Hindawi
Complexity
Volume 2021, Article ID 9426547, 14 pages
https://doi.org/10.1155/2021/9426547

Research Article

The Reciprocal Influence Criterion: An Upgrade of the Information Quality Ratio

Riccardo Rossi,[1] **Michela Gelfusa**,[1] **Filippo De Masi**,[1] **Matteo Ossidi**,[1] **and Andrea Murari**[2]

[1]*Department of Industrial Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy*
[2]*Consorzio RFX (CNR, ENEA, INFN, Universita' di Padova, Acciaierie Venete SpA), Corso Stati Uniti 4, 35127 Padova, Italy*

Correspondence should be addressed to Riccardo Rossi; riccardo.rossi.en@gmail.com

Received 23 April 2021; Revised 12 July 2021; Accepted 3 August 2021; Published 12 August 2021

Academic Editor: M. De Aguiar

Understanding and quantifying the mutual influence between systems remain crucial but challenging tasks in any scientific enterprise. The Pearson correlation coefficient, the mutual information, and the information quality ratio are the most widely used indicators, only the last two being valid for nonlinear interactions. Given their limitations, a new criterion is proposed, the reciprocal influence criterion, which is very simple conceptually and does not make any assumption about the statistics of the stochastic variables involved. In addition to being normalised as the information quality ratio, it provides a much better resilience to noise and much higher stability to the issues related to the determination of the involved probability distribution functions. A conditional version, to counteract the effects of confounding variables, has also been developed, showing the same advantages compared to the more traditional indicators. A series of systematic tests with numerical examples is reported, to compare the properties of the new indicator with the more traditional ones, proving its clear superiority in practically all respects.

1. Determining Statistical Dependence between Quantities

The investigation of the reciprocal influence between systems is a major scientific objective in practically all disciplines. Both linear and nonlinear effects can be important, the latter becoming often dominant in complex phenomena [1]. The information about systems is typically obtained with measurements affected by various forms of uncertainties. The measurands can therefore be considered as random variables, and their mutual influence is partly deterministic and partly probabilistic in nature.

Quantifying the dependence between random variables is an essential statistical task for both bivariate and multivariate data. The most popular measure of dependence between two quantities is the Pearson product-moment correlation coefficient or "Pearson's correlation coefficient" (PCC) [2]. To alleviate its limitations (see Section 2), at the beginning of the last century, various alternative indicators based on ranked variables were developed, such as the Spearman ρ or Spearman rank correlation coefficient [3] and the rank correlation coefficient or Kendall τ [4]. After the Second World War information theoretic criteria, such as mutual information, became quite popular [5]. At the dawn of the new century, a lot of work was devoted to consolidating distance correlation, an indicator that is zero only when the two variables are not correlated [6]. In the last years, the advances in genetics have also motivated the development of various techniques for the analysis of the dependence between vectors in high dimensions. For this multivariate inference, very sophisticated approaches have been proposed, ranging from projection correlation [7] to Ball covariance [8] and Brownian distance [9]. All these techniques are based on specific assumptions about the statistical properties of the vectors of random variables to consider. They are also quite involved, both conceptually and numerically.

The indicators proposed in this paper are meant to deal only with bivariate cases, for which no prior information is known about the statistics of the data, so fully general techniques are required. A conditional version to counteract the effects of confounding variables has also been developed. The indicators are also very simple from the point of view of the conceptual background, interpretation of the results, and implementation requirements.

Regarding the structure of the paper, Section 2 introduces the most consolidated indicators, typically used to assess the dependence between two variables. Section 3 discusses the main rationale behind the proposed alternative indicator: the reciprocal influence criterion. The systematic tests, comparing the properties of RIC with PCC and IQR, are reported in Sections 4 and 5, for functional and nonfunctional dependencies, respectively. The robustness against the noise of statistics, different from the Gaussian and against outliers, is discussed in Section 6. The conditional version of RIC is introduced in Section 7, before conclusions are drawn in Section 8.

2. Traditional and Alternative Criteria to Quantify the Correlations between Quantities

In order to fully appreciate the potential of the new indicators proposed in the paper, it is worth starting the discussion with a review of the traditional and alternative tools most used by practitioners for the analysis of bivariate dependence. The linear correlations between two quantities are typically calculated with the Pearson correlation coefficient (PCC). For a couple of random variables X and Y, indicating with cov as the covariance and with σ as the standard deviation, the PCC is calculated as [10]

$$\mathrm{PCC} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y}. \tag{1}$$

Even if it is very useful in practice, some limitations of the PCC, particularly, its vulnerability to noise, are very often overlooked. In any case, the detection of nonlinear correlations between two variables is a much more serious matter. To quantify nonlinear dependencies, historically, the first techniques developed were based on ranking variables (see Section 2.1). After the Second World War, indicators based on the probability distribution function (pdf) of the data have been developed; particularly, information theoretic criteria are popular (Section 2.2). Distance correlation is a much more recent development, first introduced in 2005, to remedy the problem that the Pearson correlation coefficient can be zero for dependent variables (Section 2.3).

2.1. Rank-Based Criteria. In this paper, the symbol ρ_S indicates the extension of the Pearson correlation coefficient introduced by Charles Spearman. Spearman's ρ_S is a nonparametric measure of the statistical dependence between the rankings of two variables; as such, it quantifies how well the relationship between two variables can be modelled by a monotonic function [11]. Spearman's ρ_S is calculated as the Pearson correlation between the rank values of the quantities involved and assesses how monotonic the relation between them is (whether linear or not). If there are no repeated data values, ρ_S assumes the values +1 or −1 when the two variables are perfect monotone functions of each other.

To calculate Spearman's $\rho_{S,}$ the raw data X_i and Y_i are converted into ranks first: rank_X and rank_Y; the standard deviations of these two ranked variables are indicated with the symbols σ_{rank_X} and σ_{rank_Y}. Spearman's ρ_S is then defined as

$$\rho_S = \frac{\mathrm{cov}\left(\mathrm{rank}_x, \mathrm{rank}_y\right)}{\sigma_{\mathrm{rank}_x} \sigma_{\mathrm{rank}_y}}. \tag{2}$$

The Kendall rank correlation coefficient, named after Maurice Kendall, was developed at the end of the 30 s, and it is usually indicated with the Greek letter τ. It is meant to measure the rank correlation, i.e., the similarity of the ordering of the data when ranked. Intuitively, the higher the Kendall correlation between two variables, the more similar their rank; the Kendall τ is also normalised in the sense that it ranges between 1 and −1 [11].

Mathematically, the Kendall rank correlation coefficient is calculated as

$$\tau = \frac{(\text{Number of concordant pairs}) - (\text{Number of discordant pairs})}{\binom{n}{2}}, \tag{3}$$

where the denominator is the binomial coefficient $\binom{n}{2} = n(n-1)/2$.

2.2. Information Theoretic Criteria. A widely used indicator to investigate nonlinear interactions is the mutual information, which quantifies the information shared by two systems. The Shannon or discrete version is defined as [12, 13]

$$\mathrm{MI}_{\mathrm{Shan}} = \sum\sum\left(p_{xy}\log\left(\frac{p_{xy}}{p_x p_y}\right)\right), \tag{4}$$

where p_x and p_y are the discrete probabilities of two random variables X and Y and p_{xy} is their joint probability.

