Regularized Estimation of Information via High Dimensional Canonical Correlation Analysis

Jaume Riba, Senior Member, IEEE and Ferran de Cabrera, Student Member, IEEE

Abstract

In recent years, there has been an upswing of interest in estimating information from data emerging in a lot of areas beyond communications. This paper aims at estimating the information between two random phenomena by using consolidated second-order statistics tools. The squared-loss mutual information is chosen for that purpose as a natural surrogate of Shannon mutual information. The rationale for doing so is developed for i.i.d. discrete sources -mapping data onto the simplex space-, and for analog sources -mapping data onto the characteristic space-, highlighting the links with other well-known related concepts in the literature based on local approximations of information-theoretic measures. The proposed approach gains in interpretability and scalability for its use on large datasets, providing physical interpretation to the free regularization parameters. Moreover, the structure of the proposed mapping allows resorting to Szegő’s theorem to reduce the complexity for high dimensional mappings, exhibiting strong dualities with spectral analysis. The performance of the proposed estimators is analyzed using Gaussian mixtures.

Index Terms

Data analytics, canonical correlation analysis, kernel methods, quadratic dependence measures, squared-loss mutual information, Gebelein maximal correlation, characteristic function, coherence matrix, information-theoretic learning.

I. INTRODUCTION

Entropy and mutual information, introduced by Shannon in 1948, are well-known concepts with clear operational significance in the field of information theory and communications: they establish fundamental limits in data compression and data transmission [1]. More generally, Kullback-Leibler (KL) divergence (also called relative entropy) is a dissimilarity measure between distributions, being mutual information just a particular case. In the last decades, researchers have used the concepts of entropy, mutual information and divergence in a wide class of areas beyond communications, such as data science, machine learning, neuroscience, economics, biology, language and other experimental sciences. In these areas, the aforementioned concepts have proven their utility as tools for measuring randomness, dependence and similarity of random phenomena [2], [3], substituting or working together with the conventional statistical tools of variance and covariance. As a prominent example, the field of information-theoretic learning [4] cuts across signal processing and machine learning by looking at machine learning under the umbrella of information theory. This new perspective for knowledge discovering provides guidelines for the design of nonparametric universal tools for data analytics [5]. Especially, the wish for interpretability and for understanding the learning process has become a challenging aspect in practical applications of machine learning systems due to their lack of ability to explain their actions to humans [6]. Although these models exhibit impressive capabilities, the development of tools for measuring information, which is the main motivation of this paper, can help them in reducing their vulnerability to attacks, and can provide means for diagnosis in the case of failures.

A. Main contributions, related works and overall organization

By delving into fundamental concepts of information theory and statistical signal processing, this paper aims at developing insightful tools for measuring meaningful indicators of the amount of information contained in raw data with two main objectives in mind: endowing the methods with as much interpretability as possible, thus providing insight on the selection of their free parameters; and leveraging as much as possible classical and consolidated statistical signal processing techniques based on second-order statistics.

The literature on empirical estimation of information measures and its applications is long, and a guide has been recently provided in [3]. The main contributions of this work are the following.

1) Providing a fresh view of the squared-loss mutual information surrogate for both discrete and analog sources. Its relation with other information measures is investigated and contextualized in a unified manner, focusing on the idea of local approximations.
2) Linking the problem of estimation of information with the classical problem of Canonical Correlation Analysis (CCA), which is used in many fields of statistical signal processing.

3) The proposal of an explicit universal mapping from analog sources to complex steering vectors, leading to a computationally efficient alternative to kernel-CCA (KCCA) methods and to their mechanism for regularization.

4) Providing interpretability and insight to the problem of estimating information and to its regularization.

5) The proposal of a reduced complexity approximate estimator resorting to the asymptotic behavior of Toeplitz matrices.

Concerning the related works, the presented approach is an extension of the main ideas shortly provided by the authors in [7]. Although the interest on surrogates of entropy, KL divergence and mutual information, such as Rényi entropy, Rényi divergence, f-divergence and chi-squared ($\chi^2$) divergence has a long and rich history (see [3] and references therein), its use for data analytics has been particularly focused in [4]. The idea of local approximations of information measures exposed in this paper is very similar to the linear information coupling approach proposed in [8], [5], which was used there as a tool for developing insights on otherwise intractable problems in the field of communications. Compared with the kernel-CCA approach [9] that uses the dual form of the models (kernel trick) -thus precluding its direct use in applications involving large datasets-, the proposed alternative stays in the primal model, thus gaining intuition and scalability of the overall data processing. Moreover, in [10] it is suggested that the right choice of mapping function leads to a better representation of the data in the feature space, enabling to capture as much information as possible with a reduced dimensional mapping. The proposed statistics based on a Frobenius norm of a coherence matrix is very related to the local test proposed in [11] for Gaussian vectors, with the difference that our result applies for any kind of data mapped on a specific feature space. The regularization idea based on Gaussian convolutions is inspired on [12]. Finally, the use of the Szegö’s theorem exploiting the analogy between a probability density function and a power spectral density was also explored in [13] for Kullback-Leibler divergence estimation, by using autoregressive models for the densities.

This paper is organized as follows. Section II presents a unified overview of information theoretic surrogates, providing an original and fresh description of links among different approaches, and finishing with a short outline of the proposed overall strategy. Then, Section III focuses on discrete sources and shows the fundamental link between the proposed surrogate and classical second-order statistics. Once the structure of the problem is unveiled, Section IV moves to the extension to analog sources and fresh description of links among different approaches, and finishing with a short outline of the proposed overall strategy. Concerning the related works, the presented approach is an extension of the main ideas shortly provided by the authors in [7]. Although the interest on surrogates of entropy, KL divergence and mutual information, such as Rényi entropy, Rényi divergence, f-divergence and chi-squared ($\chi^2$) divergence has a long and rich history (see [3] and references therein), its use for data analytics has been particularly focused in [4]. The idea of local approximations of information measures exposed in this paper is very similar to the linear information coupling approach proposed in [8], [5], which was used there as a tool for developing insights on otherwise intractable problems in the field of communications. Compared with the kernel-CCA approach [9] that uses the dual form of the models (kernel trick) -thus precluding its direct use in applications involving large datasets-, the proposed alternative stays in the primal model, thus gaining intuition and scalability of the overall data processing. Moreover, in [10] it is suggested that the right choice of mapping function leads to a better representation of the data in the feature space, enabling to capture as much information as possible with a reduced dimensional mapping. The proposed statistics based on a Frobenius norm of a coherence matrix is very related to the local test proposed in [11] for Gaussian vectors, with the difference that our result applies for any kind of data mapped on a specific feature space. The regularization idea based on Gaussian convolutions is inspired on [12]. Finally, the use of the Szegö’s theorem exploiting the analogy between a probability density function and a power spectral density was also explored in [13] for Kullback-Leibler divergence estimation, by using autoregressive models for the densities.

This paper is organized as follows. Section II presents a unified overview of information theoretic surrogates, providing an original and fresh description of links among different approaches, and finishing with a short outline of the proposed overall strategy. Then, Section III focuses on discrete sources and shows the fundamental link between the proposed surrogate and classical second-order statistics. Once the structure of the problem is unveiled, Section IV moves to the extension to analog sources along with insightful tools for regularization and complexity reduction. The performance of the proposed estimators is illustrated by computer simulations in Section V and VI summarizes the main conclusion of this work.

B. Notation

Column vectors: bold-faced lower case letters. Matrices: bold-faced upper case letters. $[A]_{n,m}$: element at the $n$-th row and $m$-th column of matrix $A$. $[a]_n$: $[a]_{n,1}$: $[a]$: diagonal matrix with diagonal elements $[a]_{n,n} = [a]_n$. ($.)^T$: transpose. ($.)^H$: Hermitian transpose. $\text{tr}(A)$: trace. $||.||$: Frobenius or Euclidean norm of a vector, matrix or function. $|.|$: absolute value of a complex number, or cardinality of a set. $A \in \mathbb{R}^{N \times M}$: real matrix of dimension $N \times M$. $A \in \mathbb{C}^{N \times M}$: complex matrix of dimension $N \times M$. $a \in \mathbb{R}^N$: $a \in \mathbb{R}^{N \times 1}$. $a \in \mathbb{C}^N$: $a \in \mathbb{C}^{N \times 1}$. $\mathbb{R}_+$: set of positive real numbers. $x \sim N(\mu; R)$: $x$ is a real Gaussian random vector of mean $\mu$ and covariance matrix $R$. $x \sim CN(\mu; R)$: $x$ is a complex Gaussian random vector of mean $\mu$ and covariance matrix $R$. $\mathbb{E}_p$: statistical expectation operator ($\mathbb{E}_p[f(x)] = \int f(x) dP$), where $p$ is the mass function (or density function for analog variables), and $P$ is the probability measure (or the cumulative distribution function for analog variables). $(x[l])_L = L^{-1} \sum_{l=1}^{L} x[l]$: $L$-th length sample mean operator. $1_D$: $D \times 1$ all-ones vector. $1_n$: indicator function ($1_a = 1$ if $a$ is true, and $1_a = 0$, otherwise). $A^{1/2}$: Hermitian square root matrix of the Hermitian matrix $A$. $A^{-1/2}$: Hermitian square root matrix of the Hermitian matrix $A^{-1}$. $\alpha^T$: element-wise power of a vector. $a^{T_a} = (a^\alpha)^T$. Toeplitz: Toeplitz-Hermitian matrix constructed from its first column $c$. $\odot$: Hadamard product. *: convolution operator. $\delta_{mn}$: Kronecker delta. $[x]^+$: ceiling function.

II. INFORMATION-THEORETIC MEASURES FOR DATA ANALYTICS

Shannon entropy and Kullback-Leibler (KL) divergence are fundamental quantities in information theory and its applications [1]. Some surrogates of these quantities have been proposed for data analytics with the goal of simplifying the estimation process. In this section we provide a unified rationale for the derivation of surrogate information measures (entropy, divergence and mutual information) in a way as natural as possible, along with more details concerned with the related work, and we finish with a summary of the key ideas and more concrete goals of this paper. Through all the paper, we will assume that $p_X(x)$, defined on a set $X$, is either a mass function (for discrete sources) or a square-integrable density function (for continuous sources) associated to the random variable $X$.

