rho = 10$ dB, where the contaminated PDF portrays an overlapping not present in the original PDF.

By considering $\rho \to \infty$ and $v_q^2 \to 0$, the only nonzero terms within the sum of exponential functions are those with $s_m = s'_m$, rendering the following asymptotic entropy value:

$$\lim_{\rho\to\infty, v_q^2\to 0} h_{2,v_q}(X) = \ln\left(\pi\left(2N + v_q^2\right)\right) - \ln\left(\sum_{m=0}^{M-1}\sum_{m'=0}^{M-1} p_m p_{m'} \delta_{m,m'}\right) \tag{5.126a}$$

$$= -\ln\left(\sum_{m=0}^{M-1} p_m^2\right) + \ln\left(\pi\left(2N + v_q^2\right)\right), \tag{5.126b}$$

where $\delta_{m,m'}$ is the Kronecker delta, which is equal to 1 if and only if $m = m'$, and 0 otherwise. The result is similar to that of the low SNR regime in that the second-order Rényi entropy does only depend on N. Nevertheless, in contrast with the low SNR case, now the argument of the logarithm (the IP) also depends on the symbol probabilities. In fact, $\sum_{m=0}^{M-1} p_m^2$ corresponds to the collision probability of the constellation symbols. Similarly to the case of estimating the variance in Subsection 5.2.3, the resulting IP here is also swayed by the collision probability. The minimum IP (maximum $h_{2,v_q}(X)$) is then achieved by equiprobable symbols, and it just becomes an scaled version (by $1/M$) of the IP in the case of low SNR regime, as can be see in (5.125).

Again, in this regime one could retrieve the SNR by just isolating the noise variance in (5.126), provided a very small value of kernel variance is used. This strategy is actually used in our published paper [dR19]. However, it requires a determination of the threshold in which this regime is achieved. In order to unify the methodology, this approach will not be considered (and neither the low SNR one for the same reasons), and the SNR estimation will be addressed equally by solving the NLS problem.

Lastly, intermediate values of SNR result in an underlying PDF that is neither Gaussian (or quasi-Gaussian) nor distinct components of a GMM. In this case, the second-order Rényi entropy cannot be simplified by an asymptotic analysis. This is the regime that is shown in Figure 5.12 for $\rho = 5$ dB. As it can be seen, both original and contaminated PDFs are just overlapping components of a GMM. In this case, the only option is to estimate the SNR following (5.121), which we will see that it is precisely a particularly difficult regime. Nonetheless, in terms of the choice of kernel variance, it can be seen that different regimes require different values: low SNR appreciates large values of v_q^2, high SNR calls for low v_q^2 to avoid overlapping of the GMM components, and intermediate SNR values need an intermediate choice, since higher values would also just end up with a quasi-Gaussian distribution. In short, the required kernel variance is in tone with the true SNR value.

As a final remark, this result is in agreement with the observations made in Subsection 3.3.3.1. The more intricate the underlying distribution, the more sensitive it becomes in front of the additive Gaussian noise, which is inherent in the entropy estimate utilized in this chapter. While the SNR estimator is not directly concerned about the added bias on the entropy estimate, given that it is known and taken into account in the computations, the choice of kernel variance is still relevant. If chosen carelessly, the sole process of estimating entropy may hinder the final SNR estimate.

5.4.3.2 Variance analysis

Next, we evaluate the variance of the SNR estimator, namely Var $\{\hat{\rho}\}$. The intention is not only to corroborate the performance of the estimator under different settings, but also to help decide the best kernel bandwidths for each one of them. Similar to previous sections, the knowledge of the variance is leveraged to decide the appropriate kernel bandwidth. While the computation can be cumbersome, the knowledge of a closed-form expression of the variance of the estimator proves to be useful for improving the performance of the final estimator thanks to an appropriate choice of kernel variance.

For that, we first need to analyze the cross-covaraince of the second-order Rényi estimator

$$\text{Cov}\left\{\hat{h}_{2,v_q}(X), \hat{h}_{2,v_{q'}}(X)\right\} = \\ \mathbb{E}_{f_X}\left\{\hat{h}_{2,v_q}(X)\,\hat{h}_{2,v_{q'}}(X)\right\} - \mathbb{E}_{f_X}\left\{\hat{h}_{2,v_q}(X)\right\}\mathbb{E}_{f_X}\left\{\hat{h}_{2,v_{q'}}(X)\right\} \quad (5.127)$$

for $q, q' = 1, ..., Q$. To simplify the notation, the cross-covaraince will be denoted as $\sigma_h^2\{q, q'\}$. Similarly, we define the cross-covariance of the IP as

$$\sigma_U^2\{q, q'\} = \mathbb{E}_{f_X}\left\{\hat{U}_{v_q}(X)\,\hat{U}_{v_{q'}}(X)\right\} - \mu_U\{q\}\,\mu_U\{q'\}, \quad (5.128)$$

where $\mu_U\left\{q\right\} = \mathbb{E}_{f_X}\left\{\hat{U}_{v_q}\left(X\right)\right\}$ denotes the expected value of the IP from (5.120).

In Section 5.1, the covariance of the IP is computed, but we now require the covariance of the second-order Rényi entropy estimate. For this, consider two random variables Y and Z with joint PDF $f_{Y,Z}$ and marginal PDFs f_Y and f_Z. Given a differentiable and real function $f\left(x\right)$, the covariance to be determined is as follows

$$\text{Cov}\left\{f\left(Y\right), f\left(Z\right)\right\} = \mathbb{E}_{f_{Y,Z}}\left\{f\left(Y\right)f\left(Z\right)\right\} - \mathbb{E}_{f_Y}\left\{f\left(Y\right)\right\}\mathbb{E}_{f_Z}\left\{f\left(Z\right)\right\}. \tag{5.129}$$

The covariance will be approximated with the Taylor series up to the second order around the expected value of Y and Z [BHN05, Sec. 4.3.3]. The approximation yields

$$\mathbb{E}_{f_Y}\left\{f\left(Y\right)\right\} \approx f\left(\mathbb{E}_{f_Y}\left\{Y\right\}\right) + \frac{\text{Var}\left\{Y\right\}}{2}\left.\frac{\partial f\left(x\right)}{\partial x}\right|_{x=\mathbb{E}_{f_Y}\left\{Y\right\}}, \tag{5.130}$$

$$\mathbb{E}_{f_Z}\left\{f\left(Z\right)\right\} \approx f\left(\mathbb{E}_{f_Z}\left\{Z\right\}\right) + \frac{\text{Var}\left\{Z\right\}}{2}\left.\frac{\partial f\left(x\right)}{\partial x}\right|_{x=\mathbb{E}_{f_Z}\left\{Z\right\}}, \tag{5.131}$$

and

$$\begin{aligned}\mathbb{E}_{f_{Y,Z}}\left\{f\left(Y\right)f\left(Z\right)\right\} \approx & f\left(\mathbb{E}_{f_Y}\left\{Y\right\}\right)f\left(\mathbb{E}_{f_Z}\left\{Z\right\}\right)\\ &+ f\left(\mathbb{E}_{f_Y}\left\{Y\right\}\right)\frac{\text{Var}\left\{Z\right\}}{2}\left.\frac{\partial^2 f\left(x\right)}{\partial x^2}\right|_{x=\mathbb{E}_{f_Z}\left\{Z\right\}}\\ &+ f\left(\mathbb{E}_{f_Z}\left\{Z\right\}\right)\frac{\text{Var}\left\{Y\right\}}{2}\left.\frac{\partial^2 f\left(x\right)}{\partial x^2}\right|_{x=\mathbb{E}_{f_Y}\left\{Y\right\}}\\ &+ \text{Cov}\left\{Y, Z\right\}\left.\frac{\partial f\left(x\right)}{\partial x}\right|_{x=\mathbb{E}_{f_Z}\left\{Z\right\}}\left.\frac{\partial f\left(x\right)}{\partial x}\right|_{x=\mathbb{E}_{f_Y}\left\{Y\right\}}.\end{aligned} \tag{5.132}$$

By combining the previous equations with (5.129), the following approximation is obtained:

$$\begin{aligned}\text{Cov}\left\{f\left(Y\right), f\left(Z\right)\right\} \approx & \text{Cov}\left\{Y, Z\right\}\left.\frac{\partial f\left(x\right)}{\partial x}\right|_{x=\mathbb{E}_{f_Z}\left\{Z\right\}}\left.\frac{\partial f\left(x\right)}{\partial x}\right|_{x=\mathbb{E}_{f_Y}\left\{Y\right\}}\\ &- \frac{\text{Var}\left\{Y\right\}\text{Var}\left\{Z\right\}}{4}\left.\frac{\partial^2 f\left(x\right)}{\partial x^2}\right|_{x=\mathbb{E}_{f_Y}\left\{Y\right\}}\left.\frac{\partial^2 f\left(x\right)}{\partial x^2}\right|_{x=\mathbb{E}_{f_Z}\left\{Z\right\}}\end{aligned} \tag{5.133}$$

For the case of the second-order Rényi entropy, we need to substitute the expected values with

$$\mathbb{E}_{f_Y}\left\{Y\right\} = \mu_U\left\{q\right\}, \qquad \mathbb{E}_{f_Z}\left\{Z\right\} = \mu_U\left\{q'\right\}, \tag{5.134}$$

and the variances and covariance with

$$\text{Var}\left\{Y\right\} = \sigma_U^2\left\{q, q\right\}, \qquad \mathbb{E}_{f_Z}\left\{Z\right\} = \sigma_U^2\left\{q', q'\right\}, \qquad \text{Cov}\left\{Y, Z\right\} = \sigma_U^2\left\{q, q'\right\}. \tag{5.135}$$

From (5.117), f and its derivatives are

$$f\left(x\right) = -\ln\left(x\right), \qquad \frac{\partial f\left(x\right)}{\partial x} = -\frac{1}{x}, \qquad \frac{\partial^2 f\left(x\right)}{\partial x^2} = \frac{1}{x^2}, \tag{5.136}$$

which, by gathering all previous expressions, it finally yields the approximate variance of the second-order Rényi entropy estimator:

$$\sigma_h^2\left\{q, q'\right\} \approx \frac{\sigma_U^2\left\{q, q'\right\}}{\mu_U\left\{q\right\}\mu_U\left\{q'\right\}} - \frac{\sigma_U^2\left\{q, q\right\}\sigma_U^2\left\{q', q'\right\}}{4\left(\mu_U\left\{q\right\}\mu_U\left\{q'\right\}\right)^2}. \tag{5.137}$$

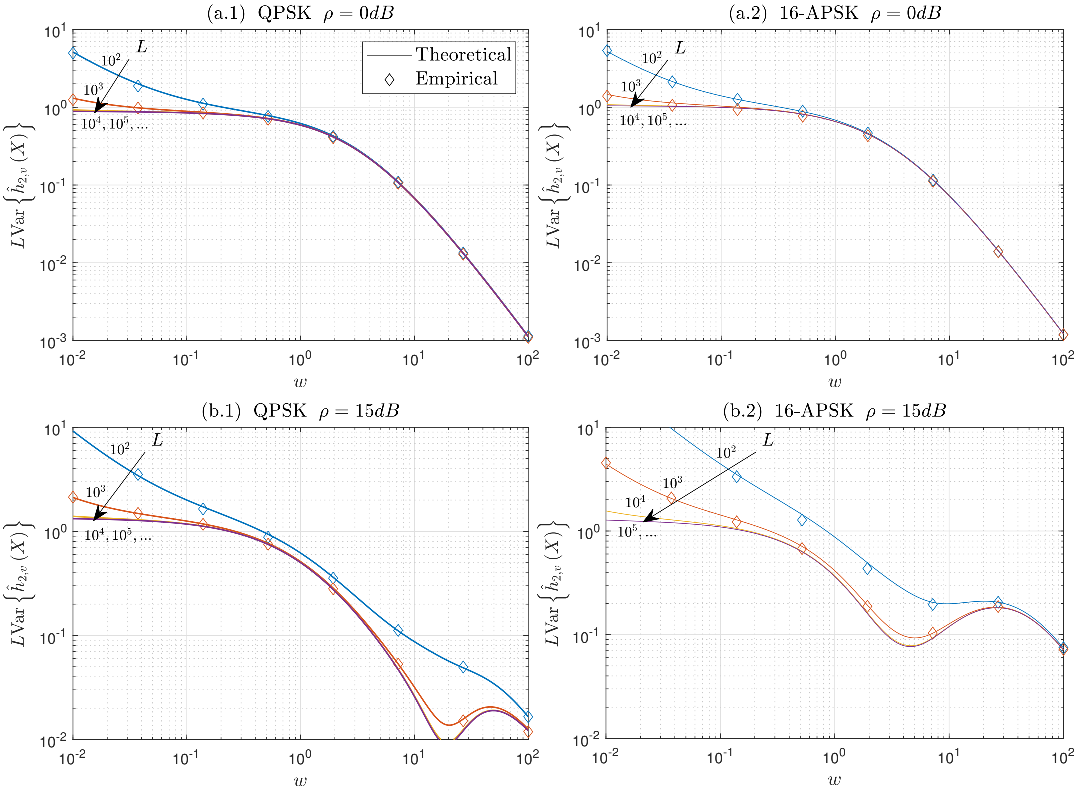


Figure 5.13: Variance amplification of the estimated second-order Rényi entropy as a function of the relative kernel variance w for two different constellations and two different values of true SNR (in dB).

The cross-covariance of the IP estimator with two different kernel bandwidths is given in Proposition 5.2 and yields

$$\sigma_U^2\left\{q, q'\right\} = \sigma_U^2\left\{q, q'\right\} = \frac{4(L-2)(a-c) + 2(b-c)}{L(L-1)}, \tag{5.138}$$

where now we have

$$a = \sum_{m,m',n=0}^{M-1} \frac{p_m p_{m'} p_n}{\pi^2\left(2N\left(v_q^2 + v_{q'}^2 + 1.5N\right) + v_q^2 v_{q'}^2\right)} \exp\left(-S\begin{bmatrix} s_m - s_{m'} \\ s_n - s_{m'} \end{bmatrix}^H \mathbf{U}_a^{-1}\begin{bmatrix} s_m - s_{m'} \\ s_n - s_{m'} \end{bmatrix}\right), \tag{5.139}$$

with

$$\mathbf{U}_a = \begin{bmatrix} 2N + v_q^2 & N \\ N & 2N + v_{q'}^2 \end{bmatrix}, \tag{5.140}$$

the term

$$b = \frac{p_m p_{m'}}{\pi^2\left(v_q v_{q'} + 2N\left(v_q + v_{q'}\right)\right)} \sum_{m=0}^{M-1}\sum_{m'=0}^{M-1} \exp\left(-\frac{S\left|s_m - s_{m'}\right|^2}{2N + \left(\frac{1}{v_q} + \frac{1}{v_{q'}}\right)^{-1}}\right), \tag{5.141}$$

and

$$c = \mu_U\{q\}\,\mu_U\left\{q'\right\}. \tag{5.142}$$

Figure 5.13 illustrates the variance of the second-order Rényi entropy estimator, i.e. the covariance with the same kernel bandwidth $\mathrm{Var}\left\{\hat{h}_{2,v_q}(X)\right\} = \sigma_h^2\{q,q\}$. Following (5.81), here it is shown as a function of the relative kernel variance, defined as

$$w = \frac{v_q^2}{N}. \tag{5.143}$$

For completeness, two different constellations with different degree of complexity are shown, the QPSK and the 16-Amplitude and Phase-Shift Keying (APSK) (specified in [DVB06]). For these, the symbols are assumed to be equiprobable with $p_m = 1/M$. Moreover, the variance is evaluated for two different values of ρ. It can be seen that the approximation performed in (5.137) corresponds to the one obtained empirically, thus confirming that is is a valid estimate. For low values of SNR, we know from Subsection 5.4.3.1 that the GMM becomes quasi-Gaussian. Consequently, the variance of the $\hat{h}_{2,v_q}(X)$ estimate resembles the one in Figure 5.3, albeit with a floor value for low values of w. When the SNR increases, the clusters do not overlap anymore and the particularities of each constellation become noticeable. Unfortunately, the sum of exponential functions added to the different symbol differences for each constellation makes the asymptotic analysis (as performed in Subsections 5.2.2.1 and 5.3.2) mathematically complex, if tractable at all. However, for any given constellation, the trend is that the variance of the estimator decreases as w increases, which is a common trait with all the entropy-based applications in this chapter given the base estimator (5.1).

Next, we measure the variance of the SNR estimator. A small error approximation will be used following (5.32). However, now the multivariate case is contemplated. Consider the vector of parameters $\boldsymbol{\alpha} = [S, N]^T$ with which the SNR is computed through the function $\rho = \xi(\boldsymbol{\alpha}) = S/N$ with $\xi : \mathbb{R}^2 \rightarrow \mathbb{R}$. The array $\boldsymbol{\alpha}$ is defined through function $\zeta : \mathbb{R}^Q \rightarrow \mathbb{R}^2$ with $\boldsymbol{\alpha} = \zeta(\mathbf{h}_2)$ and $\mathbf{h}_2 = \left[h_{2,v_1}, ..., h_{2,v_Q}\right]^T$. While ξ is known, ζ comes from the Q equations of the NLS problem in (5.121) and it is unknown. The variance of the SNR estimator is then approximated by

$$\operatorname{Var}\{\hat{\rho}\} \approx \left(\frac{\partial}{\partial \mathbf{z}} \xi(\zeta(\mathbf{z})) \bigg|_{\mathbf{z}=\mathbf{h}_2} \right)^H \boldsymbol{\Sigma}_{\mathbf{h}_2} \left(\frac{\partial}{\partial \mathbf{z}} \xi(\zeta(\mathbf{z})) \bigg|_{\mathbf{z}=\mathbf{h}_2} \right), \tag{5.144}$$

where the Jacobian for $\mathbf{z} = [z_1, ..., z_Q]^T$ is

$$\frac{\partial}{\partial \mathbf{z}} \xi(\zeta(\mathbf{z})) \bigg|_{\mathbf{z}=\mathbf{h}_2} = \left[\frac{\partial}{\partial z_1} \xi(\zeta(\mathbf{z})) \bigg|_{\mathbf{z}=\mathbf{h}_2}, ..., \frac{\partial}{\partial z_Q} \xi(\zeta(\mathbf{z})) \bigg|_{\mathbf{z}=\mathbf{h}_2} \right]^T, \tag{5.145}$$

and the covaraince matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{h}_2} = \begin{bmatrix} \sigma_h^2\{1,1\} & \sigma_h^2\{1,2\} & \cdots & \sigma_h^2\{1,Q\} \\ \sigma_h^2\{2,1\} & \sigma_h^2\{2,2\} & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_h^2\{Q,1\} & \cdots & \cdots & \sigma_h^2\{Q,Q\} \end{bmatrix}. \tag{5.146}$$

For the elements of the Jacobian, the chain rule states that

$$\frac{\partial}{\partial z_q} \xi(\zeta(\mathbf{z})) = \left(\frac{\partial \xi(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \right)^T \frac{\partial \boldsymbol{\alpha}}{\partial z_q}, \tag{5.147}$$

where

$$\frac{\partial \xi(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \left[\frac{\partial}{\partial S} \frac{S}{N}, \frac{\partial}{\partial N} \frac{S}{N} \right]^T = \left[\frac{1}{N}, \frac{-S}{N^2} \right]^T. \tag{5.148}$$

Conversely, the derivative of $\boldsymbol{\alpha} = \zeta(\mathbf{h}_2)$ cannot be directly computed given that ζ is unknown. For that, we will make use of the inverse function theorem ([Rud76], Theorem 9.42). Here, it is worth noting that the inverse function theorem is generally expressed for differentiable vector-valued functions $f : \mathbb{R}^Q \rightarrow \mathbb{R}^Q$, so the consequent Jacobian matrix is a square matrix. To avoid any associated problems with the inversion of matrices, we will fix now $Q = 2$, which translates into a nonlinear system of two unknowns and two equations with $\zeta : \mathbb{R}^2 \to \mathbb{R}^2$. An overdetermined system ($Q > 2$) may provide more robustness to the estimate through (5.121) (which can also be solved with the Levenberg–Marquardt algorithm), alleviating the choice of kernel bandwidth by just considering many more diverse v_q^2 values. However, the theoretical variance cannot then be accurately computed as with $Q = 2$. At the end, one could either optimize the two different kernel bandwidths as a function of the variance of the estimator, or to just set many different fixed kernel bandwidth values to gain in consistency with the estimated S and N. The first option is preferred since it is less computationally complex (for every kernel bandwidth a new estimate of an $L \times L$ matrix is required).