A differential version of the indicator can be formulated in terms of the probability densities f_x and f_y:

$$f_x = \frac{p_x}{\Delta x},$$
$$f_y = \frac{p_y}{\Delta y}, \qquad (5)$$
$$f_{xy} = \frac{p_{xy}}{\Delta x \, \Delta y},$$

where Δx and Δy are the dimensions of the bins. The differential mutual information, $\mathrm{MI}_{\mathrm{diff}}$, is defined as

$$\mathrm{MI}_{\mathrm{Diff}} = \int\int\left(f_{xy}\log\left(\frac{f_{xy}}{f_x f_y}\right)\mathrm{d}x\,\mathrm{d}y\right). \qquad (6)$$

It can be easily demonstrated that the two relations (4) and (6) are equivalent; the mutual information is, therefore, referred to as MI in the rest of the paper. MI has various positive properties but presents the main limitation of not being normalised. To obviate this drawback, the mutual information is typically divided by the joint entropy to obtain the so-called information quality ratio (IQR) [14]:

$$\mathrm{IQR} = \frac{\mathrm{MI}}{H(X,Y)}. \qquad (7)$$

This quantity is normalised in the sense that it assumes only values between zero and one. It is important to remember that, in equation (7), the discrete or Shannon version of the joint entropy is to be used:

$$H_{S,XY} = -\sum\sum p_{xy}\log\left(p_{xy}\right). \qquad (8)$$

The Shannon entropy is to be compared with the differential one, which can also assume negative values:

$$H_{D,XY} = -\int\int\left(f_{xy}\log\left(f_{xy}\right)\mathrm{d}x\mathrm{d}y\right). \qquad (9)$$

Equation (7) is the commonly accepted version of the information quality ratio, normally adopted because the differential version of the joint entropy can be negative, with the obvious related problems and difficulties.

Unfortunately, mainly due to the denominator, the IQR_S has some limitations, particularly, a lack of robustness to noise and a strong dependence on the choice of the bins.

2.3. Distance Correlation. The objective of distance correlation (D_{corr}) consists of quantifying the dependence between two random vectors, which do not need to have necessarily equal dimension. D_{corr} has the clear advantage, compared to the PCC, that the population distance correlation coefficient assumes a zero value only if the two random vectors are independent. Therefore, distance correlation is meant to quantify both linear and nonlinear association between two random variables or random vectors [5].

The calculation of D_{corr} requires the definition of some other preliminary quantities. Indicating with (X_k, Y_k), $k = 1, 2, \ldots, n$, a sample from a pair of real-valued or vector-valued random variables (X, Y), the elements of the n by n distance matrices $(a_{j,k})$ and $(b_{j,k})$ are all pairwise distances:

$$\begin{aligned} a_{j,k} &= \left\|X_j - X_k\right\|, \quad j,k = 1,2,\ldots,n, \\ b_{j,k} &= \left\|Y_j - Y_k\right\|, \quad j,k = 1,2,\ldots,n, \end{aligned} \qquad (10)$$

where $\|\ \|$ denotes the Euclidean norm. Defining $\overline{a}_j$ and $\overline{a}_k$as the jth row mean and the kth column mean, respectively, and with $\overline{a}$ the grand mean of the first vector mutual distance X matrix, one can then calculate all doubly centred distances (with the same notation for the Y vector matrix):

$$\begin{aligned} A_{j,k} &= a_{j,k} - \overline{a}_j - \overline{a}_k + \overline{a}, \\ B_{j,k} &= b_{j,k} - \overline{b}_j - \overline{b}_k + \overline{b}. \end{aligned} \qquad (11)$$

The desired sample distance covariance is then simply the arithmetic average of the product $A_{j,k}B_{j,k}$:

$$d\mathrm{Cov}_n^2(X,Y) := \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{j,k}\,B_{j,k}. \qquad (12)$$

Indicating with distance variance,

$$d\mathrm{Var}_n^2(X,X) := \frac{1}{n^2}\sum_{k,l=1}^{n} A_{k,l}^2, \qquad (13)$$

finally, the distance correlation is

$$d\mathrm{Cor}(X,Y) := \frac{d\mathrm{Cov}(X,Y)}{\sqrt{d\mathrm{Var}(X)d\mathrm{Var}(Y)}}. \qquad (14)$$

The main properties of dCor are that it assumes values between 0 and 1, and it is zero only if the two vectors are independent. The distance correlation software used in this work is the one published by Shen Liu [15].

3. The Reciprocal Influence Criterion: Rationale

As will be shown in the rest of the paper, all criteria summarised in the previous section have several drawbacks. In this work, a new indicator, able to quantify linear and nonlinear correlations between variables, is introduced. The new indicator, based on information theoretic quantities and named reciprocal influence criterion (RIC), has been designed to have the following properties.

3.1. Property 1. The indicator ranges from zero (no correlation) to one (perfectly correlated). The definition of correlation for this indicator is the following.

Two variables i and j are correlated when the knowledge of i helps to predict j and *vice versa*. The correlation indicator tends to one when the uncertainty of i (j) known j (i) goes to zero. Correlation must be equal to zero when the uncertainty of i (j) does not change when the j (i) is known. Note that, in this definition, symmetric equations cannot reach the value of one; as, for example, for the function $y = x^2$, since y is known, there are two valid values of x (excluded $x = 0$ and $y = 0$).

Table 1: The value of the various indicators for the cases of Figure 1.

	$y = x$	$y = x^2$	$y = e^x$	$y = \sin(3x)$
Pearson	0.99	0.03	0.93	0.01
Spearman	0.99	0.03	0.98	0.03
Kendall	0.93	0.02	0.90	0.01
Distance correlation	0.99	0.49	0.96	0.35
IQR	0.47	0.35	0.46	0.33
RIC	0.97	0.93	0.96	0.90

3.2. Property 2. The indicator does not vary as a function of the binning used for the probability density or distribution function calculations, i.e.,

$$\frac{\Delta N_{\mathrm{bin}}}{N_{\mathrm{bin}}} \gg \frac{\Delta \mathrm{RIC}}{\mathrm{RIC}}, \tag{15}$$

where ΔRIC denotes the difference between the indicator values calculated with and without the outliers.

3.3. Property 3. The indicator can be calculated indifferently for either discrete or continuous random variables.

3.4. Property 4. Small sensitivity to outliers is quantified as

$$\frac{N_{\mathrm{outliers}}}{N} \gg \frac{\Delta \mathrm{RIC}}{\mathrm{RIC}}. \tag{16}$$

The formulation of the RIC that ensures the former properties is

$$\mathrm{RIC} = 1 - \frac{A^{H_{xy}}}{A^{H_x + H_y}} = 1 - \frac{A^{H_{xy}}}{A^{H_x + H_y}} = 1 - \frac{1}{A^{\mathrm{MI}}}. \tag{17}$$

The RIC indicator is based on the ratio between the amount of information (uncertainty) shared by the two variables with respect to the sum of their individual information (uncertainty). In some ways, its definition is very similar to the IQR indicator, with the main difference that the RIC is based on the exponential of the entropy A^H. This approach allows obtaining a much more reliable indicator, which satisfies the aforementioned four desirable properties, as demonstrated in the following, and presents some additional positive qualities, namely, symmetry, asymptotic consistency, and unbiasedness, which are discussed in Appendix A.

With regard to the free parameter A, the choice of its numerical value can be optimised, depending on the nature of the data and the objectives of the investigation. The results, reported in the rest of the paper, have been obtained by setting $A = 10$, to maximise the coherence of RIC with PCC in case of linear correlations; the equations in the following are also particularised for this numerical value. A detailed discussion of the RIC behaviour with the parameter A is provided in Appendix B.

