Quantifying Total Correlations between Variables with Information Theoretic and Machine Learning Techniques

Authors: A. Murari, R.Rossi, M.Lungaroni, P.Gaudio, and M. Gelfusa

Università degli Studi di Roma Tor Vergata
Quantum Electronic and Plasma Physics research group
Department of Industrial Engineering

CONSORZIO RFX
Ricerca Formazione Innovazione

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 633053.
The views and opinions expressed herein do not necessarily reflect those of the European Commission.

Scientific Credibility

- In the last years the scientific literature has been overloaded with reports of studies, which are contradictory
- Ioannidis's 2005 paper "*Why Most Published Research Findings Are False*" has been the most downloaded technical paper from the journal PLoS Medicine. In this paper he shows that even in the 1% of the top publications in medicine, 2/3 of the studies are contradicted by others within a few years
- Various reasons for this situation:
  - Corporate takeover of public institutions
  - Decline of University independence
  - Increased complexity of the systems and phenomena to be studied.

Data Deluge

- The amount of data produced by modern societies is enormous
- JET can produce more than 55 Gbytes of data per shot (potentially about 1 Terabyte per day). Total Warehouse: almost 0.5 Petabytes
- ATLAS can produce up to about 10 Petabytes of data per year
- Hubble Space Telescope in its prime sent to earth up to 5 Gbytes of data per day
- Commercial DVD 4.7 Gbytes (Blue Ray 50 Gbytes).

These amounts of data cannot be analysed manually in a reliable way. Given the complexity of the phenomena to be studied, there is scope for the development of new tools for the assessment of the actual correlations between variables!!

Outline

I. Linear Correlations

II. Total Correlations: Information Quality Ratio

III. Neural computation: Autoencoders and Encoders

IV. Linear Correlations with Autoencoders and Encoders

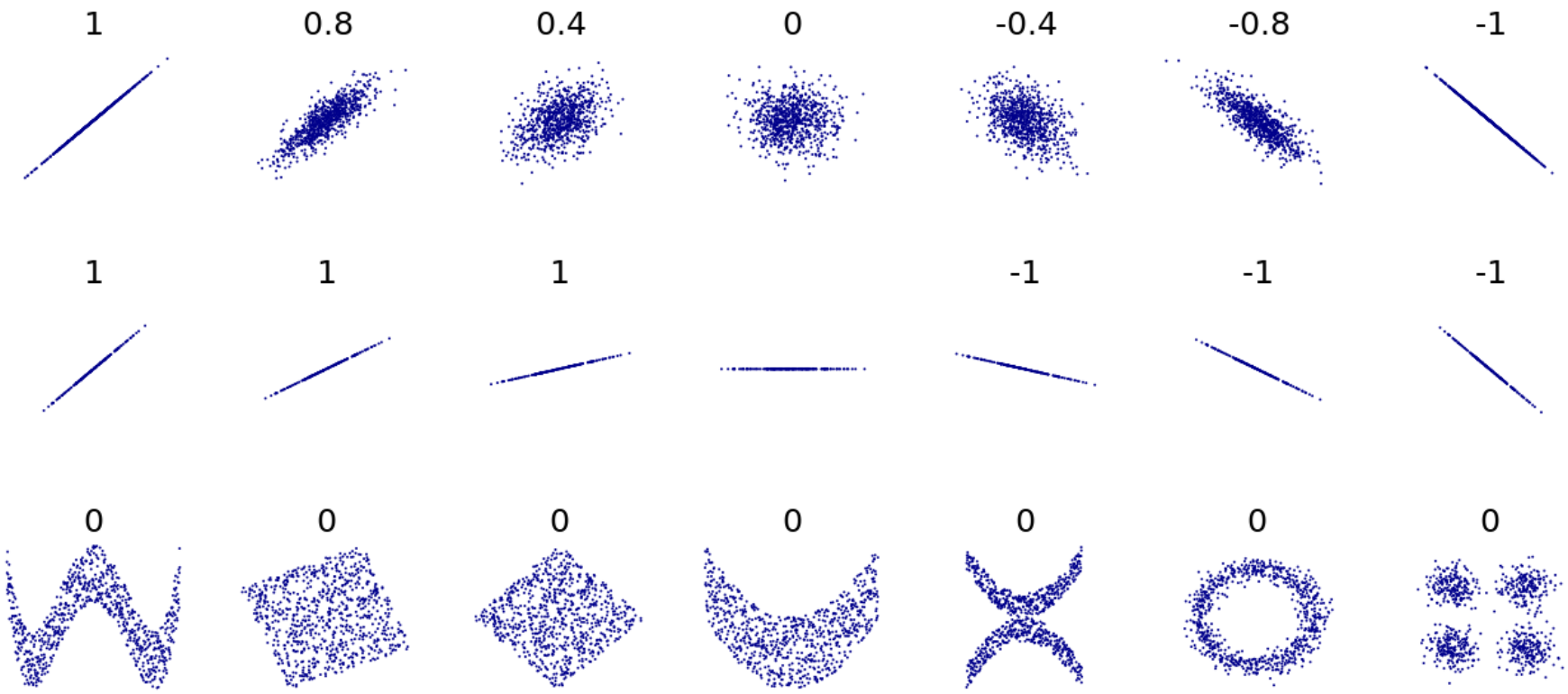
V. Total Correlations with Autoencoders and Encoders

