Accepted version of the article published in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019. DOI: 10.1109/ICASSP.2019.8682642. Available at: http://ieeexplore.ieee.org

Squared-Loss Mutual Information via High-Dimension Coherence Matrix Estimation

Ferran de Cabrera, *Student Member, IEEE* and Jaume Riba, *Senior Member, IEEE*
Signal Theory and Communications Department, Technical University of Catalonia (SPCOM/UPC)
{ferran.de.cabrera, jaume.riba}@upc.edu

***Abstract*—Squared-loss mutual information (SMI) is a surrogate of Shannon mutual information that is more advantageous for estimation. On the other hand, the coherence matrix of a pair of random vectors, a power-normalized version of the sample cross-covariance matrix, is a well-known second-order statistic found in the core of fundamental signal processing problems, such as canonical correlation analysis (CCA). This paper shows that SMI can be estimated from a pair of independent and identically distributed (i.i.d.) samples as a squared Frobenius norm of a coherence matrix estimated after mapping the data onto some fixed feature space. Moreover, low computation complexity is achieved through the fast Fourier transform (FFT) by exploiting the Toeplitz structure of the involved autocorrelation matrices in that space. The performance of the method is analyzed via computer simulations using Gaussian mixture models.**

***Index Terms*—Squared-loss mutual information, coherence matrix, canonical correlation analysis, Gaussian mixture models, characteristic function.**

I. Introduction

The estimation of information measures is an important task required in numerous signal processing and machine learning applications [1]. Although Shannon's mutual information (MI) plays an important role in information and communications theory, its estimation from finite realizations of random variables presents difficulties, complexities and small robustness to outliers due to the "log" involved in its definition. To cope with these problems, in the last years, researches have proposed some surrogate contrast functions as valuable metrics for feature selection. Among them, the so-called quadratic measures of independence, such as quadratic MI (QMI) and squared-loss MI (SMI) [2], play an important role because, when they are coupled with Parzen probability density function (PDF) estimators, they lead to kernel-based signal processing methods [3], [4]. Another prominent example of the idea is the Hilbert-Schmidt independence criterion (HSIC) [5]. Although this general idea solves some complexity issues when dealing with data, the computational complexity of kernel methods is still proportional to the squared data size [3], which compromises its direct application to large data records in a flexible manner. This issue is explored in [6], in which the computational complexity is reduced by a feature space based on an autoregressive parametrization of the PDFs, but focused on building an estimate of Shannon's entropy and Kullback-Leibler (KL) divergence, which are unidimensional information-theoretic measures.

The purpose of this paper is to show (Sec. III) that the discrete SMI can be interpreted as a squared Frobenius norm of a coherence matrix (CM) computed after assigning the symbols of each source to vectors on an arbitrary orthonormal basis. From this interpretation, and motivated by some enlightened links with other fundamental statistics used in detection, estimation and information theory, the purpose is then to show (Sec. IV) that the continuous version of the SMI can be estimated in a similar way to the discrete case after a feature map that assigns the values of each random variable to vectors of fixed dimensionality, a fact that offers significant computational complexity savings. Moreover, we propose a specific feature map based on the empirical characteristic function (CF), which provides (Sec. V) further computational savings as a result of the Toeplitz structure of the involved autocorrelation matrices.

II. Squared-Loss Mutual Information

Consider a pair of memory-less sources X and Y. For discrete sources with alphabets $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$, the SMI is defined as [2]:

$$I_s(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \left(\frac{P_{XY}(x,y) - P_X(x)P_Y(y)}{\sqrt{P_X(x)P_Y(y)}} \right)^2, \quad (1)$$

where $P_X(x) = \Pr\{X = x\}$ and $P_Y(y) = \Pr\{Y = y\}$ are the marginal probability mass functions (PMF) and $P_{XY}(x,y) = \Pr\{X = x, Y = y\}$ is the joint PMF. Similarly, for continuous sources, the SMI is given by:

$$I_s(X;Y) = \int\int \left(\frac{p_{XY}(x,y) - p_X(x)p_Y(y)}{\sqrt{p_X(x)p_Y(y)}} \right)^2 dxdy, \quad (2)$$

where $p_X(x)$ and $p_Y(y)$ are marginal PDFs and $p_{XY}(x,y)$ is the joint PDF. Both expressions can be obtained by upper-bounding the natural "log" operator in the MI expression as $\ln(x) \leq x - 1$, for which SMI becomes a quantity lower-bounded by Shannon MI (that is, $I_s(X;Y) \geq I(X;Y)$) expressed in *nats*. Whereas the ordinary MI is the Kullback-Leibler divergence from $p_{XY}(x,y)$ to $p_X(x)p_Y(y)$, SMI is the Pearson chi-squared divergence [7] and operates as a local approximation of MI [8]. The main advantage of SMI as a measure of information is the lack of the "log" operator inside the summation, which is present in the MI definition, endowing it good properties for estimation purposes.

