ECEA 2019 5th International Electronic Conference on Entropy and Its Applications 18–30 November 2019

sciforum

Conference Proceedings Paper

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

Andrea Murari [1,*], Riccardo Rossi [2], Michele Lungaroni [2], Pasquale Gaudio [2] and Michela Gelfusa [2]

[1] Consorzio RFX (CNR, ENEA, INFN, Universita' di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy
[2] Associazione EURATOM-ENEA - University of Rome "Tor Vergata", Roma, Italy
* Correspondence: andrea.murari@euro-fusion.org

Academic Editor: name
Published: date

Abstract: The increasingly sophisticated investigations of complex systems require more robust estimates of the correlations between the measured quantities. The traditional Pearson Correlation Coefficient is easy to calculate but is sensitive only to linear correlations. The total influence between quantities is therefore often expressed in terms of the Mutual Information, which takes into account also the nonlinear effects but is not normalised. To compare data from different experiments, the Information Quality Ratio is therefore in many cases of easier interpretation. On the other hand, both Mutual Information and Information Quality Ratio are always positive and therefore cannot provide information about the sign of the influence between quantities. Moreover, they require an accurate determination of the probability distribution functions of the variables involved. Since the quality and amount of data available is not always sufficient to grant an accurate estimation of the probability distribution functions, it has been investigated whether neural computational tools can help and complement the aforementioned indicators. Specific encoders and autoencoders have been developed for the task of determining the total correlation between quantities, including information about the sign of their mutual influence. Both their accuracy and computational efficiencies have been addressed in detail, with extensive numerical tests using synthetic data.

Keywords: machine learning tools; information theory; information quality ratio; total correlations; encoders; autoencoders

1. Quantifying the Influence between Variables

Causality is an essential element of human cognition. As a prerequisite to determining causal influences between quantities, their correlations have to be at least properly quantified. On the other hand, particularly when analyzing cross-sectional data, even the preliminary stage of determining the correlation between quantities can become a very challenging task. This is particularly true in the investigation of complex systems in presence of overwhelming amounts of data. Modern scientific experiments are meant to provide information about every day more complex phenomena. They can also produce very large amounts of data. JET has the potential of producing 1 Terabyte of data per day, while ATLAS can generate 10 Petabytes of data per year [1].

Given this data deluge, the need to assess accurately the relation between the various measured quantities has become more pressing. The probability that important information remains hidden in the large data warehouses is indeed very high. The other main risk resides in the possible wrong evaluation of the influence between variables, which could lead to significant blunders [2].

The most widely used tools to determine the relation between variables present some significant limitations; they either detect only the linear correlations or require large amounts of data and do not provide any hint about the directionality of the influence (see Section 2). In this contribution, it is explored to what extent specific neural networks can help in at least alleviating some of these insufficiencies; these tools are introduced in Section 3. A comprehensive series of numerical tests with synthetic data has been performed. The results are reported in Section 4 for the liner correlations. The potential of the developed techniques to quantify total correlations are reported in Section 5. Conclusions and lines of future investigation are the subject of the last section of the paper.

2. Correlation and Mutual Information between Variables

A very common and computationally efficient indicator, to quantify the linear correlations between variables, is the Pearson correlation coefficient (PCC). The PCC has been conceived to determine the bivariate correlations between two quantities in the available data sets. The definition of the PCC, traditionally indicated by the Greek letter ρ, is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, \tag{1}$$

where cov is the covariance and σ indicates the standard deviation of the variables. In the notation adopted in this paper, the variable Y is the dependent variable and X are the regressors.

The main drawback of the PCC resides in the fact that it takes into account only the linear correlations. In the case of investigations involving highly nonlinear phenomena, as it is often the case in the science of complex systems, the conclusions, obtained from the analysis of the PCC, can therefore be highly misleading. To overcome this limitation, unbiased techniques for non-linear analysis are required. Information theory provides a very powerful tool to investigate the information transfer between quantities, the so-called Mutual Information [3]. This indicator 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. The traditional symbol for the Mutual Information is $I(X,Y)$ which for discrete variables is defined as:

$$I(X,Y) = -\sum_x \sum_y P(x,y) \ln\left(\frac{P(x,y)}{P(x)P(y)}\right), \tag{2}$$

