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

Ferran de Cabrera and Jaume Riba

Technical University of Catalonia, Department of Signal Theory and Communications

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Teoria del Senyal i Comunicacions

Objectives

This work addresses the estimation of the squared-loss mutual information (SMI), focusing on:

- Define an estimate of a surrogate of the well-known mutual information that acts as a valuable metric in signal processing applications.
- Interpret the SMI as the Frobenius norm of a coherence matrix, with direct relations with other fields on information theory.
- Propose the empirical characteristic function as an effective mapping for this task.
- Reduce computational complexity by limiting the feature space dimension, avoiding the kernel methods mapping.

Introduction

The estimation of information-theoretic measures is an important task required in numerous signal processing and machine learning applications. However, the estimation of the well-known Shannon's mutual information from finite realizations is a difficult task. To cope with this problem, the squared-loss mutual information (SMI) has been proposed as a substitute metric:

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

It is worth noting that whereas Shannon's mutual information is the Kullback-Leibler divergence from $p_{XY}(x,y)$ to $p_X(x)p_Y(y)$, the SMI is the Pearson chi-squared divergence and operates as a local approximation of MI. This work shows that SMI can be estimated from independent and identically distributed samples as the 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 FFT by exploiting the Toeplitz structure of the involved autocorrelation matrices in that space.

Discrete SMI

For $[\tilde{\mathbf{p}}]_n = P_X(x_n)$, $[\tilde{\mathbf{q}}]_m = P_Y(y_m)$ and $[\tilde{\boldsymbol{J}}]_{n,m} = P_{XY}(x_n, y_m)$:

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

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

Let $\mathbf{F} \in \mathbb{C}^{N\times N}$ and $\mathbf{G} \in \mathbb{C}^{M\times M}$ be unitary matrices, then

$$\begin{aligned} I_s(X;Y) &= ||\mathbf{C}||_F^2 = ||\mathbf{F}\tilde{\mathbf{C}}\mathbf{G}^{\mathbf{H}}||_F^2 \\ &= ||\mathbf{P}^{-1/2}\left(\mathbf{J} - \mathbf{p}\mathbf{q}^{\mathbf{H}}\right)\mathbf{Q}^{-1/2}||_F^2 \end{aligned}$$

Is it a coherence matrix?

Let us construct $\mathbf{x} \in \mathcal{F}$ and $\mathbf{y} \in \mathcal{G}$ by the one-to-one mappings $\phi_X(.) : \mathcal{X} \to \mathcal{F}$ and $\phi_Y(.) : \mathcal{Y} \to \mathcal{G}$:

- We are effectively mapping the events of discrete sources onto the columns of $\boldsymbol{F}$ and $\boldsymbol{G}$.
- As a consequence $\mathbf{p} = E[\mathbf{x}]$, $\mathbf{q} = E[\mathbf{y}]$, $\mathbf{P} = E\left[\mathbf{x}\mathbf{x}^H\right]$, $\mathbf{Q} = E\left[\mathbf{y}\mathbf{y}^H\right]$, and $\mathbf{J} = E\left[\mathbf{x}\mathbf{y}^H\right]$.

Therefore, we can express

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

Fundamental links

- The divergence transition matrix of a discrete memory-less communication channel is $\mathbf{B} = [\tilde{\mathbf{p}}]^{-1/2}\tilde{\mathbf{J}}[\tilde{\mathbf{q}}]^{-1/2}$, and so:
$$\tilde{\mathbf{C}} = \mathbf{B} - \tilde{\mathbf{p}}^{1/2}\tilde{\mathbf{q}}^{H/2}$$
- The largest singular value of C, is the Hirschfeld-Gebelein-Rényi maximal correlation coefficient.
- Additionally, the following holds:
$$0 \le I_s(X;Y) \le \min(N,M) - 1$$