This work is supported by projects TEC2016-76409-C2-1-R (WINTER), Ministerio de Economia y Competividad, Spanish National Research Plan, and 2017 SGR 578 - AGAUR, Catalan Government.

III. Discrete SMI via estimation of coherence

In the discrete case, by defining the marginal probability column vectors $[\tilde{\mathbf{p}}]_n = P_X(x_n)$ and $[\tilde{\mathbf{q}}]_m = P_Y(y_m)$, and the joint probability matrix $[\tilde{\boldsymbol{J}}]_{n,m} = P_{XY}(x_n, y_m)$, the SMI given in (1) can be compactly expressed as

$$I_s(X;Y) = \sum_{n=1}^{N}\sum_{m=1}^{M} |[\tilde{\mathbf{C}}]_{n,m}|^2 = \text{tr}\left(\tilde{\mathbf{C}}^T\tilde{\mathbf{C}}\right) = ||\tilde{\mathbf{C}}||_F^2, \quad (3)$$

where $(.)^T$ denotes transpose, tr(.) denotes the trace, $||.||_F$ denotes the Frobenius norm, and matrix $\tilde{\mathbf{C}} \in \mathbb{R}^{N\times M}$ is defined as

$$\tilde{\mathbf{C}} = [\tilde{\mathbf{p}}]^{-1/2}\left(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T\right)[\tilde{\mathbf{q}}]^{-1/2}, \quad (4)$$

where $[\tilde{\mathbf{p}}]$ is the diagonal matrix with entries $\{[\tilde{\mathbf{p}}]_n\}_{1\leq n\leq N}$ and $[\tilde{\mathbf{q}}]$ is defined similarly.

Let $\mathbf{F} \in \mathbb{C}^{N\times N}$ and $\mathbf{G} \in \mathbb{C}^{M\times M}$ be unitary matrices, formed by the set of columns $\mathcal{F} = \{\mathbf{f}_n\}_{n=1,2,\ldots,N} \in \mathbb{C}^{N\times 1}$ and $\mathcal{G} = \{\mathbf{g}_m\}_{m=1,2,\ldots,M} \in \mathbb{C}^{M\times 1}$, respectively. Using trace and unitarity properties, SMI can be alternatively expressed as $I_s(X;Y) = \text{tr}(\mathbf{C}^H\mathbf{C}) = ||\mathbf{C}||_F^2$ with

$$\mathbf{C} = \mathbf{F}\tilde{\mathbf{C}}\mathbf{G}^H = \mathbf{P}^{-1/2}\left(\mathbf{J} - \mathbf{p}\mathbf{q}^H\right)\mathbf{Q}^{-1/2}, \quad (5)$$

where $(.)^H$ denotes Hermitian transpose, $\mathbf{p} = \mathbf{F}\tilde{\mathbf{p}}$, $\mathbf{q} = \mathbf{G}\tilde{\mathbf{q}}$, $\mathbf{P} = \mathbf{F}[\tilde{\mathbf{p}}]\mathbf{F}^H$, $\mathbf{Q} = \mathbf{G}[\tilde{\mathbf{q}}]\mathbf{G}^H$ and $\mathbf{J} = \mathbf{F}\tilde{\mathbf{J}}\mathbf{G}^H$. To provide an interpretation of the involved vectors and matrices in (5), let us construct the random vectors $\mathbf{x} \in \mathcal{F}$ and $\mathbf{y} \in \mathcal{G}$ from the discrete sources X and Y by the one-to-one mappings $\phi_X(.) : \mathcal{X} \to \mathcal{F}$ and $\phi_Y(.) : \mathcal{Y} \to \mathcal{G}$ defined as