A. Information potential surrogate

The Shannon entropy (in nats) is defined as

$$H(p_X) = -\mathbb{E}_{p_X} \ln p_X(x).$$

(1)
A natural surrogate of Shannon entropy can be obtained by applying the Jensen inequality. In particular, as \( \ln(.) \) is concave, we obtain

\[
H(p_X) = -\mathbb{E}_p \ln p_X(x) \geq -\mathbb{E}_{p_X} p_X(x) = H_2(p_X),
\]

(2)

with equality if and only if the source is uniform; otherwise, we get an strict inequality (\( > \)) as a consequence of the strict concavity of \( \ln(.) \). The right-hand term \( H_2(p_X) \) in (2) is just the second-order Rényi entropy, whose use for estimation via kernel methods has been explored in [4]. Second-order Rényi entropy can also be expressed as follows:

\[
H_2(p_X) = -\ln (1 - S_2(p_X)),
\]

(3)

where \( S_2(p_X) \) is the Tsallis entropy of second-order entropic index given by

\[
S_2(p_X) = 1 - V_2(p_X),
\]

(4)

being \( V_2(p_X) \) defined as the information potential:

\[
V_2(p_X) = \mathbb{E}_{p_X} p_X(x) = \|p_X\|^2.
\]

(5)

From the fundamental logarithm inequality \( \ln(1 + x) \leq x \) we can state that the second-order Tsallis entropy lower bounds the second-order Rényi entropy, that is

\[
H(p_X) \geq H_2(p_X) \geq S_2(p_X).
\]

(6)

The information potential \( V_2(p_X) \), as defined in (5), is just the squared Euclidean norm of the mass function or density function of the source, and it admits physical interpretations [4]. In the discrete case, the information potential is also called the collision probability [14] as it represents the probability that two independent outcomes of the source are equal. In the general case, the information potential has been used as a natural surrogate of (reversed sign) entropy because no logarithm is involved in its definition [4]. The information potential admits an estimation procedure from data which is much more natural than trying to estimate entropy itself. In particular, if one formulates a plug-in estimation method by first estimating the density function via the Parzen–Rosenblatt window method [15], [16], a final estimator is obtained by resorting to kernel methods [4], [17]. The information potential has also been proposed in [18] and [19] as a means to obtain robustness to outliers in the estimation of determinants of covariance matrices. For a review of plug-in methods and other methods for entropy estimation, the reader is referred to [2].

B. Chi-squared divergence surrogate

Inspired by the above rationale for the derivation of a natural surrogate of entropy, we next proceed with the divergence concept in a similar way. The KL divergence (in nats) between two probability mass or density functions \( p_X(x) \) and \( q_X(x) \), defined on the same set \( \mathcal{X} \), is given by

\[
D(p_X||q_X) = \mathbb{E}_{p_X} \ln \frac{p_X(x)}{q_X(x)}.
\]

(7)

For continuous random variables, absolute continuity of the densities with respect to each other is assumed, leading to bounded KL divergence. KL divergence is non-negative, and it is zero if and only if \( p_X(x) = q_X(x) \). A natural surrogate of KL divergence can be obtained by applying the Jensen inequality to (7). In particular, as \( \ln(.) \) is concave, we obtain

\[
D(p_X||q_X) \leq \ln \mathbb{E}_{p_X} \frac{p_X(x)}{q_X(x)} = D_{\chi^2}(p_X||q_X),
\]

(8)

with equality if and only if \( p_X(x) = q_X(x) \). For \( p_X(x) \neq q_X(x) \), i.e. for non-zero KL divergence, we get an strict inequality (\( < \)) as a consequence of the strict concavity of \( \ln(.) \). The right-hand term \( D_{\chi^2}(p_X||q_X) \) is just the second-order Rényi divergence for which some operational characterization have been provided [20]. Rényi divergence belongs to the class of \( f \)-divergences, which are useful, for example, in pattern recognition applications to identify independent components [21], as dissimilarity measures for image registration [22], and for target tracking [23]. Second-order Rényi divergence can also be expressed as follows:

\[
D_{\chi^2}(p_X||q_X) = \ln \left(1 + D_{\chi^2}(p_X||q_X)\right),
\]

(9)

where \( D_{\chi^2}(p_X||q_X) \) is the Pearson chi-squared divergence given by

\[
D_{\chi^2}(p_X||q_X) = \mathbb{E}_{p_X} \frac{p_X(x)}{q_X(x)} - 1 = \mathbb{E}_{p_X} \left( \frac{p_X(x) - q_X(x)}{\sqrt{p_X(x) q_X(x)}} \right)^2 = \frac{\| p_X - q_X \|_{\chi^2}^2}{\sqrt{q_X}},
\]

(10)

(see Appendix A). From the fundamental logarithm inequality \( \ln(1 + x) \leq x \) we can state that the Pearson chi-squared divergence upper bounds the second-order Rényi divergence, that is

\[
D(p_X||q_X) \leq D_{\chi^2}(p_X||q_X) \leq D_{\chi^2}(p_X||q_X),
\]

(11)
where the inequalities are strict (\(<\)) for non-zero divergence, and they become equality (to zero) if and only if \(p_X(x) = q_X(x)\). Looking at (11), it is worth noting that both Rényi and chi-squared divergences may be infinite even for finite KL divergence. As an example, we can mention the case of Gaussian \(p_X\) and \(q_X\) distributions with the variance of \(p_X\) being more than twice the variance of \(q_X\) (see [20], Eq. (10)). This issue is ignored in this paper because, as the focus is data analytics, the challenging problem is that of measuring divergence when it is small, as explained later on under the view of local approximations. Moreover, the empirical estimators will need to be ultimately regularized to cope with the limited data size, as detailed later on, which will lead to finite estimates for all scenarios.

As \(D_2\) is an explicit and monotonic function of \(D_{\chi^2}\) given in (9), there is no practical difference between them in terms of computational complexity from data. For this reason, especially for clarity, we will focus only on the chi-squared divergence along this paper, having in mind that a tighter upper bound \(D_2\) can be obtained from \(D_{\chi^2}\) via (9) if it were required for a particular application. Computing \(D_2\) from \(D_{\chi^2}\) may also be interesting in order to recover the additivity property of the obtained divergence measure with respect to independent (i.e. multiplicative) components in either \(p_X(x)\) or \(q_X(x)\), because while KL and Rényi divergence satisfy this property, the chi-squared divergence does not. A final consequence of the one-to-one relationship between Rényi and the adopted chi-squared divergence is that the chi-squared divergence inherits the invariance property of the Rényi divergence to nonlinear invertible transformations of the data. This follows from the more general data processing inequality (see, e.g., [24]).

We propose the use of the chi-squared divergence as defined in (10) as a natural upper bound to the KL divergence that exhibits significant computational advantages for data analytics. Its main advantage comes from the fact that no logarithm is involved on the quantity \(p_X dP_X / q_X\) that forms the integrand in the proof of (10) (see Appendix A). In any case, the logarithm is located outside the sum (or integral) if Rényi divergence is computed. This fact is what allows the use of second-order analysis techniques that are well known in statistical signal processing. The price to give entrance to these techniques is the need of mapping the data onto a high dimensional space, which constitutes the core idea explored in forthcoming sections.

1) Local approximation of KL divergence: The selection of the chi-squared divergence as a natural surrogate of KL divergence can be further reasoned by means of the following alternative rationale. Consider that \(p_X(x)\) and \(q_X(x)\) are close to each other, that is \(q_X(x) = p_X(x) + \epsilon \Delta(x)\) for some small quantity \(\epsilon\), where \(\Delta(x)\) is defined on the set \(X\) and constrained to have null area. Using the Taylor expansion of \(\ln((1 + \alpha)^{-1})\) up to the second order, i.e. \(-\alpha + \alpha^2/2 + O(\alpha^3)\), we can write the KL divergence in (7) as

\[
D(p_X || p_X + \epsilon \Delta) = E_{p_X} \ln \frac{p_X(x)}{p_X(x) + \epsilon \Delta(x)} = -\epsilon E_{p_X} \left[ \frac{\Delta(x)}{p_X(x)} \right] + \frac{1}{2} \epsilon^2 E_{p_X} \left[ \left( \frac{\Delta(x)}{p_X(x)} \right)^2 \right] + O(\epsilon^3). \tag{12}
\]

The first term is null since \(\Delta(x)\) sums up zero, which implies that

\[
D(p_X || p_X + \epsilon \Delta) = \frac{1}{2} \epsilon^2 E_{p_X} \left[ \left( \frac{\Delta(x)}{p_X(x)} \right)^2 \right] + O(\epsilon^3). \tag{13}
\]

Let us now examine the local behavior of the chi-squared divergence. Using the Taylor expansion of \((1 + \alpha)^{-1}\) up to the first order, i.e. \(1 - \alpha + O(\alpha^2)\), we can write the chi-squared divergence in (10) as

\[
D_{\chi^2}(p_X || p_X + \epsilon \Delta) = E_{p_X} \left[ \frac{(-\epsilon \Delta(x))^2}{p_X(x) [p_X(x) + \epsilon \Delta(x)]} \right] = \epsilon^2 E_{p_X} \left[ \frac{(\Delta(x))^2}{p_X(x)} \right] + \frac{1}{2} \epsilon^2 \left( 1 - \frac{\epsilon \Delta(x)}{p_X(x)} + O(\epsilon^2) \right) \left( \frac{\Delta(x)}{p_X(x)} \right)^2 + O(\epsilon^3) \tag{14}\]

From (13) & (14), the following fundamental result can be stated:

\[
D(p_X || p_X + \epsilon \Delta) = \frac{1}{2} D_{\chi^2}(p_X || p_X + \epsilon \Delta) + O(\epsilon), \tag{15}
\]

which means that half of the chi-squared divergence, that is \(\frac{1}{2} D_{\chi^2}\), constitutes a local approximation of KL divergence for close distributions. This observation is important because while \(D_{\chi^2}\) upper bounds KL divergence, \(\frac{1}{2} D_{\chi^2}\) is instead a local approximation, but not an upper bound. Note finally that, as pointed out in [8], \(D(p_X || p_X + \epsilon \Delta)\) and \(D(p_X + \epsilon \Delta || p_X)\) are considered to be equal up to first order approximation, and the same happens for \(D_{\chi^2}(p_X || p_X + \epsilon \Delta)\) and \(D_{\chi^2}(p_X + \epsilon \Delta || p_X)\).