where $P(\ )$ indicates the probability density function (pdf) of the variables in the brackets. Unfortunately, the mutual information is not a normalised quantity. On the other hand, it can be rescaled to the interval [0-1], by dividing it for the joint entropy $H(X,Y)$ defined as:

$$H(X,Y) = -\sum_x \sum_y P(x,y) \ln P(x,y), \tag{3}$$

In the literature, this new normalised quantity is called the Information Quality Ratio (IQR) [3]:

$$IQR = \frac{I(X,Y)}{H(X,Y)}, \tag{4}$$

The Information Quality Ratio is a well consolidated quantity but, as can be seen from Equations (2) and (3), the estimation of the IQR requires the calculation of the probability density functions of the variables involved. This aspect renders the analysis more demanding, compared to the PCC, both in terms of efforts and requirements about the quality of the data. Fortunately, nowadays there are quite powerful tools to obtain the pdfs from the histograms of experimental data. The techniques used in this paper belong to the family of the "Kernel Density Estimation" KDE [4], which have been deployed also in [5,9] for the analysis of experimental data from Big Physics

experiments, using the methodology described [10,11]. The outputs of these KDE tools can be considered consolidated at this stage. On the other hand, density estimation remains a delicate operation. Moreover, the IQR is always a positive quantity and therefore does not shed any light about the directionality of the information transfer and therefore of the influence between the considered quantities.

Given the previous considerations, it is reasonable to ask whether neural computation can help overcoming, or at least alleviating, the aforementioned limitations of PCC and IQR. In this perspective, in the framework of deep learning for image processing, many neural network topologies, such as encoders and autoencoders, have been recently extensively tested (see next section for a description of these networks); their properties could in principle be useful also in the analysis of the correlation between quantities [12]. In particular, it is worth assessing to what extent encoders and autoencoders can provide information about the total correlations between quantities, including directionality, and whether they can be more computationally efficient than traditional density estimators.

3. The Technology of Autoencoders and Encoders for the Assessment of Correlations

Autoencoders are feed forward neural networks with a specific type of topology, reported in Figure 1. The defining characteristic of auto encoders is the fact that the output is the same as the input [12]. They are meant to compress the input into a lower-dimensional code and then to reconstruct the output from this representation. To this end, an autoencoder consists of 3 components: encoder, code and decoder. The encoder compresses the input and produces the code, a more compact representation of the inputs. The decoder then reconstructs the input using this code only. The code constitutes a compact "summary" or "compression" of the input, which is also called the latent-space representation.

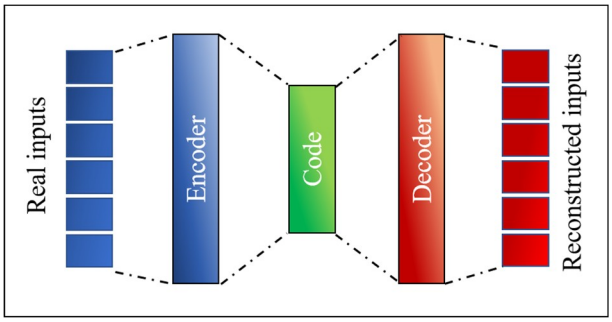

Figure 1. General topology of autoencoders.

In more detail, the input passes first through the encoder, which is a fully connected artificial Neural Network (ANN), and is translated into a code. The code is the input of the decoder, which is meant to produce an output, which is a similar as possible to the input. The decoder architecture is typically the mirror image of the encoder. Even if this condition is not an absolute requirement, it is typically the case (the actual indispensable requirement is that the dimensionality of the input and output is the same). Autoencoders can be trained via backpropagation as traditional ANNs.

Overall there are four hyperparameters that need to be to set before starting the training of an autoencoder: 1) code size i.e., the number of nodes in the middle layer 2) the number of layers 3) the number of nodes per layer 4) the loss function. All these hyperparameters have to be set to optimize the autoencoders for the present application, i.e., the assessment of the total correlation between variables (inputs). To this end, a stacked autoencoder architecture has been adopted: the layers are stacked one after another. The actual architecture of the autoencoders implemented to obtain he results discussed in the next sections is reported in Figure 2. For the investigations reported in this paper, linear activation functions have been implemented.