Then, from (5.147) and the inverse function theorem with $Q = 2$, the derivatives required in (5.147) are obtained by inverting the matrix that is composed by the derivatives of the second-order Rényi entropy expression for each parameter S and N such that

$$\begin{bmatrix} \left.\frac{\partial S}{\partial z_1}\right|_{z_1=h_{2,v_1}} & \left.\frac{\partial N}{\partial z_1}\right|_{z_1=h_{2,v_1}} \\ \left.\frac{\partial S}{\partial z_2}\right|_{z_2=h_{2,v_2}} & \left.\frac{\partial N}{\partial z_2}\right|_{z_2=h_{2,v_2}} \end{bmatrix} = \begin{bmatrix} \frac{\partial h_{2,v_1}}{\partial S} & \frac{\partial h_{2,v_1}}{\partial N} \\ \frac{\partial h_{2,v_2}}{\partial S} & \frac{\partial h_{2,v_2}}{\partial N} \end{bmatrix}^{-1}. \tag{5.149}$$

Following again the chain rule, we have

$$\frac{\partial h_{2,v_q}}{\partial S} = \frac{\partial h_{2,v_q}}{\partial V_{v_q}} \frac{\partial V_{v_q}}{\partial S} = -\frac{d_q}{\mu_U\left\{q\right\}} \tag{5.150}$$

and

$$\frac{\partial h_{2,v_q}}{\partial N} = \frac{\partial h_{2,v_q}}{\partial V_{v_q}} \frac{\partial V_{v_q}}{\partial N} = -\frac{e_q}{\mu_U\left\{q\right\}}, \tag{5.151}$$

where

$$d_q = \frac{\partial V_{v_q}}{\partial S} \tag{5.152a}$$

$$= \sum_{m=0}^{M-1} \sum_{m'=0}^{M-1} \frac{1}{M^2 \pi \left(2N + v_q\right)} \left(-\frac{\left|s_m - s_{m'}\right|^2}{2N + v_q}\right) \exp\left(-\frac{S\left|s_m - s_{m'}\right|^2}{2N + v_q}\right), \tag{5.152b}$$

and

$$e_q = \frac{\partial V_{v_q}}{\partial N} \tag{5.153a}$$

$$= \sum_{m=0}^{M-1} \sum_{m'=0}^{M-1} \frac{p_m p_{m'}}{\pi \left(2N + v_q\right)^3} \left(2S\left|s_m - s_{m'}\right|^2 - 2\left(2N + v_q\right)\right) \exp\left(-\frac{S\left|s_m - s_{m'}\right|^2}{2N + v_q}\right). \tag{5.153b}$$

Finally, by gathering the previous derivatives into (5.149) and computing (5.147), the approximate variance of the SNR estimator results in

$$\mathrm{Var}\left\{\hat{\rho}\right\} \approx \begin{bmatrix} \frac{\mu_V\{1\}(-Ne_2 - Sd_2)}{N^2(d_1e_2 - d_2e_1)} \\ \frac{\mu_V\{2\}(Ne_1 + Sd_1)}{N^2(d_1e_2 - d_2e_1)} \end{bmatrix}^H \begin{bmatrix} \sigma_h^2\left\{1,1\right\} & \sigma_h^2\left\{1,2\right\} \\ \sigma_h^2\left\{2,1\right\} & \sigma_h^2\left\{2,2\right\} \end{bmatrix} \begin{bmatrix} \frac{\mu_V\{1\}(-Ne_2 - Sd_2)}{N^2(d_1e_2 - d_2e_1)} \\ \frac{\mu_V\{2\}(Ne_1 + Sd_1)}{N^2(d_1e_2 - d_2e_1)} \end{bmatrix} \tag{5.154}$$

From the previous expression, it is relevant to mention that the variance goes to infinity for $v_1^2 = v_2^2$, given that then we have $d_1 e_2 = d_2 e_1$ in the denominators of the gradient. This case actually corresponds to the NLS problem for $Q = 1$, which can in fact be solved thanks to the normalization from (5.111) and following (5.123) and (5.124). Nonetheless, it is shown in [dR19] that this approach

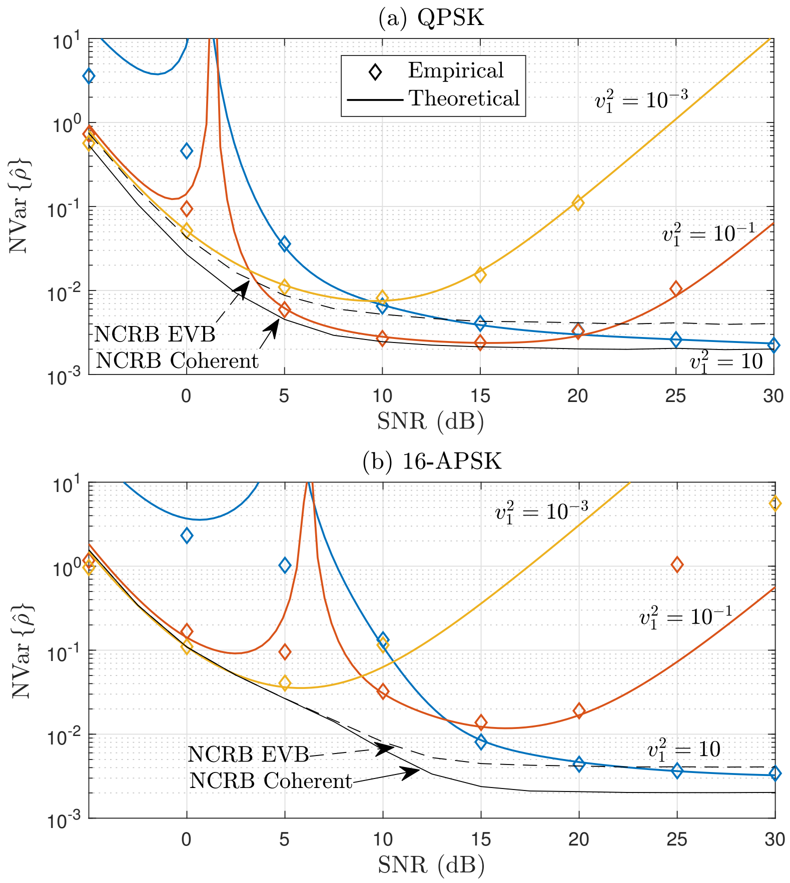


Figure 5.14: Normalized variance of the SNR estimator as a function of its true value for multiple v_1^2 and a fixed $v_2^2 = 10^3$ with $L = 500$.

induces a nonmonotonic dependence of the IP in terms of the SNR, resulting in a worse estimation for the low and medium SNR regimes. Consequently, $Q = 2$ is strongly suggested for a consistently good performance of the SNR estimator and an appropriate evaluation of the theoretical variance. From now on $Q = 2$ will be assumed unless otherwise stated.

To avoid the use of the same kernel bandwidth, we will generally impose v_2^2 to be a high enough value so that v_1^2 is always beneath. The reason to prefer a high value of v_2^2 is encountered in Figure 5.13, where larger values of relative kernel variance have lower entropy estimation variance. Later in this section we will see how the kernel variances will be chosen in base of this particular observation.

Apart from $v_1^2 = v_2^2$, another conflicting kernel variance values are the ones that yield $d_q = e_q$. That is, when the derivatives with respect to S and N coincide, the matrix from (5.149) is of rank 1, and therefore not invertible. The theoretical analysis of these v_q values is rather complex, thus this problem will be circumvented, at the time of determining the kernel variance values, by a two-step optimization with different initial guesses.

Figure 5.14 shows the comparison between the theoretical estimator variance in (5.154) and the empirically obtained by solving the NLS problem in (5.121) by means of Monte Carlo simulations with 500 iterations. This figure takes into consideration that one of the kernel variances (v_2^2) should be always greater than the other (v_1^2). Therefore, for illustration simplicity, a sufficiently high value of v_2^2 is fixed while three different values of v_1^2 are shown. The normalized variance is defined as $\mathrm{NVar}\{\hat{\rho}\} = \mathbb{E}_{f_X}\left\{(\hat{\rho} - \mathbb{E}_{f_X}\{\hat{\rho}\})^2\right\}/\rho^2$, which it is compared to the Normalized Cramér-Rao Bound (NCRB) of both EVB and coherent estimators as provided in [GLM09]. As can be seen, the theoretical variance is mostly accurate except for low SNR values with high kernel variance values, or for high SNR values with low kernel variance values. This mismatch is mainly caused by a poor resolution of the NLS problem, which does not find the global minimum for the given v_1^2 and v_2^2 values. However, it can be corrected by increasing or decreasing, respectively, the kernel variance at these SNR values. Some spikes can also be observed, exemplifying that the estimator variance greatly increases for $d_1e_2 - d_2e_1 \approx 0$, with an asymptote at $d_1e_2 = d_2e_1$. This occurrence will be addressed in the following subsection.

5.4.3.3 Kernel variance choice

Following the previous subsection, we can consider that the problem of determining the kernel variances is an optimization problem by itself. As a general rule, the best kernel variances are the ones that provide the minimum SNR estimator variance, given that the bias introduced by the kernel IP estimator is taken into consideration in the NLS problem. Since the theoretical variance has been assessed, the most direct approach is to determine v_1^2 and v_2^2 by minimizing Var $\{\hat{\rho}\}$.

Consider now that we have a relative kernel variance for each one of v_1^2 and v_2^2 such that

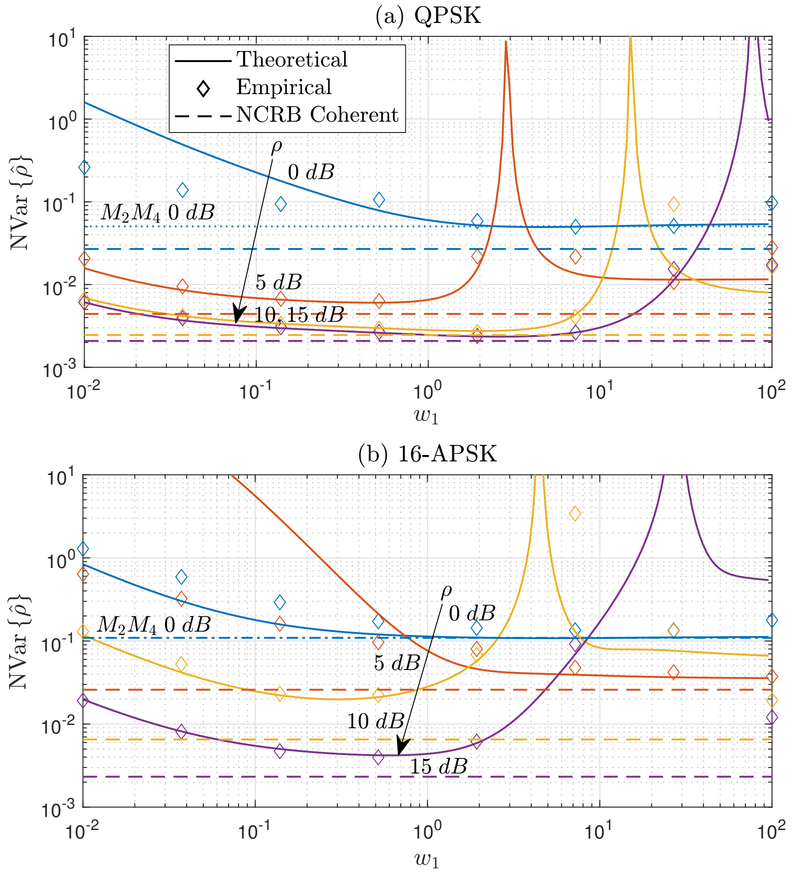
$$w_1 = \frac{v_1^2}{N}, \qquad w_2 = \frac{v_2^2}{N}. \tag{5.155}$$

Since N is unknown, the relative variance must be computed by a first estimate $\hat{N}$. Given that NLS also requires an initialization, we will benefit from the prior estimation given by the sixth-order method of moments, namely $\hat{N}_6$. Therefore, the kernel variances used for the estimation of SNR will be $v_q^2 = w_q\hat{N}_6$. Unless stated otherwise, all v_q^2 will be based on $\hat{N}_6$ in both empirical and theoretical results.

Given the $w_1^2 \neq w_2^2$ constraint, the most direct approach is to optimize the relative kernel bandwidths in a coordinate descent manner. In other words, we will use a gradient descent method for each relative kernel variance at a time, alternating between the two of them until the global minima has been reached. For convenience, we will further constrain the relative kernel bandwidth with $w_2 > w_1$ so that $w_1^2 \neq w_2^2$ is never achieved.

To get more insights on the problem, Figure 5.15 illustrates the normalized variance as a function of the relative kernel variance w_1 for fixed w_2 with different SNR values. The value of $w_2 = 10^2$ is chosen so that $w_2 > w_1$ is guaranteed for all the shown values of w_1. This figure represents a possible stage of the coordinated optimization problem, in this case on the side of w_1. Both empirical and theoretical estimator variances are shown again to corroborate that the optimized kernel variances values would endorse the desired empirical variance. Again, the empirical values are obtained through Monte Carlo simulations with 500 iterations. The overall intention is to show that the relative kernel variances can be easily optimized with a gradient descend method, albeit with a downside. The spikes that can be observed for some values of SNR correspond to the $d_q = e_q$ conundrum. These are particularly difficult to avoid, since they can disrupt the good behaviour of the gradient descend method if the initialization is either in one of its sides. To avoid this, the first step will be to perform two different gradient descend algorithms, each one with a different initialization, in order to iterate with the potentially better choice of kernel variances. It is also worth mentioning that the mismatch between the empirical and theoretical variances, mainly for low values of SNR and w_1, is again due to a poor resolution of the NLS problem, similar to Figure 5.14.

On a more general note, Figure 5.15 also shows that the entropy-based method exhibits the best performance for $\omega_1 \approx 1$ as the SNR increases. This observation confirms that the best kernel variances at high SNR, i.e. separated GMM clusters, are the ones that are at a similar order of magnitude to the noise variance. Conversely, low SNR values perform better as w_1 increases, confirming that a quasi-Gaussian distribution prefers high values of contamination to appropriately determine the true noise level. Furthermore, the empirical variance of the M_2M_4 estimator is also shown for $\rho = 0$ dB as a dashed line. As can be seen, the entropy-based estimator tends to this



(a) QPSK

(b) 16-APSK

Figure 5.15: Normalized variance of the SNR estimator as a function of its true value for multiple values of v_1^2 and fixed $v_2^2 = 10^3$.

value for an increasing kernel variance, confirming the low SNR regime study that is addressed in Subsection 5.4.3.1.

Finally, Algorithm 2 depicts the proposed coordinated descend method used for determining the relative kernel variances, and posterior v_1^2 and v_2^2, by gathering all previous insights. Here, $\min_{w_q} \{\text{Var}\{\hat{\rho}\}, w_0, A, B\}$ denotes the relative kernel variance that minimizes the SNR estimator variance with an initial guess w_0 constrained in the region $A \leq w_q \leq B$. After some parameter initialization, including a very large initial w_2, the algorithm looks for the minimum value at either sides of the spike. Once one side has been determined, the algorithm iterates w_1 and w_2 in an alternate manner until convergence is achieved. At each iteration, the $w_2 > w_1$ constraint is enforced to the regions of the optimization method. The step tolerance ϵ that will be used throughout this section will be $\epsilon = 10^{-10}$, unless otherwise stated. Generally speaking, the algorithm determines a considerable good kernel variance value with few steps, with the exception of the most difficult, medium SNR values. This behaviour will be confirmed jointly with the performance of the estimator in Subsection 5.4.5.

5.4.4 Frequency error robustness

Before proceeding further, it is worth briefly reviewing how robustness has been addressed up until this point. In Sections 5.2 (determinant of the covariance matrix) and 5.3 (generalized coherence), the sample estimators are nonrobust, and the entropy-based approach has been proposed to gain robustness. However, in this section, the scenario is different, and the SNR estimators based on the sample estimate of the moments are now the robust methods. Even more, they are EVB methods, meaning that the moment-based estimators are invariant to two common errors in digital commu-

Algorithm 2 Coordinate gradient descend method for determining the relative kernel variances.

```
$k = 1; 0 < \epsilon \ll 1; A = 10^{-4}; B = 10^{4}; w_2[0] = B$
$w_1[0] = \min\left\{\min_{w_1}\{\text{Var}\{\hat{\rho}\}, A, A, B\}, \min_{w_1}\{\text{Var}\{\hat{\rho}\}, B, A, B\}\right\}$
while $\Delta > \epsilon$
    $w_2[k] \leftarrow \min_{w_2}\{\text{Var}\{\hat{\rho}\}, w_2[k-1], w_1[k-1], B\}$
    $w_1[k] \leftarrow \min_{w_1}\{\text{Var}\{\hat{\rho}\}, w_1[k-1], A, w_2[k]\}$
    $\Delta = \max\left\{\frac{|w_1[k]-w_1[k-1]|}{w_1[k]}, \frac{|w_2[k]-w_2[k-1]|}{w_2[k]}\right\}$
    $k \leftarrow k+1$
end
```

nications channels (carrier phase and frequency errors). Thanks to this, EVB methods are widely used when the coherence time of the channel, in the sense of nonvarying carrier error offsets, is unknown, given that it is inconsequential for the EVB approach. Meanwhile, the entropy-based approach is a coherent method, and it is “only” invariant to carrier phase errors. Nevertheless, low-order moments are still not capable of discerning the intricacies of the implicit PDF, unlike an entropy measure, which makes them prone to a high SNR estimation bias for multilevel constellations. Higher-order moments provide an intermediate solution to this problem, being both invariant to carrier offsets and with better awareness of the complexity of the PDF, but they are not so easily manageable. Therefore, it is relevant to develop techniques that bestow *both* robustness in front of carrier errors and good performance with multi-level constellations, thus providing an alternative to moments-based estimators.

This subsection is devoted to the study of a method to gain robustness in front of carrier frequency errors for the entropy-based approach. Until this point, the entropy-based applications have been capitalizing on the U-statistics properties of the IP estimator. Thanks to the U-statistics standpoint, the entropy estimator is inherently unbiased, provided that the kernel variance is known and considered in the expected value, and only the variance becomes relevant. Here we develop the concept further, taking advantage of the *incomplete* U-statistics (addressed in Subsection 3.3.2), which still provide an unbiased estimate, but at the cost of a higher variance. While this approach may deteriorate the overall SNR estimate, it also may be advantageous in a scenario where the estimate window cannot be aligned with the coherence time of the channel. In fact, a common practice in estimation and detection problems in the presence of frequency errors is to limit the coherent integration of the estimator [GLS16]. This implicitly decreases the performance of the coherent estimators that require such limitations. Meanwhile, the method proposed in this subsection takes advantage of the complete length of the available data, as we will see hereunder.

First and foremost, and for the sake of clarity, let us express again the second-order Rényi entropy estimator based on kernel pairwise differences, but also expressed in terms of the consequent kernel matrix:

$$\hat{h}_{2,v_q}(X) = -\ln\left(\frac{2}{L(L-1)}\sum_{0\le i<j\le L-1} k_{v_q}(x(i)-x(j))\right) \tag{5.156a}$$

$$= -\ln\left(\frac{1}{L(L-1)}\sum_{i,j=0}^{L-1} k_{v_q}(x(i)-x(j)) - L\right) = -\ln\left(\mathbf{1}_L^T\mathbf{K}\mathbf{1}_L - L\right), \tag{5.156b}$$

where $[\mathbf{K}]_{i,j} = k_{v_q}(x(i)-x(j))$. The incomplete U-statistic approach comes from the observation that there is a strong dependence among the terms of the summation, mainly due to the repetition of samples $x(i)$ and $x(j)$. Following this rationale, the set of terms that contribute

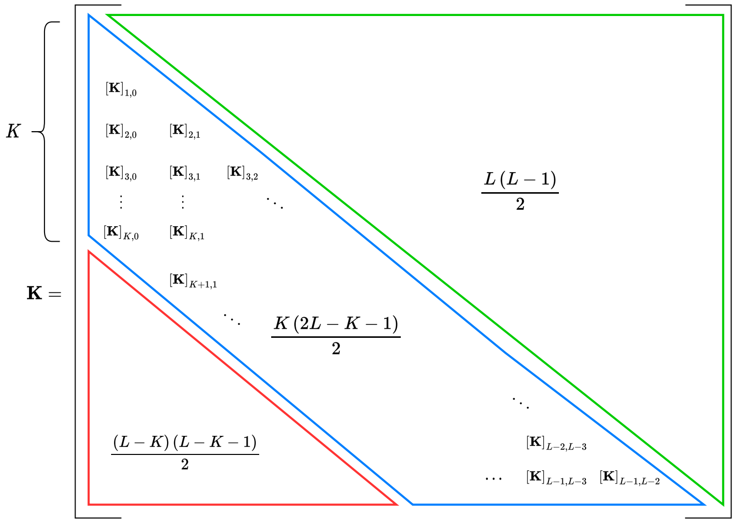


Figure 5.16: Kernel matrix composition and number of elements in each block.

mostly to the estimator are then those terms whose sample pairs do not repeat indexes:

$$\left\{k_{v_q}\left(x\left(1\right)-x\left(2\right)\right),k_{v_q}\left(x\left(3\right)-x\left(4\right)\right),...,k_{v_q}\left(x\left(L-2\right)-x\left(L-1\right)\right)\right\}. \tag{5.157}$$

The subsequent terms are also of importance to reduce the variance of the estimator, but their contribution becomes smaller the more indexes are repeated. The general approach is then to reduce the total number of pairs, $L\left(L-1\right)$, so that only the ones that are close to each other (in terms of the observation window) are used in the computation of the estimator. This new subset of terms is then considered to have negligible carrier frequency error.

Thanks to the particular structure of the kernel matrix, those terms in which we are interested are the ones closer to the main diagonal. In particular, we will determine the used terms for estimating the second-order Rényi entropy as those that are comprised within the first K diagonals. These diagonals contain the terms whose maximum index distance $j-i$ is K, consequently bounded by $1 \leq K \leq L-1$. While $K=1$ corresponds to only the first diagonal, $K=L-1$ denotes that the complete kernel matrix is being used in the estimation of the IP or the second-order Rényi entropy, and it is therefore equal to the estimator in (5.156). The subset of terms that are used to compute the second-order Rényi entropy is then composed by the pairwise differences whose indexes are within this distance. Concretely, the *incomplete* estimator then becomes

$$\hat{h}_{2,v_q,K}\left(X\right)=-\ln\left(\hat{U}_{v_q,K}\left(X\right)\right), \tag{5.158}$$

where

$$\hat{U}_{v_q,K}\left(X\right)=\frac{2}{K\left(2L-K-1\right)}\sum_{0\leq i\leq L-1}\ \sum_{i<j\leq J_i}k_{v_q}\left(x\left(i\right)-x\left(j\right)\right), \tag{5.159}$$

and $J_i=\min\left\{K+i,L-1\right\}$. In this case, the number of terms is $K\left(2L-K-1\right)/2$, which is much smaller than $L\left(L-1\right)/2$ for $K \ll L$.

Figure 5.16 illustrates the different terms used in each one of the proposed estimators. While in (5.156) the $L\left(L-1\right)/2$ elements are computed (due to symmetry), which would correspond to the upper-right block, here we propose to use only the $K\left(2L-K-1\right)/2$ terms obtained by subtracting the $\left(L-K\right)\left(L-K-1\right)/2$ terms from the complete upper/lower triangular matrix.

Next, we proceed to analyze the behaviour of the SNR estimate with a given value of K. In fact, the theoretical variance needs to be adjusted in terms of K for a precise kernel variance computation. This is a consequence of reducing the terms used in the second-order Rényi entropy

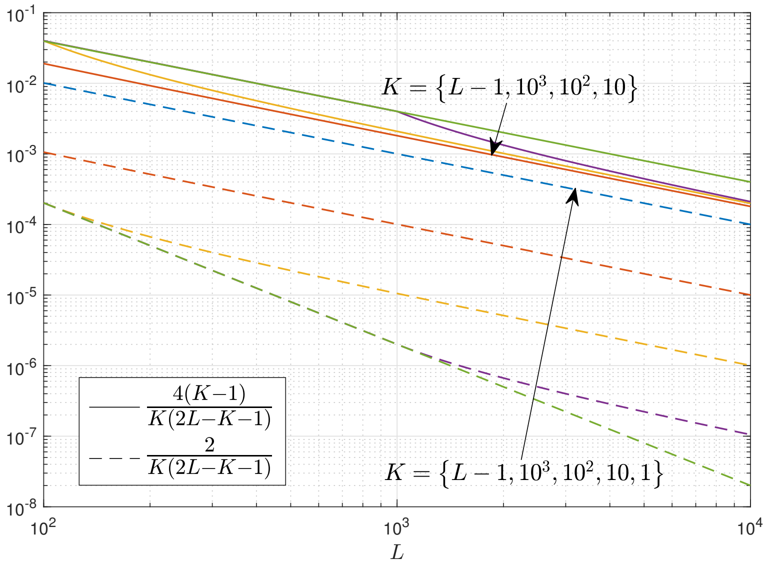


Figure 5.17: Weights associated to a, b and c of the variance of the incomplete U-statistics SNR estimator in (5.160).

in a structured manner. Therefore, the weights associated to the terms a, b and c from (5.138) must be corrected. By following a similar approach than in Appendix 7.4.3, one can determine that now the number of elements of a, described in (5.139), is $K(2L-K-1)(K-1)$, and $K(2L-K-1)/2$ in the case of b, described in (5.141). The theoretical variance of the IP estimator is then

$$\sigma^2_{U_K}\{q,q'\} = \frac{(a-c)\,K(2L-K-1)(K-1)}{\left(K(2L-K-1)/2\right)^2} + \frac{(b-c)\,K(2L-K-1)/2}{\left(K(2L-K-1)/2\right)^2} \tag{5.160a}$$

$$= \frac{4(K-1)(a-c)+2(b-c)}{K(2L-K-1)}. \tag{5.160b}$$

It can be seen now that the difference between the weights of each term is now determined by K, which should be appropriately computed. Additionally, note that for $K = L-1$, hence using all the available samples, then (5.160) and (5.138) are equal. Therefore, one can define (5.160) as the general theoretical variance of the IP, and determine K depending on the scenario.

A relevant aspect from (5.160) is the relationship between K and L. To better ascertain the behaviour of the associated weights in (5.160), Figure 5.17 shows how they behave with a fixed value of K (except for $K = L-1$) and an increasing value of L. For the cases when $K > L$, a new value for K is assigned that corresponds with the corresponding $L-1$, and that is why for $K = 10^3$, for example, only appears for $L > 10^3$. First, it can be clearly seen that all weights decrease as L increase. This means that (5.158) still benefits for an increasing value of L, even for small values of K. Furthermore, for large values of L and small K one can attain similar performance than with high K and low L. Secondly, the choice of K is more relevant for the weight associated to b than the others (given that c is governed by both weights). The result is that by decreasing K the penalty of the b terms increases, while the penalty of the a and c terms decrease, albeit not so much as with b. In fact, for $K = 1$ the weights of a are not shown, since it is directly 0.

