Hindawi
Complexity
Volume 2022, Article ID 9518303, 12 pages
https://doi.org/10.1155/2022/9518303

WILEY | Hindawi

Research Article

Complexity: Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems 2022 on the Development of Information Theoretic Model Selection Criteria for the Analysis of Experimental Data

Andrea Murari,[1] Michele Lungaroni,[2] Riccardo Rossi,[2] Luca Spolladore,[2] and Michela Gelfusa[2]

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

Correspondence should be addressed to Michele Lungaroni; michele.lungaroni@uniroma2.it

Received 21 January 2022; Accepted 20 July 2022; Published 24 August 2022

Academic Editor: M. De Aguiar

It can be argued that the identification of sound mathematical models is the ultimate goal of any scientific endeavour. On the other hand, particularly in the investigation of complex systems and nonlinear phenomena, discriminating between alternative models can be a very challenging task. Quite sophisticated model selection criteria are available but their deployment in practice can be problematic. In this work, the Akaike Information Criterion is reformulated with the help of purely information theoretic quantities, namely, the Gibbs-Shannon entropy and the Mutual Information. Systematic numerical tests have proven the improved performances of the proposed upgrades, including increased robustness against noise and the presence of outliers. The same modifications can be implemented to rewrite also Bayesian statistical criteria, such as the Schwartz indicator, in terms of information-theoretic quantities, proving the generality of the approach and the validity of the underlying assumptions.

1. Introduction to Nonfrequentist Model Selection Criteria

The promised land of modern scientific enterprises is often the formulation of robust and generally applicable mathematical models [1, 2]. The ultimate validation of any model resides in the comparison with the results of experiments or observations. In the last decades, enormous quantities of data have become available in many fields of science and engineering. The statistical inference has therefore progressively moved to centre stage. The older frequentist techniques, based on traditional significance level criteria, have been complemented by a series of Bayesian and information-theoretic criteria, in many respects more suited to managing large amounts of information.

One of the most popular model selection criteria (MSC) is the Akaike Information Criterion (AIC) [3]. The AIC can be derived from the Kullback–Leibler divergence and can be interpreted as the loss of information associated with the adoption of a model different from the exact one, generating the data. The basic idea underlying the AIC criterion resides indeed in the consideration that the less information a model loses, the higher its quality. The theoretical derivation of the AIC gives the unbiased form of the criterion [4].

$$AIC = -2\ln(L) + 2k, \tag{1}$$

where L is the likelihood of the data given the model and k is the number of estimated parameters in the model. The AIC is a metric that is minimised by the best model as a compromise between the goodness of fit (the first term) and complexity (the second term).

The general formulation of the AIC is not always easy to apply in practice as can be appreciated by a simple inspection of (1). First, in many instances, it can be impossible to

reliably calculate the likelihood. Moreover, it is well known that the number of parameters is a poor quantifier of a model complexity and it is not inherently an information-theoretic indicator. The more practical expression of the AIC, very often the one used in practice, is even more distant from its original information theoretic origin, as discussed in the next section.

The first quantity, proposed to improve the AIC, is the Gibbs–Shannon entropy H

$$H = -\sum_i p_i \log\ p_i. \tag{2}$$

The higher the value of H, the higher the uniformity of the corresponding probability distribution function (whose values are indicated with p_i). The Gibbs- Shannon entropy can improve significantly the quantification of the model complexity, as discussed in detail in Section 2.2.

The second quantity, used in the rest of the work, is the mutual information, MI.

$$MI = -\sum_x \sum_y p_{xy} \ln\left(\frac{P_{xy}}{P_x P_y}\right), \tag{3}$$

where $P_{x,y}$ is the joint pdf of the random variables X and Y. Mutual Information can play a fundamental role in determining the goodness of fit of the models, as discussed in Section 2.1.

With regard to the organization of the paper, the next section introduces the rationale and details of the proposed information-theoretic upgrades of the Akaike Information Criterion. Section 3 is devoted to a simple but challenging didactic case, meant to illustrate the effects of the modifications with an easy-to-grasp example. The family of functions and the types of noise statistics, implemented to perform a series of systematic tests, are summarised in Section 4. The results of the aforementioned tests are exemplified in Section 5 with the help of some representative cases. The extension of the approach to the Bayesian Selection criterion is covered in Section 6 before the conclusions and lines of future developments are discussed in the final section of the paper.