VI. Conclusions

Linear Correlations

Pearson correlation coefficient (PCC)

$$\rho_{X,Y} = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y}$$



Mutual Information

The so called Mutual Information can be considered a measure of the mutual dependence between two random variables X and Y; it quantifies the amount of information that can be obtained about one random variable from knowing a second random variable and includes nonlinear effects.

$$I(X,Y) = -\sum_x \sum_y P(x,y) \ln\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

The Mutual Information is not normalized: it can be devided by the joint entropy:

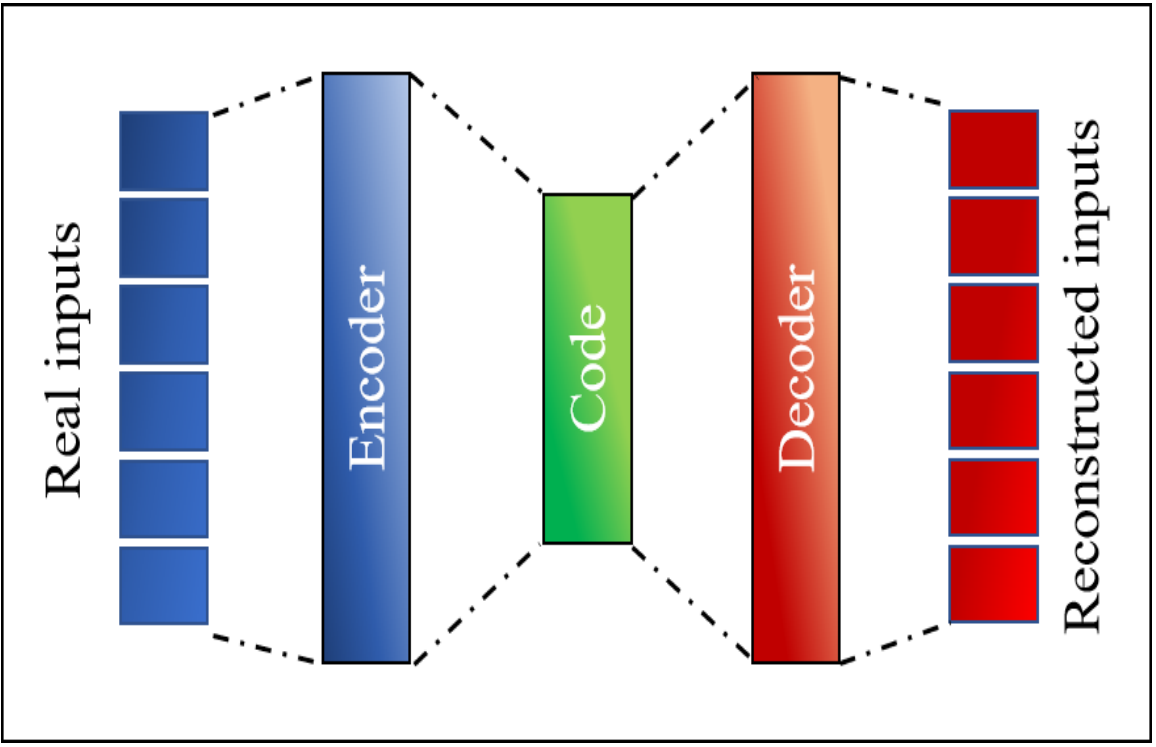
$$H(X,Y) = -\sum_x \sum_y P(x,y) \ln P(x,y)$$