These observations constitute one of the main benefits of the semi-EVB approach. The robust entropy-based method succeeds in not being limited by the coherence time of the channel by marginally sacrificing the potential performance achieved by fully coherent approaches. The choice of K does not greatly detriment the performance of the estimator, showing that the computation of repeated samples is inherently redundant [VR22]. Once again, this is a consequence of the processing of the absolute value of pair-wise differences of data instead of sample per sample.

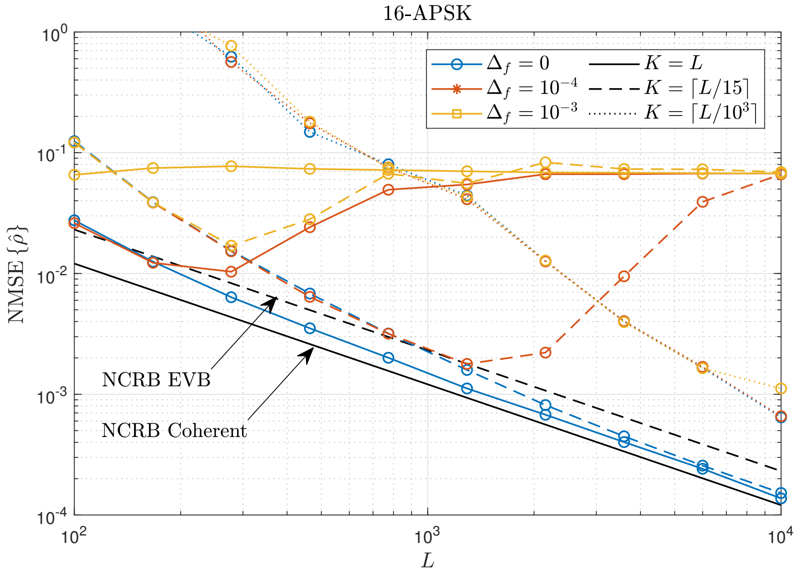


Figure 5.18: NMSE of the entropy-based SNR estimator as a function of L for multiple carrier frequency offsets and K, with 15 dB of true SNR value.

Nevertheless, given that the carrier frequency errors may vary per application and symbol block, one can determine a worst-case scenario by fixing a $\Delta_{f_{max}}$ such that $|\Delta_f| < \Delta_{f_{max}}$, and measuring the SNR with this assumption, relieving the estimator from the need of measuring Δ_f.

In order to ascertain this behaviour, the NMSE of the SNR estimator is shown in Figure 5.18. It can be seen that, even for a very restrictive choice of K, the estimator tends asymptotically to the NMSE of the complete estimator without carrier frequency errors. Consequently, the algorithm is capable of both handling synchronization errors and improving its performance as L increases as happens with EVB methods, although up to a given L depending on the value of Δ_f. For completeness, the NMSE for $K = L$ is included, showing that the estimator from (5.121) is consistent (at least for this constellation with the given SNR value).

5.4.5 Numerical results and conclusion

Finally, the performance of the proposed SNR estimator is shown by means of Monte Carlo simulations. For comparison, the following methods are also exhibited. The M_2M_4 is shown as a basis for the EVB estimators [ME94], jointly with the Kolmogorov-Smirnov test method [Fu+15]. The method of moments up to the sixth-order statistics M_6 [LM07] is also shown, as it is used as the initialization of the entropy-based NLS problem and to measure the relative kernel variances, as well as the method of moments up to the eight order, namely the M_8 [ÁLM10]. The EM estimator [Das08] is displayed for the comparison with a coherent method. The NMSE is then measured for these estimators and, since the results show both EVB and coherent approaches, we also show the NCRB of their respective methods obtained from [GLM09]. The data is composed of $L = 500$ observations for all simulations, the symbols are considered to be equiprobable with $p_m = 1/M$, and the QPSK, 16-APSK and 16-Quadrature Amplitude Modulation (QAM) constellations are exhibited. Regarding the entropy-based estimator, the case $Q = 2$ is used, the relative kernel variances are obtained through Algorithm 2, and the SNR is estimated through (5.121).

It is worth pointing out that equiprobable symbols are assumed as it is the common practice for the analyses of SNR estimators. While other distributions could be considered, the entropy-based approach performs similarly to the equiprobable case, provided that the probabilities of the symbols are known. If unknown, these should be estimated or assumed to be equiprobable (maximum entropy case), which would then contribute with an added bias in the overall estimation process. Other estimators are more robust to this facet of SNR estimation. For instance, the EM estimator

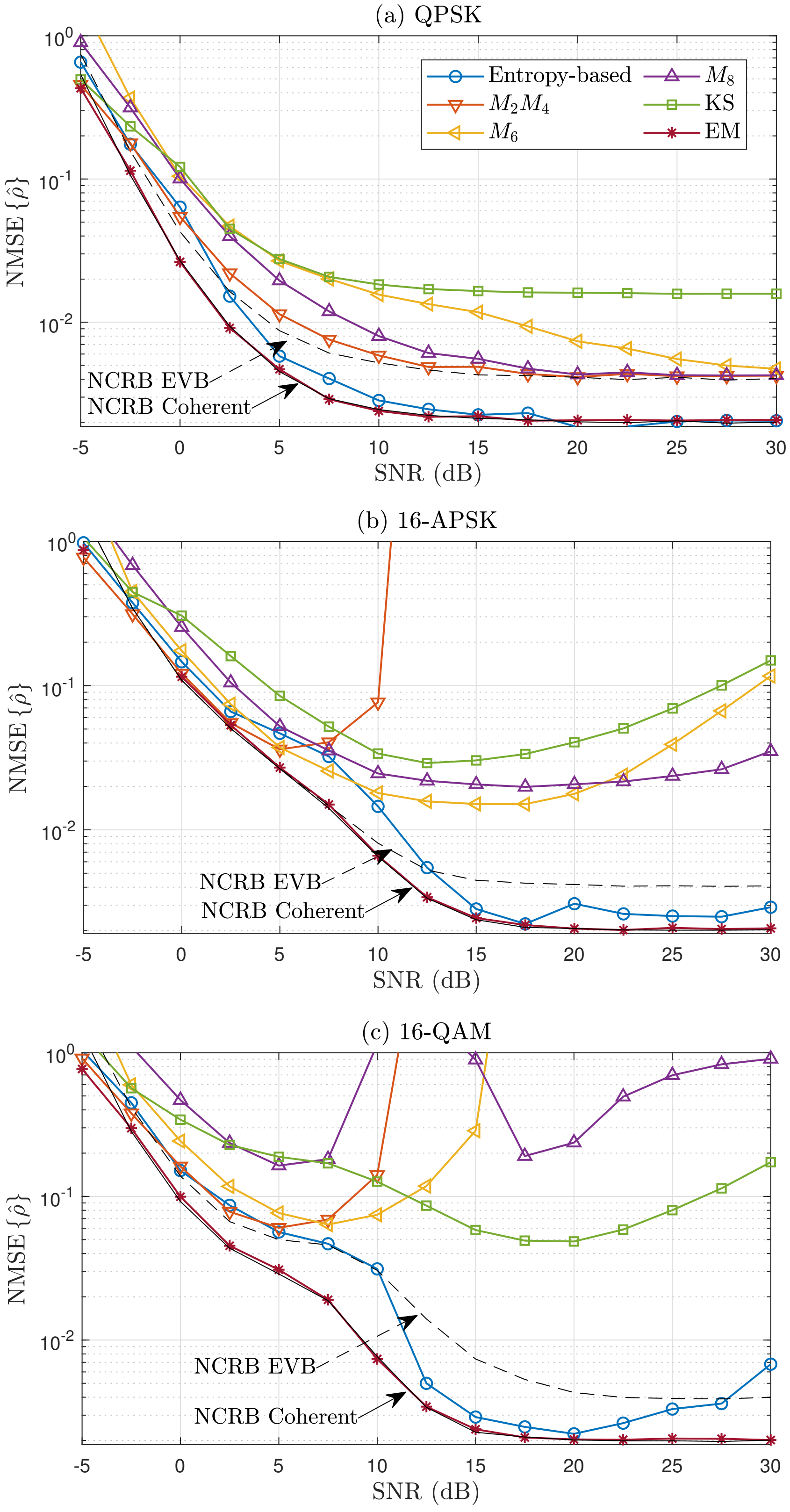


Figure 5.19: NMSE of different estimators as a function of the true SNR value for three constellations. The parameters are $L = 500$, $\phi = 0$, $\Delta_f = 0$ and $K = L$.

computes the probabilities within the algorithm itself, although it is then another point of failure if the initial guess of probabilities is not close enough to the true values. The moments-based estimators require the probability of the constellation level, which is, in principle, more robust in front of small errors of the symbol probabilities. The case of small deviations of the symbols probabilities will not be considered here and left for future work.

First, we show the aforementioned estimators under the assumption of perfect synchronization in Figure 5.19. As can be seen, the entropy-based estimator outperforms the EVB methods, generally speaking. The tendency of the entropy-based method to the M_2M_4 for low SNR values can also be appreciated, rendering the same results up to approximately $\rho = 0$ dB, depending on the constellation. At higher SNR values, the proposed method improves with respect to the EVB methods, surpassing the EVB CRB. It is only within a small interval of around the 5-10 dB SNR range in the 16-APSK constellation that the entropy-based method is slightly outperformed by some moments-based methods. It is also worth noting that the performance of the estimator degrades at very high SNR values with the 16-QAM constellation, which is mainly caused by the poor initialization of the M_6 estimate. While other methods can be used as the first guess, such as the M_8, the performance is not much improved due to this constellation being of particular difficulty to EVB methods, as can be appreciated in the figure. Regarding the Kolmogorov-Smirnov test, which is particularly focused on multilevel constellations at high SNR, the entropy-based method provides better performance and lacks the need for a computationally expensive comparison with predetermined functions. On the other hand, the EM algorithm performs close to the coherent CRB, as is expected by an iterative solution to the maximum likelihood estimation under standard conditions. It is, nonetheless, a highly computationally complex method and very dependent on the lack of synchronization errors, as it will be shown next.

Next, Figure 5.20 portrays the same SNR estimators from the previous figure, but considering synchronization errors. While EVB methods are invariant to phase and frequency offsets, the EM algorithm is hindered as the SNR increases. The entropy-based method adopts an intermediate solution by providing robustness thanks to limiting the pairwise differences. The cost of such limitation is especially acute around 10 dB for multilevel constellations, and it accentuates the problem with the 16-QAM at high SNR. From a more general point of view, although the NMSE begins to increase at high SNR, it is due to the choice of K, and it can be alleviated by further limiting the number of terms used in the entropy estimation. While it is dependent on Δ_f, the results encourage a conservative value of K to accommodate most SNR values, but it should be fine-tuned if very high SNR is expected. In summary, the entropy-based estimator performance is slightly worsened by choosing $K < L$, as expected, but it still outperforms the other methods for a wide range of SNR values, thus providing an edge in front of unknown synchronization errors.

Finally, Figure 5.21 shows the number of average iterations that requires Algorithm 2 for determining the kernel variances for each constellation shown in Figure 5.19. The number of average iterations for the case of frequency errors (thus using the incomplete U-statistics approach) is not shown for clarity of exposition, as it is roughly the same as the ones in Figure 5.21. Clearly, the algorithms struggles around the intermediate values of SNR, which is in agreement with the regimes observed in Subsection 5.4.3.1, corresponding to the one in which the signal is neither Gaussian nor distinguishable GMM components. Other than that, the number of iterations are lower at low SNR values, as it just needs to choose a sufficiently high value of kernel variance. The case of QPSK is particularly interesting, given that it requires, in average, more iterations than the multi-level constellations. This is mainly due to the fact that the variance of the estimator decreases slower than with other constellations, which can actually be seen at Figure 5.15. Since the moments-based estimators are optimal for this constellation and regime (QPSK and low SNR) in the sense of the minimum variance estimator, the entropy-based estimator tries to behave as the moments-based estimator by increasing the kernel variance indefinitely, which is only stopped when it fulfills the stopping rule. Nonetheless, generally speaking, Algorithm 2 does not require many iterations, being 6 the maximum number observed in average.

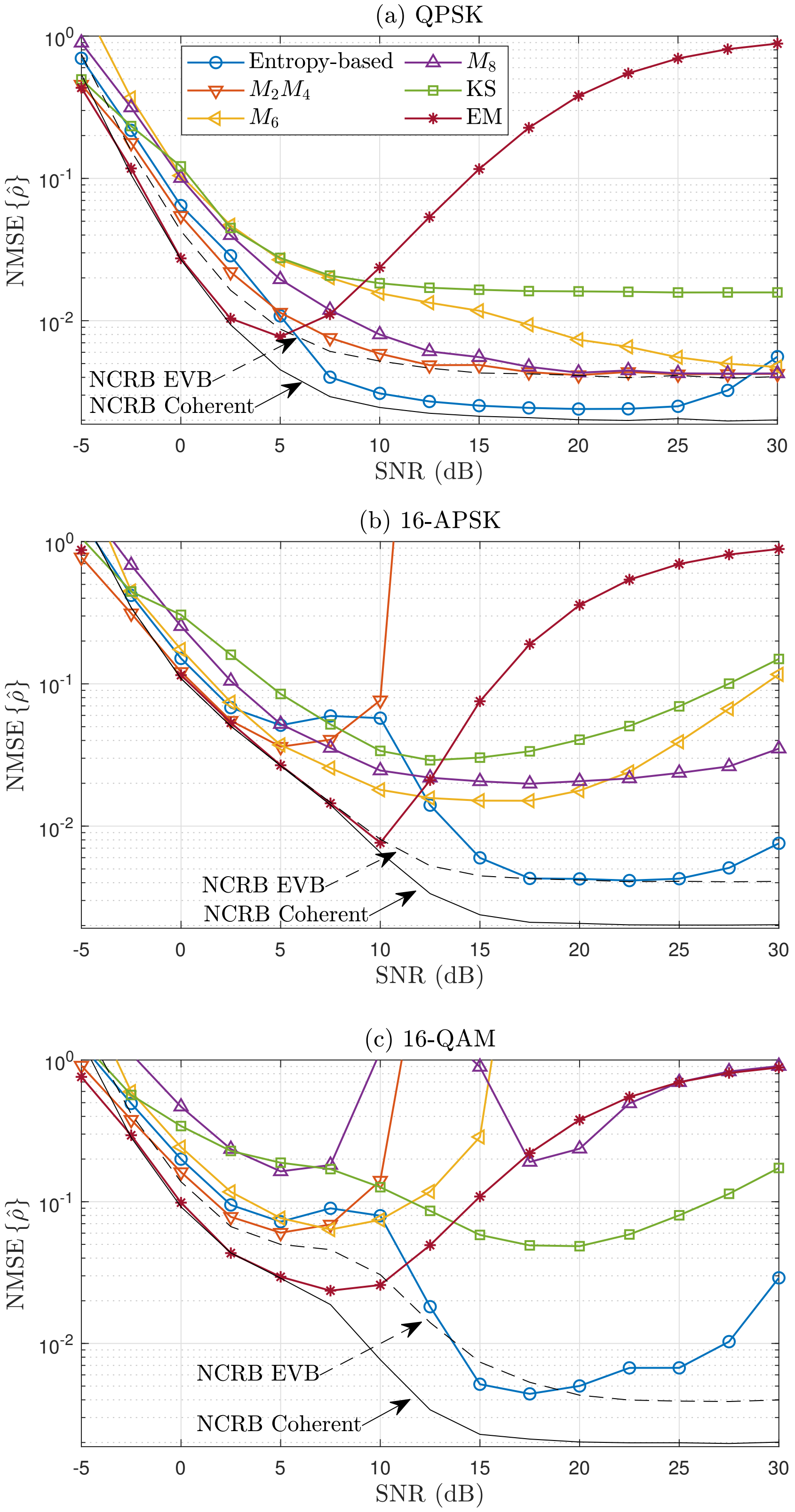


Figure 5.20: NMSE of different estimators as a function of the true SNR value for three constellations. The parameters are $L = 500$, $\Delta_f = 10^{-4}$, $K = L$, and ϕ is selected randomly between 10^{-4} and 10^{-1}.

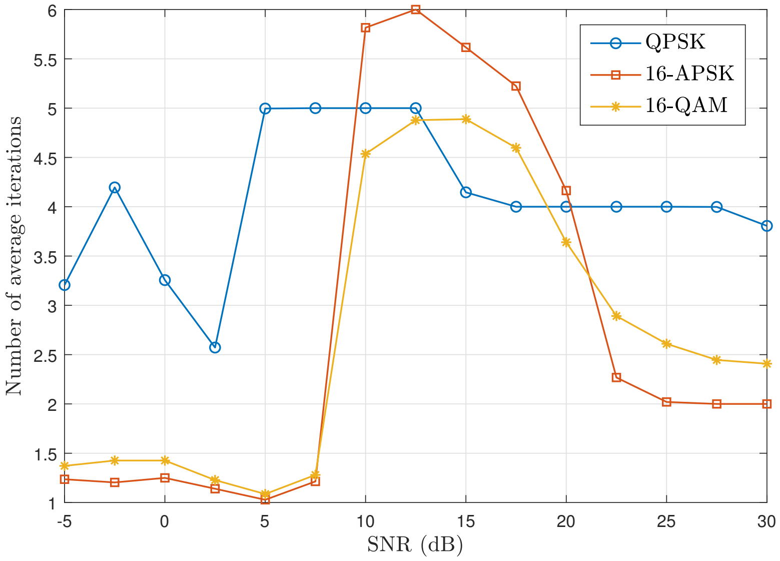


Figure 5.21: Average number of iterations for each measurement of the kernel variances in Algorithm 2 as a function of the true value of SNR for each evaluated constellation.

In conclusion, we have shown that the classical SNR estimation problem can be conducted by an information-theoretic approach. The resulting estimator is neither an EVB nor a coherent method, but a hybrid between them. At low SNR, it behaves exactly as an EVB method by increasing the kernel variance, while at high SNR it has the potential to achieve the performance of coherent methods. On the one hand, thanks to utilizing an entropy measure as a basis for the problem, the estimator is invariant in front of carrier phase errors. On the other hand, the entropy estimator itself can be modified in order to gain robustness in front of carrier frequency errors thanks to its structure based on pairwise differences. The results shown in this subsection support these statements by showing that the performance is only slightly worsened by considering the *incomplete* U-statistics approach, constituting an attractive option under synchronization errors. However, the method still deteriorates for very high SNR values, either by a bad initialization or because the effect of the carrier frequency error becomes severe. All these ideas and relations are portrayed in Figure 5.22.

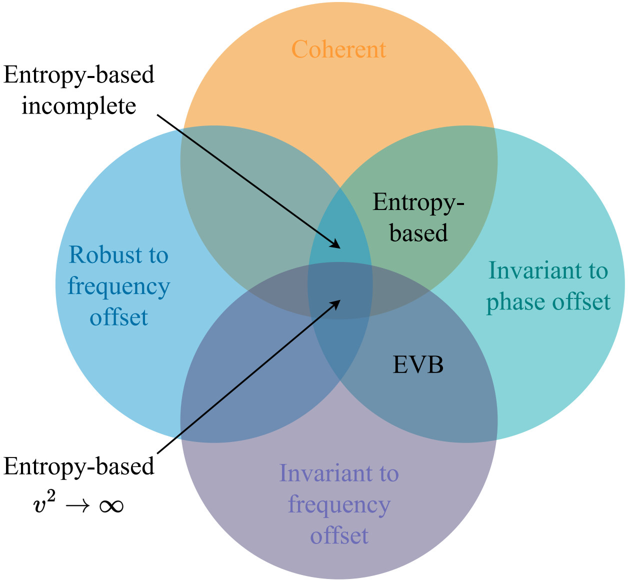


Figure 5.22: Complete Venn diagram of the entropy-based approach in relation to EVB and coherent methods.

5.5 Concluding remarks

This chapter has provided a unified rationale for solving three different problems: the estimation of the variance, the MSC, and the SNR of a digital communications channel. The motif of this joint rationale has been to substitute classical second-order signal processing techniques with a measure of entropy. This exchange has allowed us to benefit from the properties given by information-theoretic tools.

The second-order Rényi entropy has been considered for all the aforementioned estimators in order to guarantee a set of desired properties. Not only it is a quadratic measure of uncertainty, which may become simpler to estimate (Chapter 3), but the estimator used in this chapter has the particular expression of a U-statistic. This allows us to intertwine the expected value with the parameters to be estimated in a straightforward fashion. Since the U-statistic employed in each section becomes an unbiased estimate of a contaminated version of the second-order Rényi entropy, the usually restrictive bias-variance trade-off provided by the kernel estimator is relaxed, and only the variance becomes relevant. From an information-theoretic point of view, a measure of entropy becomes only sensitive to the probabilities of events, rather than to their values. Consequently, the estimators gain robustness in front of various typical adversarial effects. The sample pairwise differences of the entropy estimator reinforce this vision by dropping the dependence on the mean value of a given distribution, or by becoming invariant to fixed rotations. As a result, all the estimators derived in this chapter are particularly useful not only under standard conditions but also for scenarios where a robust estimate is required.

Furthermore, we have also shown that the entropy estimator is intrinsically related to second-order statistics. In all three problems, the estimate resembles a second-order moment estimator by just tweaking a shared parameter, the kernel bandwidth. The more the kernel bandwidth increases, the more narrow-viewed the estimators become in the sense of statistical moments. On the contrary, lower values of kernel bandwidth make the estimators more responsive to the particularities of the data, gaining awareness of the intricacies of the underlying distribution. The problem is then translated to finding the lowest possible kernel bandwidth such that the estimate becomes wary of complex data descriptors without increasing the variance of the estimator too much.

With this chapter, the core contents of the thesis have been concluded. We then proceed to finalize with the joint vision of all chapters, to analyze the contributions, and to determine possible future research lines.

Chapter 6

Conclusions

This dissertation has addressed the engagement between information measures and second-order statistics. The purpose has been to substitute variance and covariance with measures of information. However, it is undeniable that the knowledge and tools developed for the former have a much more solid trajectory. Therefore, classical second-order statistics tools have been leveraged to benefit the measurement of information, with the objective of a posterior substitution of the second-order approach itself. In the sequel, a short summary of each chapter is provided, focusing on the added value to the rationale mentioned above.

First, Chapter 1 has provided an introduction to the research topic of estimating and employing information measures, focusing on the most relevant aspects of it. Generally speaking, these are: scalability, interpretability, learning rate, and universality. These key aspects are reviewed from the perspective of the objective of the thesis.

Secondly, after some summary of required tools and known measures of linear information in Sections 2.1 and 2.2, Chapter 2 has explored the surrogates of Shannon's entropy, KL divergence, and MI that are conducive to second-order moments in Section 2.3. In this regard, the second-order Rényi entropy, the χ^2-divergence, and the SMI (or mean square contingency) have shown to be the desired surrogates for the purpose of estimation. All of these surrogates are either an upper-bound or a lower-bound of the original measures, a useful property if minimizing/maximizing these measures is required in a given application, and besides, the SMI behaves as a local approximation of the MI for close distributions. Subsection 2.3.4 explores other dependence measures or surrogates of the MI, but none of them are related to their original counterparts, nor define a proper information measure beyond the concrete case of independent random variables.