2. Model Selection Formulated in terms of Information Theoretic Quantities

Among the many indicators, for identifying the "best model" among a set of candidates, the Akaike Information Criterion AIC can be conceived originally as a pure information theoretic criterion. Unfortunately, the original formulation of the AIC criterion is typically problematic to implement in practice, particularly in applications involving complex systems and nonlinear phenomena. Both terms in the AIC present significant issues [5–7]. To bypass the practical difficulties of calculating the likelihood, the strong assumption that the data are identically distributed and independently sampled from a normal distribution is the most commonly invoked. If this traditionally called iid hypothesis is valid, it can be demonstrated that the AIC can be written (up to an additive immaterial constant depending only on the number of entries in the database) as follows:

$$AIC = n \cdot\ \ln(MSE) + 2k. \tag{4}$$

In (4), formally derived in [4], the Mean Squared Error (MSE) is calculated in terms of the residuals, the differences between the data, and the estimates of the models; in its turn n indicates the number of entries in the database.

(4) is certainly the most widely used form of AIC. On the other hand, as can be easily appreciated by inspection, the criterion is now expressed in terms of quantities, which are not information theoretic anymore. Moreover, all the statistical information content, originally in the likelihood, is reduced to the mere MSE of the residuals. The first obvious question, which comes to mind, is whether some additional statistical information about the distribution of the residuals could be taken into account, to improve the discriminatory capability of the criterion. The practical relevance of this issue is quite significant also because, in many applications, the assumptions behind (4) are clearly violated. In real life, indeed, the statistics of the noise can have a non-Gaussian distribution, memory effects can be important, and a significant number of outliers can be unavoidable. How to improve the model selection criteria in this respect is the subject of Section 2.1.

The second term in (4) is also problematic because it is well known that the number of parameters is a quite poor indicator of the complexity of a model. More sophisticated quantifiers exist, such as the VC dimension [8] and the Rademacher dimension [9], but they are often impossible to calculate for most practical functions. An alternative information theoretic and computationally simple way to calculate a model complexity is the subject of Section 2.2.

2.1. Expressing the Goodness of Fit in terms of Mutual Information. The main idea informing one of the AIC upgrades, proposed in this work, is based on the observation that the better a model, the more similar the residuals to the noise affecting the measurements. In the case of a perfect model, the residuals should present exactly the same distribution as the noise. Assuming that the noise is not correlated with the measurements, absolutely legitimate in most practical applications, this consideration can be quantified mathematically by calculating the mutual information between the model predictions and the residuals, MI_{MRes}.

$$MI_{MRes} = MI\left(y_{mod}, y_{res}\right). \tag{5}$$

The AIC can therefore be rewritten as follows:

$$AIC_{MI} = 2k + n\ \ln\left(MSE\left(1 + MI_{MRes}\right)\right). \tag{6}$$

Conceptually, (6) is to be preferred to (4) for various reasons. First, it formulates the criterion in terms of an information theoretic quantity, the mutual information. Moreover, it retains much more statistical information about the model and the residuals. At the same time, MI_{MRes} takes into account also nonlinear correlations and does not make any "a priori" assumption about the statistics of the

noise or the presence of outliers. Consequently, as shown by numerical tests, AIC_{MI} is a much more general and sensitive model selection criterion than the original AIC.