Some of the aforementioned four properties can be analytically demonstrated. First, property 1 can be proven as follows. For not correlated variables, $H_x + H_y = H_{xy}$ and $\mathrm{MI} \longrightarrow 0$; thus,

$$\lim_{\mathrm{MI}\longrightarrow 0} \mathrm{RIC} = \lim_{\mathrm{MI}\longrightarrow 0} 1 - \frac{10^{H_{D,xy}}}{10^{H_{D,x}+H_{D,y}}} = \lim_{\mathrm{MI}\longrightarrow 0} 1 - \frac{1}{10^{\mathrm{MI}}} = 0. \tag{18}$$

In the case of partially correlated quantities, RIC is always positive, and for high mutual information, it tends to one. More specifically, it is known that, for high correlation levels, the joint entropy tends to the lowest entropy of the two variables ($H_{xy} \longrightarrow \min(H_x, H_y)$) [16], which means that $\mathrm{MI} \longrightarrow \max(H_x, H_y)$. Assuming $H_y > H_x$, it follows

$$\lim_{H_{xy}\longrightarrow H_x} RIC = \lim_{H_{xy}\longrightarrow H_x} 1 - \frac{10^{H_{xy}}}{10^{H_x+H_y}} = \lim_{H_{xy}\longrightarrow H_x} 1 - \frac{10^{H_x}}{10^{H_x+H_y}} = \lim_{H_{xy}\longrightarrow H_x} 1 - \frac{1}{10^{H_y}} > 0. \tag{19}$$

Property 3 can also be proved. Indeed, writing the RIC using the differential information theoretic quantities and remembering that the sum of the entropies is equal to the sum of the mutual information and the joint entropy, it is possible to demonstrate that $\mathrm{RIC_D}$ is the same as $\mathrm{RIC_S}$:

$$\mathrm{RIC}_D = 1 - \frac{10^{H_{D,xy}}}{10^{H_{D,x}+H_{D,y}}} = 1 - \frac{1}{10^{MI}} = 1 - \frac{10^{H_{S,xy}}}{10^{H_{S,x}+H_{S,y}}} = \mathrm{RIC}_S. \tag{20}$$

Moreover, RIC is directly correlated to the mutual information value; also, property 2 is expected to be satisfied (and it is validated in Section 4).

In terms of interpretation, RIC remains an information theoretic indicator, since it relies on the mutual information. On the contrary, the monotonic transformation of (15) provides a practically normalised indicator without having to make recourse to the joint entropy. The advantages of such a reformulation will be illustrated in detail, with the help of numerical tests, in the next sections.

4. The Reciprocal Influence Criterion: Numerical Tests for Functional Dependencies

In this section, the properties of RIC are investigated with the help of a series of numerical tests with synthetic data. The reference indicators to benchmark the performance of RIC are the ones reviewed in Section 2, the most used by the practitioners. Only bivariate dependences due to functional relations are considered; nonfunctional dependencies are the subject of Section 4.1. A general overview of the results is

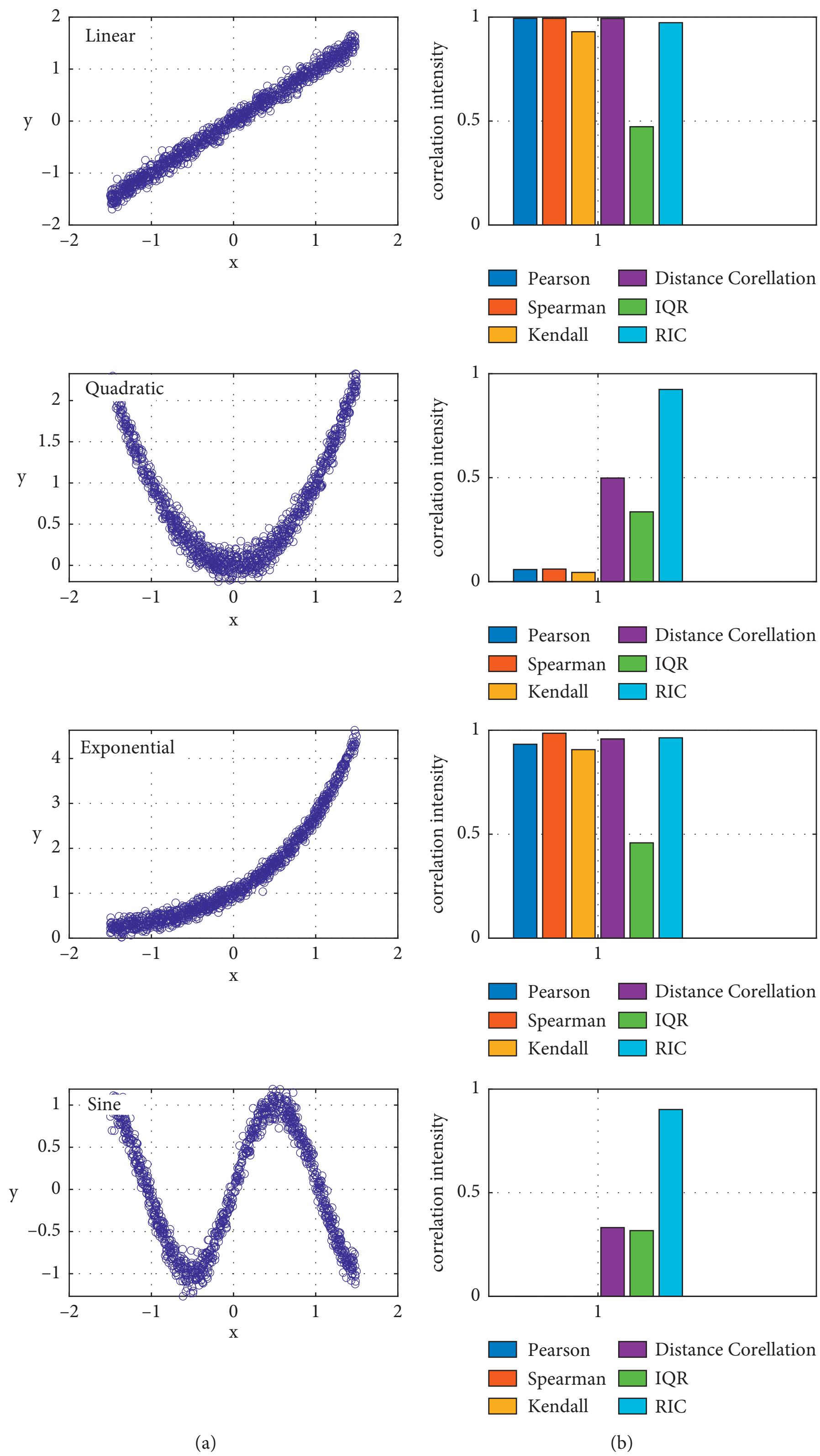

Figure 1: The main classes of functions tested. (a) Examples of synthetic data. (b) The numerical values of the indicators.

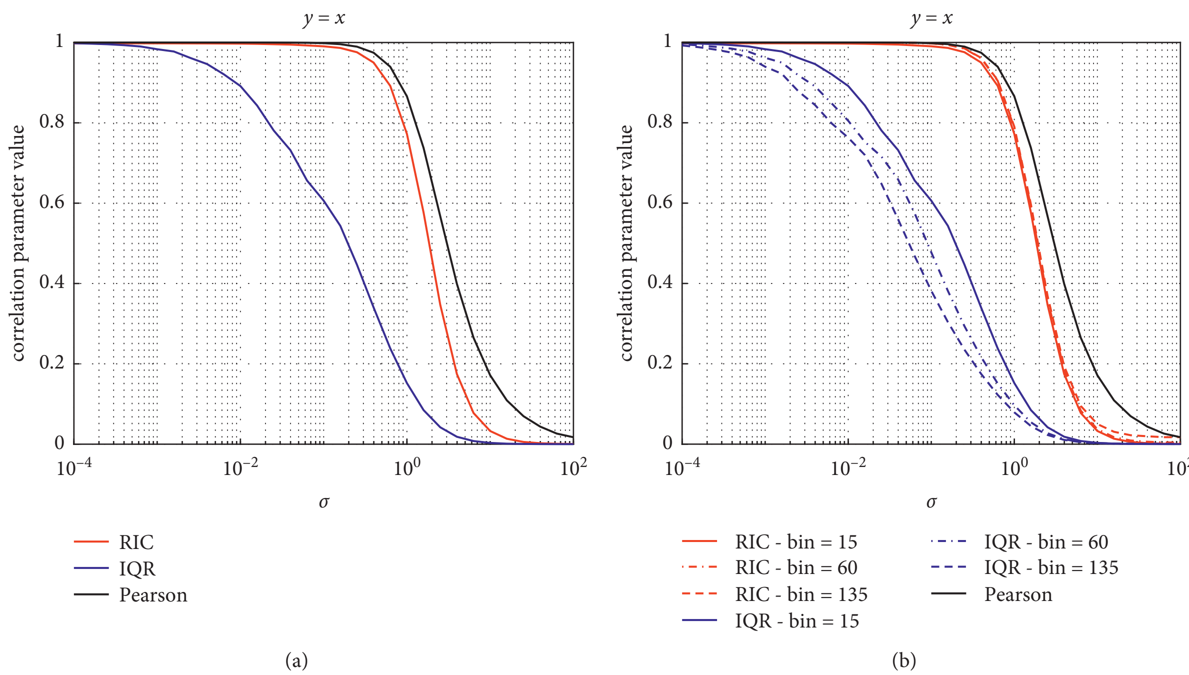

FIGURE 2: (a) Behavior of PCC, IQR, and RIC with the standard deviation of the additive noise for linear correlations between two variables: $y = x$. (b) Behavior of PCC, IQR, and RIC with the standard deviation of the additive noise when varying the binning.

provided first and some specific aspects are discussed in dedicated sections.