$$x_n \to \mathbf{f}_n \qquad y_m \to \mathbf{g}_m \ , \quad (6)$$

respectively. Clearly, the expectation theorem allows to write $\mathbf{p} = E[\mathbf{x}]$, $\mathbf{q} = E[\mathbf{y}]$, $\mathbf{P} = E\left[\mathbf{x}\mathbf{x}^H\right]$, $\mathbf{Q} = E\left[\mathbf{y}\mathbf{y}^H\right]$, and $\mathbf{J} = E\left[\mathbf{x}\mathbf{y}^H\right]$, which means that all of them can be seen as first and second order statistics of the data after a feature map as described above. Effectively, matrix $(\mathbf{J} - \mathbf{p}\mathbf{q}^H)$ is the cross-covariance matrix of the vector data mapped onto the feature space, while $\mathbf{P}$ and $\mathbf{Q}$ are the autocorrelation matrices.

Expressions (4) & (5) are relevant due to their intimate link with other fundamental results, which is worth to highlight. In particular, (4) plays an important role in Euclidean information theory [8], which results from a local approximation of MI based on information bottleneck [9] or linear information coupling. To see the link, let us write $\tilde{\mathbf{C}} = \mathbf{B} - \tilde{\mathbf{p}}^{1/2}\tilde{\mathbf{q}}^{H/2}$, where $\mathbf{B} = [\tilde{\mathbf{p}}]^{-1/2}\tilde{\mathbf{J}}[\tilde{\mathbf{q}}]^{-1/2}$ is the divergence transition matrix (DTM) [8] of a discrete memory-less communication channel, a linear map endowed with insightful geometrical interpretations. It is shown in [8] that the largest singular value of $\mathbf{B}$ is 1, with associated right and left singular vectors $\tilde{\mathbf{p}}^{1/2}$ and $\tilde{\mathbf{q}}^{1/2}$, respectively. As a direct consequence of this property, our matrix $\tilde{\mathbf{C}}$ defined in (4) verifies that $\text{rank}(\tilde{\mathbf{C}}) = \min(N,M)-1$ and, therefore, $0 \leq I_s(X;Y) \leq \min(N,M)-1$. Moreover, the second largest singular value of $\mathbf{B}$, which corresponds to the largest singular value of $\mathbf{C}$, is the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation coefficient $0 \leq \rho(X;Y) \leq 1$, an information measure that has found interesting applications in information theory [10].

Finally, expression (5) can be linked as well with other signal processing concepts. Effectively, a direct application of Woodbury identity (see Appendix) reveals that $\mathbf{C} = (\mathbf{P} - \mathbf{p}\mathbf{p}^H)^{+/2}(\mathbf{J} - \mathbf{p}\mathbf{q}^H)(\mathbf{Q} - \mathbf{q}\mathbf{q}^H)^{+/2}$ (where $+$, denoting pseudo-inverse, is required to deal with the rank deficient structure of the auto-covariance matrices $\mathbf{P} - \mathbf{p}\mathbf{p}^H$ and $\mathbf{Q} - \mathbf{q}\mathbf{q}^H$), and therefore it is a true coherence matrix. This observation tells that HGR and SMI can both be obtained through canonical correlation analysis (CCA) [11] as

$$\rho(X;Y) = \lambda_{\max}(\mathbf{C}) \quad (7)$$

$$I_s(X;Y) = \sum_{i=1}^{\min(N,M)-1} |\lambda_i(\mathbf{C})|^2, \quad (8)$$

respectively, where $0 \leq \lambda_i \leq 1$ (with $\lambda_{i+1} \leq \lambda_i$, $\forall i$) are the real, non-negative singular values of matrix $\mathbf{C}$ in (5), which coincide with those of $\tilde{\mathbf{C}}$ in (4), and $\lambda_{\min(N,M)} = \lambda_{\min}(\mathbf{C}) = 0$ for the reasons explained above. Finally, (5) is also related with the seminal work [12] where, using Wijsman's theorem, authors proved that the Frobenius norm of the sample coherence is the locally most powerful invariant test (LMPIT) for the detection of correlation in Gaussian vectors.

IV. SMI via empirical characteristic function

For the continuous case (2), our aim is to obtain an estimator still based on the Frobenius norm of a coherence matrix. Although in that case the dimension of the feature space would need to be infinite (as the alphabet size is infinite) in order to obtain the exact SMI, the main goal is to understand which is the performance/complexity trade-off achieved by using a finite feature space dimension.