2.2. Expressing the Complexity in terms of the Shannon Entropy. The other weakness in the original definition of AIC is certainly the quantification of complexity. Indeed, the simple number of parameters in a model is a very poor indicator of its flexibility and in particular of its potential to overfit (see Section 3). A possible alternative relies on the traditional idea that complexity is the middle ground between randomness and determinism. According to this view, complete randomness and perfect determinism are considered less complex than a combination of the two. This approach to complexity has a long pedigree and can be traced back to the interpretation of information as uncertainty, the concept at the basis of information theory [10]. A possible way of expressing this idea in mathematical terms is the following complexity measure C[X]:

$$C[X] = H^{\alpha}[X]D^{\beta}[X], \tag{7}$$

where H is the usual Shannon entropy and D is the distance from a uniform distribution.

$$D[X] = \sum_{1}^{N}\sum_{1}^{N}\left(p_i - \frac{1}{n}\right), \tag{8}$$

where with the usual notation, n is the number of entries in the database. The distance D reduces the estimated complexity of models, whose predictions are uniform. The entropy reduces the estimated complexity of models, whose outputs are concentrated on a few well-defined values. Conceptually, the implementation of this quantification of complexity is quite simple. The pdf of the model predictions can be inserted in (7) to obtain a simple indicator, implementing the aforementioned information theoretic interpretation of complexity.

The most delicate aspect of (7) is the choice of the exponents α and β because they contribute significantly to determining the trade-off between entropy and distance. To this end, the increments of the model predictions have been calculated as follows:

$$\text{Model}_{diff} = \left(y_{\text{model},i+1} - y_{\text{model},i}\right). \tag{9}$$

The moving averages (Mov), of the mean and standard deviation of the squared increments, are good indicators of the flexibility of a model and therefore of its potential to overfit. The normalized versions of these quantities are defined in

$$MF_1 = \frac{\sum MovST\ D\left(\text{Model}_{di\ ff}\right)^2}{n}$$
$$MF_2 = \frac{\sum Mov\text{MEAN}\left(\text{Model}_{di\ ff}\right)^2}{n}. \tag{10}$$

The ratio of the two averages calculated in (10) is

$$MF = \sqrt{\frac{MF_1}{MF_2}}. \tag{11}$$

The parameter MF increases for functions, which have stronger variations in the domain of interest and can therefore be considered more complex. Indeed, these more nervous functions would have a higher potential of overfitting the data, following the noise. This is the interpretation of the quantity MF, which is used to determine the exponents α and β.

$$\alpha = 1 + MF; \beta = 1 - MF. \tag{12}$$

Finally, the proposed final versions of the AIC expressed only in terms of the mentioned information theoretic quantities read

$$\begin{aligned} AIC_{MICx} &= n\ \ln[MSE(1 + MI)] + n\left(\ln C_x\right) \\ &= n\left(\ln\left[C_x MSE(1 + MI)\right]\right). \end{aligned} \tag{13}$$

3. A Didactic Example to Illustrate the Main Characteristics of AIC_{MICX}

To illustrate the potential and the meaning of the proposed upgrades of the AIC, an academic but challenging example, already discussed in detail in the literature [11], is described in this section. To this end, it is assumed that the actual data is generated with a polynomial function depending on 5 parameters.

$$y_{ref} = 10^{-6}x^5 - 8, 10^{-3}x^3 + 3, 10^{-2}x^2 + x - 10. \tag{14}$$

The equations, considered as possible candidate models for the data generated with (14), are reported in Table 1.

A comment about the sinusoidal functions is in place. These functions can be tuned to fit perfectly the data generated with (14) by increasing their frequency. This fact can be appreciated by inspection of the first two plots of Figure 1. If there is any noise added to the data, the sinusoidal functions, given their higher flexibility, can fit the data even better than the original equation generating it.

On the other hand, they depend only on two parameters, their amplitude and frequency. Therefore, the traditional version of the AIC would tend to prefer a well-adjusted sinusoidal model (because it would achieve lower values of both terms of the indicator). The proposed version AIC_{MICx}, on the contrary, manages to properly identify the right model, as shown in Figure 2. The plots report the differences between the AIC and AIC_{MICx} of the candidate models and the reference, the equation used to generate the data.

When these differences are positive, the reference model is the preferred one; the negative cases indicate that the criteria would have selected the wrong model. From the plots of Figure 2, it appears quite clearly that the traditional AIC would have preferred the sinusoids (particularly model 1) for various numbers of entries, whereas the AIC_{MICx} always identifies the reference model as the right one. This is achieved by taking into account the distributions of the

Table 1: The four candidate models to fit the data generated by (14).