As a related work, it is worth mentioning that local approximations of the KL divergence have been explored in [8] under the context of Linear Information Coupling (LIC) problems and Euclidean information theory, motivated by the goal of translating information theory problems into linear algebra problems, thus avoiding computational and mathematical bottlenecks. Similarly, the present paper extends this crucial idea to the analog case with the goal of translating the problem of measuring statistical dependence into much more manageable second-order analysis problems.
C. Squared-loss mutual information surrogate

Mutual information (MI) is an important concept in information theory that quantifies the statistical dependence between two random sources $X$ and $Y$, possibly defined on different sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The Shannon MI is defined as the KL divergence between the joint distribution and the product of the marginal distributions, both defined on the product set $\mathcal{X}\times\mathcal{Y}$:

$$I(X;Y) = D(p_{XY}\|p_X p_Y) = \mathbb{E}_{p_{XY}} \ln \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}. \tag{16}$$

Following the Jensen inequality upper bound idea examined in the previous subsection, a natural surrogate of mutual information for data analytics can be defined as follows:

$$I_2(X;Y) = D_2(p_{XY}\|p_X p_Y) = \ln \mathbb{E}_{p_{XY}} \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}. \tag{17}$$

The definition obtained in (17) is in agreement with the definition of Rényi mutual information proposed in [25]. However, it has to be noted that there are other possible ways to accomplish the generalization from the Rényi divergence to Rényi mutual information, most notably those suggested by Arimoto, Csiszár and Sibson (see [26] for a short review). Other definitions, as those in [27] and [28], involve a minimization process with respect to the marginals, motivated by operational interpretations.

As a particular case of [9] & [10], the second-order mutual information defined in (17) can be written as

$$I_2(X;Y) = \ln (1 + I_s(X;Y)), \tag{18}$$

where

$$I_s(X;Y) = \mathbb{E}_{p_{XY}} \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1 = \mathbb{E}_{p_{XY}} \left( \frac{p_{XY}(x,y) - p_X(x)p_Y(y)}{\sqrt{p_{XY}(x,y)p_X(x)p_Y(y)}} \right)^2 = \left\| \frac{p_{XY} - p_X p_Y}{\sqrt{p_X p_Y}} \right\|^2 \tag{19}$$

is the squared-loss mutual information (SMI) introduced in [29] for feature selection, and it is just the Pearson chi-squared divergence between the joint distribution and the product of the marginal distributions, both defined on the product set $X\times Y$.

It is also worth mentioning that the SMI can also be deduced from the so-called normalized cross-covariance operator (see [30], Eq. (9)). In this case, while constructing the Hilbert-Schmidt norm of a normalized cross-covariance operator through kernel methods, the associated explicit kernel-free integral expression corresponds to the SMI. This alternative way is particularly interesting since it establishes a clear link between second-order statistics and the SMI, provided that the data is mapped onto certain feature spaces, as we will see in the next section. Finally, note from (11) that

$$I(X;Y) \leq I_2(X;Y) \leq I_s(X;Y), \tag{20}$$

where the inequalities are strict ($<$) for dependent sources, and they become equality (to zero) if and only if the sources are independent.

1) Local approximation of Shannon mutual information: Using the local approximation of the KL divergence described in [15], if we assume the case of low dependence, that is $p_{XY}(x,y) = p_X(x)p_Y(y) + \epsilon \Delta(x,y)$ for some small quantity $\epsilon$, where $\Delta(x,y)$ is defined on the set $\mathcal{X}\times\mathcal{Y}$ and constrained to have null area, we can state that

$$I(X;Y) = \frac{1}{2} I_s(X;Y) + O(\epsilon), \tag{21}$$

as a particular case of (15), which means that half of the squared-loss mutual information, that is $\frac{1}{2} I_s$, constitutes a local approximation of Shannon mutual information for low dependence scenarios. Once again, this observation is important because while $I_s$ upper bounds mutual information, $\frac{1}{2} I_s$ is instead a local approximation, but not an upper bound. As a simple example, for the case that $p_{XY}(x,y)$ is a bivariate normal density with Pearson coefficient $\rho = \text{cov}(X,Y)/(\sigma_X\sigma_Y)$, we have $I(X;Y) = -0.5 \ln(1 - \rho^2)$ and $0.5 I_s(X;Y) = 0.5\rho^2/(1 - \rho^2)$, which are equal up to first order approximation.

2) Other quadratic dependence measures: Among other possibilities, a non-negative dependence measure that satisfies the requirement of being zero if and only if the sources are independent can also be defined as follows (see [17], Eq. (4)):

$$\xi(X;Y) = \mathbb{E}_{p_{XY}} \left( \frac{p_{XY}(x,y) - p_X(x)p_Y(y)}{\sqrt{p_{XY}(x,y)}} \right)^2 = \left\| p_{XY} - p_X p_Y \right\|^2. \tag{22}$$

Although this measure looks simpler, as it becomes the squared norm of a difference of densities, it lacks connection with Shannon mutual information, since nor inequalities nor local behavior can be stated in the way that have been established in [21] for the squared-loss mutual information measure $I_s(X;Y)$ defined in (19).
D. Summary of the key idea

After the short review and unified rationale presented above, we can summarize the goal of this paper. We have derived the surrogates (10) and (19) for divergence and mutual information, respectively, which share the following important properties.

1) The surrogates are upper bounds of information theoretic measures of well-known operational meaning, namely Kullback-Leibler divergence and Shannon mutual information. By being upper bounds, we make sure that relevant information will not be lost by using the surrogates for data analytics.

2) Half of the magnitudes measured by the surrogates are local approximations of meaningful information measures, which implies that they adopt meaningful values for the critical scenarios of close distributions for the divergence case, and small dependence regime for the mutual information case. Thus, the good local approximation property of the surrogates ensures that, in the challenging and interesting cases of measuring small information, the magnitude measured has full meaning. This behavior is lost by other quadratic measures of information.

3) The surrogates can be expressed as a second-order moment, that is, as the expectation of the squared of a random variable involving solely a ratio of densities without any logarithm. The implication is that, by designing adequate pre-conditioning of the data, classical second-order analysis techniques should be enough for estimating information.

The purpose of what follows is to propose a universal mapping strategy from the data onto a high dimensional feature space, such that the information can be extracted from that space by adequate pre-conditioning of the data, classical second-order analysis techniques should be enough for estimating information. The ultimate goal is to provide a rationale for the two-step data analytics strategy depicted in Fig. 1. Basically, the purpose of the first stage is to analyze complex dependencies between two data sources by first mapping L samples of the bivariate data onto a high-dimensional space defined on the complex field. The dimension N should be high enough to make sure that the maximum amount of complex associations potentially present on the data are captured, but it should be sufficiently small to provide reasonable computational complexity as well as regularization capabilities. After that, the second stage is based on second-order analysis techniques focused on describing linear dependencies between sets of variables. For instance, CCA ensures the previous statement and is a well known technique in the statistical signal processing field.

III. DISCRETE SOURCES: SECOND-ORDER STATISTICS ON THE SIMPLEX FEATURE SPACE

We first focus our attention on discrete sources since the key bridge to relate information with second-order statistics emerges more clearly in this case. Later, we will leverage this idea in order to smoothly generalize the concept to the more challenging case of analog sources.

Consider that X and Y are discrete random variables with alphabets \( \mathcal{X} = \{ x_n \}_{n=1,2,...,N} \) and \( \mathcal{Y} = \{ y_m \}_{m=1,2,...,M} \), respectively. Let us define the marginal probability column vectors \( \bar{p} \in \mathbb{R}_+^N \) and \( \bar{q} \in \mathbb{R}_+^M \) as \( \bar{p}_n = \Pr\{ X = x_n \} = p_X(x_n) \) for \( n = 1,2,...,N \) and \( \bar{q}_m = \Pr\{ Y = y_m \} = p_Y(y_m) \) for \( m = 1,2,...,M \). Similarly, we define the joint probability matrix \( \bar{J} \in \mathbb{R}_+^{N \times M} \) as \( \bar{J}_{n,m} = \Pr\{ X = x_n; Y = y_m \} = p_{XY}(x_n,y_m) \). Then, the SMI defined in (19) can be expressed as follows:

\[
I_s(X; Y) = \sum_{n=1}^{N} \sum_{m=1}^{M} \left[ \bar{C}_{n,m} \right]^2 = \text{tr}(\bar{C}^T \bar{C}) = ||\bar{C}||^2,
\]

where

\[
\bar{C} = [\bar{p}]^{-1/2}(\bar{J} - \bar{p}\bar{q}^T)[\bar{q}]^{-1/2}.
\]

Matrix \( \bar{C} \in \mathbb{R}_+^{N \times M} \) in (24) will be referred to as coherence matrix for the reasons explained later on, particularly due to its intimate link with the well known CCA tool in statistical signal processing. Moreover, the form of this matrix is encountered as well in the areas of information theory under the context of Linear Information Coupling (LIC) problems and Hirschfeld-Gebelein-Rényi (HGR) maximal correlation concept. As (24) contains the key ideas explored in this paper, we next provide an overview of these links along with new statements from known notions, all concerned with the problem of estimating information.
A. Relation to Linear Information Coupling (LIC) problems

Matrix $\tilde{C}$ in (24) can be expressed as follows:
\[
\tilde{C} = \left( B - \tilde{q}^{1/2} \tilde{p}^{T/2} \right)^T,
\]
where
\[
B = \left[ \tilde{q} \right]^{-1/2} \tilde{J}^T \left[ \tilde{p} \right]^{-1/2}.
\]
To provide an interpretation of (26), consider that the source corresponding to right and left singular vectors (29); iii) the squared-loss mutual information in (23) is upper bounded by $\min(\sigma_1, \sigma_2)$, where $\sigma_1 = \sigma_{\max}(B)$ is the largest singular value is equal to the second largest singular value of the divergence transition matrix in (24). Then: i) the minimum singular value is zero; ii) the largest singular value is equal to the second largest singular value of the divergence transition matrix in (29); iii) the squared-loss mutual information in (23) is upper bounded by $\min(N,M) - 1$.

Proof: According to the mentioned properties of matrix $B$, we can write its SVD as follows:
\[
B = \tilde{q}^{1/2} \tilde{p}^{T/2} + \sigma_{\max} u_{\max} v_{\max}^T + \sum_{i=3}^{\min(N,M)} \sigma_i u_i v_i^T.
\]
Now, from the expression in (25), it is clear that the SVD of matrix $\tilde{C}$ is:
\[
\tilde{C} = \sigma_{\max} u_{\max} v_{\max}^T + \sum_{i=2}^{\min(N,M)-1} \lambda_i u_{i+1} v_{i+1}^T,
\]
where $\sigma_{\max}(B) = \lambda_{\max}(C)$. In short, we find that the second largest singular value of the divergence transition matrix (a fundamental quantity in LIC problems) is equal to the largest singular value of the coherence matrix (a fundamental quantity in measuring statistical dependence). Finally, as the eigenvalues of matrix $\tilde{C}^T \tilde{C}$ are the squared modulus of the singular values of $\tilde{C}$, which are all smaller than 1 and the minimum is 0, we obtain the stated upper bound on the SMI.