4.1. Comparison Overview. A first comparison between the RIC and the other criteria has been performed for the functional dependencies linear, quadratic, sinusoidal, and exponential. The results for a series of representative cases are reported in Table 1. A graphical overview is provided in Figure 1. Random noise of Gaussian distribution, with standard deviation equal to 10% of the quantity value, has been added to all variables.

The values of Table 1 and inspection of Figure 1 reveal that RIC never performs significantly worse than the other criteria for linear dependencies. RIC starts outperforming all other indicators in the case of nonlinear functions. Moreover, as expected, RIC provides much more reliable and reasonable results whenever the functional dependence is nonmonotonic and when even the ranked methods fail miserably. Also, the indicators based on the pdf, IQR, and distance correlation show significant difficulties to provide acceptable results for the nonmonotonic dependencies. All these are general properties not only true for the examples reported but also confirmed in all cases tested.

4.2. Linear Correlations: Effects of Gaussian Noise and Binning. This section, with the help of Figure 2, is simply aimed at supporting the statement that RIC can reproduce well the values of the PCC for linear correlations. The behaviour of the other indicators is also shown for completeness. The plots of Figure 2 refer to the case of perfect linear correlation between quantities: $y = x$. On the x-axis, the standard deviation of additive noise and zero mean and sampled randomly from a Gaussian distribution is reported. The results are fully general. The RIC reproduces quite well the values of the PCC, whereas the IQR is significantly more vulnerable to both noise and binning. A similar analysis indicates that RIC is also more robust against the presence of outliers; indeed, it can tolerate about even one order of magnitude more outliers than IQR, confirming that property 4 of Section 3 is satisfied (with 5% of outliers, the average Pearson coefficient variation is about 10% and the IQR variation is 12%, while the ΔRIC is 0.7%).

4.3. Nonlinear Correlations: Effects of Gaussian Noise and Binning. The competitive advantages of RIC become even more evident in the case of nonlinear correlations. Three exemplificative cases are reported in Figure 3, in which the proposed new criterion is compared with IQR_S. The functional dependencies reported are $y = x^2$, $y = \sin(x)$, and $y = \exp(x)$.

As expected, IQR is much more sensitive to the choice of the binning and the level of noise. RIC remains stable at a value very close to 1 for a much wider range of these factors. Moreover, IQR does not output a value of 1 even for perfect correlation between the two variables. This is a consequence of the denominator not being a normalised quantity. Again, also in the case of nonlinear correlations, RIC provides much more consistent results also in the presence of a significant number of outliers (the comparative resilience is similar to the case of linear correlations).

The positive qualities of RIC, compared to PCC and IQR, are not a negligible matter in practice because, in real-life applications, the effects of the noise and the uncertainties

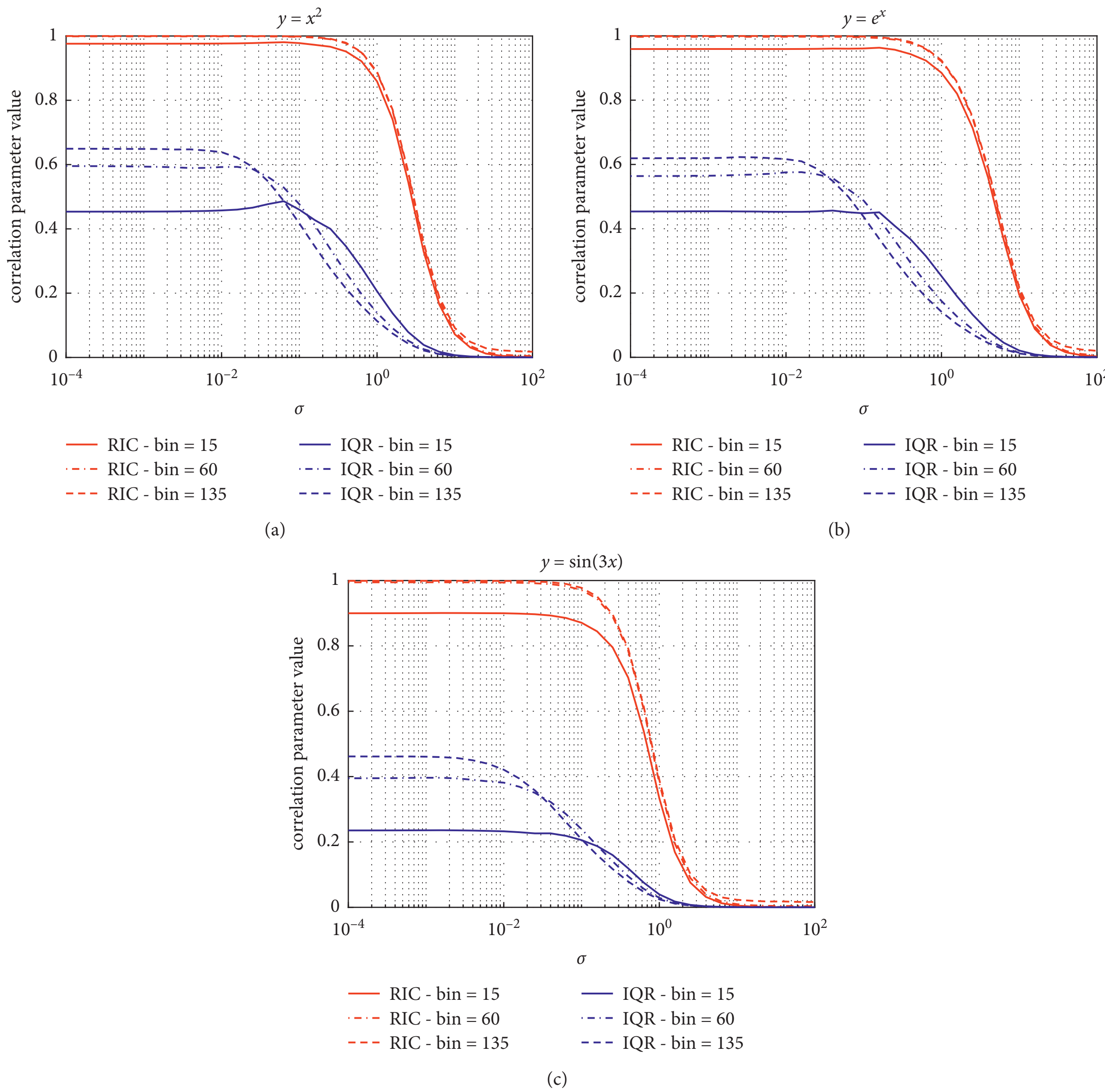

Figure 3: Behavior of PCC, IQR, and RIC with the standard deviation of the additive noise for various nonlinear dependencies. (a) $y = x2$. (b) $y = \sin\ (x)$. (c) $y = \exp\ (x)$.

about the details of the pdfs can have a strong effect on the conclusions.

It should be noted that, from the analysis performed as a function of the binning, it is clear that property 2 of Section 3 is satisfied, i.e., $\Delta N_{\text{bin}}/N_{\text{bin}} \gg \Delta\text{RIC}/\text{RIC}$.