#	Models
1	$17\ \sin(210x)$
2	$17\ \sin(209.5x)$
3	$-0.08x^2 + 1.47x - 10.38$
4	$0.75x - 10$

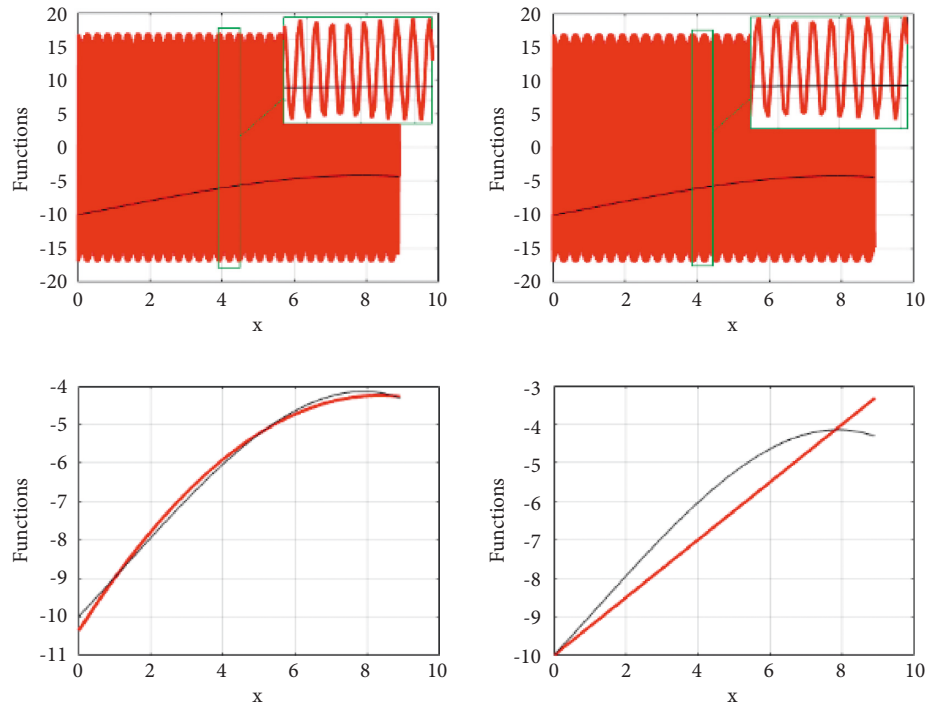

Figure 1: Black: the original data generated with (14). Red: the models of Table 1. From top left to bottom right models from 1 to 4.