Before going into detail, it is worth emphasizing that, from a completely different perspective, the core of the basic idea described in this section is well-known in the machine learning literature. Specifically, kernel methods are based on an implicit feature map onto an infinite dimensional space that relies on the "kernel trick" supported by the reproducing kernel Hilbert spaces (RKHS) theory [13]. Despite this powerful idea, regularization [3] is ultimately needed anyway in kernel signal processing to avoid overfitting caused by the excessive dimensionality. In this sense, the fixed feature map proposed in the sequel can be seen as an alternative way of regularizing the problem from scratch in order to obtain overall computational benefits, thus avoiding the need of implementing speed up techniques such as the incomplete Cholesky factorization [3].

Among different options, let us construct the random vectors $\mathbf{x} \in \mathbb{C}^{N\times 1}$ and $\mathbf{y} \in \mathbb{C}^{N\times 1}$ from the random variables X and Y by the mappings $\phi_X(.) : \mathcal{R} \to \mathbb{C}^{N\times 1}$ and $\phi_Y(.) : \mathcal{R} \to \mathbb{C}^{N\times 1}$ defined as

$$x \to e^{j\alpha\mathbf{n}x} \qquad y \to e^{j\alpha\mathbf{n}y} \ , \quad (9)$$

respectively, where $\mathbf{n} \in \mathbb{Z}^{N\times 1}$ is a vector of integers defined as $\mathbf{n} = [-K, -K+1, \cdots, K]^T$ with $N = 2K+1$ and $K \geq 1$. Based on the mapping suggested in (6) for the discrete case, the rationale is now to use the mapping from (9) for the continuous case while still estimating (5). Note that, by doing so, we are relaxing the orthogonal constraint satisfied in the discrete case to an asymptotically orthogonal constraint for

the continuous case (which means that the form (5) used with the mapping (9) will yield the true SMI for $N \to \infty$), aimed at obtaining a balance between accuracy and complexity. An insightful interpretation of (9) is that the first and second order statistics computed at the feature space are both obtained from a uniform sampling of the marginal and joint empirical CF of the original random variables, being α the sampling period. For a more in-depth rationale of (9) within the framework of independence detection, the reader is referred to [14].

In summary, assuming that L independent and identically distributed (i.i.d.) samples $\{x(l), y(l)\}_{0\leq l\leq L-1}$ from random variables X and Y are available, we obtain the $2K+1$ dimensional vector sequences $\mathbf{x}(l) = e^{j\alpha \mathbf{n} x(l)}$ and $\mathbf{y}(l) = e^{j\alpha \mathbf{n} y(l)}$ according to (9) and define the CF-based SMI as $\hat{I}_{cs}(X;Y) = ||\hat{\mathbf{C}}||_F^2$, where

$$\hat{\mathbf{C}} = \hat{\mathbf{P}}^{-1/2}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^H\right)\hat{\mathbf{Q}}^{-1/2} \tag{10}$$

is the sample CM of the mapped data, with $\hat{\mathbf{p}} = \langle \mathbf{x}(l)\rangle_L$, $\hat{\mathbf{q}} = \langle \mathbf{y}(l)\rangle_L$, $\hat{\mathbf{P}} = \left\langle \mathbf{x}(l)\mathbf{x}^H(l)\right\rangle_L$, $\hat{\mathbf{Q}} = \left\langle \mathbf{y}(l)\mathbf{y}^H(l)\right\rangle_L$, and $\hat{\mathbf{J}} = \left\langle \mathbf{x}(l)\mathbf{y}^H(l)\right\rangle_L$, where $\langle .(l)\rangle_L = L^{-1}\sum_{l=0}^{L-1} .(l)$ generally denotes the L-length sample mean operator.

V. SMI IN LARGE FEATURE SPACE DIMENSION REGIME

The main bottleneck in (10) is the inversion of autocorrelation matrices, specially for large dimension. Note that large dimensions of feature spaces are needed to discover complex non-linear associations present in the data. In this regime, a further advantage of the map (9) is that the sample autocorrelation matrices $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ have a Toeplitz structure. This can be easily seen as follows:

$$\hat{\mathbf{P}} = \left\langle e^{j\alpha \mathbf{n} x(l)} e^{-j\alpha \mathbf{n}^T x(l)}\right\rangle_L = \text{toe}\left(\hat{\mathbf{p}}_a\right), \tag{11}$$

where toe(.) denotes a Toeplitz and Hermitian matrix constructed from its first column vector, $\hat{\mathbf{p}}_a = \left\langle e^{j\alpha \mathbf{n}_a x(l)}\right\rangle_L$, and $\mathbf{n}_a \in \mathbb{Z}^{N\times 1}$ is an asymmetric vector of integers defined as $\mathbf{n}_a = [0, 1, \cdots, 2K]^T$. Similarly, $\hat{\mathbf{Q}} = \text{toe}\left(\hat{\mathbf{q}}_a\right)$, with $\hat{\mathbf{q}}_a = \left\langle e^{j\alpha \mathbf{n}_a y(l)}\right\rangle_L$.