Chapter 3 studies the applicability of kernel methods for the estimation of information measures. The identified surrogates may be expressed as the L_2-norm of some functional, but are nonlinear in nature, given that they are defined from probability distributions and not variances and covariances. In view of this, nonlinear processing is required, but one that specifically allows to translate nonlinear problems to linear ones. Kernels arise as a sought-after tool due to their capability of solving problems on a higher-dimensional feature space, where the problem may linearize itself. As a matter of fact, the unified rationale provided in Section 3.2 points out that the kernel methods for measuring information promote the estimation of surrogates of information measures that are related to second-order statistics. From the HGR coefficient, a dependence measure comprised of covariance and variances, to the SMI itself. This idea is summarized in Figure 3.1, demonstrating that the choice of surrogates from the previous chapter is aligned with the tool presented in this one. The most prominent example of this relation is then given in Section 3.3, where kernel methods are derived naturally *from*, and not *for*, an estimate of the second-order Rényi entropy. Although this estimator is not particularly novel, we have bestowed the estimate with a U-statistics point of view, which shows that the estimator is, indeed, measuring variance in the high-dimensional feature space.

In Chapter 4, a framework for estimating measures of information is developed. The objective is to oppose the kernel method standpoint through two different notions. On one side, the feature map is fully characterized, providing an explicit expression of the mapping function. Thanks to this, the feature map can be managed, and the space can be regularized with physical meaning, rather than blindly. On the other side, the dimension of the feature space itself is limited, not only to improve the computational complexity of the estimate but also to address the question of whether an infinite-dimensional space is truly required. Chapter 4 is then divided into three parts. First, in Section 4.1, the analysis is provided for discrete sources. The implications of reducing the feature space dimension are addressed here, resulting in an upper/lower-bound (entropy/dependence) of the true information measure, jointly with the direct relationship of the information measures with CCA, and partially to PCA. A key aspect of the mapping is encountered in the fact that both mapped autocorrelation and autocovariance matrices are valid for estimating information, given that these matrices are indirectly related to the probability simplex. Next, Section 4.2 provides an interlude between discrete and analog sources, determining the necessary steps, requirements, and limitations. The result is that, given that analog sources require a function space, any limitation of the dimension of the space (sampling) ensues the bound examined in the previous section. As a matter of fact, this is the price to pay for avoiding the kernel methodology, capable of spanning infinite-dimensional feature spaces. Nevertheless, the kernel approach is turned around by considering outer products instead of inner products, and the limitation of the space yields covariance and correlation matrices, whose sample estimates are known to be consistent if the data size is higher than the dimension of the feature space. Section 4.3 wraps up the previous sections by expressing the desired estimators of information from the mapped matrices. The results show that the developed estimators can compete with existing estimators while providing a low computational complexity. Furthermore, the proposed uniform sampling, which is inspired by the Gaussian-Hermite quadrature, entails a Toeplitz structure to the involved matrices, allowing both to analyze the asymptotic behaviour of the estimators and to provide an even further reduction in the overall complexity.

Lastly, Chapter 5 tackles the last facet of the dissertation, which is to substitute second-order statistics with information measures to perform a variety of tasks. Once the methodology for estimating information with a covariance measure is displayed, the objective is steered in the opposite direction. This chapter questions the advantage of considering an entropy measure for estimating the determinant of the covariance matrix in Section 5.2 (particularized to univariate random variables), the coherence in Section 5.3 (particularized to bivariate random variables), and the SNR of a digital communications channel in Section 5.4 (also particularized to univariate random variables). The first two sections have dealt with the sensibility in front of outliers of the sample covariance matrix, and the entropy-based counterpart has arisen as a robust approach to the respective problems. Entropy provides two advantages in front of second-order statistics. On the one hand, it does depend on the probability of events, and not on their value. This translates to better handling of outliers, since these are, by definition, events with low probability but with large magnitude. On the other hand, entropy also captures higher-order moments, rendering this approach a better solution for distributions that are far from the Gaussian shape. Nonetheless, the employed estimator still has an asymptotic relation with second-order statistics, enforcing the idea that entropy is more general in the sense of statistical moments. The entropy-based estimate is shown to be more effective not only than the sample estimate case, but it also performs generally better than methods specifically derived to gain robustness, such as Tyler's method. Finally, Section 5.4 provides an entropy-based SNR estimator that has desirable properties for this concrete problem. To this aim, the behaviour of the second-order Rényi entropy estimator in front of a GMM with equal variance from component to component is analyzed and subsequently employed to determine the best strategy for the purpose of determining the signal power (intervariance) and the noise power (intravariance) relationship. The resulting estimator gains invariance in front of constant rotations of the GMM, which are translated as carrier phase errors in this case, and is capable of deriving a methodology to gain robustness in front of varying rotations of the GMM, which are translated as carrier frequency errors. This is due to the pairwise processing induced by the entropy estimator, which ends up allocating the entropy-based SNR estimator as a hybrid between moment-based (generally speaking) methods and other, more complex, approaches. Moreover, the intrinsic relationship with second-order statistics is still relevant to this problem, showing that the estimate is, again, an asymptotic estimate of just the intravariance, thus equivalent to the sample variance.

In summary, the key points are as follow:

- Proposal of a finite-dimensional feature space based on the CF and outer products instead of inner products.
- Leveraging well-known signal processing techniques for improving the interpretability of the feature space.
- Estimation of the second-order Rényi entropy and the squared-loss mutual information with the previous tool.
- Interchange the Gaussian assumption to a later stage, after an entropic measure, to gain robustness.
- Take advantage of the entropy estimate based on pairwise differences of data to gain invariance, and robustness, in the problem of estimating SNR.

6.1 Future research

In view of the conclusions of this dissertation, we next proceed to identify potential topics for research and prospective extensions of the frameworks and topics developed. First, let us look to Chapter 4:

- Regarding Section 4.2, the strategy for sampling the CF is inspired by the univariate Gaussian-Hermite quadrature. Nonetheless, the required sampling is inherently bivariate (see (4.81)). Under this setting, the quadrature rule is not necessarily approximately uniform. This effect can be observed in [Jäc05]. To focus the sampling where the Gaussian distribution is concentrated can open the possibility of considering different contamination processes for each source, or to consider an "importance" sampling, where it focuses on the relevant shapes of the underlying CF. However, it should be noted that a nonuniform approach to the sampling implies the loss of the Toeplitz structure, with all that this entails.
- Again in Section 4.2, the regularization technique is inspired by the Blackman-Tukey spectral estimation and its strategy of tapering the spectral estimate, which reduces the variance of the estimator at the expense of an increased bias. From here, two potential improvements are formulated. On the one hand, a Gaussian window has used for its relation with the Gaussian convolution side, but other windows can be considered with a potential improvement on the overall estimation process. On the other hand, as rich as it is the topic of spectral estimation, other spectral estimation strategies can also be considered. For the time being, the Blackman-Tukey is rather observational than practical, being a result of the proposed regularization technique. However, the approach can be inverted, and other techniques to improve the bias-variance trade-off can be implemented. For example, Burg's or Capon's nonparametric high resolution methods [Kay88] can be studied for the encountered problem in this thesis.
- In Section 4.3, a rationale for grouping the hyper-parameters introduced by the Gaussian smoothing has been provided. However, the only parameter left p is still decoupled from the problem of estimating information. While it provides more robustness than the required choice of kernel bandwidth, the latter has a long history of rule of thumb goodness thanks to Silverman's rule. If a true data-driven model is required, a mechanism for choosing p from the data itself is required. This could be done by following similar strategies than those

in Algorithms 1 and 2, where an approximate correct value is chosen from prior estimates based on the variance of the data.

- The choice of the feature space dimension is a particularly relevant issue for the framework developed here. It does not only determine the computational complexity, but also the accuracy of the estimate. In this regard, it is of utmost interest to develop strategies for its appropriate choice. A first attempt was performed in [LCR20], with the objective of equating the problem of estimating information as a CCA problem to the determination of "active" canonical correlates. While only performed for Gaussian channels, it is an initiation to a data-driven choice of the required intrinsic dimension of the problem.

- Regarding Chapter 4 as a whole, this dissertation has addressed only the univariate case. The reason is that, by mapping from the data space to the feature space, the increase of dimensionality itself conduces to matrices. Therefore, if the input data is multivariate, the mapping does not lead up to matrices, but to multidimensional arrays. However, the posterior processing with these arrays is not trivial, more so if we take the dependence between components into consideration. The same can be said about nonindependent data, where it is not so clear how to map these dependencies into the feature space. Therefore, relevant future research is encountered in the examination of strategies for addressing both cases, either by detaching the multivariate case or data with memory so that it can be considered as virtual univariate/independent data or by performing a conjoint study of the mapping process.

Next, we focus on research lines that follow from Chapter 5:

- Similarly to the previous final point, the most straightforward extension of most of this chapter is to consider multivariate random variables. As a matter of fact, Section 2.1 is prepared for such considerations. Both determinants of the covariance matrix and generalized coherence can be developed by considering the more general, albeit intricate, case. The SNR estimator is also developed for single-input single-output channels, but most modern communications processes consider multiple-input multiple-output schemes. The extension from one to another is not trivial, since it must consider many more parameters to be estimated (the power of each link and their respective noise level). The algorithm for determining the kernel variance must compute much more values, and its scalability suffers in result. How to address these changes is certainly the next step for the SNR estimator.

- The principle of interchanging the Gaussian assumption with the entropy-based processing (Figure 5.1) is more general than the applications shown in this dissertation. The same framework can be applied to different problems that require robust signal processing. A potential research line is to explore other applications where this rationale can be useful, either in the sense of robustness or as an estimator in general.

- The extension of the SNR estimate to a wider concept of statistical channel information is also of great interest. Generally speaking, noncoherent communications can benefit from the extraction of information at the receiver side. Blind approaches to this task, such as the entropic measure, may help on the reduction of pilot contamination.

- Going into details, and considering the SNR estimator, the robustness in front of carrier frequency errors is achieved, but the rule for deciding the internal parameter K is not determined. Similar to the choice of p for the estimation of information in the previous chapter, the objective is to determine the required penalization introduced by the incomplete U-statistics estimator to achieve robustness without compromising the performance. A potential topic of research is how to optimally determine K, either directly from data observations, or by comparing with the expected model, whose number of clusters and probabilities are known under the context of the entropy-based estimator.

- Lastly, while the choice of kernel variance for the estimation of SNR provides good results, its implementation can be sometimes rather computationally heavy. Particularly, at high SNR values, if the initial estimate is not as accurate as it is in low SNR values, then the proposed algorithm tends to require more iterations, hindering the overall complexity of the estimate. A further analysis of the algorithm are in order, which would increase the general appeal of the entropy-based estimator.

Chapter 7

Appendices

7.1 Appendices of Chapter 2

7.1.1 Proof of (2.10)

First, we need to express the CF as

$$\left.\frac{\partial^{k_1+k_2+...+k_N}\varphi_X(\boldsymbol{\omega})}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}...\partial\omega_N^{k_N}}\right|_{\boldsymbol{\omega}=\mathbf{0}} = \frac{\partial^{k_1+k_2+...+k_N}}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}...\partial\omega_N^{k_N}}\left.\int_{\mathcal{X}} e^{(j\omega_1x_1+...+j\omega_Nx_N)}f_X(\mathbf{x})\,\mathrm{d}\mathbf{x}\right|_{\boldsymbol{\omega}=\mathbf{0}} \tag{7.1a}$$

$$= \frac{\partial^{k_1+k_2+...+k_N}}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}...\partial\omega_N^{k_N}}\left.\int_{\mathcal{X}} e^{j\omega_1x_1}...e^{j\omega_Nx_N}f_X(\mathbf{x})\,\mathrm{d}\mathbf{x}\right|_{\boldsymbol{\omega}=\mathbf{0}}. \tag{7.1b}$$

We begin by first expressing the integral by means of the Maclaurin series of the exponential for $n=\{1,...,N\}$. The decomposition is as follows:

$$e^{j\omega_nx_n} = \sum_{k=0}^{\infty}\frac{(j\omega_nx_n)^k}{k!}, \tag{7.2}$$

which results in the integral

$$\int_{\mathcal{X}}\sum_{k=0}^{\infty}\frac{(j\omega_1x_1)^k}{k!}...\sum_{k=0}^{\infty}\frac{(j\omega_Nx_N)^k}{k!}f_X(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathcal{X}}\prod_{n=1}^{N}\sum_{k=0}^{\infty}\frac{(j\omega_nx_n)^k}{k!}f_X(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{7.3}$$

At this point, we can split the summation between the terms whose $k=k_n$ and the ones that $k\neq k_n$. The reason behind is to separate those terms that will prevail after the k_n-th derivative of the k-th random variable at the origin, and those that will become zero. Let us express it as follows:

$$\int_{\mathcal{X}}\prod_{n=1}^{N}\sum_{k=0}^{\infty}\frac{(j\omega_nx_n)^k}{k!}f_X(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathcal{X}}\prod_{n=1}^{N}\frac{(j\omega_nx_n)^{k_n}}{k_n!}f_X(\mathbf{x})\,\mathrm{d}\mathbf{x} + O\left(\left.\omega_n^k\right|_{k\neq k_n}\right). \tag{7.4}$$

Then, the integral can be solved as follows:

$$\int_{\mathcal{X}}\prod_{n=1}^{N}\frac{(j\omega_nx_n)^{k_n}}{k_n!}f_X(\mathbf{x})\,\mathrm{d}\mathbf{x} + O\left(\left.\omega_n^k\right|_{k\neq k_n}\right)$$

$$= j^{k_1+k_2+...+k_N}\left(\int_{\mathcal{X}}x_1^{k_1}\cdots x_N^{k_N}\mathrm{d}\mathbf{x}\right)\prod_{n=1}^{N}\frac{\omega_n^{k_n}}{k_n!} + O\left(\left.\omega_n^k\right|_{k\neq k_n}\right) \tag{7.5a}$$

$$= j^{k_1+k_2+...+k_N}\mathbb{E}_{f_X}\left\{X_1^{k_1}...X_N^{k_N}\right\}\prod_{n=1}^{N}\frac{\omega_n^{k_n}}{k_n!} + O\left(\left.\omega_n^k\right|_{k\neq k_n}\right). \tag{7.5b}$$

Note that the equality is maintained for $k_n \in \mathbb{N}_0$. If $k_n = 0$, then the expectation over the n-th is not computed. Finally, we can compute the derivative by integrating the previous expression into (7.1), which results in

$$\begin{align}
&\left.\frac{\partial^{k_1+k_2+...+k_N}}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}...\partial\omega_N^{k_N}} j^{k_1+k_2+...+k_N}\mathbb{E}_{f_X}\left\{X_1^{k_1}...X_N^{k_N}\right\}\prod_{n=1}^{N}\frac{\omega_n^{k_n}}{k_n!}+O\left(\left.\omega_n^k\right|_{k\neq k_n}\right)\right|_{\boldsymbol{\omega}=\mathbf{0}} \nonumber\\
&= j^{k_1+k_2+...+k_N}\mathbb{E}_{f_X}\left\{X_1^{k_1}...X_N^{k_N}\right\}\left.\frac{\partial^{k_1+k_2+...+k_N}}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}...\partial\omega_N^{k_N}}\left(\prod_{n=1}^{N}\frac{\omega_n^{k_n}}{k_n!}+O\left(\left.\omega_n^k\right|_{k\neq k_n}\right)\right)\right|_{\boldsymbol{\omega}=\mathbf{0}} \tag{7.6a}\\
&= j^{k_1+k_2+...+k_N}\mathbb{E}_{f_X}\left\{X_1^{k_1}...X_N^{k_N}\right\}\left.\frac{\partial^{k_1+k_2+...+k_N}}{\partial\omega_1^{k_1}\partial\omega_2^{k_2}...\partial\omega_N^{k_N}}\prod_{n=1}^{N}\frac{\omega_n^{k_n}}{k_n!}\right|_{\boldsymbol{\omega}=\mathbf{0}} \tag{7.6b}\\
&= j^{k_1+k_2+...+k_N}\mathbb{E}_{f_X}\left\{X_1^{k_1}...X_N^{k_N}\right\}\left.\prod_{n=1}^{N}\frac{\partial^{k_n}}{\partial\omega_n^{k_n}}\frac{\omega_n^{k_n}}{k_n!}\right|_{\boldsymbol{\omega}=\mathbf{0}} \tag{7.6c}\\
&= j^{k_1+k_2+...+k_N}\mathbb{E}_{f_X}\left\{X_1^{k_1}...X_N^{k_N}\right\}, \tag{7.6d}
\end{align}$$

where from (7.6b) to (7.6d) one has to solve the partial derivatives in succession until it becomes 1. The terms that correspond to $k = k_n$ are can be either directly zero or have a proper derivative, although in the latter case they become zero for $\boldsymbol{\omega} = \mathbf{0}$.

7.1.2 Derivation of (2.67)

In order to properly move from (2.66) to (2.67), we need to make use of the constant -1 to complete the square in the numerator. Then, we can manipulate the expression as follows:

$$\begin{align}
D_{\chi^2}\left(f_X \parallel f_Y\right) &= \int_{\mathcal{X}}\frac{f_X^2\left(x\right)}{f_Y\left(x\right)}\mathrm{d}x - 1 \tag{7.7a}\\
&= \int_{\mathcal{X}}\frac{f_X^2\left(x\right)}{f_Y\left(x\right)}\mathrm{d}x - 2 + 1 \tag{7.7b}\\
&= \int_{\mathcal{X}}\frac{f_X^2\left(x\right)}{f_Y\left(x\right)}\mathrm{d}x - 2\int_{\mathcal{X}}\frac{f_X\left(x\right)f_Y\left(x\right)}{f_Y\left(x\right)}\mathrm{d}x + \int_{\mathcal{X}}\frac{f_Y^2\left(x\right)}{f_Y\left(x\right)}\mathrm{d}x \tag{7.7c}\\
&= \int_{\mathcal{X}}\frac{f_X^2\left(x\right) - f_X\left(x\right)f_Y\left(x\right) + f_Y^2\left(x\right)}{f_Y\left(x\right)}\mathrm{d}x \tag{7.7d}\\
&= \int_{\mathcal{X}}\frac{\left(f_X\left(x\right) - f_Y\left(x\right)\right)^2}{f_Y\left(x\right)}\mathrm{d}x, \tag{7.7e}
\end{align}$$

as it appears in (2.67).

7.1.3 Derivation of (2.69)

Consider that $f_X\left(x\right)$ and $f_Y\left(y\right) = f_X\left(x\right) + \epsilon\Delta\left(x\right)$ for some small ϵ. The function $\Delta\left(x\right)$ is also defined on the set $\mathcal{X}$ and has null area. Using the Maclaurin series of $\ln\left(1+a\right)$ up to the second-order such that $\ln\left(1+a\right) \cong -a + a^2/2 + O\left(a^3\right)$, we can write the KL divergence (2.60) as follows:

$$\begin{align}
D\left(f_X \parallel f_X + \varepsilon\Delta\right) &= \mathbb{E}_{f_X}\left\{\ln\left(\frac{f_X\left(x\right)}{f_X\left(x\right) + \varepsilon\Delta\left(x\right)}\right)\right\} \tag{7.8a}\\
&= \mathbb{E}_{f_X}\left\{-\ln\left(1 + \frac{\varepsilon\Delta\left(x\right)}{f_X\left(x\right)}\right)\right\} \tag{7.8b}
\end{align}$$

$$= \mathbb{E}_{f_X}\left\{-\frac{\varepsilon\Delta(x)}{f_X(x)} + \frac{1}{2}\left(\frac{\varepsilon\Delta(x)}{f_X(x)}\right)^2 + O\left(\varepsilon^3\right)\right\} \tag{7.8c}$$

$$= -\varepsilon\mathbb{E}_{f_X}\left\{\frac{\Delta(x)}{f_X(x)}\right\} + \frac{\varepsilon^2}{2}\mathbb{E}_{f_X}\left\{\left(\frac{\Delta(x)}{f_X(x)}\right)^2\right\} + O\left(\varepsilon^3\right). \tag{7.8d}$$

Given that the first term becomes zero due to the null area of $\Delta(x)$, then we have

$$D\left(f_X \parallel f_X + \varepsilon\Delta\right) \cong \frac{\varepsilon^2}{2}\mathbb{E}_{f_X}\left\{\left(\frac{\Delta(x)}{f_X(x)}\right)^2\right\} + O\left(\varepsilon^3\right). \tag{7.9}$$

Let us now proceed with the same approach but with the χ^2-divergence. Consider now the Maclaurin series expansion of $(1-a)^{-1}$ up to the second order, i.e. $1 + a + O\left(a^2\right)$. Then, we can write (2.67) as

$$D_{\chi^2}\left(f_X \parallel f_X + \varepsilon\Delta\right) = \mathbb{E}_{f_X}\left\{\frac{\left(f_X(x) - \left(f_X(x) + \varepsilon\Delta(x)\right)\right)^2}{f_X(x)\left(f_X(x) + \varepsilon\Delta(x)\right)}\right\} \tag{7.10a}$$

$$= \mathbb{E}_{f_X}\left\{\frac{\varepsilon^2\Delta^2(x)}{f_X(x)\left(f_X(x) + \varepsilon\Delta(x)\right)}\right\} \tag{7.10b}$$

$$= \mathbb{E}_{f_X}\left\{\left(\frac{\varepsilon\Delta(x)}{f_X(x)}\right)^2\left(1 + \frac{\varepsilon\Delta(x)}{f_X(x)}\right)^{-1}\right\} \tag{7.10c}$$

$$= \mathbb{E}_{f_X}\left\{\left(\frac{\varepsilon\Delta(x)}{f_X(x)}\right)^2\left(1 - \left(-\frac{\varepsilon\Delta(x)}{f_X(x)}\right)\right)^2\right\} \tag{7.10d}$$

$$= \mathbb{E}_{f_X}\left\{\left(\frac{\varepsilon\Delta(x)}{f_X(x)}\right)^2\left(1 - \frac{\varepsilon\Delta(x)}{f_X(x)} + O\left(\varepsilon^2\right)\right)\right\} \tag{7.10e}$$

$$= \varepsilon^2\mathbb{E}_{f_X}\left\{\left(\frac{\varepsilon\Delta(x)}{f_X(x)}\right)^2\right\} + O\left(\varepsilon^3\right). \tag{7.10f}$$

By joining the results of (7.9) and (7.10), we can state the following result:

$$D\left(f_X || f_X + \epsilon\Delta\right) = \frac{1}{2}D_{\chi^2}\left(f_X || f_X + \epsilon\Delta\right) + O(\epsilon^3), \tag{7.11}$$

as it is expressed in (2.69).

7.2 Appendices of Chapter 3

7.2.1 Integral of the product of two Gaussian functions

Let us define the general Gaussian functions

$$g_{\sigma_1}\left(x - \mu_1\right) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{\left(x + \mu_1\right)^2}{2\sigma_1^2}\right), \tag{7.12}$$

$$g_{\sigma_2}\left(x - \mu_2\right) = \frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{\left(x + \mu_2\right)^2}{2\sigma_2^2}\right). \tag{7.13}$$

We want to solve

$$I = \int_{-\infty}^{\infty} g_{\sigma_1}\left(x + \mu_1\right) g_{\sigma_2}\left(x + \mu_2\right) \mathrm{d}x = \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty}\exp\left(-\frac{\left(x + \mu_1\right)^2}{2\sigma_1^2} - \frac{\left(x + \mu_2\right)^2}{2\sigma_2^2}\right)\mathrm{d}x \tag{7.14a}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty}\exp\left(-\frac{\sigma_2^2\left(x+\mu_1\right)^2+\sigma_1^2\left(x+\mu_2\right)^2}{2\sigma_1^2\sigma_2^2}\right)\mathrm{d}x. \tag{7.14b}$$

Let us take the denominator of the argument of the exponential and complete the square:

$$-a\left(x-b\right)^2+c=-\sigma_2^2\left(x+\mu_1\right)^2-\sigma_1^2\left(x+\mu_2\right)^2, \tag{7.15}$$

$$a=\left(\sigma_1^2+\sigma_2^2\right),\qquad b=-\frac{\sigma_1^2\mu_2+\sigma_2^2\mu_1}{\sigma_1^2+\sigma_2^2},\qquad c=\frac{\sigma_1^2\sigma_2^2\left(\mu_2-\mu_2\right)^2}{\sigma_1^2+\sigma_2^2}. \tag{7.16}$$

We then have

$$I=\frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty}\exp\left(-\frac{\left(\sigma_1^2+\sigma_2^2\right)}{2\sigma_1^2\sigma_2^2}\left(x-b\right)^2-\frac{\left(\mu_2-\mu_1\right)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}\right)\mathrm{d}x \tag{7.17a}$$

$$=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{\left(\mu_2-\mu_1\right)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{\left(\sigma_1^2+\sigma_2^2\right)}{2\sigma_1^2\sigma_2^2}\left(x-b\right)^2\right)\mathrm{d}x. \tag{7.17b}$$

The remaining Gaussian integral is now solvable with $\int_{-\infty}^{\infty}\exp\left(-a\left(x-b\right)^2\right)\mathrm{d}x=\sqrt{\pi/a}$, which results in

$$I=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{\left(\mu_2-\mu_1\right)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}\right)\sqrt{\pi\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}=\frac{1}{\sqrt{2\pi\left(\sigma_1^2+\sigma_2^2\right)}}\exp\left(-\frac{\left(\mu_2-\mu_1\right)^2}{2\left(\sigma_1^2+\sigma_2^2\right)}\right) \tag{7.18a}$$

$$=g_\sigma\left(\mu_2-\mu_1\right), \tag{7.18b}$$

with $\sigma=\sqrt{\left(\sigma_1^2+\sigma_2^2\right)}$. In order to obtain (3.73), we juts need to fix $\sigma_1=\sigma_2=h$, $\mu_1=x_i$ and $\mu_2=x_j$.