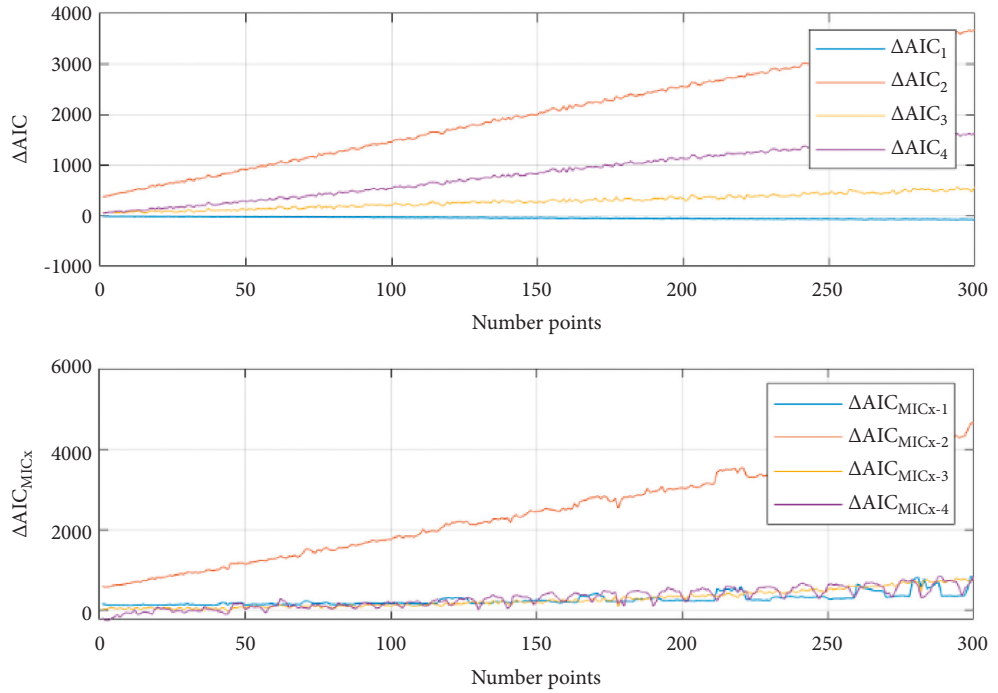

Figure 2: Comparison of the discriminating power of the traditional AIC and the proposed AIC_{MICx}. On the (y) axis the difference between the indicators for the various models and the reference one used to generate the data, is reported. On the (x) axis a scan in the number of entries.

residuals and by better estimating the complexity of the models. The details about the comparison, between the traditional AIC and the new version proposed in this paper, are fully documented in Appendix A for the specific example reported in this section.

4. The Main Functional Classes and Noise Statistics for Practical Applications

To assess the performance of the alternative AIC model selection criterion proposed in Section 2, a series of systematic numerical tests have been performed. The analysis is focussed mainly on four classes of models that cover the most widely used in practice. They are the classes of polynomials, power laws, power laws multiplied by a squashing term, and exponential functions. In the rest of the paper, only the results for bidimensional functions (of the form $z = f(x, y)$) are discussed, because they are susceptible of clear visualization, which helps illustrating the properties of the criterion. The extension to a larger number of variables is straightforward and does not pose any conceptual difficulty. Therefore, the considerations and conclusions reported have to be assumed valid also in higher dimensions. For the reader's convenience, the mathematical form of the aforementioned models is reported in the left column of Table 2.

Significant attention has been devoted to noise statistics. Three of the most relevant distribution functions have been tested: Gaussian, uniform, and multi-Gaussian [12]. Again for the reader's convenience, the mathematical formulation of these types of noise is summarised in the right column of Table 2, together with the parameter values valid for the runs reported in the rest of the paper. Since in practice very often the presence of outliers in the data cannot be excluded, the robustness of the proposed upgrade of the AIC in this respect has also been verified. This has been achieved by randomly adding to the synthetic data values sampled from a Gaussian distribution of small variance but nonzero mean (see the entry called Asymmetric noise in Table 2 for a precise mathematical definition).

Table 2: The main families of functions tested and the statistics of the additive noise.

Families of functions	Additive noise applied
Polynomials $y = a_0 x^{b_0} + a_1 x^{b_1} + a_n x^{b_n}$	*Uniform Noise* $\mu = \pm 10$ until ± 50
Power Laws $y = a_0 x^{b_0}, x^{b_1}, x^{b_n}$	*Traditional Gaussian Noise* $\mu = 0$ range of $\sigma = \pm 10$ until ± 50
Power Laws with Squashing term $y = a_0 x^{b_0}, x^{b_1}, x^{b_2} \frac{1}{1+exp(-a_n x^{b_n})}$	*Multi-Gaussian Noise* $\mu_i = 0$ range of $\sigma_i = \pm 10$ until $\pm 50 \forall i = 1..n$
Exponentials $y = a_0 x^{b_0} \exp(a_n x^{b_n})$	*Asymmetric Noise* $\mathcal{N}_1$: $\mu_1 = 0$ and $\sigma_1 = 10$; $\mathcal{N}_2$: $\mu_2 \neq 0$ and $\sigma_2 = 30$; with $\mu_2 = 2(\sigma_1 + \sigma_2)/100, \overline{f(x)}$ Ratio between $\mathcal{N}_1, \mathcal{N}_2 \Rightarrow 0.75$ until 0.95

5. Representative Results of Numerical Tests

As mentioned, a systematic series of tests with synthetic data has been performed to assess the competitive advantage of the proposed version of the AIC. All the combinations of cases summarised in Section 4 have been investigated. The new version $\mathrm{AIC}_{\mathrm{MICx}}$ has always proved to have better discriminatory capabilities than the traditional AIC. In practice, this means that $\mathrm{AIC}_{\mathrm{MICx}}$ at least provides better separation between the right model (the one used to generate the data) and its wrong competitors. This has proved to occur for any type of function, noise statistics, and levels of outliers. In general, the more severe the conditions, the higher the level of noise or outliers, and the better the $\mathrm{AIC}_{\mathrm{MICx}}$ performance compared to the traditional AIC. In some cases, as the one already discussed in Section 3, only the $\mathrm{AIC}_{\mathrm{MICx}}$ can converge on the right model.

