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

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 $[\hat{p}] = p_X(x_m), [\hat{q}] = p_Y(y_m)$ and $[J] = p_{XY}(x_m,y_m)$,

$$ \hat{C} = [\hat{p}]^{1/2} (J - [\hat{p}] [\hat{q}])^{1/2} [\hat{q}]^{-1/2} $$

$$ I_s(X;Y) = \sum_{i=1}^{N \times M} |C_i|^2 $$

Fundamental links

- The divergence transition matrix of a discrete memory-less communication channel is $B = [\hat{p}]^{1/2} [J]^{-1/2} [\hat{q}]^{-1/2}$, and so:
  $$ \hat{C} = B - [\hat{p}]^{1/2} [J]^{1/2} [\hat{q}]^{-1/2} $$

- The largest singular value of $C$ is the Hirschfeld-Gebelein-Renyi maximal correlation coefficient.

- Additionally, the following holds:
  $$ 0 \leq I_s(X;Y) \leq \min(N,M) - 1 $$

Empirical characteristic function

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

$$ x(l) \rightarrow \left[ e^{i\rho [x(l)]} \right] \quad y(l) \rightarrow \left[ e^{i\rho [y(l)]} \right] $$

which leads to

$$ \hat{I}_s(X;Y) = \left| \hat{P} \right|^{1/2} (J - \hat{p} \hat{q})^{1/2} $$

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

SMI in high feature space dimension

Note that

$$ \hat{P} = \left[ e^{i\mu_0 \hat{x}(l)} | \hat{e}^{i\mu_1 \hat{x}(l)} \right]_L = \text{toe} (\hat{p}_a) $$

for $\hat{p}_a = \left( e^{i\mu_0 \hat{x}(l)} \right)_L$ and $\hat{n}_a = [0, 1, \cdots, 2K]^T$.

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

Therefore, let us express an asymptotic approximation:

$$ \hat{I}_{\text{sm}}(X;Y) = \left| \langle \hat{p} \rangle^{1/2} U \left( J - \hat{p} \hat{q} \right) U^H \langle \hat{q} \rangle^{1/2} \right| $$

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

Simulation results

<table>
<thead>
<tr>
<th>Uncorrelated model</th>
<th>Correlated model</th>
<th>LSMI</th>
<th>SMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \rho ]</td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
</tbody>
</table>

Figure 1: Mean estimated SMI vs the coherence factor $\rho$ 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


Acknowledgements

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

Contact Information

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

<table>
<thead>
<tr>
<th>Department of Teoria del Senyal i Comunicacions</th>
</tr>
</thead>
</table>
| Departament de Teoria del Senyal
| National Research Plan, and 2017 SGR 578 - AGAUR, Catalan Government |
| This work is supported by projects TEC2016-76409-C2-1-R (WINTER), Ministerio de Economía y Competitividad, Spanish National Research Plan, and 2017 SGR 578 - AGAUR, Catalan Government |
| Contact Information |
| Email: {fernan.de.cabrera, jaume.riba}@upc.edu |
| DS-{119, 116} |
| UPC Campus Nord, C/Jordi Girona 1-3, 08034 Barcelona |