7.3 Appendices of Chapter 4

7.3.1 Proof of Proposition 4.1

From (4.10) we have:

$$\left\|\frac{1}{L}\mathbf{X}\mathbf{X}^H\right\|_{\mathrm{F}}^2=\operatorname{tr}\left(\frac{1}{L^2}\mathbf{X}\mathbf{X}^H\mathbf{X}\mathbf{X}^H\right)=\operatorname{tr}\left(\frac{1}{L^2}\mathbf{F}^H\mathbf{F}\mathbf{D}\mathbf{D}^T\mathbf{F}^H\mathbf{F}\mathbf{D}\mathbf{D}^T\right) \tag{7.19a}$$

$$=\operatorname{tr}\left(\frac{1}{L^2}\mathbf{F}^H\mathbf{F}\hat{\mathbf{P}}\mathbf{F}^H\mathbf{F}\hat{\mathbf{P}}\right), \tag{7.19b}$$

where $\hat{\mathbf{P}}=\mathbf{D}\mathbf{D}^T/L$ from (4.6). We define the projection matrix $\mathbf{A}=\mathbf{F}^H\mathbf{F}\in\mathbb{C}^{N\times N}$, which is an idempotent matrix whose singular values $\sigma_n\left(\mathbf{A}\right)$ are composed by $\operatorname{rank}\left(\mathbf{A}\right)=N'$ ones, and $N-N'$ zeros. Then we have

$$\left\|\frac{1}{L}\mathbf{X}\mathbf{X}^H\right\|_{\mathrm{F}}^2=\operatorname{tr}\left(\frac{1}{L^2}\mathbf{A}\mathbf{D}\mathbf{D}^T\mathbf{A}\mathbf{D}\mathbf{D}^T\right)=\operatorname{tr}\left(\mathbf{A}\mathbf{P}\mathbf{A}\mathbf{P}\right)=\operatorname{tr}\left(\mathbf{B}\mathbf{B}\right), \tag{7.20}$$

where $\mathbf{B}=\mathbf{A}\mathbf{P}$ is a square, but not necessarily Hermitian, matrix.

From the Cauchy-Schwarz inequality we can gather that

$$\operatorname{tr}\left(\mathbf{B}\mathbf{B}\right)\leq\left\|\mathbf{B}^H\right\|_{\mathrm{F}}\left\|\mathbf{B}\right\|_{\mathrm{F}}=\left\|\mathbf{B}\right\|_{\mathrm{F}}^2=\operatorname{tr}\left(\mathbf{B}^H\mathbf{B}\right), \tag{7.21}$$

where the equality is held for Hermitian matrices. Therefore, from (7.20) we have

$$\left\|\frac{1}{L}\mathbf{X}\mathbf{X}^H\right\|_{\mathrm{F}}^2 \leq \mathrm{tr}\left(\mathbf{B}^H\mathbf{B}\right) = \mathrm{tr}\left(\mathbf{P}\mathbf{A}^H\mathbf{A}\mathbf{P}\right) = \mathrm{tr}\left(\mathbf{P}\mathbf{P}\mathbf{A}\right) = \mathrm{tr}\left(\mathbf{P}^2\mathbf{A}\right), \tag{7.22}$$

where $\mathbf{P}^2 = \mathbf{P}\mathbf{P}$ denotes a diagonal matrix whose elements are squared. The Von Neumann trace inequality, provided that the involved matrices are positive semi-definite, states the following:

$$\mathrm{tr}\left(\mathbf{P}^2\mathbf{A}\right) \leq \sum_{n=1}^{N}\sigma_n\left(\mathbf{P}^2\right)\sigma_n\left(\mathbf{A}\right) = \sum_{n=1}^{N}p_n^2\sigma_n\left(\mathbf{A}\right) \leq \sum_{n=1}^{N}p_n^2, \tag{7.23}$$

given that $\sigma_n\left(\mathbf{A}\right) \leq 1$. By gathering all the previous inequalities we can finally express

$$\left\|\frac{1}{L}\mathbf{X}\mathbf{X}^H\right\|_{\mathrm{F}}^2 \leq \mathrm{tr}\left(\mathbf{P}^2\mathbf{F}^H\mathbf{F}\right) \leq \sum_{n=1}^{N}\sigma_n\left(\mathbf{P}^2\right)\sigma_n\left(\mathbf{A}\right) \leq \sum_{n=1}^{N}p_n^2. \tag{7.24}$$

Clearly, for $N' = N$, then $\mathbf{F}$ is a unitary matrix with $\mathbf{F}^H\mathbf{F} = \mathbf{A} = \mathbf{I}_N$, and all the equalities are met. Otherwise, for $N' < N$, the inequalities are strict. In the case of the collision entropy

$$\hat{H}_2\left(X\right) = -\ln\left\|\frac{1}{L}\mathbf{X}\mathbf{X}^H\right\|_{\mathrm{F}}^2 \tag{7.25}$$

the logarithm is a monotonic function that preserves the inequalities, while the reverse sign turns around the inequalities, finally proving (4.11) with

$$\hat{H}_2\left(X\right) \geq -\ln\sum_{n=1}^{N}p_n^2 = \hat{\hat{H}}_2\left(X\right). \tag{7.26}$$

7.3.2 Proof of Lemma 4.2

From (4.32) we have

$$\|\hat{\mathbf{C}}\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\hat{\mathbf{R}}_x^{-1/2}\hat{\mathbf{C}}_{xy}\hat{\mathbf{R}}_y^{-1/2}\hat{\mathbf{R}}_y^{-H/2}\hat{\mathbf{C}}_{xy}^H\hat{\mathbf{R}}_x^{-H/2}\right) = \mathrm{tr}\left(\hat{\mathbf{C}}_{xy}\hat{\mathbf{R}}_y^{-1}\hat{\mathbf{C}}_{xy}^H\hat{\mathbf{R}}_x^{-1}\right). \tag{7.27}$$

The correlation and covariance matrices are as follows:

$$\hat{\mathbf{R}}_x = \frac{1}{L}\mathbf{X}\mathbf{X}^H = \mathbf{F}\left[\hat{\mathbf{p}}\right]\mathbf{F}^H, \tag{7.28a}$$

$$\hat{\mathbf{R}}_y = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H = \mathbf{G}\left[\hat{\mathbf{q}}\right]\mathbf{G}^H, \tag{7.28b}$$

$$\hat{\mathbf{C}}_{xy} = \frac{1}{L}\mathbf{X}\mathbf{P}_{\mathbf{1}}^{\perp}\mathbf{Y}^H = \mathbf{F}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\mathbf{G}^H. \tag{7.28c}$$

Then, we have

$$\|\hat{\mathbf{C}}\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\mathbf{F}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\mathbf{G}^H\left(\mathbf{G}\left[\hat{\mathbf{q}}\right]\mathbf{G}^H\right)^{-1}\mathbf{G}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)^T\mathbf{F}^H\left(\mathbf{F}\left[\hat{\mathbf{p}}\right]\mathbf{F}^H\right)^{-1}\right). \tag{7.29}$$

Given that the mapping matrices $\mathbf{F}$ and $\mathbf{G}$ are nonsingular, we can directly express

$$\|\hat{\mathbf{C}}\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\mathbf{F}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\left[\hat{\mathbf{q}}\right]^{-1}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)^T\left[\hat{\mathbf{p}}\right]^{-1}\mathbf{F}^{-1}\right), \tag{7.30}$$

which, from the circularity of the trace, it directly follows that

$$\|\hat{\mathbf{C}}\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\mathbf{F}^{-1}\mathbf{F}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\left[\hat{\mathbf{q}}\right]^{-1}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)^T\left[\hat{\mathbf{p}}\right]^{-1}\right), \tag{7.31a}$$

$$= \mathrm{tr}\left(\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\left[\hat{\mathbf{q}}\right]^{-1}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)^T\left[\hat{\mathbf{p}}\right]^{-1}\right). \tag{7.31b}$$

Therefore, it is the same estimator as in (4.30), as we wanted to show.

7.3.3 Proof of Theorem 4.1

Consider the following properties, which will be used along the proof:

$$\hat{\mathbf{p}}^T \mathbf{1}_N = \hat{\mathbf{q}}^T \mathbf{1}_M = 1, \qquad [\hat{\mathbf{p}}]\, \mathbf{1}_N = \hat{\mathbf{p}}, \qquad [\hat{\mathbf{q}}]\, \mathbf{1}_M = \hat{\mathbf{q}}, \tag{7.32}$$

$$\hat{\mathbf{J}} \mathbf{1}_M = \hat{\mathbf{p}}, \qquad \mathbf{1}_N^T \hat{\mathbf{J}} = \hat{\mathbf{q}}^T. \tag{7.33}$$

Since $\mathbf{1}_N$ and $\mathbf{1}_M$ are the left and right singular vectors of matrix $\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T$ associated to its null singular value, then

$$\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right) \mathbf{1}_M = \mathbf{0}_N, \tag{7.34a}$$

$$\mathbf{1}_N^T \left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right) = \mathbf{0}_M^T. \tag{7.34b}$$

Consequently, from (7.31), we can write

$$\|\hat{\mathbf{C}}\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\left([\hat{\mathbf{q}}]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_M \mathbf{1}_M^T\right)\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)^T\left([\hat{\mathbf{p}}]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_N \mathbf{1}_N^T\right)\right) \quad \forall \beta \in \mathbb{R}. \tag{7.35}$$

From the Woodbury matrix identity, we have

$$\left([\hat{\mathbf{p}}]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_N \mathbf{1}_N^T\right)^{-1} = [\hat{\mathbf{p}}] - \frac{[\hat{\mathbf{p}}]\, \mathbf{1}_N \mathbf{1}_N^T [\hat{\mathbf{p}}]}{\beta + \mathbf{1}_N^T [\hat{\mathbf{p}}]\, \mathbf{1}_N} = [\hat{\mathbf{p}}] - \beta \hat{\mathbf{p}}\hat{\mathbf{p}}^T, \tag{7.36a}$$

$$\left([\hat{\mathbf{q}}]^{-1} + \frac{\beta}{1-\beta}\mathbf{1}_M \mathbf{1}_M^T\right)^{-1} = [\hat{\mathbf{q}}] - \frac{[\hat{\mathbf{q}}]\, \mathbf{1}_M \mathbf{1}_M^T [\hat{\mathbf{q}}]}{\beta + \mathbf{1}_M^T [\hat{\mathbf{q}}]\, \mathbf{1}_M} = [\hat{\mathbf{q}}] - \beta \hat{\mathbf{q}}\hat{\mathbf{q}}^T. \tag{7.36b}$$

In the asymptotic case these matrices become

$$\lim_{\beta \to 1} \left([\hat{\mathbf{p}}] - \beta \hat{\mathbf{p}}\hat{\mathbf{p}}^T\right) \mathbf{1}_N = \mathbf{0}_N, \tag{7.37a}$$

$$\lim_{\beta \to 1} \left([\hat{\mathbf{q}}] - \beta \hat{\mathbf{q}}\hat{\mathbf{q}}^T\right) \mathbf{1}_M = \mathbf{0}_M. \tag{7.37b}$$

As a result, both matrices in (7.36) share with matrix $\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T$ the same singular vectors associated to the null singular value. For $\beta \to 1$ these matrices also become sample covariance matrices, such that

$$\lim_{\beta \to 1} [\hat{\mathbf{p}}] - \beta \hat{\mathbf{p}}\hat{\mathbf{p}}^T = \hat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T, \tag{7.38a}$$

$$\lim_{\beta \to 1} [\hat{\mathbf{q}}] - \beta \hat{\mathbf{q}}\hat{\mathbf{q}}^T = \hat{\mathbf{Q}} - \hat{\mathbf{q}}\hat{\mathbf{q}}^T, \tag{7.38b}$$

which yield

$$\lim_{\beta \to 1} \|\hat{\mathbf{C}}\|_{\mathrm{F}}^2 = \mathrm{tr}\left(\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)\left(\hat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T\right)^{-1}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T\right)^T\left(\hat{\mathbf{Q}} - \hat{\mathbf{q}}\hat{\mathbf{q}}^T\right)^{-1}\right). \tag{7.39}$$

Therefore, the equality with the SMI in (4.38) can be achieved by computing autocovariance matrices instead of autocorrelation matrices, and using $N' = N - 1$ and $M' = M - 1$ for the limiting case of $\beta = 1$. Regarding the mapping matrices $\mathbf{F}$ and $\mathbf{G}$, one can follow a similar procedure to Appendix 7.3.2.

7.3.4 Taylor expansion of the cross-covariance in the feature space

Consider

$$\mathrm{Cov}\left\{\tilde{z}_{x'}\left(\omega_1\right),\tilde{z}_{y'}\left(\omega_2\right)\right\}=G\left(\omega_1\right)G\left(-\omega_2\right)\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right)-\varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right). \quad (7.40)$$

Consider the window functions $G\left(\omega_1\right)$ and $G\left(-\omega_2\right)$ to have a bandwidth ν that define a finite support of the CFs. For ν close to 0, we can consider delta-like CF, which yields the following second-order Taylor expansion of the CFs around the origin:

$$\begin{aligned}\mathrm{Cov}\left\{\tilde{z}_{x'}\left(\omega_1\right),\tilde{z}_{y'}\left(-\omega_2\right)\right\}&\approx\varphi_{X,Y}\left(0,0\right)-\varphi_X\left(0\right)\varphi_Y\left(0\right)\\&+\left[\begin{array}{cc}\omega_1&-\omega_2\end{array}\right]\nabla\varphi+\frac{1}{2}\left[\begin{array}{cc}\omega_1&-\omega_2\end{array}\right]\mathbf{H}_\varphi\left[\begin{array}{c}\omega_1\\-\omega_2\end{array}\right]\end{aligned} \quad (7.41)$$

The gradient $\nabla\varphi$ corresponds to

$$\nabla\varphi=\left[\begin{array}{c}\left.\frac{\partial}{\partial\omega_1}\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right)-\varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right)\right|_{\omega_1,\omega_2=0}\\\left.\frac{\partial}{\partial\omega_2}\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right)-\varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right)\right|_{\omega_1,\omega_2=0}\end{array}\right], \quad (7.42)$$

which, following (2.13), we have

$$\nabla\varphi=\left[\begin{array}{c}j\mathbb{E}_{f_X}\left\{X\right\}-\varphi_Y\left(-\omega_2\right)j\mathbb{E}_{f_X}\left\{X\right\}|_{\omega_2=0}\\j\mathbb{E}_{f_Y}\left\{Y\right\}-\varphi_X\left(\omega_1\right)j\mathbb{E}_{f_Y}\left\{Y\right\}|_{\omega_1=0}\end{array}\right]=\left[\begin{array}{c}0\\0\end{array}\right]. \quad (7.43)$$

Similarly, the elements of the Hessian matrix correspond to

$$\begin{aligned}\left[\mathbf{H}_\varphi\right]_{1,2}&=\left.\frac{\partial^2}{\partial\omega_1\omega_2}\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right)-\varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right)\right|_{\omega_1,\omega_2=0} && (44a)\\&=-\mathbb{E}_{f_{X,Y}}\left\{XY\right\}-\left(-j\mathbb{E}_{f_X}\left\{X\right\}j\mathbb{E}_{f_Y}\left\{Y\right\}\right) && (44b)\\&=-\mathbb{E}_{f_{X,Y}}\left\{XY\right\}-\mathbb{E}_{f_X}\left\{X\right\}\mathbb{E}_{f_Y}\left\{Y\right\} && (44c)\\\left[\mathbf{H}_\varphi\right]_{2,1}&=\left.\frac{\partial^2}{\partial\omega_2\omega_1}\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right)-\varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right)\right|_{\omega_1,\omega_2=0}=\left[\mathbf{H}_\varphi\right]_{1,2} && (45a)\\\left[\mathbf{H}_\varphi\right]_{1,1}&=\left.\frac{\partial^2}{\partial\omega_1^2}\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right)-\varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right)\right|_{\omega_1,\omega_2=0} && (46a)\\&=-\mathbb{E}_{f_X}\left\{X^2\right\}-\left.\left(-\mathbb{E}_{f_X}\left\{X^2\right\}\varphi_Y\left(-\omega_2\right)\right)\right|_{\omega_2=0}=0 && (46b)\\\left[\mathbf{H}_\varphi\right]_{2,2}&=\left.\frac{\partial^2}{\partial\omega_2^2}\left(\varphi_{X,Y}\left(\omega_1,-\omega_2\right)-\varphi_X\left(\omega_1\right)\varphi_Y\left(-\omega_2\right)\right)\right|_{\omega_1,\omega_2=0} && (47a)\\&=-\mathbb{E}_{f_Y}\left\{Y^2\right\}-\left.\left(-\mathbb{E}_{f_Y}\left\{Y^2\right\}\varphi_X\left(\omega_1\right)\right)\right|_{\omega_1=0}=0. && (47b)\end{aligned}$$

By gathering the previous expressions with (7.41), we then have

$$\mathrm{Cov}\left\{\tilde{z}_{x'}\left(\omega_1\right),\tilde{z}_{y'}\left(\omega_2\right)\right\}\approx\omega_1\omega_2\left(\mathbb{E}_{f_{X,Y}}\left\{XY\right\}+\mathbb{E}_{f_X}\left\{X\right\}\mathbb{E}_{f_Y}\left\{Y\right\}\right), \quad (7.48)$$

as it is written in (4.71).

7.3.5 Second-order Rényi entropy of a GMM

Consider the PDF

$$f_X\left(x\right)=\sum_{m=0}^{M-1}\frac{p_m}{\sqrt{2\pi\sigma_m^2}}\exp\left(-\frac{\left(x-\mu_m\right)^2}{2\sigma_m^2}\right). \quad (7.49)$$

The IP is then expressed as

$$V(X) = \int_{-\infty}^{\infty} f_X^2(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \left(\sum_{m=0}^{M-1} \frac{p_m}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2} \right) \right)^2 \mathrm{d}x \tag{7.50a}$$

$$= \int_{-\infty}^{\infty} \sum_{m'=0}^{M-1} \sum_{m=0}^{M-1} \frac{p_{m'} p_m}{2\pi\sqrt{\sigma_{m'}^2 \sigma_m^2}} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2} - \frac{(x-\mu_{m'})^2}{2\sigma_{m'}^2} \right) \mathrm{d}x. \tag{7.50b}$$

Due to the separability of the integrals, we just need to compute

$$A_{m,m'} = \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu_m)^2}{2\sigma_m^2} - \frac{(x-\mu_{m'})^2}{2\sigma_{m'}^2} \right) \mathrm{d}x \tag{7.51a}$$

$$= \int_{-\infty}^{\infty} \exp\left(\frac{-x^2\left(\sigma_{m'}^2 + \sigma_m^2\right) + 2x\left(\mu_m\sigma_{m'}^2 + \mu_{m'}\sigma_m^2\right) - \mu_m^2\sigma_{m'}^2 - \mu_{m'}^2\sigma_m^2}{2\sigma_{m'}^2\sigma_m^2} \right) \mathrm{d}x \tag{7.51b}$$

$$= \int_{-\infty}^{\infty} \exp\left(\frac{-\left(\sigma_{m'}^2 + \sigma_m^2\right)(x-a)^2 - \sigma_{m'}^2\sigma_m^2\left(\mu_m - \mu_{m'}\right)^2\left(\sigma_{m'}^2 + \sigma_m^2\right)^{-1}}{2\sigma_{m'}^2\sigma_m^2} \right) \mathrm{d}x \tag{7.51c}$$

$$= \exp\left(\frac{-\left(\mu_m - \mu_{m'}\right)^2}{2\left(\sigma_{m'}^2 + \sigma_m^2\right)} \right) \int_{-\infty}^{\infty} \exp\left(\frac{-\left(\sigma_{m'}^2 + \sigma_m^2\right)(x-a)^2}{2\sigma_{m'}^2\sigma_m^2} \right) \mathrm{d}x, \tag{7.51d}$$

where the last steps are obtained by completing the square, similar to (7.15), and

$$a = \frac{\left(\mu_m\sigma_{m'}^2 + \mu_{m'}\sigma_m^2\right)}{\sigma_{m'}^2 + \sigma_m^2}. \tag{7.52}$$

The remaining integral term can be solved by the Gaussian integral such that

$$\int_{-\infty}^{\infty} \exp\left(\frac{-\left(\sigma_{m'}^2 + \sigma_m^2\right)(x-a)^2}{2\sigma_{m'}^2\sigma_m^2} \right) \mathrm{d}x = \sqrt{\frac{2\pi\sigma_{m'}^2\sigma_m^2}{\left(\sigma_{m'}^2 + \sigma_m^2\right)}}. \tag{7.53}$$

By gathering (7.51) and (7.53) we have

$$A_{m,m'} = \sqrt{\frac{2\pi\sigma_{m'}^2\sigma_m^2}{\left(\sigma_{m'}^2 + \sigma_m^2\right)}} \exp\left(\frac{-\left(\mu_m - \mu_{m'}\right)^2}{2\left(\sigma_{m'}^2 + \sigma_m^2\right)} \right), \tag{7.54}$$

which leads up to

$$V(X) = \sum_{m'=0}^{M-1} \sum_{m=0}^{M-1} \frac{p_{m'} p_m}{2\pi\sqrt{\sigma_{m'}^2\sigma_m^2}} A_{m,m'} \tag{7.55a}$$

$$= \sum_{m'=0}^{M-1} \sum_{m=0}^{M-1} \frac{p_{m'} p_m}{2\pi\sqrt{\sigma_{m'}^2\sigma_m^2}} \sqrt{\frac{2\pi\sigma_{m'}^2\sigma_m^2}{\left(\sigma_{m'}^2 + \sigma_m^2\right)}} \exp\left(\frac{-\left(\mu_m - \mu_{m'}\right)^2}{2\left(\sigma_{m'}^2 + \sigma_m^2\right)} \right) \tag{7.55b}$$

$$= \sum_{m'=0}^{M-1} \sum_{m=0}^{M-1} \frac{p_{m'} p_m}{\sqrt{2\pi\left(\sigma_{m'}^2 + \sigma_m^2\right)}} \exp\left(\frac{-\left(\mu_m - \mu_{m'}\right)^2}{2\left(\sigma_{m'}^2 + \sigma_m^2\right)} \right), \tag{7.55c}$$

concluding the computation.