In the rest of this section, some relevant examples of the performed tests are reported. They have to be considered absolutely representative of the vast majority of systematic investigations performed.

In the first case discussed in the following, the model generating the data consists of a power law multiplied by a squashing term. The importance and popularity of power laws are difficult to overstate. Self-similarity can result in many quantities presenting a power law trend. Power laws are also particularly important for the investigation of scalings. On the other hand, power law monomials can be too rigid and the multiplication by a squashing factor can provide some additional flexibility. The function implemented to generate the synthetic data is reported in the last row of Table 3. The other rows of the same table report the alternative models. The synthetic data generated with the reference model of Table 3 is shown in Figure 3, together with the functions constituting the alternative models. Two different levels of Gaussian additive noise are shown; corresponding to a standard deviation of 15% and 30% of the synthetic data averaged amplitude. As can be derived by simple inspection of the plots, $\mathrm{AIC}_{\mathrm{MICx}}$ not only increases the separation between the models, compared to the traditional AIC, but it also allows identifying the equation generating the data. Indeed whereas, for some numbers of entries and 30% of added noise, the AIC of the candidate models can be lower than the reference one, the $\mathrm{AIC}_{\mathrm{MICx}}$ always identifies the model generating the data as the best; this can be seen by noticing that the values of the $\mathrm{AIC}_{\mathrm{MICx}}$ differences, with respect to the best model, are always positive.

The discriminatory power of $\mathrm{AIC}_{\mathrm{MICx}}$ is even higher in the case of high noise. This fact is exemplified by the following example, in which the generating model belongs to the class of exponential functions. The alternative models are reported in Table 4, whose last row reports the equation used to generate the data. In addition to Gaussian noise, with a standard deviation of 30% and 60% of the synthetic data averaged amplitude, some concentrated high noise has also been added, according to the relations specified in the last row of Table 2. The better performance of $\mathrm{AIC}_{\mathrm{MICx}}$ compared to the traditional AIC can be easily recognised by

Table 3: Power law plus a squashing term.

#	Models	k
1	$1.6810^4 \sin(x_1/x_2^{4.18})$	4
2	$3x_2 \exp(-x_3^{9.48})$	4
3	$17.87(x_1/x_2^{0.45})^{0.47}$	4
4	$3.5x_1^{0.4}x_2^{0.8}$	3
ref	$\mathbf{2x_1^{0.6}x_2^{1.1}1/1 + \exp(-2x_3^{1.5})}$	**6**

The value is shown in bold because it is the reference model.