5. The Reciprocal Influence Criterion: Numerical Tests for Nonfunctional Dependencies

The cases treated in this section, to exemplify the properties of RIC for nonfunctional dependencies (with additive Gaussian noise of mean equal to zero and standard deviation equal to 0.1), are shown in Figure 4. These types of dependencies are quite involved and difficult to resolve. They are fully nonlinear and they cannot even be represented by functions. For all these cases, RIC performs significantly better than all other indicators. A synthetic overview of the results is reported in Table 2.

Inspection of Table 2 and Figure 4 reveals that the RIC criterion is always higher than the others by a factor. A part of the case of the rhomboid dependence always provides a value of 0.8 or higher, whereas the other indicators are closer to zero. The RIC, therefore, provides a much more reliable indication that there is a strong correlation between the two variables involved. Even the two other most sophisticated criteria, the IQR and distance correlation, perform significantly worse for all the examples investigated.

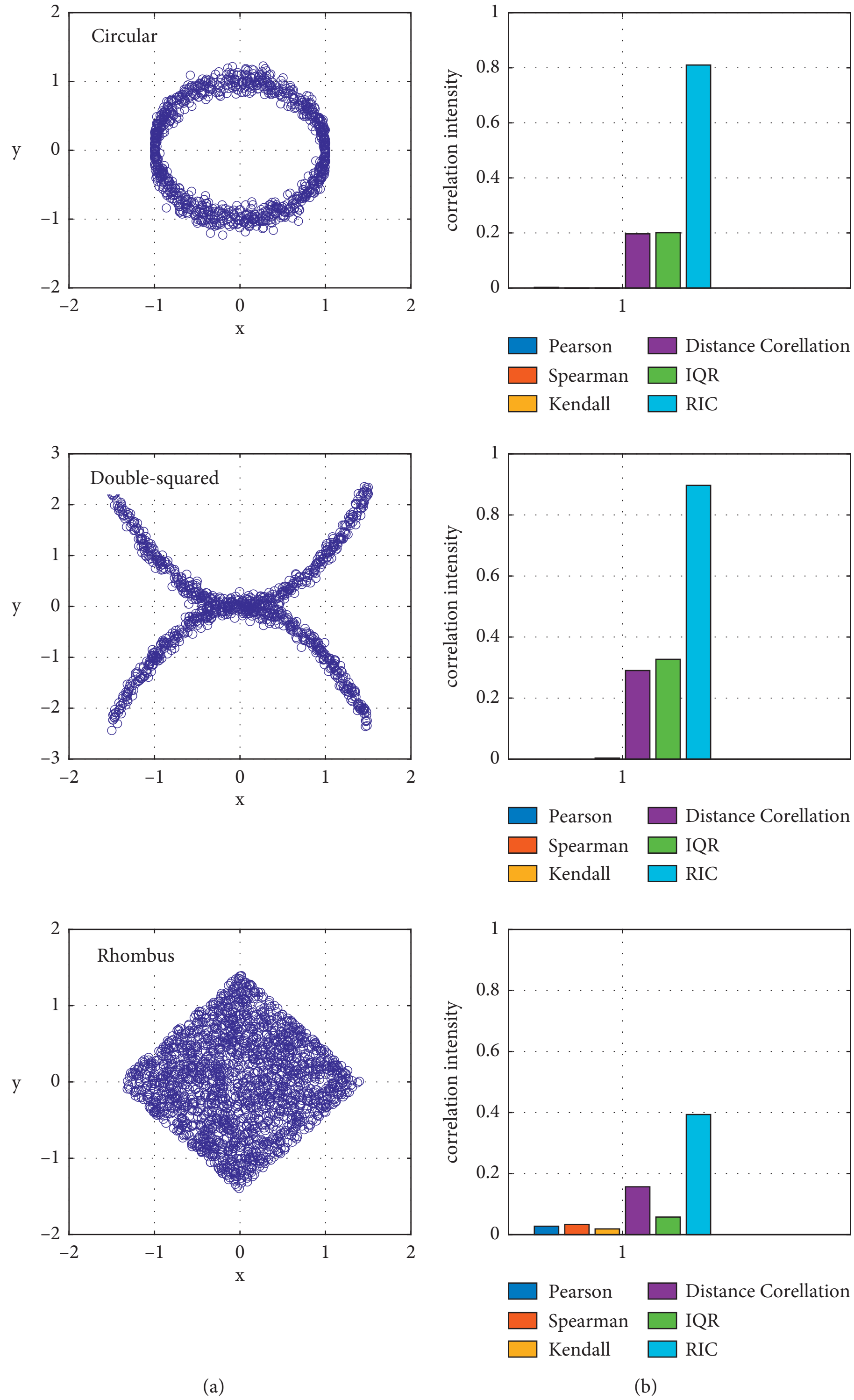

Figure 4: Main examples of the main functional dependencies investigated. (a) Examples of synthetic data. (b) The numerical values of the indicators.

6. Robustness to Noise of Different Statistics and Outliers

The signals and data acquired in many scientific disciplines are typically affected by noise. The assumption of Gaussian statistics is often justified, but there are also other important types of noise of great practical and theoretical importance. Two of the most relevant distributions are certainly the Poisson and gamma.

Poisson distribution:

Table 2: The value of the various indicators for the case of Figure 4.

	Circular	Double squared	Rhombus
Pearson	0.00	0.00	0.03
Spearman	0.00	0.00	0.04
Kendall	0.00	0.00	0.02
Distance correlation	0.19	0.31	0.15
IQR	0.21	0.31	0.05
RIC	0.80	0.89	0.38