7.3.6 Proof of Lemma 4.3

From Theorem 2.5, the weak condition of Szegö's theorem establishes that a Hermitian-Toeplitz matrix behaves asymptotically as a circulant matrix if the Hermitian sequence t_n is square-integrable for $N \to \infty$. Consequently, following (2.22) and the diagonalization in (2.17), $\mathbf{T}$ is asymptotically diagonalizable by $\mathbf{W}\mathbf{T}\mathbf{W}^H$, whose diagonal is then composed by the discrete Fourier transform of t_n. In addition, the discrete Fourier transform of a Hermitian sequence g_n can be written as

$$\mathcal{F}\left\{g_n\right\} = \sum_{n=-(N-1)}^{N-1} g_n e^{-j2\pi f n} = g_0 + 2\mathrm{Re}\left\{\sum_{n=1}^{N-1} g_n e^{-j2\pi f n}\right\} = 2\mathrm{Re}\left\{\sum_{n=0}^{N-1} g_n e^{-j2\pi f n}\right\} - g_0, \tag{7.56}$$

where $g_n = t_n\left(1 - n/N\right)$ and $g_0 = 1$. We can then write

$$\mathbf{g} = \mathbf{t} \odot \frac{\mathbf{v}_{\mathrm{a}}}{N}, \tag{7.57}$$

where $[\mathbf{g}]_n = g_n$. Given that $[\mathbf{W}]_{n,n'} = \frac{1}{\sqrt{N}} e^{-j2\pi n n'/N}$, and sampling $f = n'/N$ for $n' = 0, ..., N-1$, then the column vector constructed from (7.56) is as follows:

$$2\mathrm{Re}\left(\sqrt{N}\mathbf{W}\left(\mathbf{t} \odot \frac{\mathbf{v}_{\mathrm{a}}}{N}\right)\right) - \mathbf{1}_N = \frac{2}{\sqrt{N}}\mathrm{Re}\left(\mathbf{W}\left(\mathbf{t} \odot \mathbf{v}_{\mathrm{a}}\right)\right) - \mathbf{1}_N, \tag{7.58}$$

where $\sqrt{N}$ is added to compensate the normalization factor from $\mathbf{W}$, as written (4.135).

7.3.7 Proof of Theorem 4.2

Given that the Frobenius norm is invariant in front of unitary transformations, we have

$$\hat{I}_{\mathrm{s}}\left(X;Y\right) = \left\|\hat{\mathbf{R}}_{x'}^{-1/2}\hat{\mathbf{C}}_{x'y'}\hat{\mathbf{R}}_{y'}^{-1/2}\right\|_{\mathrm{F}}^2 \tag{7.59a}$$

$$= \left\|\left(\mathbf{W}\hat{\mathbf{R}}_{x'}\mathbf{W}^H\right)^{-1/2}\mathbf{W}\hat{\mathbf{C}}_{x'y'}\mathbf{W}^H\left(\mathbf{W}\hat{\mathbf{R}}_{y'}\mathbf{W}^H\right)^{-1/2}\right\|_{\mathrm{F}}^2. \tag{7.59b}$$

Following Lemma 4.3, we can state that

$$\lim_{N\to\infty}\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\sum_{n'=0}^{N-1}\left|\left[\mathbf{W}\hat{\mathbf{R}}_{x'}\mathbf{W}^H\right]_{n,n'} - \left[\mathrm{diag}\left\{\frac{2}{\sqrt{N}}\mathrm{Re}\left(\mathbf{W}\left(\hat{\mathbf{p}}_a \odot \mathbf{v}_a\right)\right) - \mathbf{1}_N\right\}\right]_{n,n'}\right|^2} = 0 \tag{7.60a}$$

$$\lim_{N\to\infty}\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\sum_{n'=0}^{N-1}\left|\left[\mathbf{W}\hat{\mathbf{R}}_{y'}\mathbf{W}^H\right]_{n,n'} - \left[\mathrm{diag}\left\{\frac{2}{\sqrt{N}}\mathrm{Re}\left(\mathbf{W}\left(\hat{\mathbf{q}}_a \odot \mathbf{v}_a\right)\right) - \mathbf{1}_N\right\}\right]_{n,n'}\right|^2} = 0, \tag{7.60b}$$

given that $[\hat{\mathbf{q}}_a]_0 = [\hat{\mathbf{p}}_a]_0 = 1$. Let

$$\hat{\mathbf{A}} = \left(\mathbf{W}\hat{\mathbf{R}}_{x'}\mathbf{W}^H\right)^{-1/2}\mathbf{E}\left(\mathbf{W}\hat{\mathbf{R}}_{y'}\mathbf{W}^H\right)^{-1/2} \tag{7.61}$$

as in (4.139), and

$$\hat{\mathbf{B}} = \hat{\mathbf{P}}'^{-1/2}\mathbf{E}\hat{\mathbf{Q}}'^{-1/2} \tag{7.62}$$

as in (4.140), where $\mathbf{E} = \mathbf{W}\hat{\mathbf{C}}_{x'y'}\mathbf{W}^H$. Given (7.60), we can write

$$\lim_{N\to\infty}\frac{1}{\sqrt{N}}\left\|\left(\mathbf{W}\hat{\mathbf{R}}_{x'}\mathbf{W}^H\right)^{-1/2}\mathbf{E}\left(\mathbf{W}\hat{\mathbf{R}}_{y'}\mathbf{W}^H\right)^{-1/2} - \hat{\mathbf{P}}'^{-1/2}\mathbf{E}\hat{\mathbf{Q}}'^{-1/2}\right\|_{\mathrm{F}} = 0, \tag{7.63}$$

which simplifies to

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \left\| \hat{\mathbf{A}} - \hat{\mathbf{B}} \right\|_{\mathrm{F}} = 0 \tag{7.64}$$

as written in (4.141).

Finally, (4.142) is a direct consequence of (4.141) and Corollary 2.5.1, which at the same time, also follow from Theorem 2.5. Specifically, given

$$\hat{I}_{\mathrm{s}}(X;Y) = \|\hat{\mathbf{A}}\|_{\mathrm{F}}^2 = \sum_{n=0}^{N-1} \sigma_n^2\left(\hat{\mathbf{A}}\right), \quad \hat{I}_{\mathrm{as}}(X;Y) = \|\hat{\mathbf{B}}\|_{\mathrm{F}}^2 = \sum_{n=0}^{N-1} \sigma_n^2\left(\hat{\mathbf{B}}\right), \tag{7.65}$$

we can write

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \left(\sigma_n^2\left(\hat{\mathbf{A}}\right) - \sigma_n^2\left(\hat{\mathbf{B}}\right)\right) = \lim_{N\to\infty} \frac{1}{N} \left(\sum_{n=0}^{N-1} \sigma_n^2\left(\hat{\mathbf{A}}\right) - \sum_{n=0}^{N-1} \sigma_n^2\left(\hat{\mathbf{B}}\right)\right) = 0, \tag{7.66}$$

which yields

$$\lim_{N\to\infty} \hat{I}_{\mathrm{s}}(X;Y) = \lim_{N\to\infty} \hat{I}_{\mathrm{as}}(X;Y). \tag{7.67}$$

7.4 Appendices of Chapter 5

7.4.1 Information potential of a multivariate complex GMM

Consider the PDF

$$f_X(\mathbf{x}) = \sum_{m=0}^{M-1} \frac{p_m}{\pi^N |\boldsymbol{\Sigma}_m|} \exp\left(-(\mathbf{x} - \boldsymbol{\mu}_m)^H \boldsymbol{\Sigma}_m^{-1} (\mathbf{x} - \boldsymbol{\mu}_m)\right). \tag{7.68}$$

The IP is then expressed as

$$V(X) = \int_{\mathcal{X}} \left(\sum_{m=0}^{M-1} \frac{p_m}{\pi^N |\boldsymbol{\Sigma}_m|} \exp\left(-(\mathbf{x} - \boldsymbol{\mu}_m)^H \boldsymbol{\Sigma}_m^{-1} (\mathbf{x} - \boldsymbol{\mu}_m)\right)\right)^2 \mathrm{d}\mathbf{x} \tag{7.69a}$$

$$= \sum_{m=0}^{M-1} \sum_{m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^{2N} |\boldsymbol{\Sigma}_m| |\boldsymbol{\Sigma}_{m'}|} \int_{\mathcal{X}} \exp\left(-(\mathbf{x} - \boldsymbol{\mu}_m)^H \boldsymbol{\Sigma}_m^{-1} (\mathbf{x} - \boldsymbol{\mu}_m) - (\mathbf{x} - \boldsymbol{\mu}_{m'})^H \boldsymbol{\Sigma}_{m'}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{m'})\right) \mathrm{d}\mathbf{x} \tag{7.69b}$$

$$= \sum_{m,m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^{2N} |\boldsymbol{\Sigma}_m| |\boldsymbol{\Sigma}_{m'}|} \exp\left(-(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'})^H (\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'})^{-1} (\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'})\right) A_{m,m'}, \tag{7.69c}$$

where

$$A_{m,m'} = \int_{\mathcal{X}} \exp\left(-\left(\mathbf{x} - \boldsymbol{a}_{m,m'}\right)^H \left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_{m'}^{-1}\right) \left(\mathbf{x} - \boldsymbol{a}_{m,m'}\right)\right) \mathrm{d}\mathbf{x} \tag{7.70}$$

and

$$\boldsymbol{a}_{m,m'} = \boldsymbol{\Sigma}_{m'} (\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'})^{-1} \boldsymbol{\mu}_m + \boldsymbol{\Sigma}_m (\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'})^{-1} \boldsymbol{\mu}_{m'}. \tag{7.71}$$

The integral of $A_{m,m'}$ yields the following result for any $\mathbf{a}_{m,m'}$:

$$A_{m,m'} = \int_{\mathcal{X}} \exp\left(-\left(\mathbf{x} - \boldsymbol{a}_{m,m'}\right)^H \left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_{m'}^{-1}\right) \left(\mathbf{x} - \boldsymbol{a}_{m,m'}\right)\right) \mathrm{d}\mathbf{x} \tag{7.72a}$$

$$= \int_{\mathcal{X}} \frac{\pi^N \left|\left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_{m'}^{-1}\right)\right|^{-1}}{\pi^N \left|\left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_{m'}^{-1}\right)\right|^{-1}} \exp\Big(-\left(\mathbf{x} - \boldsymbol{a}_{m,m'}\right)^H \boldsymbol{\Gamma}^{-1} \left(\mathbf{x} - \boldsymbol{a}_{m,m'}\right)\Big)\, \mathrm{d}\mathbf{x} \tag{7.72b}$$

$$= \pi^N \left|\left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_{m'}^{-1}\right)\right|^{-1} = \frac{\pi^N}{\left|\left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_{m'}^{-1}\right)\right|}. \tag{7.72c}$$

By gathering (7.69) and (7.72) we finally have

$$V(X) = \sum_{m,m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^{2N} |\boldsymbol{\Sigma}_m| |\boldsymbol{\Sigma}_{m'}|} \frac{\pi^N}{\left|\left(\boldsymbol{\Sigma}_m^{-1} + \boldsymbol{\Sigma}_{m'}^{-1}\right)\right|}$$

$$\exp\Big(-\left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\right)^H \left(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'}\right)^{-1} \left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\right)\Big) \tag{7.73a}$$

$$\sum_{m,m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^N \left|\left(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'}\right)\right|} \exp\Big(-\left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\right)^H \left(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_{m'}\right)^{-1} \left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\right)\Big). \tag{7.73b}$$

7.4.2 Proof of Proposition 5.1

This proof actually follows closely the derivations from Appendix 7.4.1. However, here we particularize by letting Z to be a distribution that will latter encompass the difference between two samples.

The distribution of Z is

$$f_Z(\mathbf{z}) = \sum_{q=0}^{Q-1} \frac{1}{\pi^N |\mathbf{W}_q|} p_q \exp\Big(-\left(\mathbf{z} - \mathbf{a}_q\right)^H \mathbf{W}_q^{-1} \left(\mathbf{z} - \mathbf{a}_q\right)\Big). \tag{7.74}$$

We want to compute

$$\mathbb{E}_{f_Z}\left\{g_{\mathbf{V}}(\mathbf{z})\right\} = \int_{\mathcal{Z}} \sum_{q=0}^{Q-1} \frac{1}{\pi^N |\mathbf{V}|} \frac{1}{\pi^N |\mathbf{W}_q|} \tag{7.75a}$$

$$p_q \exp\Big(-\left(\mathbf{z} - \mathbf{a}_q\right)^H \mathbf{W}_q^{-1} \left(\mathbf{z} - \mathbf{a}_q\right)\Big) \exp\left(-\mathbf{z}^H \mathbf{V}^{-1} \mathbf{z}\right) \mathrm{d}\mathbf{z}. \tag{7.75b}$$

Due to the separability of the integrals we just need to solve the following integral:

$$A_q = \int_{\mathcal{Z}} p_q \exp\Big(-\left(\mathbf{z} - \mathbf{a}_q\right)^H \mathbf{W}_q^{-1} \left(\mathbf{z} - \mathbf{a}_q\right) - \mathbf{z}^H \mathbf{V}^{-1} \mathbf{z}\Big)\, \mathrm{d}\mathbf{z} \tag{7.76a}$$

$$= p_q \int_{\mathcal{Z}} \exp\Big(-\left(\mathbf{z} - \mathbf{b}_q\right)^H \left(\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right) \left(\mathbf{z} - \mathbf{b}_q\right) - \mathbf{a}_q^H \left(\mathbf{V} + \mathbf{W}_q\right)^{-1} \mathbf{a}_q\Big)\, \mathrm{d}\mathbf{z} \tag{7.76b}$$

$$= p_q \exp\Big(-\mathbf{a}_q^H \left(\mathbf{V} + \mathbf{W}_q\right)^{-1} \mathbf{a}_q\Big) \int_{\mathcal{Z}} \exp\Big(-\left(\mathbf{z} - \mathbf{b}_q\right)^H \left(\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right) \left(\mathbf{z} - \mathbf{b}_q\right) dz\Big)\, \mathbf{z}, \tag{7.76c}$$

where $\mathbf{b}_q = \Big(\mathbf{V}\left(\mathbf{V} + \mathbf{W}_q\right)^{-1}\Big) \mathbf{a}_q$. The integral yields the following result for any $\mathbf{b}_q$:

$$B_q = \int_{\mathcal{Z}} \exp\Big(-\left(\mathbf{z} - \mathbf{b}_q\right)^H \left(\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right) \left(\mathbf{z} - \mathbf{b}_q\right)\Big)\, \mathrm{d}\mathbf{z} \tag{7.77a}$$

$$= \frac{\pi^N \left|\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right|^{-1}}{\pi^N \left|\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right|^{-1}} \int_{\mathcal{Z}} \exp\Big(-\left(\mathbf{z} - \mathbf{b}_q\right)^H \left(\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right) \left(\mathbf{z} - \mathbf{b}_q\right)\Big)\, \mathrm{d}\mathbf{z} \tag{7.77b}$$

$$= \pi^N \left|\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right|^{-1} = \frac{\pi^N}{\left|\mathbf{V}^{-1} + \mathbf{W}_q^{-1}\right|}. \tag{7.77c}$$

By grouping the constants that we left outside the integrals we lastly have

$$\begin{aligned}
\mathbb{E}_{f_Z}\left\{g_{\mathbf{V}}\left(\mathbf{z}\right)\right\} &= \sum_{q=0}^{Q-1} \frac{1}{\pi^N\left|\mathbf{V}\right|} \frac{1}{\pi^N\left|\mathbf{W}_q\right|} A_q && (7.78a)\\
&= \sum_{q=0}^{Q-1} \frac{1}{\pi^N\left|\mathbf{V}\right|} \frac{1}{\pi^N\left|\mathbf{W}_q\right|} p_q \exp\left(-\mathbf{a}_q^H\left(\mathbf{V}+\mathbf{W}_q\right)^{-1}\mathbf{a}_q\right) B_q && (7.78b)\\
&= \sum_{q=0}^{Q-1} \frac{1}{\pi^N\left|\mathbf{V}\right|} \frac{1}{\pi^N\left|\mathbf{W}_q\right|} \frac{\pi^N}{\left|\mathbf{V}^{-1}+\mathbf{W}_q^{-1}\right|} p_q \exp\left(-\mathbf{a}_q^H\left(\mathbf{V}+\mathbf{W}_q\right)^{-1}\mathbf{a}_q\right) && (7.78c)\\
&= \sum_{q=0}^{Q-1} \frac{1}{\pi^N\left|\mathbf{V}+\mathbf{W}_q\right|} p_q \exp\left(-\mathbf{a}_q^H\left(\mathbf{V}+\mathbf{W}_q\right)^{-1}\mathbf{a}_q\right) && (7.78d)
\end{aligned}$$

as it appears in (5.6).

7.4.3 Proof of Proposition 5.2

The expression to be determined is

$$\operatorname{Cov}\left\{\hat{U}_{\mathbf{V}_1}\left(X\right)\hat{U}_{\mathbf{V}_2}\left(X\right)\right\} = \frac{1}{\left(L\left(L-1\right)/2\right)^2} \sum_{0\leq i\leq L-1} \sum_{i<j\leq J_i} d\left(i,i',j,j'\right) - c\left(i,i',j,j'\right), \qquad (7.79)$$

where

$$\begin{aligned}
d\left(i,i',j,j'\right) &= \mathbb{E}_{f_X}\left\{g_{\mathbf{V_1}}\left(\mathbf{x}\left(i\right)-\mathbf{x}\left(i'\right)\right) g_{\mathbf{V}_2}\left(\mathbf{x}\left(j\right)-\mathbf{x}\left(j'\right)\right)\right\}, && (7.80)\\
c\left(i,i',j,j'\right) &= \mathbb{E}_{f_X}\left\{g_{\mathbf{V_1}}\left(\mathbf{x}\left(i\right)-\mathbf{x}\left(i'\right)\right)\right\} \mathbb{E}_{f_X}\left\{g_{\mathbf{V}_2}\left(\mathbf{x}\left(j\right)-\mathbf{x}\left(j'\right)\right)\right\}. && (7.81)
\end{aligned}$$

Under this setting, we can identify three different cases:

1. $L\left(L-1\right)\left(L-2\right)\left(L-3\right)/4$ terms such that $i\neq j$ and $i'\neq j'$ that do not contribute to the variance analysis due to the i.i.d. assumption. In these cases, we directly have $d\left(i,i',j,j'\right) - c\left(i,i',j,j'\right) = 0$.

2. $L\left(L-1\right)\left(L-2\right)/2$ terms with $i=j$ and $i'\neq j'$, and $L\left(L-1\right)\left(L-2\right)/2$ terms with $i\neq j$ and $i'=j'$. Both of these cases yield the same result and we need to solve

$$d\left(i,i',j,i'\right) = a = \mathbb{E}_{f_X}\left\{g_{\mathbf{V_1}}\left(\mathbf{x}\left(i\right)-\mathbf{x}\left(i'\right)\right) g_{\mathbf{V}_2}\left(\mathbf{x}\left(j\right)-\mathbf{x}\left(i'\right)\right)\right\} \qquad (7.82)$$

and

$$c\left(i,i',j,i'\right) = c = \mathbb{E}_{f_X}\left\{g_{\mathbf{V_1}}\left(\mathbf{x}\left(i\right)-\mathbf{x}\left(i'\right)\right)\right\} \mathbb{E}_{f_X}\left\{g_{\mathbf{V}_2}\left(\mathbf{x}\left(j\right)-\mathbf{x}\left(i'\right)\right)\right\}. \qquad (7.83)$$

Clearly, c is known from Corollary 5.1.1, which results in

$$\begin{aligned}
c =& V\left(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}+\mathbf{V}_1/2\right) V\left(X \mid p_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}+\mathbf{V}_2/2\right) =\\
& \sum_{m,m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^N\left|2\boldsymbol{\Sigma}+\mathbf{V}_1\right|} \exp\left(-\left(\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\right)^H\left(2\boldsymbol{\Sigma}+\mathbf{V}_1\right)^{-1}\left(\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\right)\right) \qquad (7.84a)\\
& \sum_{m,m'=0}^{M-1} \frac{p_m p_{m'}}{\pi^N\left|2\boldsymbol{\Sigma}+\mathbf{V}_2\right|} \exp\left(-\left(\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\right)^H\left(2\boldsymbol{\Sigma}+\mathbf{V}_2\right)^{-1}\left(\boldsymbol{\mu}_m-\boldsymbol{\mu}_{m'}\right)\right) \qquad (7.84b)
\end{aligned}$$

On the other hand, a can be obtained through Lemma 5.1 with $\gamma = 0.5$, $R = R' = M^2$, $\boldsymbol{W} = 2\boldsymbol{\Sigma}$, $\mathbf{a}_r = \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}$,$\mathbf{a}_{r'} = \boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}$ and $n' = m'$, which resolves into

$$a = \frac{1}{\pi^{2N} \left\| \begin{bmatrix} 2\boldsymbol{\Sigma} + \mathbf{V}_1 & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & 2\boldsymbol{\Sigma} + \mathbf{V}_2 \end{bmatrix} \right\|} \sum_{m=0}^{M-1} \sum_{m'=0}^{M-1} \sum_{n=0}^{M-1} \sum_{n'=0}^{M-1}$$
$$p_m p_{m'} p_n p_{n'} \exp\left(- \begin{bmatrix} \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'} \\ \boldsymbol{\mu}_n - \boldsymbol{\mu}_{m'} \end{bmatrix}^H \begin{bmatrix} 2\boldsymbol{\Sigma} + \mathbf{V}_1 & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & 2\boldsymbol{\Sigma} + \mathbf{V}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'} \\ \boldsymbol{\mu}_n - \boldsymbol{\mu}_{m'} \end{bmatrix} \right). \tag{7.85}$$

Since there are M duplicate terms given $n' = m'$, a is then simplified as it appears in (5.13).

3. Lastly, we have $L\left(L-1\right)/2$ terms with $i = j$ and $i' = j'$. For these, we have $c\left(i, i', i, i'\right) = c$, which is the same as in in (7.84), and

$$b = \frac{1}{\pi^{2N} \left\| \begin{bmatrix} 2\boldsymbol{\Sigma} + \mathbf{V}_1 & 2\boldsymbol{\Sigma} \\ 2\boldsymbol{\Sigma} & 2\boldsymbol{\Sigma} + \mathbf{V}_2 \end{bmatrix} \right\|} \sum_{m=0}^{M-1} \sum_{m'=0}^{M-1}$$
$$p_m p_n \exp\left(- \begin{bmatrix} \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'} \\ \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'} \end{bmatrix}^H \begin{bmatrix} 2\boldsymbol{\Sigma} + \mathbf{V}_1 & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & 2\boldsymbol{\Sigma} + \mathbf{V}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'} \\ \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'} \end{bmatrix} \right), \tag{7.86}$$

which is again obtained through Lemma 5.1 with $\gamma = 1$, $R = R' = M$, $\boldsymbol{W} = 2\boldsymbol{\Sigma}$, $\mathbf{a}_r = \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}$ and $\mathbf{a}_{r'} = \boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}$.

Finally, by gathering all previous terms we have

$$\operatorname{Cov}\left\{ \hat{U}_{\mathbf{V}_1}\left(X\right) \hat{U}_{\mathbf{V}_2}\left(X\right) \right\} = 2\frac{L\left(L-1\right)\left(L-2\right)/2}{\left(L\left(L-1\right)/2\right)^2}\left(a-c\right) + \frac{L\left(L-1\right)/2}{\left(L\left(L-1\right)/2\right)^2}\left(b-c\right) \tag{7.87a}$$

$$= \frac{4\left(L-2\right)}{L\left(L-1\right)}\left(a-c\right) + \frac{2}{L\left(L-1\right)}\left(b-c\right) \tag{7.87b}$$

$$= \frac{4\left(L-2\right)}{L\left(L-1\right)}a + \frac{2}{L\left(L-1\right)}b - \frac{4\left(L-2\right)+2}{L\left(L-1\right)}c, \tag{7.87c}$$

as it appears in (5.12).