B. Relation to Canonical Correlation Analysis (CCA)

The above matrix $\tilde{C}$ in (24) has the form of a coherence matrix and, therefore, it turns out that the squared-loss mutual information $I_{sl}(X; Y)$ can be directly related with the standard CCA method [32]. This connection of ideas is relevant since CCA is an important tool applied in many fields of signal processing and machine learning, so it should not be surprising if, eventually, that notion becomes fundamental as well for the problem of estimating information. Interestingly, the connection with CCA that will be unveiled in the sequel is a direct consequence of the fact that no logarithm is present in the definition of the squared-loss mutual information surrogate proposed in this paper.

To see that bridge, let us try to express matrix $C$ as a function of second-order statistics computed from the available data consisting of a sequence of $L$ i.i.d. pairs $\{x(l), y(l)\} \in \mathcal{X} \times \mathcal{Y}$ for $l = 1, 2, \ldots, L$. Let $\hat{p}$, $\hat{q}$ and $\hat{J}$ be estimates of the
marginal and joint mass functions. From \(^\{23,24\}\), we define a plug-in estimator of the SMI as \(\hat{I}_s(X;Y) = ||\hat{C}||_F^2\), where 
\[
\hat{C} = [\hat{\mathbf{p}}]^{-1/2}(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^H)[\hat{\mathbf{q}}]^{-1/2}. 
\]
Let us define the full-rank data matrices \(\mathbf{D}_x (N \times L)\) and \(\mathbf{D}_y (M \times L)\) as follows:
\[
[\mathbf{D}_x]_{n,l} = 1_{x(l)=x_n}, \quad [\mathbf{D}_y]_{n,l} = 1_{y(l)=y_n}. \tag{32}
\]
These data matrices are the result of a one-to-one mapping process from the elements of the sources to the canonical basis of dimension equal to the set cardinality. Clearly, the mass function estimates required by the plug-in estimator of SMI can be computed through first and second-order statistics as follows:
\[
\hat{\mathbf{p}} = \frac{1}{L}\mathbf{D}_x \mathbf{1}, \quad \hat{\mathbf{q}} = \frac{1}{L}\mathbf{D}_y \mathbf{1},
\]
\[
[\hat{\mathbf{p}}] = \frac{1}{L}\mathbf{D}_x \mathbf{D}_x^H, \quad [\hat{\mathbf{q}}] = \frac{1}{L}\mathbf{D}_y \mathbf{D}_y^H,
\]
\[
\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^T = \frac{1}{L}\mathbf{D}_x \hat{\mathbf{P}}_1 \mathbf{D}_y^H, \tag{33}
\]
where \(\hat{\mathbf{P}}_1 = \mathbf{I} - 11^T/L\) is the projection matrix onto the orthogonal space spanned by \(\mathbf{1}\). As a result, the mass function estimates required in the computation of the SMI are just the two sample mean vectors, the two autocorrelation matrices and the cross-covariance matrix.

**Theorem 2.** Preliminary link SMI-CCA: Let \(\mathbf{X} \in \mathbb{C}^{N \times L}\) and \(\mathbf{Y} \in \mathbb{C}^{M \times L}\) be data matrices obtained as \(\mathbf{X} = \mathbf{F}\mathbf{D}_x\) 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 full-rank mapping matrices (code-books). The estimated squared-loss mutual information based on a plug-in estimator is given by the Frobenius norm of a sample coherence matrix, that is:
\[
||\hat{\mathbf{C}}||_F^2 = \hat{I}_s(X;Y), \tag{34}
\]
where
\[
\hat{\mathbf{C}} = \hat{\mathbf{R}}_{xy}^{-1/2}\hat{\mathbf{C}}_{xy}\hat{\mathbf{R}}_{xy}^{-1/2}, \tag{35}
\]
being \(\hat{\mathbf{R}}_{xy} = \mathbf{X}\mathbf{X}^H/L\) and \(\hat{\mathbf{R}}_{xy} = \mathbf{Y}\mathbf{Y}^H/L\) the sample autocorrelation matrices and \(\hat{\mathbf{C}}_{xy} = \hat{\mathbf{X}}\hat{\mathbf{P}}_1\mathbf{Y}^H/L\) the sample cross-covariance matrix. In particular, a sufficient condition for (34) is that \(\mathbf{F} = \mathbf{I}_N\) and \(\mathbf{G} = \mathbf{I}_M\), which implies mapping the data onto the orthonormal canonical basis.

**Proof:** See Appendix B.

Although Thm. 2 sets the link between the SMI surrogate and second-order statistics, matrix \(\hat{\mathbf{C}}\) in (35) is not (apparently) a coherence matrix as that required by CCA, because autocorrelation instead of covariances are involved. However, the following theorem establishes the full link with CCA:

**Theorem 3.** Full link SMI-CCA: Let \(\mathbf{X} \in \mathbb{C}^{N' \times L}\) \((N' < N)\) and \(\mathbf{Y} \in \mathbb{C}^{M' \times L}\) \((M' < M)\) be data matrices obtained as \(\mathbf{X} = \mathbf{F}\mathbf{D}_x\) 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 full-rank mapping matrices (code-books). Let us define the small-size sample coherence matrix as
\[
\hat{\mathbf{C}}_{N',M'} = \hat{\mathbf{C}}_{x}^{-1/2}\hat{\mathbf{C}}_{xy}\hat{\mathbf{C}}_{y}^{-1/2}, \tag{36}
\]
being \(\hat{\mathbf{C}}_x = \hat{\mathbf{X}}\hat{\mathbf{P}}_1\mathbf{X}^H/L\) and \(\hat{\mathbf{C}}_y = \hat{\mathbf{Y}}\hat{\mathbf{P}}_1\mathbf{Y}^H/L\) the sample covariance matrices and \(\hat{\mathbf{C}}_{xy} = \hat{\mathbf{X}}\hat{\mathbf{P}}_1\mathbf{Y}^H/L\) the sample cross-covariance matrix. Then:
\[
||\hat{\mathbf{C}}_{N',M'}||_F^2 \leq \hat{I}_s(X;Y). \tag{37}
\]
In particular, a sufficient condition for the equality in (37) is that \(N' = N - 1\), \(M' = M - 1\) and that the columns of \(\mathbf{F}\) and \(\mathbf{G}\) are given by the \((N - 1)\)-simplex and by the \((M - 1)\)-simplex, respectively.

**Remark 1.** As a result of Thm. 3 we conclude that the coherence matrix required for estimating the SMI through its Frobenius norm can be estimated either by (35) or (36), provided that the Moore-Penrose inverse is generally used to cope with the rank-deficient case \(33\).

**Proof:** See Appendix C.

The physical meaning of matrices \(\mathbf{F}\) and \(\mathbf{G}\) in both Thms. 2 and 3 is that their columns contain the vectors to which the events of sources \(X\) and \(Y\) are mapped, respectively. The implication of Thm. 3 is that, as \(\mathbf{C}_{N-1,M-1}\) in (36) is just the \(1\)The data matrices are assumed full-rank for clarity, implying that \(L\) is sufficiently large such that \((x_n, y_m) \in \{x(l), y(l)\}_{l=1:L}\) for all \(n = 1 : N\) and \(1 : M\). Note that \([\hat{\mathbf{p}}]\) and \([\hat{\mathbf{q}}]\) are therefore invertible under this assumption. The issue of rank-deficient data matrices will be specifically addressed later on.
sample coherence matrix required in CCA, the squared-loss mutual information can be expressed just as the sum of the squared canonical correlations:

$$\hat{I}_s (X; Y) = \sum_{i=1}^{\min(N,M)-1} \lambda_i^2 (\mathcal{C}).$$  \hspace{1cm} (38)

Note also that, since a coherence matrix is invariant under linear invertible transforms, the code-books used for the SMI computation are irrelevant, provided that linearly independent vectors (columns of $F$ and $G$) are used. Otherwise, if the dimension of the space spanned after the mapping of $X$ is smaller than required (i.e. $N' < N - 1$ and/or $M' < M - 1$), the contribution of the smallest canonical correlations may be lost. The minimum dimension for the mapping of a source to vectors is therefore equal to the cardinality minus one. Moreover, the theorem also states implicitly that using higher dimension (i.e. $N' > N - 1$ and/or $M' > M - 1$) will yield a low-rank structure on $C_u$ and/or $C_y$. This idea will take a fundamental role in the process of leveraging all these notions to the analog case. In short, Fig. 2 illustrates the stated notion behind Thm. 3 binary data can be mapped to 1-dimensional points in the set $\{-1, 1\}$; ternary data can be mapped to 2-dimensional points in the set $\{[1,0], [-0.5, \sqrt{3}/2], [-0.5, -\sqrt{3}/2]\}$, and so on.

### C. Relation to Gebelein Maximal Correlation

Finally, the SMI measure can be linked with another important notion. Let $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ be scalar representations of the sources $X$ and $Y$, respectively. For $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation between sources $X$ and $Y$ is defined as [34], [35]:

$$\rho(X; Y) = \sup_{f, g : \mathbb{E}_{p_X} f = \mathbb{E}_{p_Y} g = 0, \mathbb{E}_{p_X} f^2 = \mathbb{E}_{p_Y} g^2 = 1} \mathbb{E}_{p_X Y} [f(x)g(y)] = \sup_{f,g} \frac{\sigma_{fg}}{\sqrt{\sigma_f^2 \sigma_g^2}},$$  \hspace{1cm} (39)

where

$$\sigma_{fg} = \mathbb{E}_{p_X Y} [(f(x) - \mathbb{E}_{p_X} f(x))(g(y) - \mathbb{E}_{p_Y} g(y))],$$

$$\sigma_f^2 = \mathbb{E}_{p_X} ((f(x) - \mathbb{E}_{p_X} f(x))^2),$$

$$\sigma_g^2 = \mathbb{E}_{p_Y} ((g(y) - \mathbb{E}_{p_Y} g(y))^2),$$

and the supremum in (39) is taken over all Borel functions $f$ and $g$. The HGR maximal correlation $\rho(X; Y)$ represents the maximal Pearson coefficient that can be obtained after mapping the events of the sources to reals, and it has found numerous applications in information theory and statistics (see [36] and reference therein). Recently, the HGR has been proposed as a practical and more relevant surrogate of mutual information in the field of security and privacy [37].