The Information Quality Ratio (IQR) is the best normalized (0-1) indicator to use:

$$IQR = \frac{I(X,Y)}{H(X,Y)}$$

Neural computation: Autoencoders

Autoencoders are feed forward neural networks with a specific type of topology, reported in the Figure.

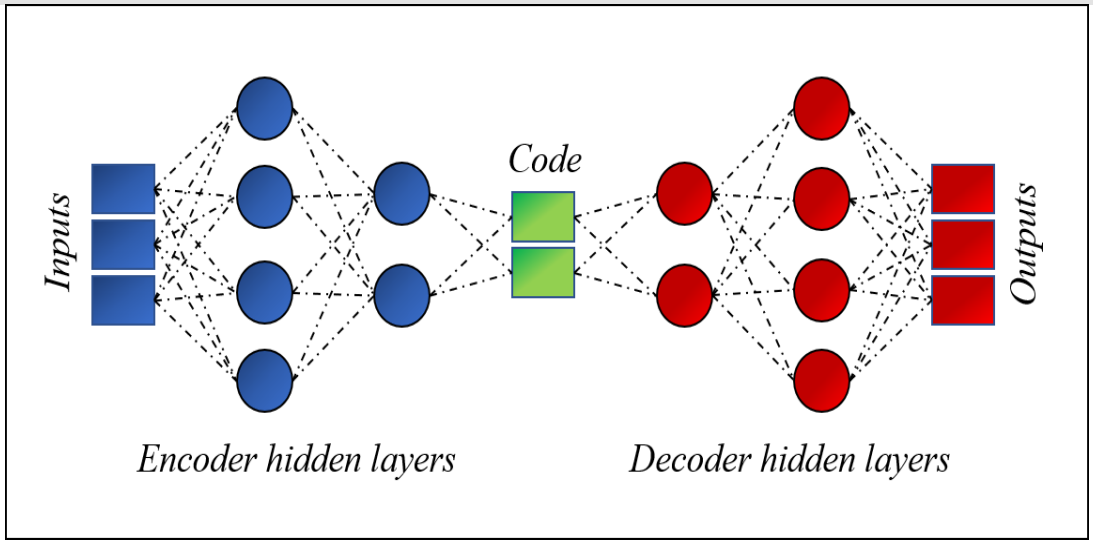


For correlations, the outputs are the same as the inputs.
In the case of regression, the output is the set of dependent variables.

The defining characteristic of auto encoders is that the output is the same as the input. They are meant to compress the input into a lower-dimensional *code* and then to reconstruct the output from this representation.

Conclusions

The actual architecture of the autoencoders used to obtain the results presented in the following is reported on the right.



The basic elements of the proposed method, to obtain the correlations (linear or total), consists of adopting the architecture of the Figure and then of reducing the neurons in the intermediate layer until the autoencoder does not manage to reproduce the outputs properly (starting with a number of neurons equal to the number of inputs).

The weights of the input out coefficients can be written in matrix form as:

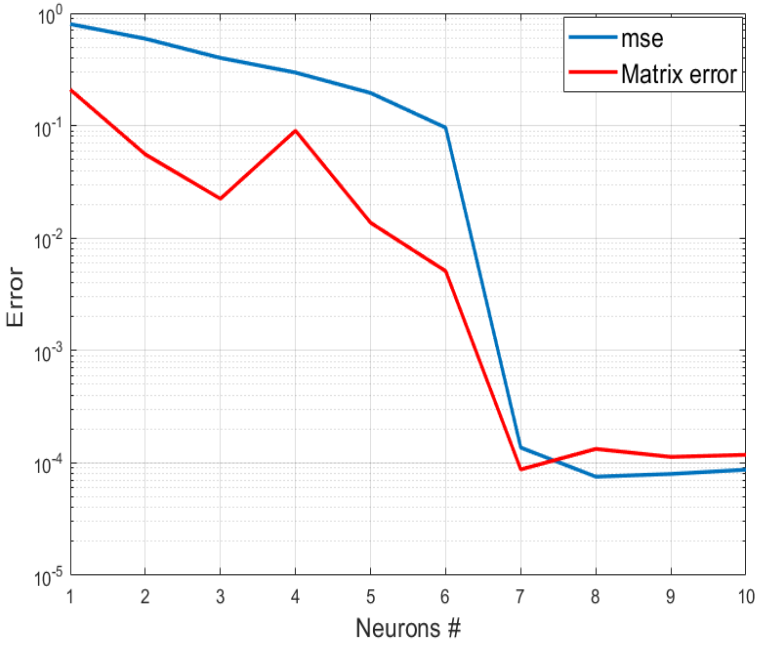
$$\boldsymbol{W} = \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix}$$

Normalization

The weights can be manipulated to obtain normalized coefficients (values 1 on the diagonal) as follows:

$$\Lambda_{i,j} = \sqrt{\frac{2W_{i,j}W_{j,i}}{W_{i,i}^2 + W_{j,j}^2}}$$

Example: A set of 10 different variables have been generated: $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ are independent from each other. The remaining variables have been generated with the relations: $x_8 = cost\ x_1; x_9 = cost\ x_2; x_{10} = cost\ x_3$.



Example

The Λ matrix agrees perfectly with the one reporting the Pearson Correlation Coefficients.

				Pearson						
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.00	0.00	-0.02	0.01	0.03	0.00	0.00	1.00	0.00	-0.02
x_2	0.00	1.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	0.00
x_3	-0.02	0.00	1.00	0.00	-0.01	0.00	0.01	-0.02	0.00	1.00
x_4	0.01	0.02	0.00	1.00	0.00	-0.02	0.01	0.01	0.02	0.00
x_5	0.03	0.00	-0.01	0.00	1.00	-0.01	0.00	0.03	0.00	-0.01
x_6	0.00	0.00	0.00	-0.02	-0.01	1.00	-0.01	0.00	0.00	0.00
x_7	0.00	0.00	0.01	0.01	0.00	-0.01	1.00	0.00	0.00	0.01
x_8	1.00	0.00	-0.02	0.01	0.03	0.00	0.00	1.00	0.00	-0.02
x_9	0.00	1.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	0.00
x_{10}	-0.02	0.00	1.00	0.00	-0.01	0.00	0.01	-0.02	0.00	1.00