The Toeplitz structure provides two main advantages to reduce complexity. On the one hand, it turns out that second order statistics $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ of the mapped data can be constructed solely from their (extended) first order statistics through (11). On the other hand, assuming that marginal PDFs are square integrable, i.e. $\int p_X^2(x)dx < \infty$ and $\int p_Y^2(y)dy < \infty$, then Parseval's theorem implies that CFs are also square integrable and, therefore, $\lim_{w\to\pm\infty} E\left[e^{jwx}\right] = \lim_{w\to\pm\infty} E\left[e^{jwy}\right] = 0$, which ensures the off-diagonal decay in $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$. Under these conditions, Szegö's theorem [15] establishes that, for large dimension, Toeplitz matrices are asymptotically diagonalizable by the unitary Fourier matrix, and its eigenvalues asymptotically behave like samples of the Fourier transform of its entries. This fact naturally motivates a frequency-domain analysis to reduce complexity in the large-dimension regime, similar to the traditional analysis of large stationary correlated processes. In particular, as the Frobenius norm is invariant under unitary transformations, we propose a computationally effective approximation of the CF-based SMI for the case of very large N as $\hat{I}_{acs}(X;Y) = ||\hat{\mathbf{C}}'||_F^2$, where

$$\hat{\mathbf{C}}' = [\hat{\mathbf{p}}']^{-1/2}\mathbf{U}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^H\right)\mathbf{U}^H[\hat{\mathbf{q}}']^{-1/2}, \tag{12}$$

with $\mathbf{U}$ the unitary Fourier matrix, $\hat{\mathbf{p}}' = \text{diag}(\mathbf{U}\hat{\mathbf{P}}\mathbf{U}^H)$, and $\hat{\mathbf{q}}' = \text{diag}(\mathbf{U}\hat{\mathbf{Q}}\mathbf{U}^H)$, and diag(.) denotes a vector formed by the main diagonal entries of a matrix, i.e. $[\text{diag}(.)]_n = [.]_{n,n}$.

Finally, inspired on the duality that this problem holds with classical spectral estimation theory [16], vectors $\hat{\mathbf{p}}'$ and $\hat{\mathbf{q}}'$ can be efficiently computed as Fourier transforms of "windowed" empirical CF samples:

$$\hat{\mathbf{p}}' = 2\sqrt{N}\text{Re}\left[\mathbf{U}^H\left(\hat{\mathbf{p}}_a \odot \mathbf{w}\right)\right] - 1 \tag{13a}$$

$$\hat{\mathbf{q}}' = 2\sqrt{N}\text{Re}\left[\mathbf{U}^H\left(\hat{\mathbf{q}}_a \odot \mathbf{w}\right)\right] - 1, \tag{13b}$$

where $\mathbf{w}$ is a unilateral triangular window with elements $\{[\mathbf{w}]_n\}_{0\leq n\leq N-1} = 1 - n/N$ and $\odot$ denotes the Hadamard product. Note that (12), (13a) and (13b) can be computed very efficiently by means of the fast Fourier transform (FFT) algorithm, which is crucial in the large-dimension regime. Interestingly, (13a) and (13b) can be seen as Blackman-Tukey [16] PDF estimators from finite samples of the empirical CFs, a dual version of the classical nonparametric spectral estimation problem from finite samples of the empirical autocorrelation function of stationary data.

VI. SIMULATION RESULTS

To check the validity of the SMI estimator, we use the Gaussian mixture models (GMM) recently proposed in [14]. Correlated and uncorrelated data is modeled as $\boldsymbol{z}_c \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\rho)$ and $\boldsymbol{z}_u \sim 0.5\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\rho) + 0.5\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{-\rho})$, respectively, with $\boldsymbol{\Sigma}_\rho = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. It is not difficult to show that the true SMI for these models is $I_{s_c}(X;Y) = \rho^2/(1-\rho^2)$ and $I_{s_u}(X;Y) = \rho^4/(1-\rho^4)$, which provides a means for performance evaluation.