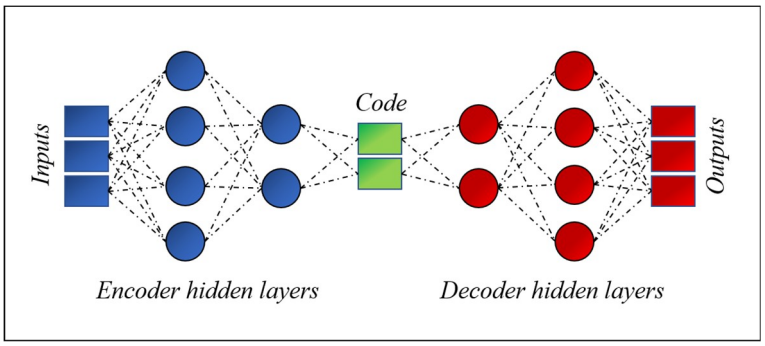

Figure 2. Architecture of the autoencoders used in the present work.

The architecture of the autoencoders is well suited to the determination of the total correlation between quantities. Another important type of correlation is the one between regressors and dependent variables. This is the case of regression, a very relevant case for both scientific and engineering studies. For this type of application, the topology of the so-called encoders is sufficient. Encoders can be thought of as just the first part of the autoencoder (see Figure 1) with the code of the same dimensionality as the dependent variable space.

The basic elements of the proposed method, to obtain the correlations (linear or total), consists of adopting the architecture of Figure 2 (or of the simple encoder for the case of regression) 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). After identifying the dimensionality of the latent space, the minimum number of neurons in the intermediate layer, for which the inputs are properly reconstructed at the output, a specific manipulation of the weights allows obtaining the required information. The operations of the weights differ depending on whether the linear or nonlinear correlations are investigated and therefore the details are provided in the next two sections.

4. The Technology of Autoencoders and Encoders for the Assessment of Correlations

Even if the main objective of the work consists of finding an alternative method to quantify the total correlations between quantities, a first analysis of the linear correlations is worth the effort. Being able to reproduce the PCC is useful to grasp the main elements of the approach and also to increase the confidence in the results. On the other hand, the PCC can be strongly affected by the noise present in the data; assessing whether neural computational tools can help in this respect is therefore quite valuable, particularly for scientific applications such as thermonuclear plasmas, whose measurements present quite high levels of uncertainties.

To fix the ideas, let's consider two examples involving three variables x_1, x_2 and x_3. In the first case the correlation between the first two variables is of value unity ($x_2 = \cos t x_1$); in the second case the correlation is reduced to 0.8 by the presence of random noise ($x_2 = \cos t x_1 + random$), whose standard deviation is 20% of the actual signal standard deviation. In both examples the dimensionality of the latent space is 2. The matrix of the weights of the autoencoders can be expressed in matrix form as:

$$\boldsymbol{W} = \begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix}, \tag{5}$$

To obtain normalized coefficients, so that the correlation coefficient of a variable with itself is 1, it is necessary to define e new matrix Λ, whose coefficients are:

$$\Lambda_{i,j} = \sqrt{\frac{2W_{i,j}W_{j,i}}{W_{i,i}^2+W_{j,j}^2}}, \tag{6}$$

The tables reported in Figure 3 provide a comparison of the correlation coefficients calculated with the PCC and with the proposed method of the autoencoders. As can be concluded by simple inspection of the numerical values in these tables, the proposed technique manages to exactly reproduce the PCC estimates in the case of absence of noise. When applied to noisy entries, the autoencoder seems to provide a better estimate of the expected off diagonal correlation coefficients. This is an interesting point which will be analyzed more extensively at the end of this section.

Case 1

R:	1.000	1.000	0.000
	1.000	1.000	0.000
	0.000	0.000	1.000
Λ:	1.000	1.000	0.000
	1.000	1.000	0.000
	0.000	0.000	1.000

Case 2

R:	1.000	0.711	0.002
	0.711	1.000	0.004
	0.002	0.004	1.000
Λ:	1.000	0.819	0.007
	0.819	1.000	0.000
	0.007	0.000	1.000

Figure 3. Comparison of correlation coefficients for the two cases described in the text.

A series of numerical tests has been performed to prove the generality of the conclusions obtained for simple cases. For examples, it has been verified that the approach remains valid in case of problems of larger dimensionality. An example of 10 variables is reported in the following. 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: $x_8 = \cos t x_1$; $x_9 = \cos t x_2$; $x_{10} = \cos t x_3$.