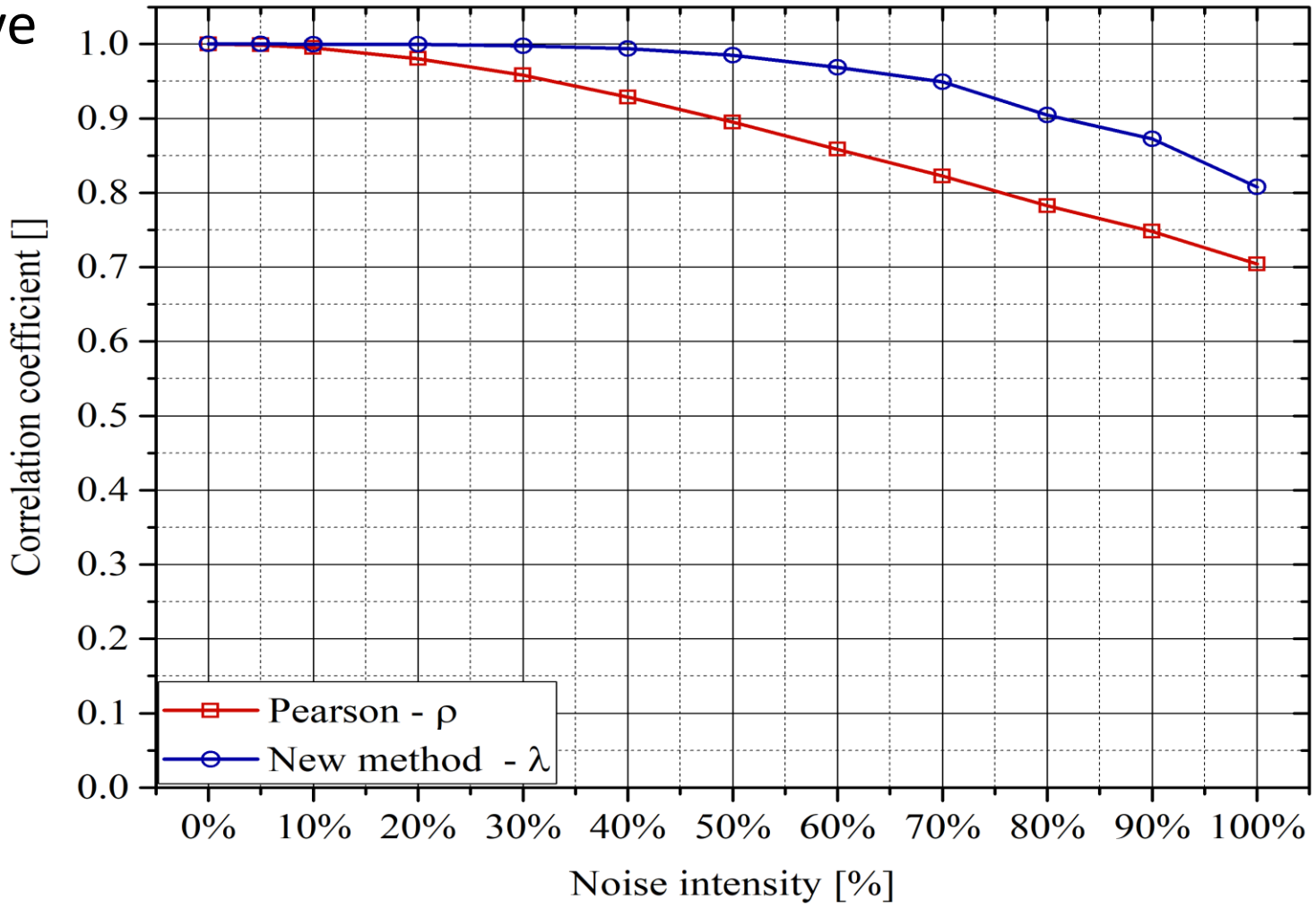
					Lambda - 7 neurone					
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.00	0.00	0.01	0.00	0.00	0.00	0.00	1.00	0.00	0.00
x_2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.01
x_3	0.01	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.01	1.00
x_4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
x_5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
x_6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
x_7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
x_8	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
x_9	0.00	1.00	0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.00
x_{10}	0.00	0.01	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

The case presented belong to the batteries of tests performed without noise.

Noise dependence

The approach of the Autoencoders is much more robust against noise (Gaussian in the figure)

Representative case



Total Correlations

Total correlations can have a different dependence in different region of the parameter space. The integration of local dependencies is proposed as a global indicator:

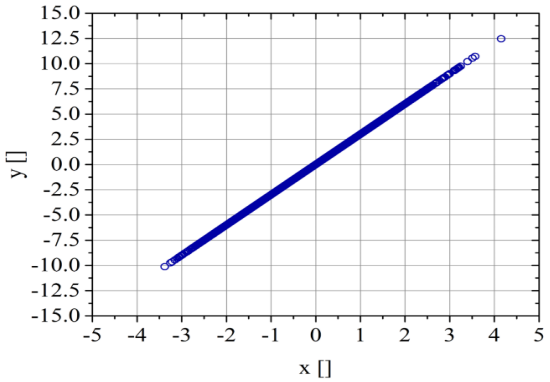
$$\rho_{int} = \frac{1}{\Delta x} \int |\rho(x)| dx$$