For sources with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, the problem of estimating the HGR maximal correlation coefficient can be reformulated as follows. Let $u \in \mathbb{R}^N$ and $v \in \mathbb{R}^M$ be the vectors containing the reals towards which the events of sources $X$ and $Y$ are mapped, respectively, that is $[u]_n = f(x_n)$ for $n = 1 : N$ and $[v]_m = g(y_m)$ for $m = 1 : M$. Then, from a sequence of $L$ i.i.d. pairs $\{x(l), y(l)\}$ we obtain the pairs of $L$-th length samples $\{u^H D_x, v^H D_y\}$. Clearly,

$$\hat{\rho}(X; Y) = \max_{u, v} \frac{u^H \hat{C}_{xy} v}{\sqrt{u^H C_u u} \sqrt{v^H C_v v}},$$  \hspace{1cm} (41)

which is given by the maximum singular value of the empirical coherence matrix $\hat{C}$, implying that $0 \leq \hat{\rho}(X; Y) \leq 1$. In contrast, according to what we have shown in this paper, the SMI is given by the sum of the squares of all potentially non-zero singular values of the coherence matrix, as seen in (38). Therefore, apart from the best single mapping to reals that the HGR notion provides, the SMI looks as well to other mappings to canonical coordinates of the coherence, thus becoming more sensitive to complex hidden relationships between the observed data. To illustrate these ideas, Fig. 3 compares the two
Fig. 3. Half of the ratio between the squared-loss MI and MI (left) and between the squared HGR and MI (right) for random discrete memory-less channels with random input distributions, for different $N (= M)$ values of alphabet sizes.

measures of information in what concerns to their relation with mutual information by testing discrete memory-less channels generated randomly with input distribution also generated randomly. It is seen that, in contrast with the HGR measure, the SMI exhibits a consistent behavior at the small dependence regime, in the sense of its much smaller dispersion around the value of 1 of the ratio between half of the HGR measure and the true MI, as well as its much less sensitivity to the alphabet size of the sources. It should be noted that, under the presented unified vision, the HGR maximal correlation coefficient can be seen as an extreme case of (badly) measuring the SMI by mapping the sources to reals (or a dimension spanned by the mapping equal to 1) instead of mapping them onto linearly independent vectors or onto the simplex, and it represents the maximum loss of information with respect to the true MI.

IV. ANALOG SOURCES: SECOND-ORDER STATISTICS ON THE CHARACTERISTIC FEATURE SPACE

For discrete sources, we have shown that estimating the SMI surrogate of mutual information via second-order statistics entails the mapping of events onto a vectorial space spanning a minimum dimension equal to the source cardinality minus one. But analog sources are of infinite dimension, so an infinite dimensional mapping is in principle required to retain all the information. This key idea, informally stated in Cover’s theorem on the separability of patterns [38], is well known in the field of machine learning. In particular, kernel methods have the ability (called kernel trick [39]) of implicitly using linear algebra on high (infinite) dimensional spaces without the necessity of explicitly visiting that huge space.

Leaving the kernel paradigm, in this section we propose an explicit mapping of analog sources onto a space of finite dimension on the complex field. The motivation is two-fold: on the one hand, the explicit mapping allows the use of the standard CCA as seen for discrete sources; on the other hand, the fixed dimension acts itself as a regularizer of the problem. The advantage is that, while the computation of scalar products on the implicit high dimensional spaces is very direct by using kernel methods, it is not so clear how to implement the inversion of matrices as those required by CCA. Although kernelized versions of CCA (KCCA) have been proposed (see [9], [40]) in the context of several signal processing and machine learning applications (e.g. blind source separation and nonlinear channel identification/equalization [41]), they involve costly inversion of big Gram matrices of kernel dot products between all data pairs, thus requiring strategies for decreasing complexity, such as the incomplete Cholesky factorization [39]. In addition, kernel methods ultimately need to be regularized to avoid overfitting. In that sense, the alternative based on an explicit mapping proposed in the sequel can be seen as procedure for regularizing the problem from the beginning, providing interpretability and computational complexity savings.

2“A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated” (T. M. Cover).
A. Dependence, correlation and characteristic function

To motivate the mapping that will be finally proposed in [48], let us write the marginal and joint characteristic functions (CF) (defined as the Fourier transform of the PDFs with sign reversal in the complex exponential) of a pair of analog sources $X$ and $Y$ as follows:

$$
\varphi_X(\omega_1) = \int p_X(x)e^{j\omega_1 x}dx = \mathbb{E}_{p_X}[z_1],
$$

$$
\varphi_Y(\omega_2) = \int p_Y(y)e^{j\omega_2 y}dy = \mathbb{E}_{p_Y}[z_2],
$$

$$
\varphi_{XY}(\omega_1,\omega_2) = \int p_{XY}(x,y)e^{j(\omega_1 x + \omega_2 y)}dxdy = \mathbb{E}_{p_{XY}}[z_1z_2],
$$

(42)

where $z_1 = e^{j\omega_1 X}$ and $z_2 = e^{j\omega_2 Y}$ are complex random variables obtained from $X$ and $Y$ through a nonlinear mapping. Clearly, if $X$ and $Y$ are independent, then $\varphi_{XY}(\omega_1,\omega_2) = \mathbb{E}_{p_X}[z_1]\mathbb{E}_{p_Y}[z_2] = \varphi_X(\omega_1)\varphi_Y(\omega_2)$ for all possible values of $\omega_1$ and $\omega_2$, implying that $z_1$ and $z_2$ are uncorrelated random variables. Note that the converse is also true: if $z_1$ and $z_2$ are uncorrelated for all possible values of $\omega_1$ and $\omega_2$, then $X$ and $Y$ are independent, since the condition $\varphi_{XY}(\omega_1,\omega_2) = \varphi_X(\omega_1)\varphi_Y(\omega_2)$ implies that $p_{XY}(x,y) = p_X(x)p_Y(y)$ as a result of the bijective property of the Fourier transform. It is important to emphasize that the converse statement above guarantees that any kind of statistical dependence between $X$ and $Y$ will be “manifested” as correlation for some values of $\omega_1$ and $\omega_2$, which means that the set of complex exponential functions is not restrictive for the problem of independence detection via second-order statistics. In other words, independence detection can be formulated as a problem of correlation detection (see [42]) by resorting to the characteristic function, provided that sufficient number of values of $\omega_1$ and $\omega_2$ are explored.

From this observation, two natural questions arise for the problem of SMI estimation: how many points of $\omega$ and $\omega_2$ for correlation analysis need to be explored? How small the separation between the explored points needs to be? To answer to these questions, we next propose a finite support for regularization (Sec. IV-B) and a uniform sampling (Secs. IV-B&IV-C) of the characteristic function, which further yield to an efficient estimation approach (Sec. IV-D).

B. Regularization through Gaussian convolutions

It is well known that the problem of estimating differential entropy and mutual information needs to be regularized [2]. In the sequel, we propose a regularization approach based on the properties of the characteristic function. The core idea is the concept of Gaussian convolutions, which has been recently proposed in [12] in the framework of differential entropy estimation as a means to achieve the parametric rate of convergence (w.r.t. the sample size) for distributions belonging to any nonparametric class. In the context of this paper, the approach has the additional advantage of providing a clear physical meaning to the proposed estimators of information, as seen in the sequel.

Consider that sources $X$ and $Y$ are in fact contaminated by independent zero-mean additive Gaussian sources $V_x$ and $V_y$ with known smoothing variance $\sigma^2$ and PDF $p_V$:

$$
x'(l) = x(l) + v_x(l), \quad y'(l) = y(l) + v_y(l).
$$

The purpose is now to estimate the contaminated information between the virtual sources $x'(l)$ and $y'(l)$ using the data obtained from the actual sources $x(l)$ and $y(l)$, which are still accessible. By doing so, a natural regularization of the problem is achieved, as seen in the sequel.

Since the PDF of the sum of independent random variables is the convolution of densities, that is $p_{X'}(x) = p_X(x) * p_V(x)$ and $p_{Y'}(y) = p_Y(y) * p_V(y)$, the CF is just the product of CFs of each term, that is $\varphi_{X'}(\omega) = \varphi_X(\omega)\varphi_V(\omega)$ and $\varphi_{Y'}(\omega) = \varphi_Y(\omega)\varphi_V(\omega)$, where

$$
\varphi_V(\omega) = e^{-\sigma^2\omega^2/2}
$$

is the CF of both $V_x$ and $V_y$. The key point is that the Gaussian shape has an effective support, which allows focusing on a finite interval given by $|\omega| \leq \omega_{\text{max}} = k\sigma^{-1}$, typically with $k = 2.5$. Consequently, as $|\varphi_X(\omega)| \leq 1$ and $|\varphi_Y(\omega)| \leq 1$, the CFs of the contaminated sources $X'$ and $Y'$ become both roughly zero for $|\omega| > \omega_{\text{max}}$ as well. The higher is $\sigma^2$, the stronger is the smoothing effect caused on the PDFs, and the smaller is the effective support of the CFs, exhibiting an insightful duality with the classical spectral estimation problem. Note that, by the general data processing inequality for $f$-divergences (see [43] and references therein), the additive perturbation in both sources regularizes the problem by decreasing and bounding the amount of mutual information to be measured, yielding to a negative bias contribution to the estimators as verified later on with computer simulations. In summary, estimates of the CFs of the contaminated sources can be obtained by just tapering the sample mean estimators as follows:

$$
\hat{\varphi}_{X'}(\omega) = \left\langle e^{j\omega x(l)} \right\rangle_L \varphi_V(\omega), \quad \hat{\varphi}_{Y'}(\omega) = \left\langle e^{j\omega y(l)} \right\rangle_L \varphi_V(\omega).
$$

(45)
Once the effective support of the empirical CFs is fixed, consider a uniform sampling in the ω domain with sampling period α. As CFs and PDFs are Fourier pairs, the sampling of CFs implies a periodic extension of the PDFs, such that the implicit density of X becomes

\[ p_X(x) = \sum_k (p_X \ast p_k) \left( x - \frac{2\pi k}{\alpha} \right), \]

and similarly for Y. The smaller is α, the smaller is the aliasing effect in (46), so the sampling period α can be roughly determined as the inverse of the expected dynamic range of the PDFs of the sources, that is, \( \alpha = 1/(q\sigma_x) \), typically with \( q = 3 \).

Assuming a CF support of \( 2\omega_{\text{max}} \), this yields a number of sampling points of the CFs given by

\[ N = 2 \left[ \frac{\omega_{\text{max}}}{\alpha} \right] + 1 = 2 \left[ kq\frac{\sigma_x}{\sigma} \right] + 1, \]

where an odd value of \( N \) is imposed for clarity in forthcoming developments. It is worth noting that, as the Gaussian window minimizes the uncertainty principle, a commonly used rationale in the classical time-frequency analysis, an additive Gaussian perturbation will therefore minimize the effective support of the contaminated characteristic function (i.e. the effective dimension of the feature space) for a given smoothing variance \( \sigma^2 \), which further supports the rationale for using the tool of Gaussian convolutions as a natural regularizer in the specific methodology explored in this paper. The interpretation of (47) is that of moving the problem to a finite parametric representation of the PDFs, which originally belong to a nonparametric class. Then, as the implicit number of parameters of the problem becomes finite, the SMI estimation problem will turn out to be consistent.