Empirical characteristic function

Assume L i.i.d. samples $\{x(l), y(l)\}_{0\le l\le L-1}$. Then, the mapping

$$\mathbf{x}(l) \to \begin{bmatrix} e^{j\alpha(-K)x(l)} \\ \\ e^{j\alpha Kx(l)} \end{bmatrix} \quad \mathbf{y}(l) \to \begin{bmatrix} e^{j\alpha(-K)y(l)} \\ \\ e^{j\alpha Ky(l)} \end{bmatrix}$$

which leads to

$$\hat{I}_{cs}(X;Y) = ||\hat{\mathbf{P}}^{-1/2}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^{\mathbf{H}}\right)\hat{\mathbf{Q}}^{-1/2}||_F^2$$

with the sample means $\hat{\mathbf{p}} = \langle\mathbf{x}(l)\rangle_L$, $\hat{\mathbf{q}} = \langle\mathbf{y}(l)\rangle_L$, $\hat{\mathbf{P}} = \left\langle\mathbf{x}(l)\mathbf{x}^H(l)\right\rangle_L$, $\hat{\mathbf{Q}} = \left\langle\mathbf{y}(l)\mathbf{y}^H(l)\right\rangle_L$, and $\hat{\mathbf{J}} = \left\langle\mathbf{x}(l)\mathbf{y}^H(l)\right\rangle_L$.

SMI in high feature space dimension

Note that

$$\hat{\mathbf{P}} = \left\langle e^{j\alpha\mathbf{n}x(l)}e^{-j\alpha\mathbf{n}^Tx(l)}\right\rangle_L = \mathrm{toe}\left(\hat{\mathbf{p}}_a\right)$$

for $\hat{\mathbf{p}}_a = \left\langle e^{j\alpha\mathbf{n}_a x(l)}\right\rangle_L$ and $\mathbf{n}_a = [0, 1, \cdots, 2K]^T$.

- For large dimension, Szegö's theorem establishes that Toeplitz matrices are asymptotically diagonalizable by the unitary Fourier matrix.

Therefore, let us express an asymptotic approximation:

$$\hat{I}_{acs}(X;Y) = ||[\hat{\mathbf{p}}']^{-1/2}\mathbf{U}\left(\hat{\mathbf{J}} - \hat{\mathbf{p}}\hat{\mathbf{q}}^H\right)\mathbf{U}^H[\hat{\mathbf{q}}']^{-1/2}||_F^2$$

with $\mathbf{U}$ the unitary Fourier matrix, $\hat{\mathbf{p}}' = \mathrm{diag}(\mathbf{U}\hat{\mathbf{P}}\mathbf{U}^H)$, and $\hat{\mathbf{q}}' = \mathrm{diag}(\mathbf{U}\hat{\mathbf{Q}}\mathbf{U}^H)$.

Simulation results

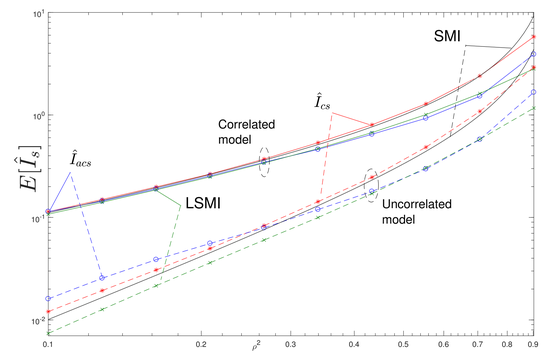

Figure 1: Mean estimated SMI vs the coherence factor ρ of the covariance matrix of a bivariate Gaussian distribution.

Conclusion

We have shown that we can measure the SMI after mapping the values of each random variables to vectors of fixed dimensionality. From this observation, two implications are explored: the estimator is based on a coherence matrix, a well-known statistic with multiple uses on signal processing [1], and computational savings thanks to the limitation of the feature space. Unlike the typical cross-validation approach with kernels as plug-in estimates of the PDF, the parameters selection is based on dual ideas from spectral analysis.

References

[1] 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*, 59(4):2128–2141, Apr. 2013.

[2] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52 (12 pages), 2009.

[3] S. L. Huang, C. Suh, and L. Zheng. Euclidean information theory of networks. *IEEE Trans. on Inf. Theory*, 61(12):6795–6814, Dec. 2015.

[4] 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.

Acknowledgements

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

Contact Information

- Email: {ferran.de.cabrera, jaume.riba}@upc.edu
- D5-{119, 116}
- UPC Campus Nord, C/Jordi Girona 1-3, 08034 Barcelona