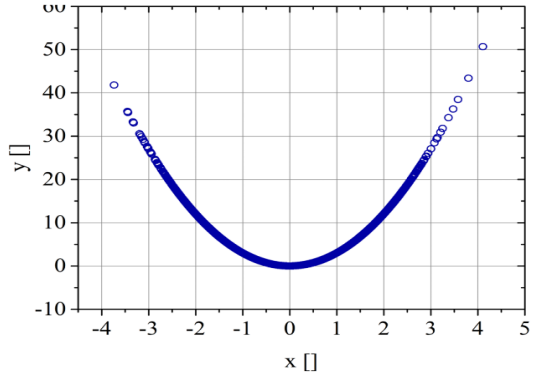
A second indicator is useful to determine the direction of the mutual influence. It is called mononicity and it is defined as:

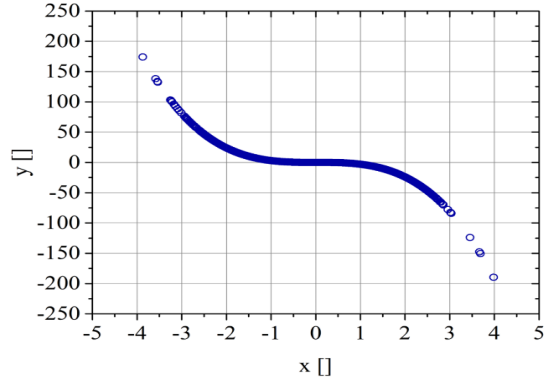
$$M_{int} = \frac{1}{\Delta x} \int sign\big(\rho(x)\big) dx$$

Total Correlations

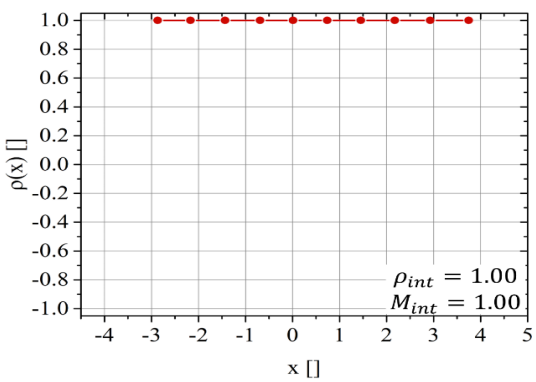
Data

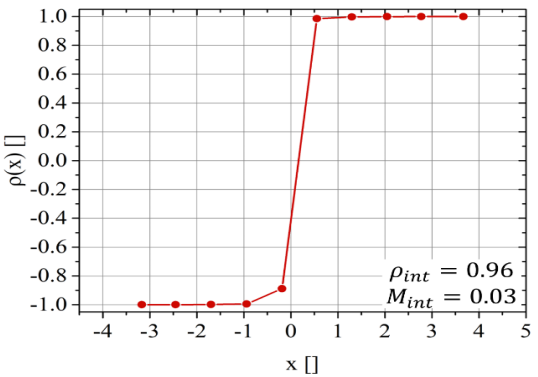


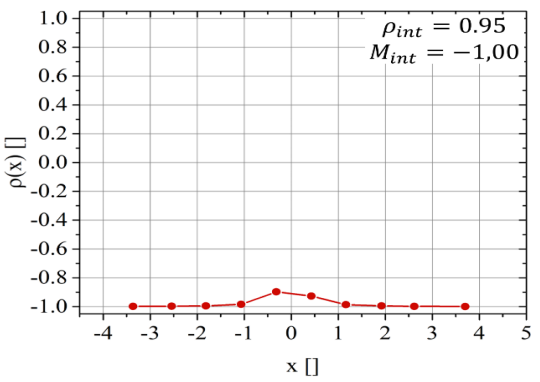




Correlation







The two global indicators proposed characterise quite well the mutual relation between two variables.