The sampling period α in (9) is chosen in order to minimize the aliasing induced by sampling the CF, which should be lower than the inverse of the standard deviation of the data. The data is normalized to unit variance to ensure that the sampling period is the same for any given ρ, which is set to $\alpha = 0.1$. On the other hand, N needs to be sufficiently large in order to reduce the implicit smearing of the PDF. Specifically, we want the mapping with overall support αN on the CF to minimize the smearing, whose impact is dominant on the narrowest direction of the joint PDF and is characterized by the smallest eigenvalue of the covariance matrix $\lambda_{\min}(\boldsymbol{\Sigma}_\rho)$. In general we will choose the minimum feature space dimension $N_{\min}$ in order to handle the worst case in terms of PDF resolution, which corresponds to the largest effective support of the CF. Therefore, $N_{\min}$ depends on the maximum $|\rho|$ of the data being measured. As a result we have $N_{\min} = 1/(\alpha\lambda_{\min}(\boldsymbol{\Sigma}_\rho))$, being $\lambda_{\min}(\boldsymbol{\Sigma}_\rho) = 1 - |\rho|_{\max}$.

Figure 1 shows the mean value of the estimated SMIs for different ρ^2 by means of Monte Carlo simulations, compared with the least squares mutual information estimator (LSMI)

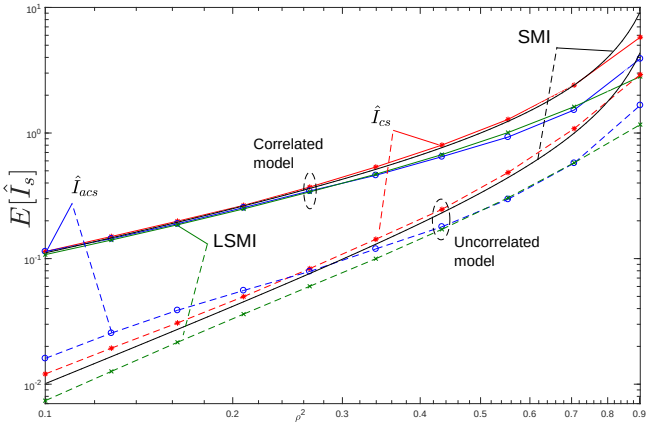

Fig. 1. Mean of the estimated SMI vs. ρ^2 for both models using $\hat{I}_{cs}$ $(N = 75)$, $\hat{I}_{acs}$ $(N = 151)$ and LSMI for $L = 10^5$. The true SMI is also shown.

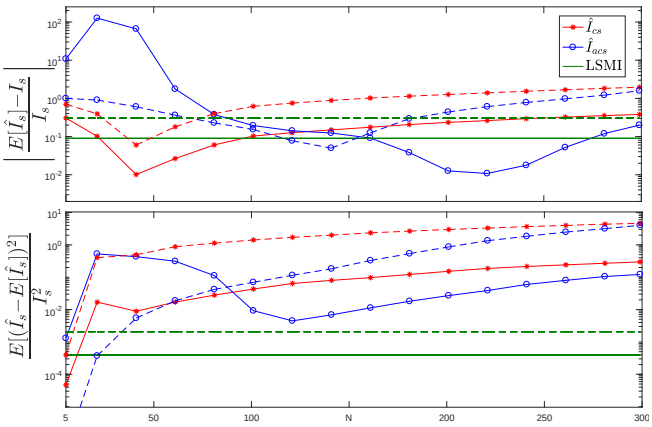

Fig. 2. Normalized bias and variance of $\hat{I}_{cs}$ and $\hat{I}_{acs}$ vs. N for $\rho^2 = 0.25$, jointly with the LSMI for $L = 10^5$. Continuous and dashed lines represent the correlated and uncorrelated models, respectively.

[2]. The LSMI kernel bandwidth is selected through cross-validation as in [17], and the regularization parameter is 10^{-4}. In general, a bias can be observed that is accentuated for increasing values of ρ due to the impact of the implicit smearing, which shows the difficulty of estimating the SMI when it is large. However, $\hat{I}_{cs}$ provides a better estimate than the LSMI, specifically in this context of high SMI. Regarding the Szegö's approach, note that $\hat{I}_{acs}$ is evaluated with a higher value of N than that used for $\hat{I}_{cs}$ in order to observe the desired behavior, thus providing a fair comparison.