7.4.4 Derivation of (5.49)

For real-valued data, the distribution is the following:

$$f_X\left(x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2} \right), \tag{7.88}$$

whose MIP estimator becomes

$$\hat{W}\left(X\right) = \frac{2}{L(L-1)} \sum_{0 \le i < j \le L-1} \kappa_v\left(x(i), x(j)\right), \tag{7.89}$$

where

$$\kappa_v\left(x(i), x(j)\right) = \exp\left(-\frac{\left(x(i) - x(j)\right)^2}{2v^2} \right). \tag{7.90}$$

By following similar steps to the ones in Corollary 5.1.1, the following expectation can be obtained:

$$\mathbb{E}_{f_X}\left\{\kappa_v\left(x(i), x(j)\right)\right\} = \sqrt{\frac{v^2}{2\sigma^2 + v^2}}. \tag{7.91}$$

By following Proposition 5.2 and (5.22) the variance of this estimator corresponds to

$$\text{Var}\left\{\hat{W}\left(X\right)\right\} = \frac{4\left(L-2\right)a + 2b - \left(4\left(L-2\right)+2\right)c}{L\left(L-1\right)}, \tag{7.92}$$

where

$$a = \frac{w}{\sqrt{\left(w+1\right)\left(w+3\right)}}, \qquad b = \frac{w}{\sqrt{\left(w+4\right)w}}, \qquad c = \frac{w}{w+2}. \tag{7.93}$$

Consequently, the difference between the complex and the real-valued cases is encountered in the values of a, b and c.

From (7.91), the variance estimator is now defined as

$$\hat{\sigma}^2 = \xi_v\left(\hat{W}\left(X\right)\right) = \frac{v^2}{2}\left(\frac{1}{\hat{W}^2\left(X\right)} - 1\right), \tag{7.94}$$

which yields the following approximate estimator variance by following similar steps than those in (5.34) and (5.36):

$$\widetilde{\text{Var}}\left\{\hat{\sigma}^2\right\} \approx \text{Var}\left\{\hat{W}\left(X\right)\right\}\frac{\left(w+2\right)^3}{w}. \tag{7.95}$$

Regarding the asymptotic performance, we have from (5.40) that the corresponding large data size behaviour of the relative variance for the real case is

$$\lim_{L\to\infty} L\widetilde{\text{Var}}\left\{\hat{\sigma}^2\right\} = 4\left(a-c\right)\frac{\left(w+2\right)^3}{w} = 4\left(w+2\right)^2\left(\frac{w+2}{\sqrt{\left(w+1\right)\left(w+3\right)}} - 1\right), \tag{7.96}$$

whose limits are

$$2 \leq \lim_{L\to\infty} L\widetilde{\text{Var}}\left\{\hat{\sigma}^2\right\} \leq 2.4752. \tag{7.97}$$

The inequality that needs to be fulfilled for the asymptotic case to hold is the same, which is

$$\left(b-c\right) < \frac{1}{\alpha}2\left(L-2\right)\left(a-c\right). \tag{7.98}$$

By isolating L, the minimum data size now becomes

$$L > \frac{\alpha}{2}\left(\frac{\left(w+2-\sqrt{w}\sqrt{w+2}\right)\sqrt{w+3}\sqrt{w+1}}{\left(w+2-\sqrt{w+1}\sqrt{w+3}\right)\sqrt{w+4}\sqrt{w}}\right) + 2, \tag{7.99}$$

which, by assuming a sufficiently low value of w it yields the following simplification:

$$\sqrt{w} > \frac{\alpha}{L}\left(\frac{\sqrt{3}}{2\left(2-\sqrt{3}\right)}\right), \tag{7.100}$$

as we wanted to show.

7.4.5 Expectation of the sample variance under the contaminated model

The sample variance estimator, under the U-statistics standpoint in Example 10.2, is written as follows:

$$\hat{\sigma}_{\mathrm{S}}^2 = \frac{2}{L(L-1)} \sum_{0 \leq i < j \leq L-1} \frac{1}{2} \left| x_\varepsilon\left(i\right) - x_\varepsilon\left(j\right) \right|^2 . \tag{7.101}$$

Given that every term within the summation is i.i.d., the expectation of the estimate is:

$$\mathbb{E}_{f_{X_\varepsilon}} \left\{ \hat{\sigma}_{\mathrm{S}}^2 \right\} = \frac{1}{2} \mathbb{E}_{f_{X_\varepsilon}} \left\{ \left| x_\varepsilon\left(i\right) - x_\varepsilon\left(j\right) \right|^2 \right\} . \tag{7.102}$$

From the model in (5.50), and given that the additive outlier process is independent from X, we then have the following:

$$\begin{aligned}
\mathbb{E}_{f_{X_\varepsilon}} \left\{ \hat{\sigma}_{\mathrm{S}}^2 \right\} &= \frac{1}{2} \left(\mathbb{E}_{f_X} \left\{ \left| x\left(i\right) - x\left(j\right) \right|^2 \right\} + \mathbb{E}_{p_{Y_\varepsilon}} \left\{ \left| z\left(i\right) y\left(i\right) - z\left(j\right) y\left(j\right) \right|^2 \right\} \right) &\text{(7.103a)} \\
&= \sigma^2 + \frac{1}{2} \left(\mathbb{E}_{p_{Y_\varepsilon}} \left\{ \left| z\left(i\right) y\left(i\right) \right|^2 \right\} - 2\mathrm{Real} \left\{ \mathbb{E}_{p_{Y_\varepsilon}} \left\{ z\left(i\right) z\left(j\right) y\left(i\right) y^*\left(j\right) \right\} \right\} \right) &\text{(7.103b)} \\
&= \sigma^2 + \mathbb{E}_{p_Y} \left\{ \left| y\left(i\right) \right|^2 \right\} \mathbb{E}_{p_Z} \left\{ \left| z\left(i\right) \right|^2 \right\} - \left| \mathbb{E}_{p_Y} \left\{ y\left(i\right) \right\} \right|^2 \left| \mathbb{E}_{p_Z} \left\{ z\left(i\right) \right\} \right|^2 &\text{(7.103c)} \\
&= \sigma^2 + \varepsilon \left(\sigma_y^2 + \mu_y^2 \right) - \varepsilon^2 \mu_y^2 = \sigma^2 + \varepsilon \left(\sigma_y^2 + \mu_y^2 \left(1 - \varepsilon\right) \right) , &\text{(7.103d)}
\end{aligned}$$

as it is written in (5.52).

Bibliography

[ABH18] J. M. Amigó, S. G. Balogh, and S. Hernández. “A brief review of generalized entropies”. In: *Entropy* 20.11 (2018), p. 813.

[Ach+16] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi. “Estimating Rényi entropy of discrete distributions”. In: *IEEE Transactions on Information Theory* 63.1 (2016), pp. 38–56.

[AH84] J. C. Aubuchon and T. P. Hettmansperger. “A note on the estimation of the integral of $f^2(x)$”. In: *Journal of Statistical Planning and Inference* 9.3 (1984), pp. 321–331.

[ÁLM10] M. Álvarez-Díaz, R. López-Valcarce, and C. Mosquera. “SNR estimation for multilevel constellations using higher-order moments”. In: *IEEE Transactions on Signal Processing* 58.3 (2010), pp. 1515–1526.

[AN17] N. Asendorf and R. R. Nadakuditi. “Improved Detection of Correlated Signals in Low-Rank-Plus-Noise Type Data Sets Using Informative Canonical Correlation Analysis (ICCA)”. In: *IEEE Transactions on Information Theory* 63.6 (2017), pp. 3451–3467.

[Ari77] S. Arimoto. “Information measures and capacity of order α for discrete memoryless channels”. In: *Topics in information theory* (1977).

[Aro50] N. Aronszajn. “Theory of reproducing kernels”. In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[AS64] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Vol. 55. US Government printing office, 1964.

[Bac22] F. Bach. “Information theory with kernel methods”. In: *arXiv preprint arXiv: 2202.08 545* (2022).

[Bae22] J. C. Baez. “Rényi entropy and free energy”. In: *Entropy* 24.5 (2022).

[Bao+19] Z. Bao, J. Hu, G. Pan, and W. Zhou. “Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case”. In: *The Annals of Statistics* 47.1 (2019), pp. 612–640.

[Bei+97] J. Beirlant, E. J. Dudewicz, L. Györfi, E. C. Van der Meulen, et al. “Nonparametric entropy estimation: An overview”. In: *International Journal of Mathematical and Statistical Sciences* 6.1 (1997), pp. 17–39.

[BGV92] B. E. Boser, I. M. Guyon, and V. N. Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[Bha46] A. Bhattacharyya. “On a Measure of Divergence between Two Multinomial Populations”. In: *Sankhy: The Indian Journal of Statistics (1933-1960)* 7.4 (1946), pp. 401–406.

[BHC07] J. Benesty, Y. Huang, and J. Chen. “Time delay estimation via minimum entropy”. In: *IEEE Signal Processing Letters* 14.3 (2007), pp. 157–160.

[BHN05] H. Benaroya, S. M. Han, and M. Nagurka. *Probability Models in Engineering and Science*. Vol. 192. CRC press, 2005.

[Bin12] N. H. Bingham. "Szegö's theorem and its probabilistic descendants". In: *Probability Surveys* 9 (2012), pp. 287–324.

[BJ02] F. R. Bach and M. I. Jordan. "Kernel independent component analysis". In: *Journal of machine learning research* 3.Jul (2002), pp. 1–48.

[BLG08] M. B. Blaschko, C. H. Lampert, and A. Gretton. "Semi-supervised laplacian regularization of kernel canonical correlation analysis". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2008, pp. 133–145.

[BLK97] M. Borga, T. Landelius, and H. Knutsson. *A Unified Approach to PCA, PLS, MLR and CCA*. Tech. rep. Linköping University, Department of Electrical Engineering, 1997.

[Blo76] G. Blom. "Some properties of incomplete U-statistics". In: *Biometrika* 63.3 (1976), pp. 573–580.

[BMA14] F. Bellili, A. Methenni, and S. Affes. "Closed-Form CRLBs for SNR Estimation From Turbo-Coded BPSK-, MSK-, and Square-QAM-Modulated Signals". In: *IEEE Transactions on Signal Processing* 62.15 (2014), pp. 4018–4033.

[BN06] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*. Vol. 4. Springer, 2006.

[Bor98] M. Borga. "Learning multidimensional signal processing". PhD thesis. Linköping University Electronic Press, 1998.

[BR78] M. Ben-Bassat and J. Raviv. "Rényi's entropy and the probability of error". In: *IEEE Transactions on Information Theory* 24.3 (1978), pp. 324–331.

[BT11] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

[BV00] J.-F. Bercher and C. Vignat. "Estimating the entropy of a signal with applications". In: *IEEE transactions on signal processing* 48.6 (2000), pp. 1687–1694.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Cam65] L. L. Campbell. "A coding theorem and Rényi's entropy". In: *Information and control* 8.4 (1965), pp. 423–429.

[CGS95] D. Cochran, H. Gish, and D. Sinno. "A geometric approach to multiple-channel signal detection". In: *IEEE Transactions on Signal Processing* 43.9 (1995), pp. 2049–2057.

[Che15] S. Chen. "Optimal bandwidth selection for kernel density functionals estimation". In: *Journal of Probability and Statistics* 2015 (2015).

[CK19] X. Chen and K. Kato. "Randomized incomplete U-statistics in high dimensions". In: *The Annals of Statistics* 47.6 (2019), pp. 3127–3156.

[CL54] H. Chernoff and E. L. Lehmann. "The use of maximum likelihood estimates in χ^2 tests for goodness of fit". In: *The Annals of Mathematical Statistics* 25.3 (1954), pp. 579–586.

[Col19] J.-F. Collet. "An Exact Expression for the Gap in the Data Processing Inequality for f-Divergences". In: *IEEE Transactions on Information Theory* 65.7 (2019), pp. 4387–4391.

[Cov65] T. M. Cover. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition". In: *IEEE transactions on electronic computers* EC-14.3 (1965), pp. 326–334.

[CR18] F. de Cabrera and J. Riba. “A novel formulation of Independence Detection based on the Sample Characteristic Function”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. 2018, pp. 2608–2612.

[CRV17] F. de Cabrera, J. Riba, and G. Vázquez. “Entropy-based covariance determinant estimation”. In: *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE. 2017, pp. 1–5.

[CS00] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge university press, 2000.

[Csi08] I. Csiszár. “Axiomatic characterizations of information measures”. In: *Entropy* 10.3 (2008), pp. 261–273.

[Csi95] I. Csiszár. “Generalized cutoff rates and Rényi’s information measures”. In: *IEEE Transactions on Information Theory* 41.1 (1995), pp. 26–34.

[Csö85] S. Csörg. “Testing for independence by the empirical characteristic function”. In: *Journal of Multivariate Analysis* 16.3 (1985), pp. 290–299.

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 2006.

[CW08] E. J. Candès and M. B. Wakin. “An introduction to compressive sampling”. In: *IEEE signal processing magazine* 25.2 (2008), pp. 21–30.

[Das00] A. DasGupta. “Best constants in Chebyshev inequalities with various applications”. In: *Metrika* 51.3 (2000), pp. 185–200.

[Das08] A. Das. “NDA SNR estimation: CRLBs and EM based estimators”. In: *TENCON 2008 - IEEE Region 10 Conference*. 2008.

[Dav79] P. Davis. *Circulant Matrices*. Monographs and textbooks in pure and applied mathematics. Wiley, 1979.

[DDR17] T. Dao, C. M. De Sa, and C. Ré. “Gaussian quadrature for kernel features”. In: *Advances in neural information processing systems* 30 (2017).

[Din+13] G. Ding, Q. Wu, Y.-D. Yao, J. Wang, and Y. Chen. “Kernel-based learning for statistical signal processing in cognitive radio networks: Theoretical foundations, example applications, and future directions”. In: *IEEE Signal Processing Magazine* 30.4 (2013), pp. 126–136.

[Don06] D. L. Donoho. “Compressed sensing”. In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[Don16] X. Dong. “The gravity dual of Rényi entropy”. In: *Nature communications* 7.1 (2016), pp. 1–6.

[dR19] F. de Cabrera and J. Riba. “Entropy-Based Non-Data-Aided SNR Estimation”. In: *53rd Asilomar Conference on Signals, Systems, and Computers*. 2019, pp. 731–735.

[DV00] G. A. Darbellay and I. Vajda. “Entropy expressions for multivariate continuous distributions”. In: *IEEE Transactions on Information Theory* 46.2 (2000), pp. 709–712.

[DV99] G. A. Darbellay and I. Vajda. “Estimation of the information by an adaptive partitioning of the observation space”. In: *IEEE Transactions on Information Theory* 45.4 (1999), pp. 1315–1321.

[DVB06] D. V. B. (DVB). Second Generation Framing Structure, Channel Coding and Modulation Systems for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications. Standard. ETSI: Sophia Antipolis, France: ETSI EN 300 307 V1.1.2, 2006.

[EH14] T. van Erven and P. Harremoës. “Rényi Divergence and Kullback-Leibler Divergence”. In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820.

[Eld+17] Y. C. Eldar, A. O. Hero III, L. Deng, J. Fessler, J. Kovacevic, H. V. Poor, and S. Young. “Challenges and open problems in signal processing: Panel discussion summary from ICASSP 2017 [panel and forum]”. In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 8–23.

[EP02] D. Erdogmus and J. C. Príncipe. “An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems”. In: *IEEE Transactions on Signal Processing* 50.7 (2002), pp. 1780–1786.

[Erd02] D. Erdogmus. “Information theoretic learning: Rényi’s entropy and its applications to adaptive system training”. PhD thesis. University of Florida, 2002.

[FB14] S. Fehr and S. Berens. “On the conditional Rényi entropy”. In: *IEEE Transactions on Information Theory* 60.11 (2014), pp. 6801–6810.

[FBG07] K. Fukumizu, F. R. Bach, and A. Gretton. “Statistical consistency of kernel canonical correlation analysis.” In: *Journal of Machine Learning Research* 8.2 (2007).

[Fig+12] C. Figuera, J. L. Rojo-Álvarez, M. Wilby, I. Mora-Jiménez, and A. J. Caamano. “Advanced support vector machines for 802.11 indoor location”. In: *Signal Processing* 92.9 (2012), pp. 2126–2136.

[Fig+14] C. Figuera, Ó. Barquero-Pérez, J. L. Rojo-Álvarez, M. Martínez-Ramón, A. Guerrero-Curieses, and A. J. Caamano. “Spectrally adapted Mercer kernels for support vector nonuniform interpolation”. In: *Signal Processing* 94 (2014), pp. 421–433.

[FM77] A. Feuerverger and R. A. Mureika. “The empirical characteristic function and its applications”. In: *The annals of Statistics* (1977), pp. 88–97.

[Fu+15] Y. Fu, J. Zhu, S. Wang, and Z. Xi. “Reduced complexity SNR estimation via Kolmogorov-Smirnov test”. In: *IEEE Communications Letters* 19.9 (2015), pp. 1568–1571.

[Fuk+07] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. “Kernel measures of conditional dependence”. In: *Advances in neural information processing systems* 20 (2007).

[Gap08] W. Gappmair. “Cramer-Rao Lower Bound for Non-Data-Aided SNR Estimation of Linear Modulation Schemes”. In: *IEEE Transactions on Communications* 56.5 (2008), pp. 689–693.

[GC87] H. Gish and D. Cochran. “Invariance of the magnitude-squared coherence estimate with respect to second-channel statistics”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35.12 (1987), pp. 1774–1776.

[GC88] H. Gish and D. Cochran. “Generalized coherence”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1988, pp. 2745–2748.

[GE03] I. Guyon and A. Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[GG10] A. Gretton and L. Györfi. “Consistent nonparametric tests of independence”. In: *The Journal of Machine Learning Research* 11 (2010), pp. 1391–1423.

[GKC19] B. Ghojogh, F. Karray, and M. Crowley. “Eigenvalue and generalized eigenvalue problems: Tutorial”. In: *arXiv preprint arXiv:1903.11240* (2019).

[GLM09] W. Gappmair, R. López-Valcarce, and C. Mosquera. “Cramér-Rao lower bound and EM algorithm for envelope-based SNR estimation of nonconstant modulus constellations”. In: *IEEE Transactions on Communications* 57.6 (2009), pp. 1622–1627.

[GLM10] W. Gappmair, R. López-Valcarce, and C. Mosquera. "Joint NDA estimation of carrier frequency/phase and SNR for linearly modulated signals". In: *IEEE Signal Processing Letters* 17.5 (2010), pp. 517–520.

[GLS16] D. Gómez-Casco, J. A. López-Salcedo, and G. Seco-Granados. "Generalized integration techniques for high-sensitivity GNSS receivers affected by oscillator phase noise". In: *IEEE Statistical Signal Processing Workshop (SSP)*. 2016.

[GM48] R. E. Greenwood and J. Miller. "Zeros of the Hermite polynomials and weights for Gauss' mechanical quadrature formula". In: *Bulletin of the American Mathematical Society* 54.8 (1948), pp. 765–769.

[GN18] G. Govaert and M. Nadif. "Mutual information, phi-squared and model-based co-clustering for contingency tables". In: *Advances in data analysis and classification* 12.3 (2018), pp. 455–488.

[Gol+19] Z. Goldfeld, K. Greenewald, J. Weed, and Y. Polyanskiy. "Optimality of the plug-in estimator for differential entropy estimation under Gaussian convolutions". In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2019, pp. 892–896.

[Gol+20] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy. "Convergence of smoothed empirical measures with applications to entropy estimation". In: *IEEE Transactions on Information Theory* 66.7 (2020), pp. 4368–4391.

[GPC06] K. Ghartey, A. Papandreou-Suppappola, and D. Cochran. "Time-varying techniques for multisensor signal detection". In: *IEEE transactions on signal processing* 54.9 (2006), pp. 3353–3362.

[Gra+06] R. M. Gray et al. "Toeplitz and Circulant Matrices: A Review". In: *Foundations and Trends in Communications and Information Theory* 2.3 (2006), pp. 155–239.

[Gra11] R. M. Gray. *Entropy and Information Theory*. Springer Science & Business Media, 2011.

[Gre+05a] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. "Measuring statistical dependence with Hilbert-Schmidt norms". In: *International conference on algorithmic learning theory*. Springer. 2005, pp. 63–77.

[Gre+05b] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. "Kernel Methods for Measuring Independence". In: *Journal of Machine Learning Research* 6 (2005), pp. 2075–2129.

[Gre+05c] A. Gretton et al. "Kernel constrained covariance for dependence measurement". In: *International Workshop on Artificial Intelligence and Statistics*. PMLR. 2005, pp. 112–119.

[Gre+07] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. "A kernel statistical test of independence". In: *Advances in neural information processing systems* 20 (2007).

[GS02] A. L. Gibbs and F. E. Su. "On Choosing and Bounding Probability Metrics". In: *International Statistical Review / Revue Internationale de Statistique* 70.3 (2002), pp. 419–435.

[GS58] U. Grenander and G. Szegö. *Toeplitz forms and their applications*. Univ of California Press, 1958.

[GSS14] A. Guntuboyina, S. Saha, and G. Schiebinger. "Sharp Inequalities for f-Divergences". In: *IEEE Transactions on Information Theory* 60.1 (2014), pp. 104–121.

[HA17] C. L. Haley and M. Anitescu. "Optimal bandwidth for multitaper spectrum estimation". In: *IEEE Signal Processing Letters* 24.11 (2017), pp. 1696–1700.

[Har28] R. V. Hartley. “Transmission of information”. In: *Bell System technical journal* 7.3 (1928), pp. 535–563.

[HH18] A. R. Heravi and G. A. Hodtani. “Where does minimum error entropy outperform minimum mean square error? A new and closer look”. In: *IEEE Access* 6 (2018), pp. 5856–5864.

[Hir35] H. O. Hirschfeld. “A Connection between Correlation and Contingency”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 31.4 (1935), pp. 520–524.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge university press, 2012.

[HKS15] M. Hafez, T. Khattab, and H. M. H. Shalaby. “Blind SNR Estimation of Gaussian-Distributed Signals in Nakagami Fading Channels”. In: *IEEE Transactions on Wireless Communications* 14.7 (2015), pp. 3509–3518.

[HLE05] A. Hegde, T. Lan, and D. Erdogmus. “Order Statistics Based Estimator for Rényi’s Entropy”. In: *2005 IEEE Workshop on Machine Learning for Signal Processing*. IEEE. 2005, pp. 335–339.

[HM93] P. Hall and S. C. Morton. “On the estimation of entropy”. In: *Annals of the Institute of Statistical Mathematics* 45.1 (1993), pp. 69–88.

[HO00] A. Hyvärinen and E. Oja. “Independent component analysis: algorithms and applications”. In: *Neural Networks* 13.4 (2000), pp. 411–430.

[Hoe48] W. Hoeffding. “A Class of Statistics with Asymptotically Normal Distribution”. In: *The Annals of Mathematical Statistics* 19.3 (1948), pp. 293–325.

[Hor04] Y. Horikawa. “Comparison of support vector machines with autocorrelation kernels for invariant texture classification”. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 1. IEEE. 2004, pp. 660–663.

[Hot36] H. Hotelling. “Relations Between Two Sets of Variates”. In: *Biometrika* 28.3/4 (1936), pp. 321–377.

[HSS04] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. “Canonical correlation analysis: An overview with application to learning methods”. In: *Neural computation* 16.12 (2004), pp. 2639–2664.

[HSS08] T. Hofmann, B. Schölkopf, and A. J. Smola. “Kernel methods in machine learning”. In: *The annals of statistics* 36.3 (2008), pp. 1171–1220.