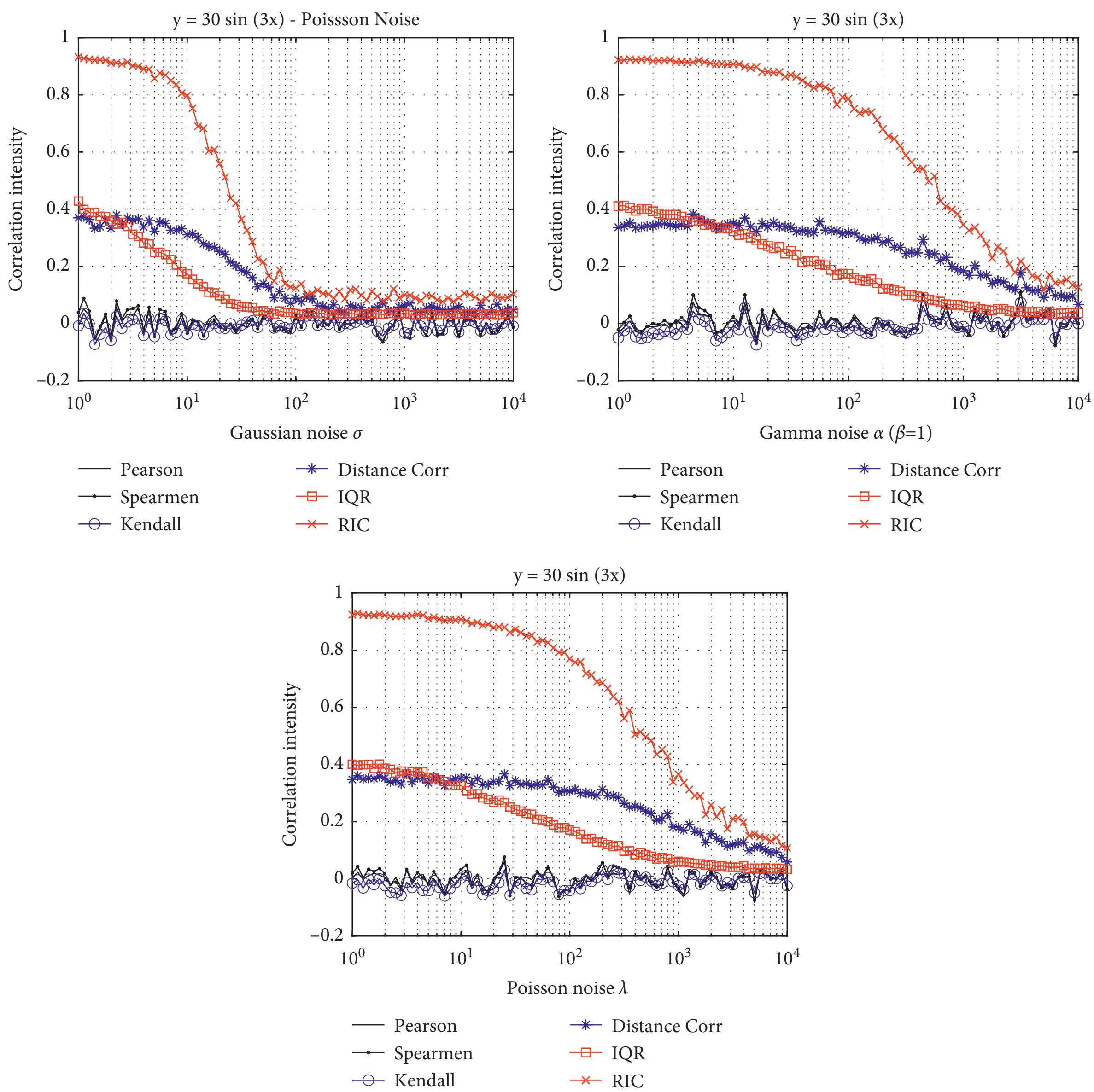

Figure 5: Comparison of RIC and the other criteria for nonlinear correlations, $y = 30\ \sin\ (3x)$, in which the signals are affected by additive noise of different statistics.

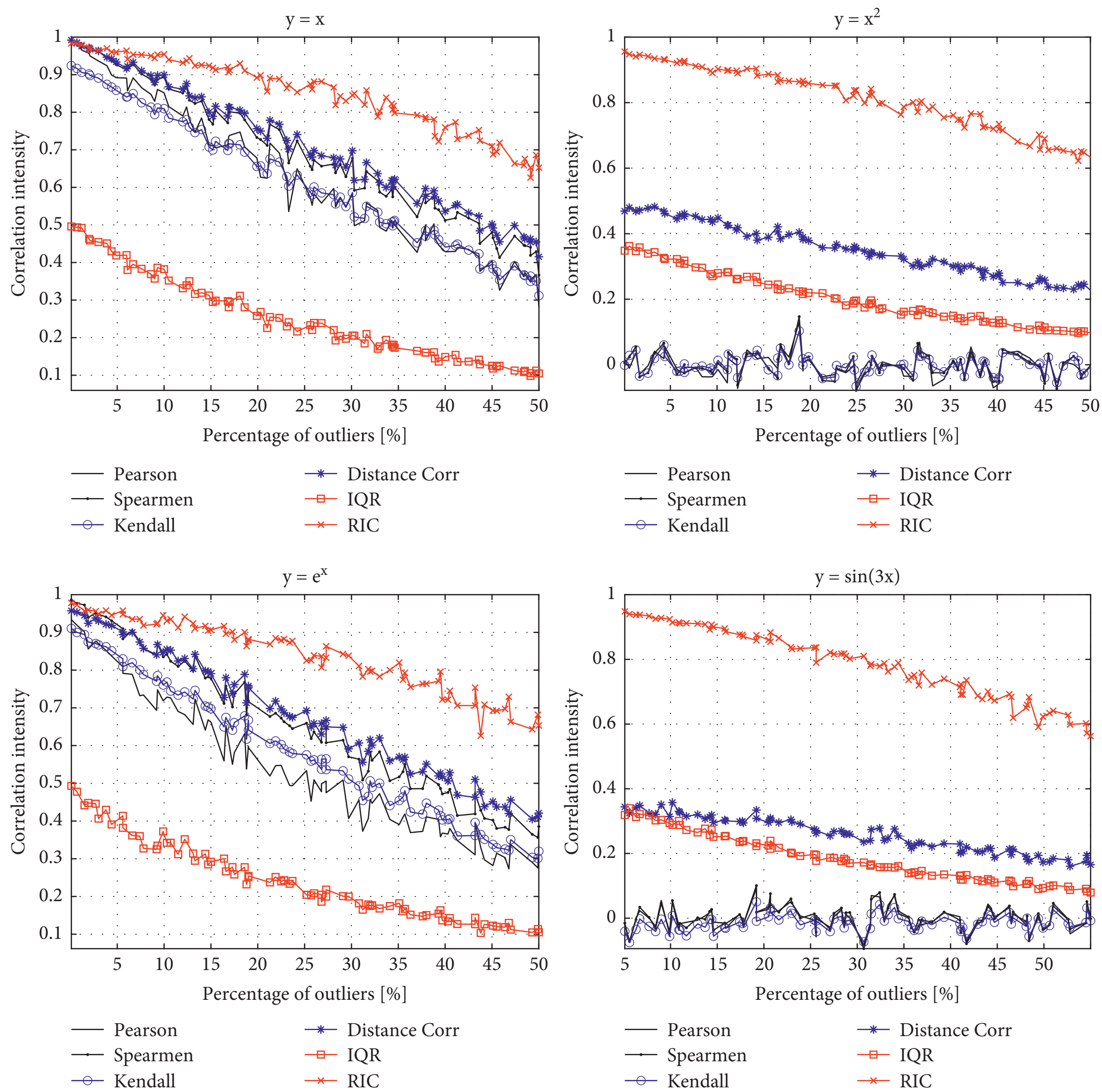

FIGURE 6: Robustness of RIC and other criteria against outliers. The x-axis reports the percentage of outliers generated randomly using a Gaussian distribution of large standard deviation.

$$f(x) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad \forall n \in \mathbb{N}. \tag{21}$$

Gamma distribution:

$$f(x) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad \forall x > 0, k, \theta \in \mathbb{N}. \tag{22}$$

Figure 5 reports some comparative examples of the performance of the various indicators for these two distributions. The cases reported are full representatives of a series of systematic tests performed to investigate this point. In general, as for the Gaussian distribution reported in Figure 5 as a reference, RIC is also much less affected by additive noise of different distributions.

Another potential source of data contamination, of great practical relevance, is the presence of outliers. If the user is aware of the problem and has some prior information about the statistics of the outliers, some measures to remedy the situation can be taken before applying the dependence indicators [17, 18]. These measures belong to the family of robust statistics and can be quite effective. On the contrary, it is not always the case that the practitioner is aware of the issue and therefore investigating the robustness of the various dependence criteria to outliers remains a significant subject. Some representative results of a series of systematic tests, performed to assess this aspect, are shown in Figure 6 for the main classes of functional dependencies. In these cases, the outliers are generated as random Gaussian points with a mean equal to zero and standard deviation