Figure 2 shows the normalized bias and normalized variance for both models as a function of N. On the one hand, $\hat{I}_{cs}$ provides a reduced bias for $N > N_{\min}$ needed to ensure sufficiently small smearing at $\rho = 0.5$, which in this case corresponds to $N_{\min} = 20$. On the other hand, the variance of $\hat{I}_{cs}$ is increased for higher N and, since the measurements are strictly overestimating the true SMI at low ρ values, the bias is also increased. As a consequence of this effect, the bias can be reduced by increasing L. Although LSMI provides less variance, its regularization parameter is very sensitive in terms of bias due to overfitting, which is usually selected through cross-validation. On the contrary, $\hat{I}_{cs}$ admits a more insightful selection of the algorithm parameters, directly guided from classical concepts shared by spectral analysis such as smearing and leakage. Concerning to $\hat{I}_{acs}$, higher N tend to approximate both bias and variance to those of $\hat{I}_{cs}$, hence confirming the asymptotic behavior. However, the optimal N in terms of bias is higher than the $N_{\min}$ for $\hat{I}_{cs}$ since the Toeplitz matrices need to be sufficiently large in order to ensure a good approximate of the diagonalization through the Fourier transform.

VII. Conclusions

This paper derived an estimator of the squared-loss mutual information (SMI) from i.i.d. data. The advantage of the estimator is twofold: a lower computational complexity due to the limitation of the feature space dimension, and a computationally efficient algorithm based on the Szegö's theorem and on the fast Fourier transform. Both approaches presented in this paper provide lower bias at the cost of an increased variance, but their parameters can be selected based on dual ideas from spectral analysis, thus avoiding the need for parameter tuning through cross-validation. As a future work, the bias and variance of the estimator can be handled by the means of a tapering function over the characteristic function space, which would lead to a more controllable estimate for any feature space dimension selection.

Appendix

From (3) and (4), and using trace properties,

$$I_s(X;Y) = \operatorname{tr}\left((\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)[\tilde{\mathbf{q}}]^{-1}(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)^T[\tilde{\mathbf{p}}]^{-1}\right). \quad (14)$$

Let $\mathbf{0}_D$ and $\mathbf{1}_D$ be $D \times 1$ vectors with all-zeros and all-ones. As $\tilde{\mathbf{q}}^T\mathbf{1}_M = 1$ (unit-area of PDF) and $\tilde{\mathbf{J}}\mathbf{1}_M = \tilde{\mathbf{p}}$ (joint PDF marginalization), we have $(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)\mathbf{1}_M = \mathbf{0}_N$ and $\mathbf{1}_N^T(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T) = \mathbf{0}_M^T$. Since $\mathbf{1}_N^T$ and $\mathbf{1}_M$ are the left and right singular vectors of $(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)$ associated to a null singular value,

$$I_s(X;Y) = \operatorname{tr}\Big((\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)\left([\tilde{\mathbf{q}}]^{-1} + \varepsilon\mathbf{1}_M\mathbf{1}_M^T\right) \times(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)^T\left([\tilde{\mathbf{p}}]^{-1} + \varepsilon\mathbf{1}_N\mathbf{1}_N^T\right)\Big) \quad (15)$$

for any ε. Using Woodbury identity and that $[\tilde{\mathbf{q}}]\mathbf{1}_M = \tilde{\mathbf{q}}$,

$$\left([\tilde{\mathbf{q}}]^{-1} + \varepsilon\mathbf{1}_M\mathbf{1}_M^T\right)^{-1} = [\tilde{\mathbf{q}}] - \frac{[\tilde{\mathbf{q}}]\mathbf{1}_M\mathbf{1}_M^T[\tilde{\mathbf{q}}]}{\frac{1}{\epsilon} + \mathbf{1}_M^T[\tilde{\mathbf{q}}]\mathbf{1}_M} = [\tilde{\mathbf{q}}] - \frac{\tilde{\mathbf{q}}\tilde{\mathbf{q}}^T}{\frac{1}{\epsilon} + 1}, \quad (16)$$

and similarly for $\left([\tilde{\mathbf{p}}]^{-1} + \varepsilon\mathbf{1}_N\mathbf{1}_N^T\right)^{-1}$. Now, for $\varepsilon \to \infty$,