[HSZ15] S.-L. Huang, C. Suh, and L. Zheng. “Euclidean Information Theory of Networks”. In: *IEEE Transactions on Information Theory* 61.12 (2015), pp. 6795–6814.

[Hu+13] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. “Learning theory approach to minimum error entropy criterion”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 377–397.

[Hua+14] L. Huang, Y. Xiao, H. C. So, and J. Fang. “Accurate performance analysis of Hadamard ratio test for robust spectrum sensing”. In: *IEEE Transactions on Wireless Communications* 14.2 (2014), pp. 750–758.

[Hua+19] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng. “On universal features for high-dimensional learning and inference”. In: *arXiv preprint arXiv:1911.09105* (2019).

[Hub64] P. J. Huber. “Robust Estimation of a Location Parameter”. In: *Ann. Math. Statist.* 35.4 (1964), pp. 73–101.

[HZ12] S.-L. Huang and L. Zheng. “Linear information coupling problems”. In: *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE. 2012, pp. 1029–1033.

[Jäc05] P. Jäckel. “A note on multivariate Gauss-Hermite quadrature”. In: *London: ABN-Amro. Re* (2005).

[Jan84] S. Janson. “The asymptotic distributions of incomplete U-statistics”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 66.4 (1984), pp. 495–505.

[Jan91] A. Janssen. “Optimality property of the Gaussian window spectrogram”. In: *IEEE transactions on signal processing* 39.1 (1991), pp. 202–204.

[JB95] D. L. Jones and R. G. Baraniuk. “An adaptive optimal-kernel time-frequency representation”. In: *IEEE Transactions on Signal Processing* 43.10 (1995), pp. 2361–2371.

[Jen09] R. Jenssen. “Kernel entropy component analysis”. In: *IEEE transactions on pattern analysis and machine intelligence* 32.5 (2009), pp. 847–860.

[JH13] U. Javed and S. A. Hassan. “SNR estimation for non-coherent M-FSK receivers in Rice fading environment”. In: *IEEE communications letters* 17.9 (2013), pp. 1786–1789.

[Joe89] H. Joe. “Estimation of entropy and other functionals of a multivariate density”. In: *Annals of the Institute of Statistical Mathematics* 41.4 (1989), pp. 683–697.

[Jol02] I. Jolliffe. *Principal Component Analysis*. New York, NY: Springer-Verlag, 2002.

[KAS14] N. Klausner, M. R. Azimi-Sadjadi, and L. L. Scharf. “Detection of spatially correlated time series from a network of sensor arrays”. In: *IEEE transactions on signal processing* 62.6 (2014), pp. 1396–1407.

[Kay88] S. M. Kay. *Modern Spectral Estimation: Theory and Application*. Prentice Hall, 1988.

[Kay93] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., 1993.

[Kay98] S. Kay. “Model-based probability density function estimation”. In: *IEEE Signal Processing Letters* 5.12 (1998), pp. 318–320.

[KG03] M. Kuss and T. Graepel. *The geometry of kernel canonical correlation analysis*. Tech. rep. Max Planck Institute for Biological Cybernetics, 2003.

[Kim18] Y.-S. Kim. “Low complexity estimation method of Rényi entropy for ergodic sources”. In: *Entropy* 20.9 (2018), p. 657.

[KL51] S. Kullback and R. A. Leibler. “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[KMT12] S. Kumar, M. Mohri, and A. Talwalkar. “Sampling methods for the Nyström method”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 981–1006.

[KN04] S. Kotz and S. Nadarajah. *Multivariate t-distributions and their Applications*. Cambridge University Press, 2004.

[KSG04] A. Kraskov, H. Stögbauer, and P. Grassberger. “Estimating mutual information”. In: *Physical review E* 69.6 (2004), p. 066138.

[KT88] J. T. Kent and D. E. Tyler. “Maximum likelihood estimation for the wrapped Cauchy distribution”. In: *Journal of Applied Statistics* 15.2 (1988), pp. 247–254.

[LCR20] C. A. López, F. de Cabrera, and J. Riba. “Estimation of Information in Parallel Gaussian Channels via Model Order Selection”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 5675–5679.

[LHS13] D. Lopez-Paz, P. Hennig, and B. Schölkopf. “The randomized dependence coefficient”. In: *Advances in neural information processing systems* 26 (2013).

[Lie90] E. H. Lieb. “Integral bounds for radar ambiguity functions and Wigner distributions”. In: *Journal of mathematical physics* 31.3 (1990), pp. 594–599.

[Lim+15] W. Lim, M. Kim, H. Park, and K. Jung. “Double Nyström method: An efficient and accurate Nyström scheme for large-scale data sets”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1367–1375.

[LLC09] E. Liitiäinen, A. Lendasse, and F. Corona. “On the statistical estimation of Rényi entropies”. In: *2009 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE. 2009, pp. 1–6.

[LM07] R. López-Valcarce and C. Mosquera. “Sixth-order statistics-based non-data-aided SNR estimation”. In: *IEEE Communications Letters* 11.4 (2007), pp. 351–353.

[LP19a] A. Lapidoth and C. Pfister. “Two Measures of Dependence”. In: *Entropy* 21.8 (2019).

[LP19b] K. Li and J. C. Príncipe. “No-trick (treat) kernel adaptive filtering using deterministic features”. In: *arXiv preprint arXiv:1912.04530* (2019).

[LP20] K. Li and J. C. Príncipe. “Fast estimation of information theoretic learning descriptors using explicit inner product spaces”. In: *arXiv preprint arXiv:2001.00265* (2020).

[LPS08] N. Leonenko, L. Pronzato, and V. Savani. “A class of Rényi information estimators for multidimensional densities”. In: *The Annals of Statistics* 36.5 (2008), pp. 2153–2182.

[Luk63] E. Lukacs. “Applications of Characteristic Functions in Statistics”. In: *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)* 25.2 (1963), pp. 175–188.

[Luk70] E. Lukacs. *Characteristic Functions*. Griffin, 1970.

[LV01] A. Leshem and A.-J. Van der Veen. “Multichannel detection of Gaussian signals with uncalibrated receivers”. In: *IEEE Signal Processing Letters* 8.4 (2001), pp. 120–122.

[LV06] F. Liese and I. Vajda. “On Divergences and Informations in Statistics and Information Theory”. In: *IEEE Transactions on Information Theory* 52.10 (2006), pp. 4394–4412.

[Mak19] A. Makur. “Information Contraction and Decomposition”. PhD thesis. Massachusetts Institute of Technology, 2019.

[Mar+19] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons, 2019.

[MDW16] M. Mohammadkarimi, O. A. Dobre, and M. Z. Win. “Non-Data-Aided SNR Estimation for Multiple Antenna Systems”. In: *2016 IEEE Global Communications Conference (GLOBECOM)*. 2016, pp. 1–5.

[ME94] R. Matzner and F. Englberger. “An SNR estimation algorithm using fourth-order moments”. In: *Proceedings of 1994 IEEE International Symposium on Information Theory*. 1994.

[Mer09] J. Mercer. “Functions of positive and negative type, and their connection the theory of integral equations”. In: *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character* 209.441-458 (1909), pp. 415–446.

[MRL95] Y.-I. Moon, B. Rajagopalan, and U. Lall. “Estimation of mutual information using kernel density estimators”. In: *Physical Review E* 52.3 (1995), p. 2318.

[Mua+17] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. “Kernel mean embedding of distributions: A review and beyond”. In: *Foundations and Trendső in Machine Learning* 10.1-2 (2017), pp. 1–141.

[MXZ06] C. A. Micchelli, Y. Xu, and H. Zhang. “Universal Kernels.” In: *Journal of Machine Learning Research* 7.12 (2006).

[NK11] G. R. Naik and D. K. Kumar. “An overview of independent component analysis and its applications”. In: *Informatica* 35.1 (2011).

[Nos+21] M. Noshad, J. Choi, Y. Sun, A. Hero, and I. D. Dinov. “A data value metric for quantifying information content and utility”. In: *Journal of big Data* 8.1 (2021), pp. 1–23.

[Nut81] A. Nuttall. “Invariance of distribution of coherence estimate to second-channel statistics”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.1 (1981), pp. 120–122.

[NW06] J. Nocedal and S. Wright. *Numerical Optimization*. Springer-Verlag New York, 2006.

[NZ03] S. Nadarajah and K. Zografos. “Formulas for Rényi information and related measures for univariate distributions”. In: *Information Sciences* 155.1-2 (2003), pp. 119–138.

[OUE08] U. Ozertem, I. Uysal, and D. Erdogmus. “Continuously Differentiable Sample-Spacing Entropy Estimation”. In: *IEEE Transactions on Neural Networks* 19.11 (2008), 1978 –1984.

[Par62] E. Parzen. “On estimation of a probability density function and mode”. In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076.

[PB00] D. R. Pauluzzi and N. C. Beaulieu. “A comparison of SNR estimation techniques for the AWGN channel”. In: *IEEE Transactions on Communications* 48.10 (2000), pp. 1681–1691.

[PB04] F. Pérez-Cruz and O. Bousquet. “Kernel methods and their potential use in signal processing”. In: *IEEE Signal Processing Magazine* 21.3 (2004), pp. 57–65.

[Pea04] K. Pearson. *On the Theory of Contingency and its Relation to Association and Normal Correlation*. Vol. 1. Dulau and Company London, UK, 1904.

[Pez+04] A. Pezeshki, L. Scharf, M. Azimi-Sadjadi, and M. Lundberg. “Empirical canonical correlation analysis in subspaces”. In: *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.* Vol. 1. 2004, 994–997 Vol.1.

[Pha04] D.-T. Pham. “Fast algorithms for mutual information based independent component analysis”. In: *IEEE Transactions on Signal Processing* 52.10 (2004), pp. 2690–2700.

[Pin+17] F. du Pin Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy. “Principal inertia components and applications”. In: *IEEE Transactions on Information Theory* 63.8 (2017), pp. 5011–5038.

[PL96] A. Pagès-Zamora and M. Lagunas. “Joint probability density function estimation by spectral estimate methods”. In: *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 5. 1996, 2936–2939 vol. 5.

[PLH11] J. C. Príncipe, W. Liu, and S. Haykin. *Kernel Adaptive Filtering: a Comprehensive Introduction*. John Wiley & Sons, 2011.

[Pop17] A. Popoff. *Fundamentals of Signal Processing in Metric Spaces with Lattice Properties: Algebraic Approach*. CRC Press, 2017.

[PP02] A. Papoulis and S. U. Pillai. *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education, 2002.

[PPS10] D. Pál, B. Póczos, and C. Szepesvári. “Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs”. In: *Advances in Neural Information Processing Systems* 23 (2010).

[PPW17] F. du Pin Calmon, Y. Polyanskiy, and Y. Wu. “Strong data processing inequalities for input constrained additive noise channels”. In: *IEEE Transactions on Information Theory* 64.3 (2017), pp. 1879–1892.

[Prí10] J. C. Príncipe. *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*. New York Springer, 2010.

[Pri58] R. Price. "A useful theorem for nonlinear devices having Gaussian inputs". In: *IRE Transactions on Information Theory* 4.2 (1958), pp. 69–72.

[PSQ13] S. Park, E. Serpedin, and K. Qaraqe. "Gaussian assumption: The least favorable but the most useful". In: *IEEE Signal Processing Magazine* 30.3 (2013), pp. 183–186.

[PT01] V. Popovici and J.-P. Thiran. "Higher order autocorrelations for pattern classification". In: *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*. Vol. 3. IEEE. 2001, pp. 724–727.

[PXP06] A. R. Paiva, J.-W. Xu, and J. C. Príncipe. "Kernel principal components are maximum entropy projections". In: *International Conference on Independent Component Analysis and Signal Separation*. Springer. 2006, pp. 846–853.

[Ram+09] D. Ramírez, J. Vía, I. Santamaría, and P. Crespo. "Entropy and Kullback-Leibler divergence estimation based on Szegö's theorem". In: *17th European Signal Processing Conference*. 2009, pp. 2470–2474.

[Ram+10] D. Ramírez, J. Vía, I. Santamaría, and L. L. Scharf. "Detection of spatially correlated Gaussian time series". In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5006–5015.

[Ram+11] D. Ramírez, G. Vázquez-Vilar, R. López-Valcarce, J. Vía, and I. Santamaría. "Detection of rank-P signals in cognitive radio networks with uncalibrated multiple antennas". In: *IEEE Transactions on Signal Processing* 59.8 (2011), pp. 3764–3774.

[Ram+13] D. Ramírez, J. Vía, I. Santamaría, and L. L. Scharf. "Locally most powerful invariant tests for correlation and sphericity of Gaussian vectors". In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 2128–2141.

[RC19] J. Riba and F. de Cabrera. "A proof of de Bruijn identity based on generalized Price's theorem". In: *IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2019, pp. 1–5.

[Rén07] A. Rényi. *Probability Theory*. Courier Corporation, 2007.

[Rén59] A. Rényi. "On measures of dependence". In: *Acta Mathematica Academiae Scientiarum Hungarica* 10 (1959), pp. 441–451.

[Rén61] A. Rényi. "On measures of entropy and information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California. 1961.

[Rén65] A. Rényi. "On the foundations of information theory". In: *Revue de l'Institut International de Statistique* (1965), pp. 1–14.

[Res+11] 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". In: *Science* 334.6062 (2011), pp. 1518–1524.

[Roj+18] J. L. Rojo-Álvarez, M. Martínez-Ramón, J. Munoz-Mari, and G. Camps-Valls. *Digital Signal Processing with Kernel Methods*. John Wiley & Sons, 2018.

[Ros56] M. Rosenblatt. "Remarks on some nonparametric estimates of a density function". In: *The annals of mathematical statistics* (1956), pp. 832–837.

[RR07] A. Rahimi and B. Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems* 20 (2007).

[RS15] N. J. Roseveare and P. J. Schreier. “Model-order selection for analyzing correlation between two data sets using CCA with PCA preprocessing”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5684–5687.

[Rud76] W. Rudin. *Principles of Mathematical Analysis*. Vol. 3. McGraw-Hill New York, 1976.

[Rud90] W. Rudin. *Fourier Analysis on Groups*. John Wiley & Sons, Ltd, 1990.

[RVV10] J. Riba, J. Villares, and G. Vázquez. “A nondata-aided SNR estimation technique for multilevel modulations exploiting signal cyclostationarity”. In: *IEEE Transactions on Signal Processing* 58.11 (2010), pp. 5767–5778.

[SAP10] J. A. Suykens, C. Alzate, and K. Pelckmans. “Primal and dual model representations in kernel-based learning”. In: *Statistics Surveys* 4 (2010), pp. 148–183.

[Sar16] S. Sarbu. “On Rényi’s entropy estimation with one-dimensional Gaussian kernels”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 4408–4412.

[Sav+15] V. Savic, E. G. Larsson, J. Ferrer-Coll, and P. Stenumgaard. “Kernel methods for accurate UWB-based ranging with reduced complexity”. In: *IEEE Transactions on Wireless Communications* 15.3 (2015), pp. 1783–1793.

[SBP15] Y. Sun, P. Babu, and D. P. Palomar. “Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions”. In: *IEEE Transactions on Signal Processing* 63.12 (2015), pp. 3096–3109.

[SC04] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

[Sej+13] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. “Equivalence of distance-based and RKHS-based statistics in hypothesis testing”. In: *The annals of statistics* (2013), pp. 2263–2291.

[Ser09] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

[Set+11] S. Seth, M. Rao, I. Park, and J. C. Príncipe. “A unified framework for quadratic measures of independence”. In: *IEEE Transactions on Signal Processing* 59.8 (2011), pp. 3624–3635.

[Sev+07] J. F. Sevillano, I. Vélez, M. Leyh, S. Lipp, A. Irizar, and F. Fontán. “In-service SNR estimation without symbol timing recovery for QPSK data transmission systems”. In: *IEEE Transactions on Wireless Communications* 6.9 (2007), pp. 3202–3207.

[Sha48] C. E. Shannon. “A mathematical theory of communication”. In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[Sib69] R. Sibson. “Information radius”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 14.2 (1969), pp. 149–160.

[Sil86] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Vol. 26. CRC Press, 1986.

[Skó15] M. Skórski. “Shannon entropy versus Rényi entropy from a cryptographic viewpoint”. In: *IMA International Conference on Cryptography and Coding*. Springer. 2015, pp. 257–274.

[SM00] L. Scharf and C. Mullis. “Canonical coordinates and the geometry of inference, rate, and capacity”. In: *IEEE Transactions on Signal Processing* 48.3 (2000), pp. 824–831.

[SN12] J. F. Silva and S. Narayanan. “Complexity-regularized tree-structured partition for mutual information estimation”. In: *IEEE transactions on information theory* 58.3 (2012), pp. 1940–1952.

[SP09] S. Seth and J. C. Príncipe. “On speeding up computation in information theoretic learning”. In: *2009 International Joint Conference on Neural Networks*. IEEE. 2009, pp. 2883–2887.

[SR09] G. J. Székely and M. L. Rizzo. “Brownian distance covariance”. In: *The annals of applied statistics* 3.4 (2009), pp. 1236–1265.

[SRB07] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. “Measuring and testing dependence by correlation of distances”. In: *The annals of statistics* 35.6 (2007), pp. 2769–2794.

[SS00] A. J. Smola and B. Schökopf. “Sparse Greedy Matrix Approximation for Machine Learning”. In: *Proceedings of the Seventeenth International Conference on Machine Learning*. ICML ’00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 911–918.

[SS04] P. Stoica and Y. Selen. “Model-order selection: a review of information criterion rules”. In: *IEEE Signal Processing Magazine* 21.4 (2004), pp. 36–47.

[SS10] P. J. Schreier and L. L. Scharf. *Statistical Signal Processing of Complex-valued Data: The Theory of Improper and Noncircular Signals*. Cambridge university press, 2010.

[SS11] E. M. Stein and R. Shakarchi. *Fourier Analysis: An Introduction*. Vol. 1. Princeton University Press, 2011.

[SSA10] L. M. Silva, J. M. de Sá, and L. A. Alexandre. “The MEE principle in data classification: A perceptron-based analysis”. In: *Neural computation* 22.10 (2010), pp. 2698–2728.

[SSB+02] B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

[SSM98] B. Schölkopf, A. Smola, and K.-R. Müller. “Nonlinear component analysis as a kernel eigenvalue problem”. In: *Neural computation* 10.5 (1998), pp. 1299–1319.

[Sta59] A. J. Stam. “Some inequalities satisfied by the quantities of information of Fisher and Shannon”. In: *Information and Control* 2.2 (1959), pp. 101–112.

[Suz+09] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. “Mutual information estimation reveals global associations between stimuli and biological processes”. In: *BMC bioinformatics* 10.1 (2009), pp. 1–12.

[SV16] I. Sason and S. Verdú. “f-Divergence Inequalities”. In: *IEEE Transactions on Information Theory* 62.11 (2016), pp. 5973–6006.

[SW98] T. Summers and S. Wilson. “SNR mismatch and online estimation in turbo decoding”. In: *IEEE Transactions on Communications* 46.4 (1998), pp. 421–423.

[Sza14] Z. Szabó. “Information Theoretical Estimators Toolbox”. In: *Journal of Machine Learning Research* 15.9 (2014), pp. 283–287.

[TBA16] M. S. Tabass, G. M. Borzadaran, and M. Amini. “Rényi entropy in continuous case is not the limit of discrete case”. In: *Mathematical Sciences and Applications e-Notes* 4.1 (2016), pp. 113–117.

[TK10] S. T. Tokdar and R. E. Kass. “Importance sampling: a review”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1 (2010), pp. 54–60.

[Tsa88] C. Tsallis. “Possible generalization of Boltzmann-Gibbs statistics”. In: *Journal of statistical physics* 52.1 (1988), pp. 479–487.

[Tyl87] D. E. Tyler. “A distribution-free M-estimator of multivariate scatter”. In: *The annals of Statistics* (1987), pp. 234–251.

[TZ15] N. Tishby and N. Zaslavsky. “Deep learning and the information bottleneck principle”. In: *2015 IEEE Information Theory Workshop (ITW)*. 2015, pp. 1–5.

[TZI15] S. Tridenski, R. Zamir, and A. Ingber. “The Ziv-Zakai-Rényi Bound for Joint Source-Channel Coding”. In: *IEEE Transactions on Information Theory* 61.8 (2015), pp. 4293–4315.

[Vae10] S. V. Vaerenbergh. *Kernel Methods for Nonlinear Identification, Equalization and Separation of Signals*. Universidad de Cantabria, 2010.

[Vas76] O. Vasicek. “A test for normality based on sample entropy”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 38.1 (1976), pp. 54–59.

[Ver15] S. Verdú. “α-mutual information”. In: *2015 Information Theory and Applications Workshop (ITA)*. 2015, pp. 1–6.

[Von18] J. Von Neumann. *Mathematical Foundations of Quantum Mechanics: New Edition*. Princeton university press, 2018.

[VR22] M. Vilà and J. Riba. “A Test for Conditional Correlation Between Random Vectors Based on Weighted U-Statistics”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 5792–5796.

[VV05] J. Villares and G. Vázquez. “Second-order parameter estimation”. In: *IEEE transactions on signal processing* 53.7 (2005), pp. 2408–2420.

[VZB17] I. Valero-Toranzo, S. Zozor, and J.-M. Brossier. “Generalization of the de Bruijn Identity to General ϕ-Entropies and ϕ-Fisher Informations”. In: *IEEE Transactions on Information Theory* 64.10 (2017), pp. 6743–6758.

[Wac+05] M. P. Wachowiak, R. Smolikova, G. D. Tourassi, and A. S. Elmaghraby. “Estimation of generalized entropies with sample spacing”. In: *Pattern Analysis and Applications* 8.1 (2005), pp. 95–101.

[Wau+07] X. Wautelet, C. Herzet, A. Dejonghe, J. Louveaux, and L. Vandendorpe. “Comparison of EM-based algorithms for MIMO channel estimation”. In: *IEEE Transactions on communications* 55.1 (2007), pp. 216–226.

[WK85] M. Wax and T. Kailath. “Detection of signals by information theoretic criteria”. In: *IEEE Transactions on acoustics, speech, and signal processing* 33.2 (1985), pp. 387–392.

[WKV09] Q. Wang, S. R. Kulkarni, and S. Verdú. *Universal Estimation of Information Measures for Analog Sources*. Vol. 5:3. Foundations, trends in Communications, and Information Theory, 2009.

[WS00] C. Williams and M. Seeger. “Using the Nyström method to speed up kernel machines”. In: *Advances in neural information processing systems* 13 (2000).

[WSW20] W. Wang, Y. Shen, and Y. Wang. “Low-complexity non-data-aided SNR estimation for multilevel constellations”. In: *IEEE Communications Letters* 24.1 (2020), pp. 113–116.

[Xu+08] J.-W. Xu, A. R. C. Paiva, I. Park, and J. C. Principe. “A Reproducing Kernel Hilbert Space Framework for Information-Theoretic Learning”. In: *IEEE Transactions on Signal Processing* 56.12 (2008), pp. 5891–5902.

[Xu+22] X. Xu, S.-L. Huang, L. Zheng, and G. W. Wornell. “An information theoretic interpretation to deep neural networks”. In: *Entropy* 24.1 (2022), p. 135.

[YXS16] Z. Yang, L. Xie, and P. Stoica. “Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution”. In: *IEEE Transactions on Information Theory* 62.6 (2016), pp. 3685–3701.

[ZC06] S. K. Zhou and R. Chellappa. “From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.6 (2006), pp. 917–929.

[Zou+12] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma. “Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts”. In: *IEEE Signal Processing Magazine* 29.4 (2012), pp. 61–80.

[ZZJ04] L. Zhang, W. Zhou, and L. Jiao. “Wavelet support vector machine”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34.1 (2004), pp. 34–39.

List of Figures