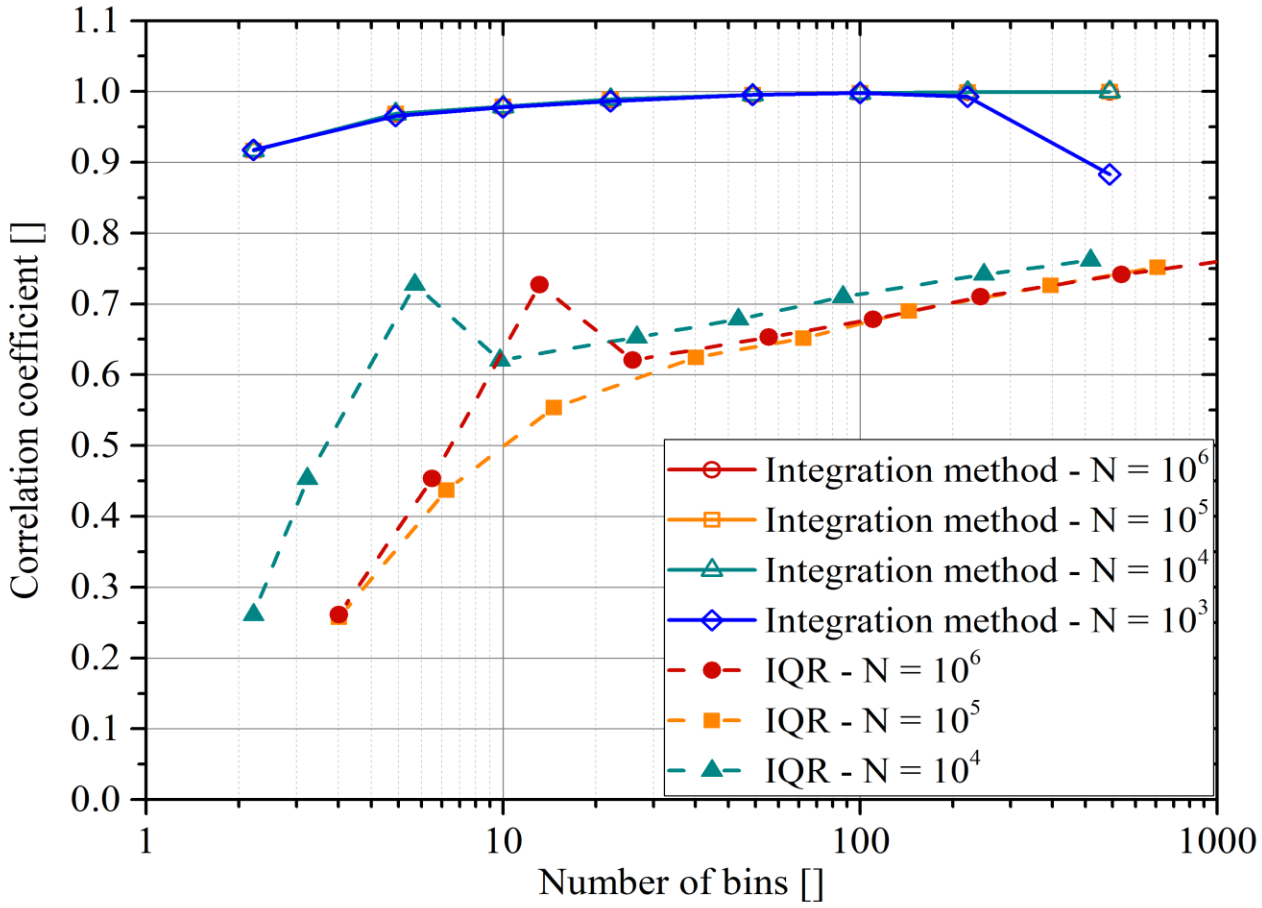
Top: linear dependence $\rho_{int} = 1$ $M_{int} = 1$.

Middle: quadratic dependence $\rho_{int} = 0.96$ $M_{int} = 0.03$.

Bottom: cubic dependence $\rho_{int} = 0.95$ $M_{int} = -1$

Total Correlations

The proposed methodology based on autoencoders seem to work much better than the IQR. It is less sensitive to the details of the binning and requires less data.



Total Correlations

The use of autoencoders and encoders has provided very interesting results.

- For the determination of the linear correlations between quantities, the proposed method provides the same values as the PCC but it is significantly more robust against the effects of additive random noise.
- To investigate the total correlations between quantities, the combined used of the integrated correlation coefficient and the monotonicity has proved to be much more informative and more robust than the IQR.

With regard to future development, the technique for the investigation of the total correlations needs to be extended to the case of more variables with an accurate assessment of the effects of the noise.

Thank You for Your Attention!

QUESTIONS?