$$(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)[\tilde{\mathbf{q}}]^{-1} = \lim_{\varepsilon\to\infty}(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)\left([\tilde{\mathbf{q}}]^{-1} + \varepsilon\mathbf{1}_M\mathbf{1}_M^T\right) \quad (17)$$
$$= \lim_{\beta\to 1}(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)\left([\tilde{\mathbf{q}}] - \beta\tilde{\mathbf{q}}\tilde{\mathbf{q}}^T\right)^{-1} = (\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)\left([\tilde{\mathbf{q}}] - \tilde{\mathbf{q}}\tilde{\mathbf{q}}^T\right)^+,$$

where $+$ is introduced because $\left([\tilde{\mathbf{q}}] - \tilde{\mathbf{q}}\tilde{\mathbf{q}}^T\right)$ has a null eigenvalue with eigenvector $\mathbf{1}_M$, and similarly for $(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)^T[\tilde{\mathbf{p}}]^{-1}$. Then,

$$I_s(X;Y) = \operatorname{tr}\Big((\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)\left([\tilde{\mathbf{q}}] - \tilde{\mathbf{q}}\tilde{\mathbf{q}}^T\right)^+ \times(\tilde{\mathbf{J}} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T)^T\left([\tilde{\mathbf{p}}] - \tilde{\mathbf{p}}\tilde{\mathbf{p}}^T\right)^+\Big), \quad (18)$$

and, as $\mathbf{F}$ and $\mathbf{G}$ from (5) are complex and unitary, we obtain

$$I_s(X;Y) = \operatorname{tr}\Big((\mathbf{J} - \mathbf{p}\mathbf{q}^H)\left(\mathbf{P} - \mathbf{p}\mathbf{p}^H\right)^+ \times(\mathbf{J} - \mathbf{p}\mathbf{q}^H)^H\left(\mathbf{Q} - \mathbf{q}\mathbf{q}^H\right)^+\Big). \quad (19)$$

REFERENCES

[1] Q. Wang, S. R. Kulkarni, and S. Verdú, *Universal estimation of information measures for analog sources*, N. P. Inc., Ed. Foundations and trends in Communications and Information Theory, 2009.

[2] 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, p. S52 (12 pages), 2009.

[3] J. C. Príncipe, *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives*. NewYork: Springer, 2010.

[4] S. Seth, M. Rao, I. Park, and J. C. Príncipe, "A unified framework for quadratic measures of independence," *IEEE Trans. on Signal Process.*, vol. 59, no. 8, pp. 3624–3635, Aug. 2011.

[5] A. Gretton, A. Smola, O. Bousquet, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," Max Planck Institute for Biological Cybernetics, Tübingen, Technical report, 2005.

[6] D. Ramírez, J. Vía, I. Santamaría, and P. Crespo, "Entropy and Kullback-Leibler divergence estimation based on Szegö's theorem," *17th European Signal Processing Conference, EUSIPCO*, 2009.

[7] J. Sainui and M. Sugiyama, "Direct approximation of quadratic mutual information and its application to dependence-maximization clustering," *IEICE Transactions on Information and Systems*, vol. E96-D, no. 10, pp. 2282–2285, Oct. 2012.

[8] S. L. Huang, C. Suh, and L. Zheng, "Euclidean information theory of networks," *IEEE Trans. on Inf. Theory*, vol. 61, no. 12, pp. 6795–6814, Dec. 2015.

[9] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, (Allerton House, UIUC, Illinois, USA)*, 1999.

[10] A. Makur, "A study of local approximations in information theory," Master Thesis, June 2013.

[11] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, Dec. 1936.

[12] D. Ramírez, J. Vía, I. Santamaría, and 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.

[13] B. Schölkopf and A. J. Smola, *Learning with kernels*. MIT Press, 2002.

[14] F. de Cabrera and J. Riba, "A novel formulation of independence detection based on the sample characteristic function," *26th European Signal Processing Conference, EUSIPCO*, Sep. 2018.

[15] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*. Foundations and trends in Communications and Information Theory, 2006.

[16] S. M. Kay, *Modern spectral estimation: theory and application*. Prentice-Hall, 1988.

[17] M. Sugiyama and T. Suzuki, "Least-squares independence test," *IEICE Transactions on Information and Systems*, vol. E94-D, no. 6, pp. 1333–1336, 2011.