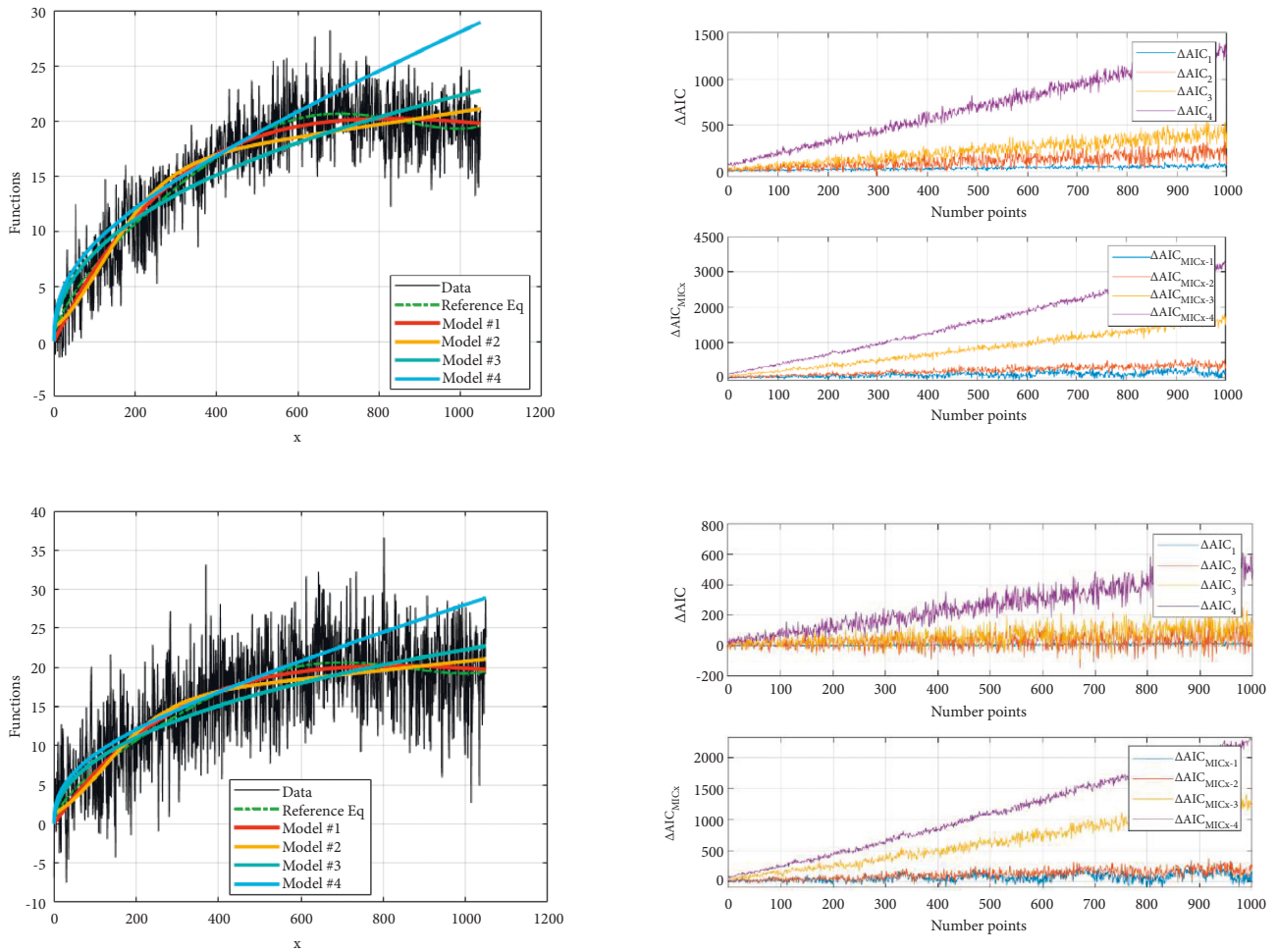

Figure 3: Model selection performances for two levels of additive noise: 15% top and 30% bottom. For each level of noise, the top plots show in black the synthetic data generated with the reference equation of Table 2. The coloured curves are the various candidate models and in dashed point green is the reference one. The bottom plots are the comparison of AIC and AIC_{MICx} results in terms of the difference with respect to the exact reference model.

Table 4: Power law plus a squashing term.

#	Models	k
1	$0.4x^{0.2}\exp(x)$	4
2	$0.8\ \exp(x) - 0.4\,x^2)$	5
3	$3x^2/1 + \exp(-0.1x)$	5
4	$0.5x^3 + 2x$	4
ref	$\mathbf{0.6x\ exp(x^{0.6})}$	**4**

The value is shown in bold because it is the reference model.

inspection of the plots in Figure 4. Indeed, the separation between the alternative models and the right one is much larger for the AIC_{MICx} than for the traditional AIC (the reader should please consider also the different scales of the plots in Figure 4).