As can be seen in the plots of Figure 4, the autoencoder manages to clearly identify the dimensions of the latent space. The minimum number of neurons in the intermediate layer for, which the autoencoder manages to reproduce the inputs with minimal errors, is 7. The matrix Λ of the correlation coefficients reproduces also exactly the one obtained with the application of the PCC, as shown in Figure 5.

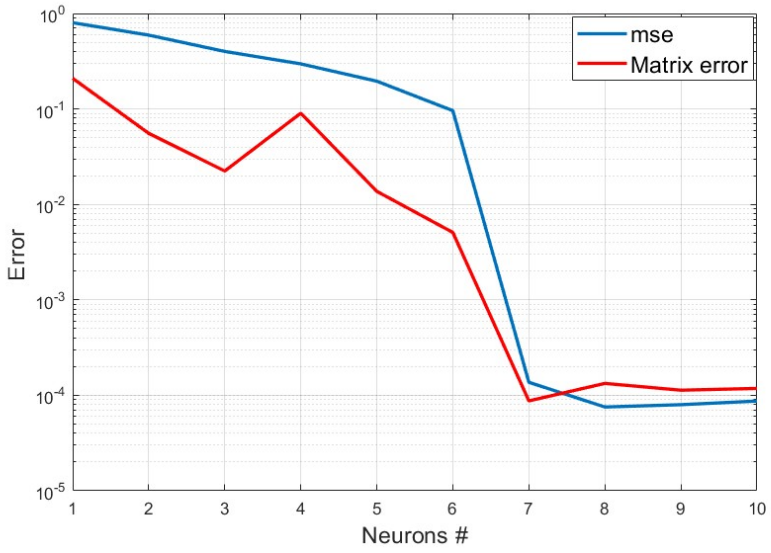

Figure 4. Trend of the errors in the reconstruction of the input data with the dimensionality of the intermediate layer in the autoencoder.

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

(**a**)

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

(**b**)

Figure 5. Left: the PCC for a set of 10 variables correlated as specified in the text. Right: the correlation coefficients obtained with the proposed method of the autoencoders.

Systematic tests have also been performed to assess the robustness of the proposed approach to additive noise. Gaussian noise of various amplitude has been added to the variables. It is important to notice that the method based on the autoencoders has proved more resilient that the traditional PCC. In general, for linear correlations, the Pearson coefficient start declining for the standard deviation of the noise of the order of 20% than the amplitude of the signal, whereas the matrix Λ remains stable up to at least 60%of additive noise. A typical dependence of the off-diagonal terms of the matrix Λ and the traditional PCC, versus the percentage of noise, is shown in the plots of Figure 6.

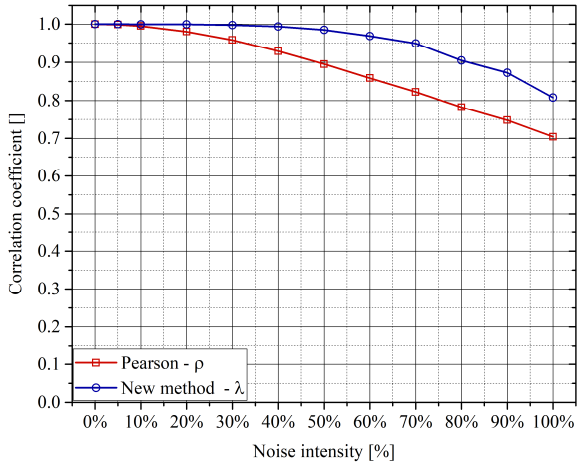

Figure 6. Trend of the off-diagonal term of the matrix Λ and the PCC versus the percentage of additive Gaussian noise. The noise intensify is calculated as the standard deviation of the noise divided by the standard deviation of the variable amplitude.

4. Numerical Tests for Total Correlations

The quantification of the total correlation between measurements poses some additional problems. First of all, if the influence between the variables includes nonlinear effects, the level of correlation depends on the range of the variables themselves. A simple example is a parabolic dependence, which is not only constant but whose sign also depends on the range of the variables. To address this issue, the procedure devised for assessing the linear correlations with the autoencoders has to be modified as follows. The determination of the latent space is the same. Once this step is completed, it is necessary to subdivide the range of the variables in sufficiently small intervals; on every one of these intervals the local correlation coefficient can be calculated as described

in the previous section. The integral of these local correlation coefficients is the integral correlation indicated with int:

$$\rho_{int} = \frac{1}{\Delta x}\int |\rho(x)| dx, \tag{7}$$

Of course, as the IQR, this indicator does not provide any information about the sign of the correlation. To quantify this aspect, a good indicator is the monotonicity of the correlation, which can be defined as:

$$M_{int} = \frac{1}{\Delta x}\int sign\big(\rho(x)\big) dx, \tag{8}$$

To exemplify the potential of the proposed approach, Figure 7 reports the local correlation coefficient for a linear, a quadratic and a cubic dependence. For the first case, as expected for a linear dependence, both the integral correlation coefficient and the monotonicity have a value of one. For the quadratic case, the int is again practically unitary, whereas the monotonicity is almost zero. In the cubic case the ρ_{int} is against unity and the monotonicity -1.

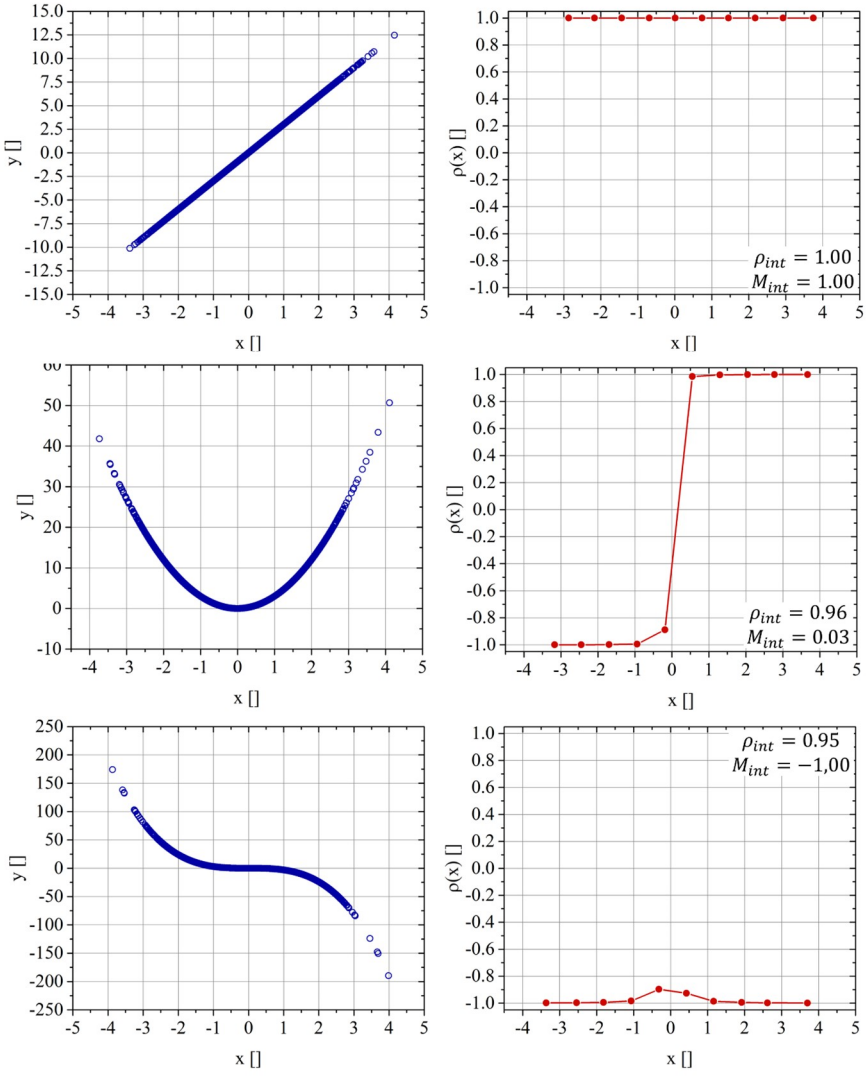

Figure 7. Top: two linearly dependent variables (left) and the relative local correlation coefficient ρ (right). Middle: two quadratic dependent variables (left) and the relative local correlation coefficient ρ (right). Bottom: two variables with a cubic negative dependence (left) and the relative local correlation coefficient ρ (right). The integral values of the correlation coefficient and of the monotonicity are reported in the inserts.

The combination of the integrated correlation coefficient and the monotonicity is therefore much more informative than the simple IQR. The int represents very well the actual dependence between the variables; on the other hand, the monotonicity provide information about the direction and the constancy of the mutual influence; negative signs of the monotonicity indicate an inverse dependence and low values the fact that the mutual dependence changes sign over the domain of the variables.