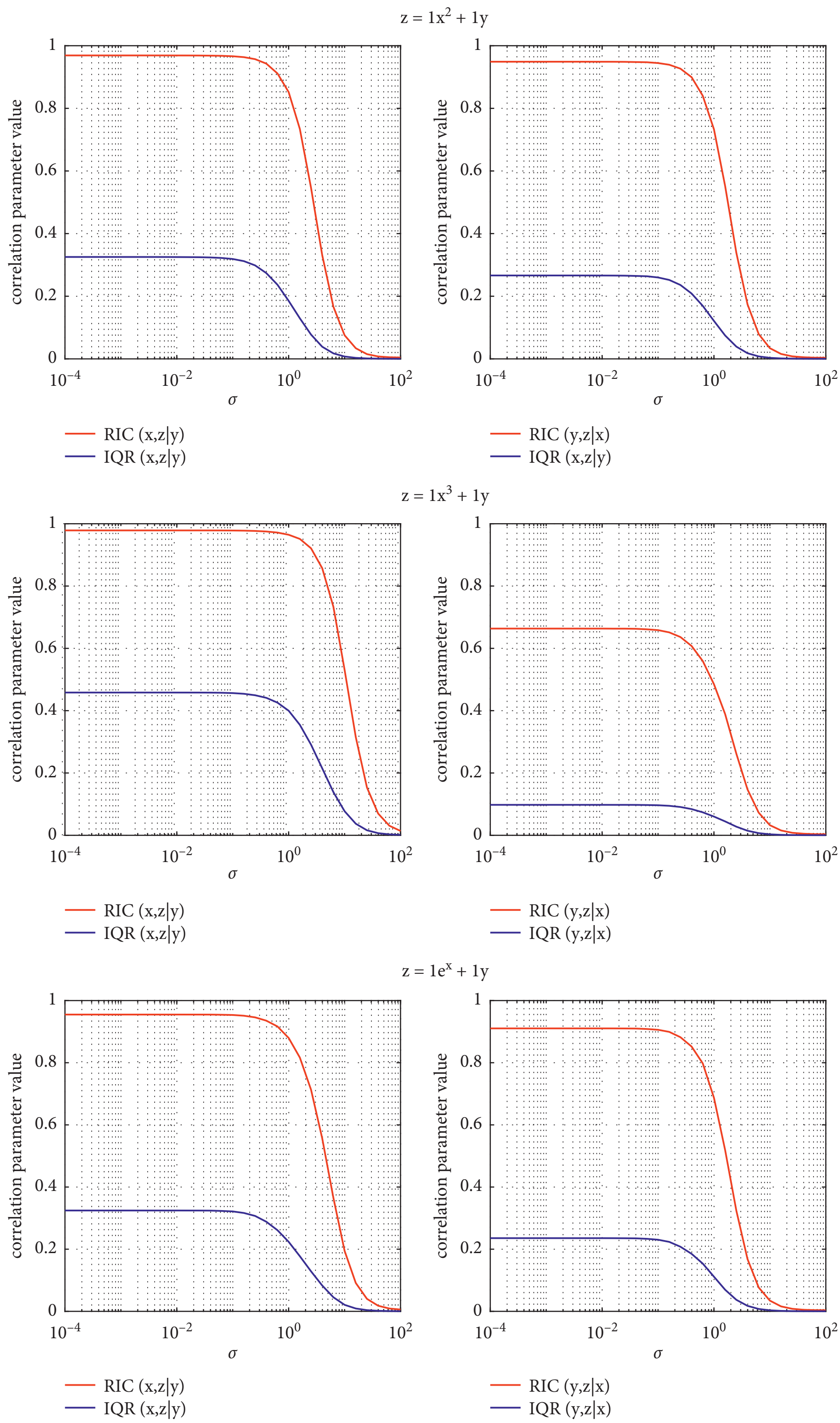

Figure 7: Comparison of $\mathrm{RIC}_{\mathrm{cond}}\,(X, Z\,|\,Y)$ and $\mathrm{IQR}_{\mathrm{cond}}\,(X, Z\,|\,Y)$ for various types of mutual influences in the presence of confounders.

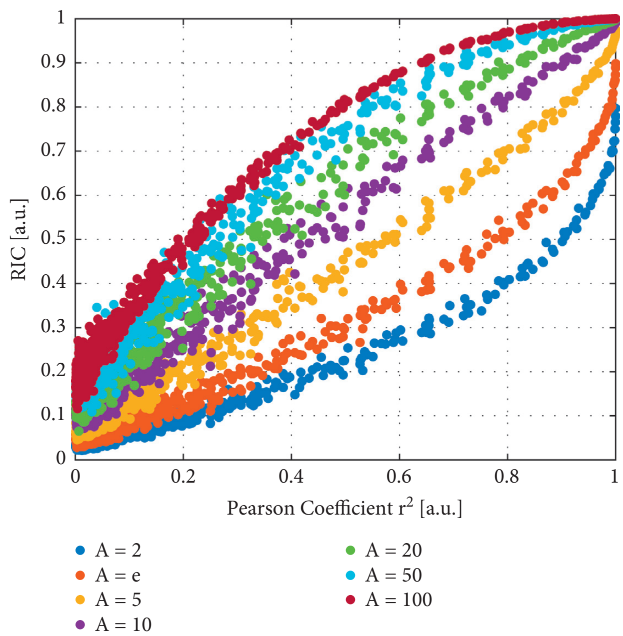

Figure 8: Trends of RIC versus the squared Pearson correlation coefficient for various values of A.

comparable to the range of the function. As expected, the most vulnerable indicator is the PCC. The rank-based criteria and the distance correlation are slightly more insensitive; the outliers must be relatively high both in number and amplitude to affect the ranking, before they have a detrimental effect on the values of these indicators. In any case, even in the presence of outliers, RIC remains the most robust criterion.

7. The Conditional Version of the Reciprocal Influence Criterion

In many applications of correlation analysis, one fundamental objective consists of determining the mutual influence between variables in the presence of confounding factors. It is therefore natural to investigate the potential of a conditional version of RIC:

$$\mathrm{RIC}_{\mathrm{cond}}(X, Z|Y) = 1 - \frac{1}{10^{\mathrm{MI}_{\mathrm{cond}}(X,Z|Y)}}. \tag{23}$$

In the following plots, the performances of $\mathrm{RIC}_{\mathrm{cond}}$ are compared with the ones of a conditional version of IQR:

$$\mathrm{IQR}_{\mathrm{cond}}(X, Z|Y) = \frac{\mathrm{MI}_{\mathrm{cond}}(X, Z|Y)}{H_{XZ}}. \tag{24}$$

The plots of Figure 7 report only the cases of nonlinear correlations ($z = x^2 + y$, $z = x^3 + y$, $z = e^x + y$), but the same conclusions apply also to linear effects.

In addition, in this application, RIC provides much better resilience to noise, and it is less dependent on the choice of the binning to determine the pdfs of the quantities involved. Moreover, even for very low levels of uncertainty, in the limit of no noise, it manages to identify more clearly the mutual correlations actually at play.

8. Discussion and Conclusions

To quantify the mutual influence between quantities, a new indicator has been introduced, the reciprocal influence criterion. A conditional version to separate the effects of confounding factors has also been devised. RIC reproduces the results of PCC in the case of linear correlations but is more robust against the influence of additive noise and outliers. In the case of nonlinear influences, RIC outperforms not only the ranked criteria and the distance covariance but also the information theoretic indicators such as the IQR in many respects; it provides more interpretable results and is more robust against noise and less sensitive to the choice of the binning of the pdfs involved. All these competitive advantages can be quite important in practice.

Other aspects, not to be neglected in the perspective of a wide application of the proposed indicator, are the fact that RIC is conceptually very simple, easy to implement, and fully general, in the sense that it does not rely on specific assumptions about the properties of the stochastic variables involved. In terms of requirements on the data, of course, enough examples must be available to properly calculate the pdfs, but again RIC is more parsimonious than the other indicators, which require estimating the probability distribution functions of the quantities involved.

Regarding future developments, it is planned to investigate whether alternative versions of the entropy and, therefore, of the derived quantities, can help improve the performance of RIC [19–22]. Furthermore, additional formulations, more suited to the investigation of actual causal relations than simple correlations, are also under consideration [23–26]. In terms of practical applications, some of the most immediate range from the investigation of synchronization experiments and disruptions in thermonuclear fusion [27–37] to the refinement of measurement techniques and data analysis methods in earth sciences and plasma physics [38–40].

Appendix

A. Additional Useful Properties of the Reciprocal Influence Criterion

This appendix is devoted to showing how RIC satisfies some basic properties, which are desirable of any correlation criterion. They are symmetry, asymptotic consistency, and unbiasedness (independence from offset).

Symmetry: RIC (X, Y) is equal to RIC (Y, X). This is obviously true, being the mutual information symmetric. Indeed,

$$\mathrm{RIC}(X,Y) = 1 - \frac{1}{10^{\mathrm{MI}(X,Y)}} = 1 - \frac{1}{10^{\mathrm{MI}(Y,X)}} = \mathrm{RIC}(Y,X). \tag{A.1}$$

Asymptotical consistency: this property means basically that the indicator assumes appropriate values when the MI decreases toward zero or becomes very large. Also, in this respect, RIC behaves very satisfactorily. The indicator ranges from 0 to 1 and increases monotonically with MI, as can be derived directly from equations (19) and (20).