C. Second-order statistics on the characteristic space and SMI estimate

Given the physical sense of the proposed regularization, we propose (see [7]) a uniform symmetric and finite sampling of \( \omega_1 \) and \( \omega_2 \) to define mappings \( \phi_X(.) : \mathbb{R} \to \mathbb{C}^N \) and \( \phi_Y(.) : \mathbb{R} \to \mathbb{C}^N \) as

\[ x \to x = e^{j\alpha x} \quad y \to y = e^{j\alpha y}, \]

respectively, where \( n \in \mathbb{Z}^{N \times 1} \) is a vector of integers defined as \( n = [-K,-K+1,\ldots,K]^T \) with \( N = 2K+1 \). To appreciate the rationale, note that if one lets \( \alpha \to 0 \) and \( N \to \infty \) simultaneously in such a way that \( N\alpha \to \infty \) as well, for instance \( \alpha = O(N^{-1/2}) \), we are then mapping the sources onto asymptotically orthogonal vectors, which ensures that the SMI estimate that we developed for discrete sources (based now on CCA performed on the new spaces) will be asymptotically unbiased, according with Thms. 2 and 3. Note that the required feature space dimension is determined by (47), which explains why using a finite dimension acts as a natural regularization of the problem.

Consider a sequence of \( L \) i.i.d. pairs \( \{x(l), y(l)\} \in \mathbb{R}^2 \) for \( l = 1, 2, \ldots, L \). Using the mappings defined in (48), we obtain the pair of vector sequences \( \{x(l), y(l)\} \in \mathbb{C}^{N \times 2} \) in the feature space and construct the data matrices \( X \in \mathbb{C}^{N \times L} \) and \( Y \in \mathbb{C}^{N \times L} \) as follows:

\[ [X]_{:l} = x(l), \quad [Y]_{:l} = y(l). \]

From Thm. 2 and Remark 1, the SMI for analog sources can be finally estimated as

\[ \hat{I}_s(X; Y) = ||C||^2, \]

with

\[ \hat{C} = \hat{R}_{xx'}^{-1/2} \hat{C}_{x'y'} \hat{R}_{y'y}^{-1/2}. \]

Note that, following the concept of Gaussian convolutions, the sample autocorrelations and the cross-covariance in (51) refer to the contaminated sources \( X' \) and \( Y' \), for which the result in (45) is used in the sequel to compute both the cross-covariance and the autocorrelation matrices.

On the one hand, concerning the cross-covariance in (51), we clearly obtain from (45) that:

\[ \hat{C}_{x'y'} = \left\langle e^{j\alpha x(l)} e^{-j\alpha y(l)} \right\rangle_L \circ (ww^T) - \hat{P}q^H \]

where the weighted first-order statistics are

\[ \hat{P} = \left\langle e^{j\alpha x(l)} \right\rangle_L \circ w, \quad \hat{q} = \left\langle e^{j\alpha y(l)} \right\rangle_L \circ w \]

and the symmetric tapering vector is defined as

\[ [w]_n = \varphi_V((n-K)\alpha) = e^{-\sigma^2 \alpha^2 (n-K)^2/2} \]

for \( n = 0, 1, \ldots, N-1 \).
On the other hand, the elements of the sample autocorrelation matrices in (51) can be expressed as \([\hat{R}_x]_{n,m} = \langle e^{j\alpha(n-m)x(l)} \rangle_L \varphi_V(\alpha(n-m))\) and \([\hat{R}_y]_{n,m} = \langle e^{j\alpha(n-m)y(l)} \rangle_L \varphi_V(\alpha(n-m))\) for \(n, m = 0, 1, \ldots, 2K\), which endows them with a Toeplitz structure. As a result, we can construct them as follows:

\[
\hat{R}_{x'} = \text{Toe} (\hat{p}_a), \quad \hat{R}_{y'} = \text{Toe} (\hat{q}_a),
\]

where \(\hat{p}_a\) and \(\hat{q}_a\) are defined as the extended weighted first-order statistics,

\[
\hat{p}_a = \langle e^{j\alpha n_s x(l)} \rangle_L \odot w_a, \quad \hat{q}_a = \langle e^{j\alpha n_s y(l)} \rangle_L \odot w_a,
\]

with \(n_s = [0, 1, \ldots, N - 1]^T\) and the asymmetric tapering vector is defined as

\[
[w_a]_n = \varphi_V(n\alpha) = e^{-\sigma^2\alpha^2 n^2/2}
\]

for \(n = 0, 1, \ldots, N - 1\).

As a final remark, note that the regularization technique proposed above differs from the classical regularization technique used in KCCA based on diagonal loading of autocorrelation matrices [9]. Although both techniques succeed in solving the rank deficient issue, the proposed regularization based on tapering provides physical interpretation to the overall effect on the final estimate.

\[D. \text{ Large feature space dimension regime approximation}\]

The Toeplitz structure of \(\hat{R}_{x'}\) and \(\hat{R}_{y'}\) can be further exploited for the computation of the inverses in (51). Szegö’s theorem (see [44], [45]) establishes that, for large dimension, a Toeplitz matrix is asymptotically diagonalizable by the unitary Fourier matrix, and its eigenvalues asymptotically behave like samples of the Fourier transform of its entries. The most general and relaxed assumption that guarantees the behavior stated in Szegö’s theorem is that the columns of the matrices are square-integrable for \(N \rightarrow \infty\). This condition is clearly ensured by the tapering operation in (56) & (57). Effectively, as the Gaussian taper in (57) is square-integrable for any \(\sigma^2 > 0\) and the sample CFs are upper-bounded, that is \(|\langle e^{j\alpha n x(l)} \rangle_L| \leq 1\) and \(|\langle e^{j\alpha n y(l)} \rangle_L| \leq 1\), then the sample vectors \(\hat{p}_a\) and \(\hat{q}_a\) become square-integrable for \(N \rightarrow \infty\). This fact motivates a frequency-domain tool to reduce complexity by leveraging the approximate diagonalization of the involved Toeplitz matrices after a Fourier transform. In particular, the following theorem sets the required theoretical framework:

**Theorem 4.** Let \(t_n \in \mathbb{C}\) be an Hermitian sequence such that \(t_0 = 1\) and \(\lim_{N \rightarrow \infty} \sum_{n=0}^{N-1} |t_n|^2 < \infty\). Let us define vector \(t \in \mathbb{C}^N\) and Hermitian-Toeplitz matrix \(T \in \mathbb{C}^{N \times N}\) as \([t]_n = t_n\) and \(T = \text{Toe} (t)\), respectively. Let \(U \in \mathbb{C}^{N \times N}\) be the unitary Fourier matrix. Then

\[
\lim_{N \rightarrow \infty} \left\{ [UTU^H]_{n,m} - \left( 2\sqrt{N} \text{Re} \left( [U^H (t \odot v)]_n \right) - 1 \right) \delta_{nm} \right\} = 0
\]

for \(n, m = 1, \ldots, N - 1\), where \(v\) is a unilateral triangular window with elements \([v]_n = 1 - n/N\).

**Proof:** See [45] for detailed proofs concerning the limit behavior. In addition, in (58) we have used that the Fourier transform of an Hermitian sequence \(g_n\) can be written as \(\sum_{n=-\infty}^{\infty} g_n e^{-j2\pi wn} = g_0 + 2\text{Re} \left( \sum_{n=1}^{\infty} g_n e^{-j2\pi wn} \right)\) and \(g_n = t_n (1 - n/N)\) with \(g_0 = 1\).

Consider now the SMI estimate proposed in (50) & (51), which can be expressed as follows:

\[
\hat{I}_s (X; Y) = \left\| \hat{R}_{x'}^{-1/2} \hat{C}_{x'y'} \hat{R}_{y'}^{-1/2} \right\|^2.
\]

(59)

As the Frobenious norm is invariant under unitary transforms, we can write

\[
\hat{I}_s (X; Y) = \left\| (U \hat{R}_x U^H)^{-1/2} U \hat{C}_{x'y'} U^H (U \hat{R}_y U^H)^{-1/2} \right\|^2.
\]

(60)

Thm. 4 states that the transformed autocorrelation matrices \(U \hat{R}_x U^H\) and \(U \hat{R}_y U^H\) in (60) are asymptotically diagonal, which allows the formulation of an approximate and computationally efficient SMI estimator for the large dimension regime in the following manner:

\[
\hat{I}_{as} (X; Y) = \left\| [\hat{p}]^{-1/2} U \hat{C}_{x'y'} U^H [\hat{q}]^{-1/2} \right\|^2,
\]

(61)

where, as \([\hat{p}_a]_0 = [\hat{q}_a]_0 = 1\), we can use (56) & (57) to write the final transformed vectors as:

\[
\hat{p}' = 2\sqrt{N} \text{Re} \left( U^H (\hat{p}_a \odot v) \right) - 1,
\]

\[
\hat{q}' = 2\sqrt{N} \text{Re} \left( U^H (\hat{q}_a \odot v) \right) - 1.
\]

(62)

The main advantage of the proposed approximate estimator in (61) is that matrix inverses are avoided and only element-wise inverses are required.
V. Numerical Results

The performance of the proposed estimators, and the impact of their free parameters, is evaluated by means of Monte Carlo simulations. We measure the mean and variance of the estimated amount of information using the Gaussian Mixture Model (GMM) proposed in [42], which is illustrated in Fig. 4. The data is normalized such that \( E[p_X|x] = E[p_Y|y] = 0, \) \( E[p_X|x^2] = E[p_Y|y^2] = 1. \) The usefulness of this model lies on the fact that \( E[p_{XY}|xy] = 0 \) for any value of MI, thus forcing the estimators to discover dependence from uncorrelated data. To test the estimators, the true value of \( I_s(X;Y) \) is obtained by a genie-aided estimator based on empirical averaging [2] under the knowledge of the estimators to discover dependence from uncorrelated data.