In terms of comparison with the IQR, Figures 8 summarizes a typical trend with the number of bins and the number of the entries in the database. As can be conclude from simple inspection of the plot, ρ_{int} provides a much better estimate of the correlation level between the independent and dependent variables (the actual value in the synthetic data is 1). The integrated correlation coefficient is also much more robust against the choice of the bins and the number of entries, two factors which affect a lot the IQR that is based on the details of the pdf.

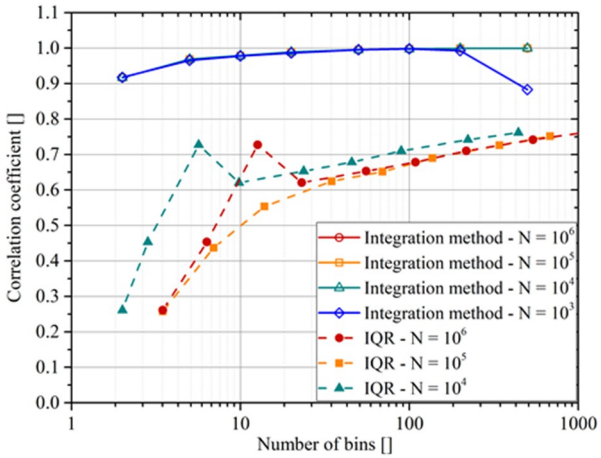

Figure 8. Comparison of the int and the IQR for the negative cubic dependence (third case of Figure 7). The x axis reports the number of bins and N is the number of generated points used to calculate the indicators.

4. Conclusions

An approach to the identification of the mutual influence between variables, using neural computational tools, has been proposed. The developed technique has been validated with a series of systematic tests with synthetic data. 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 than the IQR. The int reflects quite well the actual dependence between quantities. The monotonicity provides very valuable information about the constancy of the mutual influence over the investigated domain. The int is also less sensitive to the details of parameters, mainly the number of bins, required to calculate IQR. The int is also less demanding in terms of quantity and quality of the data required, to provide reliable estimates of the mutual influence between quantities.

With regard to future development, the technique for the investigation of the total correlations needs to be extended to the case of more variables. Conceptually this is not a problem. The main question remains the requirements in terms of amounts of data. The needs in terms of data amounts will obviously depend also on the quality of the inputs. This study will have therefore to be complemented with an accurate assessment of the effects of the noise. In any case, the one-dimensional studies and the preliminary indications of multi-dimensional tests unequivocally indicate that the proposed approach based on the autoencoders can handle much better sparse and noisy data.

References

1. T.H. Davenport, DJ Patil Data scientist Harvard business review, 2012, perso.esiee.fr
2. B. Goldacre "Bad science" Forth Estate, London 2008
3. D. J. C. MacKay "Information Theory, Inference and Learning Algorithms", Cambridge University Press, Sep 2003
4. B.W.Silverman, "Density Estimation for Statistics and Data Analysis", Chapmans & Hall, 1986
5. Murari, A. et al. "Non-power law scaling for access to the H-mode in tokamaks via symbolic regression" (2013) Nuclear Fusion, 53 (4), art. no. 043001, DOI: 10.1088/0029-5515/53/4/043001
6. Murari A. et al., P. "Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form" (2015) Plasma Physics and Controlled Fusion, 57 (1), art. no. 014008, DOI: 10.1088/0741-3335/57/1/014008
7. Murari A. et al "A new approach to the formulation and validation of scaling expressions for plasma confinement in tokamaks" 2015 Nuclear Fusion, Volume 55, Number 7 https://doi.org/10.1088/0029-5515/55/7/073009
8. Murari A. et al, "Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities" Nuclear Fusion 56, 2, 26005, DOI: 10.1088/0029-5515/56/2/026005
9. Murari A. et al "Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions" 2013 Nucl. Fusion 53 033006, doi.org/10.1088/0029-5515/53/3/033006
10. Schmid M and Lipson H, Science, Vol 324, April 2009
11. Koza J.R., "Genetic Programming: On the Programming of Computers by Means of Natural Selection". MIT Press, Cambridge, MA, USA (1992).
12. I. Goodfellow, Y. Bengio, Aaron Courville "Deep Learning" (Adaptive Computation and Machine Learning Series) the MIT Press, London 2017Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.