Independence to offset (unbiased): this additional very important property assures that a constant offset or bias in the data does not affect the results and can be written as RIC $(X+a,\ Y+b) = \mathrm{RIC}\ (X,\ Y)\ \forall a,\ b \in P$. Also, this property is a direct consequence of the unbiased nature of the mutual information. Indeed,

$$\begin{aligned}\mathrm{RIC}(X+a,Y+b) &= 1 - \frac{1}{10^{\mathrm{MI}(X+a,Y+b)}} = 1 - \frac{1}{10^{\mathrm{MI}(X,Y)}} \\ &= \mathrm{RIC}(X,Y).\end{aligned} \tag{A.2}$$

B. The Choice of Parameter A

In general, the RIC indicator is defined as

$$\mathrm{RIC} = 1 - \frac{A^{H_{xy}}}{A^{H_x+H_y}} = 1 - \frac{1}{A^{\mathrm{MI}}}. \tag{B.1}$$

In this appendix, the behaviour of RIC with respect to A is discussed. The parameter A can indeed be optimised depending on the situation and the objectives of the analysis. This degree of freedom can become handy in various applications. The reason why A has been set equal to 10, to obtain the results presented in his work, is that, for this choice, RIC produces quite well the trends of the PCC in case the correlation between the two quantities analysed is linear. Figure 8 shows how the RIC values vary as a function of the linear correlations between x and y and A.

For $A = 10$, RIC varies almost linearly with squared PCC, and the consistency with this very popular indicator can be considered a positive quality in most applications.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Disclosure

Rossi and Gelfusa are the co-first authors. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data, in the writing of the manuscript, and in the decision to publish the results.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

All authors have contributed equally to this study.

References

[1] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/correlation Analysis for the Behavioural Sciences*, Psychology Press, Hove, East Sussex, UK, 3rd edition, 2002.

[2] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, no. 347, pp. 240–242, 1895.

[3] C. Spearman, "The proof and measurement of association between two things," *American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[4] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–89, 1938.

[5] S. Guiasu, *Information Theory with Applications*, McGraw-Hill, New York, NY, USA, 1977.

[6] G. . J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing independence by correlation of distances," *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[7] L. Zhu, K. Xu, R. Li, and W. Zhong, "Projection correlation between two random vectors," *Biometrika*, vol. 104, no. 4, pp. 829–843, 2017.

[8] W. Pan, "Ball covariance: a generic measure of dependence in banach space," *Journal of the American Statistical Association*, vol. 115, pp. 2020–529.

[9] G. J. Székely and M. L. Rizzo, "Brownian distance covariance," *Annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[10] SPSS Tutorials: Pearson Correlation, 2017, https://libguides.library.kent.edu/SPSS/PearsonCorr.

[11] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-by-Step Approach*, Wiley, Hoboken, NJ, USA, 2014.

[12] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, UK, 2003.

[13] C. Arndt, *Information Measures, Information and its Description in Science and Engineering*, Springer Series: Signals and Communication Technology, 2004.

[14] T. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, NY, USA, 2nd edition, 2006.

[15] R. Wilcox, "Introduction to robust estimation and hypothesis testing," *Statistical Modeling and Decision Science*, Elsevier/Academic Press, Amsterdam, Netherlands, 3rd edition, 2012.

[16] P. Baudot, M. Tapia, D. Bennequin, and J.-M. Goaillard, "Topological information data analysis," *Entropy*, vol. 21, no. 9, p. 869, 2019.

[17] S. Liu, *Distance Correlation*, 2020, https://www.mathworks.com/matlabcentral/fileexchange/39905-distance-correlation.

[18] T. P. Hettmansperger and J. W. McKean, *Robust Nonparametric Statistical Methods, Kendall's Library of Statistics*, John Wiley & Sons, Inc., New York, NY, USA, 1998.

[19] F. Nielsen and R. Nock, "A closed-form expression for the Sharma-Mittal entropy of exponential families," *Journal of Physics A: Mathematical and Theoretical*, vol. 45, no. 3, Article ID 032003, 2011.

[20] A. Rényi, "On measures of information and entropy," in *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pp. 547–561, Berkeley, CA, USA, 1961.

[21] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1-2, pp. 479–487, 1988.

[22] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics*, Springer, Berlin, Germany, 2009.

[23] M. C. Romano, M. Thiel, J. Kurths, and W. von Bloh, "Multivariate recurrence plots," *Physics Letters A*, vol. 330, no. 3-4, pp. 214–223, 2004.

[24] N. Marwan, M. Carmenromano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Physics Reports*, vol. 438, no. 5-6, pp. 237–329, 2007.

[25] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, pp. 461–464, 2000.

[26] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[27] A. Murari, M. Lungaroni, E. Peluso et al., "On the use of transfer entropy to investigate the time horizon of causal influences between signals," *Entropy*, vol. 20, no. 9, p. 627, 2018.

[28] A. Murari, E. Peluso, M. Gelfusa et al., "Application of transfer entropy to causality detection and synchronization experiments in tokamaks," *Nuclear Fusion*, vol. 56, no. 2, Article ID 026006, 2016.

[29] A. Murari, "How to assess the efficiency of synchronization experiments in tokamaks," *Nuclear Fusion*, vol. 56, Article ID 076008, 2016.

[30] A. Murari, T. Craciunescu, E. Peluso, E. Lerche, and M. Gelfusa, "On efficiency and interpretation of sawteeth pacing with on-axis ICRH modulation in JET," *Nuclear Fusion*, vol. 57, no. 12, Article ID 126057, 2017.

[31] E. Lerche, M. Lennholm, I. S. Carvalho et al., "Sawtooth pacing with on-axis ICRH modulation in JET-ILW," *Nuclear Fusion*, vol. 57, no. 3, Article ID 036027, 2017.

[32] L. Garzotti, "Investigating pellet ELM triggering physics using the new small size pellet launcher at JET," in *Proceedings of the 37th EPS Conference on Controlled Fusion and Plasma Physics*, European Physical Society, Dublin, Ireland, June 2010, https://ocs.ciemat.es/EPS2010PAP/pdf/P2.131.pdf.

[33] P. C. De Vries, M. Baruzzo, G. M. D. Hogeweij et al., "The influence of an ITER-like wall on disruptions at JET," *Physics of Plasmas*, vol. 21, no. 5, Article ID 056101, 2014.

[34] J. Paméla, F. Romanelli, M. L. Watkins et al., "The JET programme in support of ITER," *Fusion Engineering and Design*, vol. 82, no. 5-14, pp. 590–602, 2007.

[35] J. Pamela, "Overview of JET results," *Nuclear Fusion*, vol. 43, no. 12, 2003.

[36] A. Murari, P. Boutot, J. Vega et al., "Clustering based on the geodesic distance on Gaussian manifolds for the automatic classification of disruptions," *Nuclear Fusion*, vol. 53, Article ID 033006, 2013.

[37] A. Murari, F. Pisano, J. Vega et al., "Extensive statistical analysis of ELMs on JET with a carbon wall," *Plasma Physics and Controlled Fusion*, vol. 56, no. 11, Article ID 114007, 2014.

[38] F. P. Orsitto, A. Boboc, P. Gaudio et al., "Mutual interaction of Faraday rotation and Cotton-Mouton phase shift in JET polarimetric measurements," *Review of Scientific Instruments*, vol. 81, no. 10, Article ID 10D533, 2010.

[39] C. Bellecci, M. Francucci, P. Gaudio et al., "Application of a CO_2 dial system for infrared detection of forest fire and reduction of false alarm," *Applied Physics B*, vol. 87, no. 2, pp. 373–378, 2007.

[40] F. P. Orsitto, "Analysis of Faraday rotation in JET polarimetric measurements," *Plasma Physics and Controlled Fusion*, vol. 53, no. 3, 2011.