Fig. 5 shows the mean of the proposed SMI estimators as a function of small \( (I_s(X,Y) \in [0,0.1]) \) and moderate \( (I_s(X,Y) \in (0.1,1]) \) values of true SMI, for different values of the dimension-variance pair \((\sigma^2, N)\) and data size \( L.\) The feature space dimension \( N \) is fixed from \( \sigma^2 \) by [47] using \( k = 3 \) and \( q = 2.5.\) Clearly, the linearity range of all curves increases as \( \sigma^2 \) decreases (and \( N \) increases accordingly) in the large SMI regime by increasing the ceiling value, with the price of additionally increasing the SMI floor at the small SMI regime. For a given pair \((\sigma^2, N)\), that floor gets inversely proportional to \( L.\) In short, the small dependence regime is the data limited regime, and the strong dependence regime is the dimension limited regime. In order to compensate the floor level at small data regime, a reduced bias estimator \( \hat{I}_s(X;Y) - I_s(X,Y) \) is also shown, where \( Y_{ind} \) is independent data identically distributed as \( Y \) and obtained by circularly shifting the data sequence associated to \( Y \) by \( j \) positions with \( j \neq 0 \) and \( j \neq L.\) In this way, the residual biases associated to estimates of theoretically null squared canonical correlations are reduced, thus improving the impact on the overall bias at the small data regime for a sufficiently small (big) smoothing variance \( \sigma^2 \) (dimension \( N \)), and regardless of the kind of data statistics. Finally, to validate the regularization based on Gaussian convolutions, another genie-aided estimate computed from truly contaminated data with independent additive Gaussian noises of variance \( \sigma^2 \) is also shown. As expected, the proposed estimators become asymptotically close from above to the contaminated SMI value as \( L \to \infty, \) which provides a physical interpretation of the negative bias that emerges at the moderate dependence regime.

Fig. 6 depict the bias and variance vs. \( L.\) of the proposed estimator \( \hat{I}_s(X,Y) \) along with its reduced bias version \( \hat{I}_s(X,Y) - I_s(X,Y) \) for the selection of the perturbation variance, the estimator uses the classical Silverman’s rule [46], [47] derived in the context of nonparametric kernel density functionals estimation, which is known to provide consistent results for small dimensional data [4]. According with this rule, the perturbation variance is set to monotonically decrease with \( L \) as \( \sigma^2 = p/\left(L^{2.5}\right), \) being \( p \) a free parameter, which is shown to provide a good bias-variance trade-off at all data-size regimes as well as consistency of the estimate for \( L \to \infty.\) This relation between data-size and perturbation variance can also be encountered in the context of spectral density estimation after minimizing the MSE with respect to the taper bandwidth [48], recalling the resemblance between the perturbation based on Gaussian convolutions and the spectral estimation problem. For clarity, the rationale for using this rule also in the context of estimating information is sketched in Appendix D. It can be seen that the reduced bias estimator is especially effective at small values of \( L,\) approximating the performance of the original method as \( L \) increases, at the cost of providing a slightly higher variance. The least squares mutual information estimator (LSMI) [29] is also shown, whose parameters are selected through cross-validation. For completeness, two MI estimators are also tested: one is based on adaptively partitioning the observation space [49], and the other consists on measuring entropy through the k-nearest neighbor algorithm [50] with a single neighbor. Note that, although the true values of SMI and MI differ (both measured through the genie-aided estimator), the comparison between SMI and MI estimators is fair since normalized bias and variance are used as performance indicators.
Fig. 5. Mean of the estimated SMI (up) and reduced-bias estimators (down) as a function of the true SMI for $\alpha = 1/3$, showing the role of $\sigma^2$ and $L$, with $N = 2 \lceil 7.5/\sigma \rceil + 1$.

Fig. 6. Normalized bias and normalized variance of the estimated SMI as a function of data size $L$ for $\alpha = 1/3$, $\sigma^2 = p/L^{2/5}$ and $N = 2 \lceil 7.5/\sigma \rceil + 1$. (a): SMI=1, MI=0.4, $p = 0.1$. (b): SMI=0.1, MI=0.05, $p = 0.25$. 
Finally, the performance of the approximate frequency-domain estimator is shown in Fig. 7 in terms of bias. It can be seen that, as the dimension increases, the performance of the approximate estimator converges to that of the original estimator, provided that a nonzero smoothing variance is used, with the advantage of a significantly reduced computational load. Note that the greater is the smoothing variance, the faster is the convergence rate of the frequency-domain estimator to the original performance, at the expense of an increased negative bias.

VI. CONCLUSION

In this paper, we have derived estimators of the degree of dependence of a pair of i.i.d. data, which are based solely on second-order statistics computed after mapping the data onto a complex space with higher dimension. The use of second-order statistics is possible as a result of selecting a surrogate of mutual information that is a quadratic measure of dependence.

In particular, it is shown that the squared-loss mutual information used in the field of machine learning corresponds to a second-order statistics based on the Frobenius norm of a coherence matrix, which is known to be directly linked with the standard CCA tool. The selected squared-loss surrogate has the property of upper-bounding Shannon mutual information. Moreover, it behaves as a local approximation of twice the Shannon mutual information, at the expense of an increased negative bias.

Finally, some pending issues are left for future work, such as the extension of the estimator to the case of data with memory, as proposed for instance in [51], and a data-dependent dimensionality reduction strategy prior to CCA, for which some preliminary results based on the minimum description length principle have recently been provided in [52].

VII. APPENDICES

Appendix A: Derivation of (10).

Defining $P_X$ as the probability measure, we have

$$ D_{\chi^2}(p_X||q_X) = \int \frac{p_X}{q_X} dP_X - 1 = \int \left( \frac{p_X}{q_X} - 2 \right) dP_X + 1 = \int \left( \frac{p_X^2}{q_X} - 2p_X q_X \right) dP_X \left. \right|_{p_X q_X} + \int \left( \frac{q_X^2}{p_X q_X} \right) dP_X $$

$$ = \int \left( \frac{p_X^2 - 2p_X q_X + q_X^2}{p_X q_X} \right) dP_X = \int \left( \frac{p_X q_X}{\sqrt{p_X q_X}} \right)^2 dP_X = \mathbb{E}_{p_X} \left( \frac{p_X(x) - q_X(x)}{\sqrt{p_X(x) q_X(x)}} \right)^2, $$

(63)

as written in [10].
Appendix B: Proof of Theorem 2

From (35), we have
\[ ||\hat{C}||^2 = \text{tr} \left( \hat{C}_{xy} \hat{R}_y^{-1} \hat{C}_{yy} \hat{R}_x^{-1} \right), \]  
(64)
where, using (33), \( \hat{R}_y = XX^H / L = F[\hat{p}]F^H \), \( \hat{R}_y = YY^H / L = G[\hat{q}]G^H \) and \( \hat{C}_{xy} = X (P_1^H Y) / L = F(\hat{J} - \hat{\beta} \hat{q}^T)G^H \).

Plugging them on (64), we have
\[ ||\hat{C}||^2 = \text{tr} \left( F(\hat{J} - \hat{\beta} \hat{q}^T) \hat{G}^H (G[\hat{q}]G^H)^{-1} G(\hat{J} - \hat{\beta} \hat{q}^T)^T F^H (F[\hat{p}]F^H)^{-1} \right) \]  
(65)
Using that \( F \) and \( G \) are invertible, we get \( ||\hat{C}||^2 = \text{tr} \left( F(\hat{J} - \hat{\beta} \hat{q}^T)[\hat{q}]^{-1} (\hat{J} - \hat{\beta} \hat{q}^T)^T [\hat{p}]^{-1} F^{-1} \right) \). Finally, the circularity of trace allows writing
\[ ||\hat{C}||^2 = \text{tr} \left( (\hat{J} - \hat{\beta} \hat{q}^T)[\hat{q}]^{-1} (\hat{J} - \hat{\beta} \hat{q}^T)^T [\hat{p}]^{-1} \right) = ||\hat{C}||^2, \]  
(66)
as we wanted to show.

Appendix C: Proof of Theorem 3

The following properties are used for the proof: \( \hat{q}^T 1_N = \hat{q}^T 1_M = 1 \), \( \hat{J} 1_M = \hat{p} \), \( 1_N^T \hat{J} = \hat{q}^T \), \( [\hat{p}] 1_N = \hat{p} \) and \( [\hat{q}] 1_M = \hat{q} \).

Clearly,
\[ \hat{J} - \hat{\beta} \hat{q}^T \]  
(67)
which means that \( 1_N \) and \( 1_M \) are left and right singular vectors of matrix \( (\hat{J} - \hat{\beta} \hat{q}^T) \), respectively, associated to its null singular value. From (66), then we can write
\[ ||\hat{C}||^2 = \text{tr} \left( (\hat{J} - \hat{\beta} \hat{q}^T)[\hat{q}]^{-1} + \frac{\beta}{1-\beta} 1_M 1_M^T \right) (\hat{J} - \hat{\beta} \hat{q}^T)^T [\hat{p}]^{-1} + \frac{\beta}{1-\beta} 1_N 1_N^T \),  
(68)
for any \( \beta \). From the Woodbury identity, we can write
\[ \left( \frac{\beta}{1-\beta} 1_M 1_M^T \right)^{-1} = [\hat{q}] - \frac{[\hat{q}] 1_M^T [\hat{q}]^T}{\beta^2 + \beta 1_M^T [\hat{q}]^T} = [\hat{q}] - \beta \hat{q} \hat{q}^T \]  
(69)
Clearly,
\[ \lim_{\beta \to 1} \left( \frac{\beta}{1-\beta} 1_M 1_M^T \right) 1_M = 0_M, \]  
(70)
\[ \lim_{\beta \to 1} \left( \frac{\beta}{1-\beta} 1_N 1_N^T \right) 1_N = 0_N, \]  
(71)
which means that these two matrices, which are sample covariance matrices for \( \beta \to 1 \), share with matrix \( \hat{J} - \hat{\beta} \hat{q}^T \) (see (67)) the same singular vectors associated to null singular value. Therefore, for the limiting case of \( \beta = 1 \), rank reduction using full-rank matrices \( F \in \mathbb{C}^{N' \times N} \) (with \( N' = N - 1 \)) and \( G \in \mathbb{C}^{M' \times M} \) (with \( M' = M - 1 \)) is possible, constrained to computing covariance instead of autocorrelation matrices from data, which proves the equality with the SMI. For \( N' < N - 1 \) and/or \( M' < M - 1 \), however, the smallest singular values will be lost, which proves the inequality.

Appendix D: Perturbation variance setting

For large \( L \), the bias and variance of the SMI estimator are given by
\[ \text{bias}(\hat{I}_s) = -O(\sigma^2) + O(\sigma^{-1} L^{-1}) \]  
(71)
\[ \text{var}(\hat{I}_s) = O(\sigma^{-1} L^{-1}). \]  
(72)
The term \( O(\sigma^2) \geq 0 \) is a result of the data processing inequality and the consistent (with \( L \)) terms \( O(\sigma^{-1} L^{-1}) \) decrease with \( \sigma \) as a result of (47). Both have also been approximately confirmed by simulations for a wide range of scenarios. As the mean squared error is \( \text{mse}(\hat{I}_s) = \text{bias}^2(\hat{I}_s) + \text{var}(\hat{I}_s) \), the condition \( \lim_{L \to \infty} \sigma^2 = \lim_{L \to \infty} \sigma^{-1} L^{-1} = 0 \) is required to yield \( \lim_{L \to \infty} \text{mse}(\hat{I}_s) = 0 \), which moves to choosing \( \sigma \) as a monotonically decreasing function of \( L \) such that \( \sigma^{-1} L^{-1} \) is also monotonically decreasing. Let us adopt a power law \( \sigma = O(L^{-\gamma}) \), for which the condition \( 0 < \gamma < 1 \) guarantees the desired convergence given by
\[ \text{mse}(\hat{I}_s) = O(L^{-\min[4\gamma,1-\gamma]}). \]  
(73)
Then, the value of $\gamma$ can finally be optimized by the following MiniMax rule:

$$\gamma = \arg \max_{\gamma} \min \{4\gamma, 1 - \gamma\} = \frac{1}{5}$$

(74)

similarly as the Silverman’s rule for kernel smoothing, which implies setting the perturbation variance as $\sigma^2 = p/(L^{2/5})$, being $p$ the new relative free parameter of the estimator.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: Wiley, 2006.
[2] Q. Wang, S. R. Kulkarni, and S. Verdú, “Universal estimation of information measures for analog sources.” Foundations and trends in Communications and Information Theory, 2009, no. 5-3.
[3] S. Verdú, “Empirical estimation of information measures: A literature guide,” Entropy, 21(8), 2019.
[4] J. C. Principe, *Information Theoretic Learning: Rényi’s Entropy and Kernel Perspectives*. New York: Springer, 2010.
[5] K.-C. Chen, S.-L. Huang, L. Zheng, and H. V. Poor, “Communication theoretic data analytics,” IEEE Journal on Selected Areas in Communications, vol. 33, no. 4, pp. 663–A, Apr. 2015.
[6] Y. C. Eldar, A. O. Hero III, I. 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],” IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 8–23, Nov. 2017.
[7] F. de Cabrer and J. Ribas, “Squared-loss mutual information via high-dimension coherence matrix estimation,” in IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP), 2019, pp. 5142–5146.
[8] S.-L. Huang, C. Suh, and L. Zheng, “Euclidean information network of theories,” IEEE Trans. Inf. Theory, vol. 61, no. 12, pp. 6795–6814, 2015.
[9] F. R. Bach and M. I. Jordan, “Kernel independence component analysis,” J. Mach. Learn. Res., vol. 3, pp. 1–48, 2002.
[10] H. Bier, J. Bohmann, and K. Müller, “On relevant dimensions in kernel feature spaces,” Journal of Machine Learning Research, vol. 9, pp. 1875 –1908, Jun. 2008.
[11] D. Ramírez, J. Via, I. Santamaría, and L. L. Scharf, “Locally most powerful invariant tests for correlation and sphericity of Gaussian vectors,” IEEE Trans. Inf. Theory, vol. 59, no. 4, pp. 2128–2141, Apr. 2013.
[12] Z. Goldfeld, K. Greenwald, J. Weed, and Y. Polyaniskiy, “Optimality of the plug-in estimator for differential entropy estimation under gaussian convolutions,” in IEEE International Symposium on Information Theory (ISIT), 2019, pp. 892–896.
[13] D. Ramírez, J. Via, I. Santamaría, and P. Crespo, “Entropy and Kullback-Leibler divergence estimation based on Szegő’s theorem,” in European Signal Processing Conference, EUSIPCO, 2009.
[14] S. Fehr and S. Berens, “On the conditional Rényi entropy,” IEEE Trans. Inf. Theory, vol. 60, no. 11, pp. 6801, 6810 2014.
[15] E. Parzen, “Empirical estimation of information measures: A literature guide,” Entropy, 21(8), 2019.
[16] M. Rosenblatt, “Remarks on some nonparametric estimates of a density function,” Ann. Math. Statist., no. 27, pp. 832–837, 1956.
[17] S. Seth, M. Rao, I. Park, and J. C. Principe, “A unified framework for quadratic measures of independence,” in IEEE International Symposium on Information Theory, 2017.
[18] F. de Cabrer, J. Ribas, and G. Vázquez, “Robust estimation of the magnitude squared coherence based on kernel signal processing,” in Proc. of the 51st Asilomar Conference on Signals, Systems and Computers, Pacific Grove (CA), IEEE, November 2017.
[19] ——, “Entropy-based covariance determinant estimation,” in IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2017.
[20] T. van Erven and P. Harremoës, “Rényi divergence and Kullback-Leibler divergence,” IEEE Trans. Inf. Theory, vol. 60, no. 7, pp. 3797–3820, July 2014.
[21] Y. Bao and H. Krim, “Rényi entropy based divergence measures for ICA,” in IEEE Stat. Signal Process. Workshop (SSP), vol. 28, 2003, pp. 565–568.
[22] J. P. Pluim, J. Maintz, and M. A. Viergever, “f-information measures in medical image registration,” IEEE Trans. Inf. Theory, vol. 23, no. 12, pp. 1508–1516, Dec. 2004.
[23] C. Kreucher, K. Kastella, and A. O. Hero, “Multi-target sensor management using alpha-divergence measures,” in Information Processing in Sensor Networks, S. B. Heidelberg, Ed., 2003.
[24] F. Liese and I. Vajda, “On divergences and informations in statistics and information theory,” IEEE Trans. Inf. Theory, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.
[25] S. Tridenski, R. Zamir, and A. Ingber, “The Ziv-Zakai-Rényi bound for joint source-channel coding,” IEEE Trans. Inf. Theory, vol. 61, no. 8, pp. 4293–4315, Aug. 2015.
[26] S. Verdú, “eo-mutual information,” in Information Theory and Applications Workshop (ITA), 2015, pp. 1–6.
[27] M. Tomamichel and M. Hayashi, “Operational interpretation of Rényi information measures via composite hypothesis testing against product and Markov distributions,” IEEE Trans. Inf. Theory, vol. 64, no. 2, pp. 1064–1082, 2018.
[28] A. Lapidoth and C. Pfister, “Two measures of dependence,” Entropy, no. 21, pp. 781, Aug. 2019.
[29] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese, “Mutual information estimation reveals global associations between stimuli and biological processes,” BMC Bioinformatics, vol. 10, no. 1, pp. S52 (12 pages), 2009.
[30] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, “Kernel measures of conditional dependence,” in Advances in Neural Information Processing Systems, pp. 489–496, 2008.
[31] S.-L. Huang and L. Zheng, “Linear information coupling problems,” in IEEE Int. Symp. on Information Theory (ISIT), 2015, pp. 74–78.
[32] H. Bretz, “Relations between two sets of variates,” Biometrika, vol. 28, no. 3–4, pp. 369–377, Dec. 1936.
[33] A. Peszlski, L. Scharf, M. R. Azimi-Sadjadi, and M. Lundberg, “Empirical canonical correlation analysis in subspaces,” in 38th Asilomar Conf. Signals, Syst. Comput., vol. 1, Nov. 2004, pp. 994–997.
[34] A. Rényi, “On measures of dependence,” in Acta Mathematica Academiae Scientiarum Hungarica, vol. 10, no. 3–4, 1959, pp. 441–451.
[35] N. Papadatos and T. Xifara, “A simple method for obtaining the maximal correlation coefficient and related characterizations,” Journal of Multivariate Analysis, no. 118, pp. 102–114, 2013.
[36] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, “On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Covet,” [Online]. Available: https://arxiv.org/abs/1304.6133, 2013.
[37] C. T. Li and A. E. Gamal, “Maximal correlation secrecy,” IEEE Trans. on Inf. Theory, vol. 64, no. 5, pp. 3916–3926, May 2018.
[38] S. Haykin, *Neural networks and learning machines*, 3rd ed. Prentice-Hall, 2009.
[39] J. L. Rojo-Alvarez, M. Martinez-Marton, J. Muñoz-Mari, and G. Camps-Valls, *Digital Signal Processing with Kernel Methods*, Wiley-Blackwell, Ed. IEEE Press, 2018.
[40] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” Technical Report CSD-TR-03-02, Royal Holloway University of London, 2003.
[41] S. Van Vaerenbergh, J. Via, and I. Santamaría, “Blind identification of SIMO Wiener systems based on kernel canonical correlation analysis,” IEEE Transactions on Signal Processing, vol. 61, no. 9, pp. 2219–2230, 2013.
[42] 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), 2018, pp. 2608–2612.

[43] J.-F. Collet, “An exact expression for the gap in the data processing inequality for f-divergences,” IEEE Trans. Inf. Theory, vol. 65, no. 7, pp. 4387–4391, July 2019.

[44] U. Grenander and G. Szegö, Toeplitz forms and their applications. Berkeley: Univ. Calif. Press, 1958.

[45] R. Gray, Toeplitz and circulant matrices: a review. Foundations and trends in Communications and Information Theory, 2006.

[46] B. Silverman, Density estimation for statistics and data analysis. Chapman and Hall, London, 1986.

[47] S. Chen, “Optimal bandwidth selection for kernel density functionals estimation (https://doi.org/10.1155/2015/242683),” Journal of probability and statistics, vol. 2015, no. ID 242683, pp. 1–22, 2015.

[48] C. L. Haley and M. Anitescu, “Optimal bandwidth for multitaper spectrum estimation,” in IEEE Signal Processing Letters, vol. 24, no. 11, pp. 1696-1700, Nov. 2017.

[49] G. A. Darbellay and I. Vajda, “Estimation of the information by an adaptive partitioning of the observation space,” IEEE Trans. Inf. Theory, vol. 45, no. 4, pp. 1315–1321, May 1999.

[50] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” Physical Review E., id. 066138 (https://arxiv.org/abs/cond-mat/0305641), vol. 69, 2004.

[51] R. Malladi, D. H. Johnson, G. P. Kalamangalam, N. Tandon, and B. Aazhang, “Mutual information in frequency and its application to measure cross-frequency coupling in epilepsy,” in IEEE Transactions on Signal Processing, vol. 66, no. 11, Jun. 2018, pp. 3008–3023.

[52] C. A. López, F. de Cabrera, and J. Riba, “Estimation of information in parallel Gaussian channels via model order selection,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.