6. Extension to Bayesian Model Selection

It is worth noting that the same modifications proposed for the AIC can be applied also to the Bayesian information criterion (BIC) [13]. BIC is based on Bayesian theory and has been designed to maximize the posterior probability of a model given the data. BIC is again a cost function and therefore it is also an indicator to be minimised. The BIC's most general form is

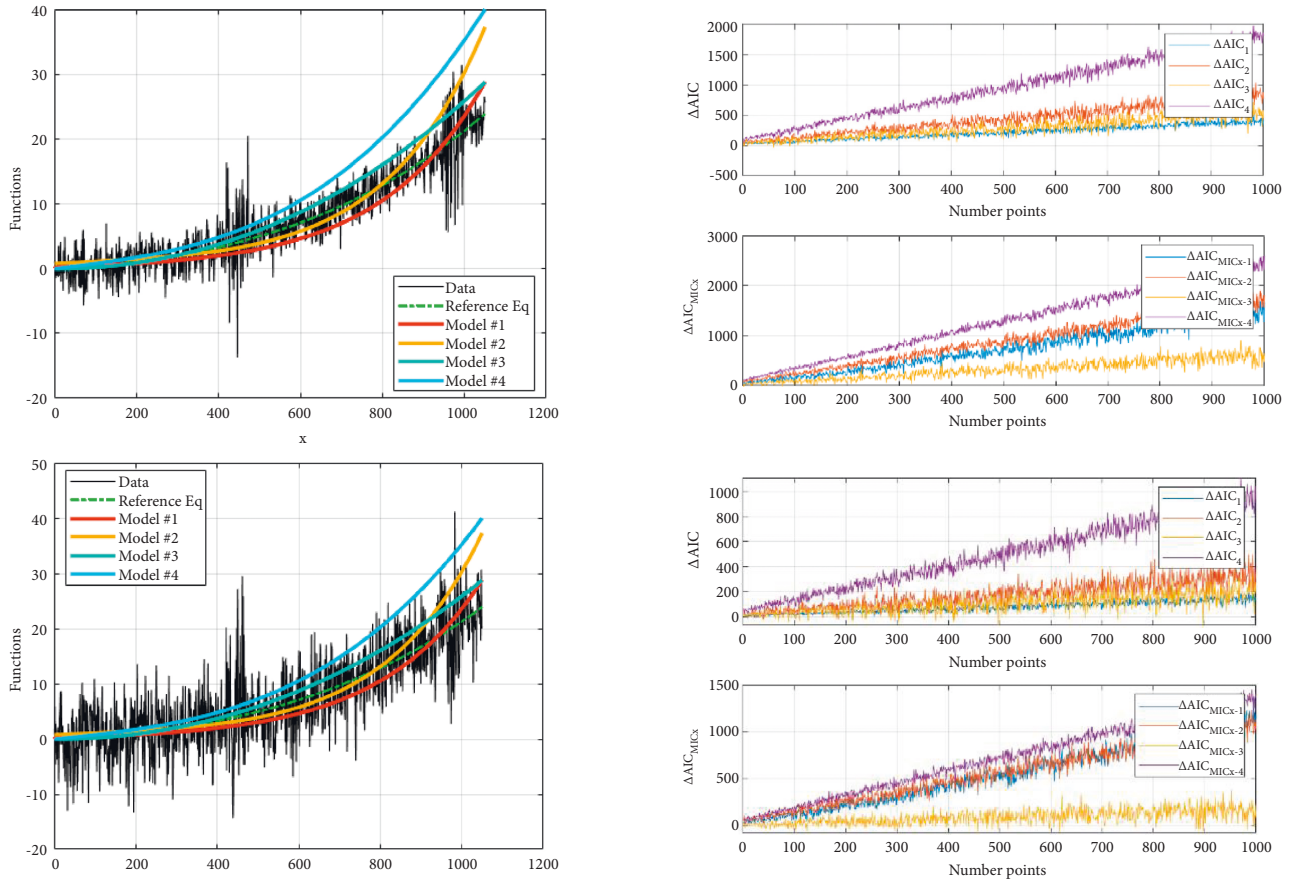

FIGURE 4: Model selection performances for two levels of additive noise: 30% top and 60% bottom. For each level of noise, the top plots show in black the synthetic data generated with the reference equation of Table 4. The coloured curves are the various candidate models and in dashed point green is the reference one. The bottom plots are the comparison of AIC and AIC_{MICx} results in terms of the difference with respect to the exact reference model.

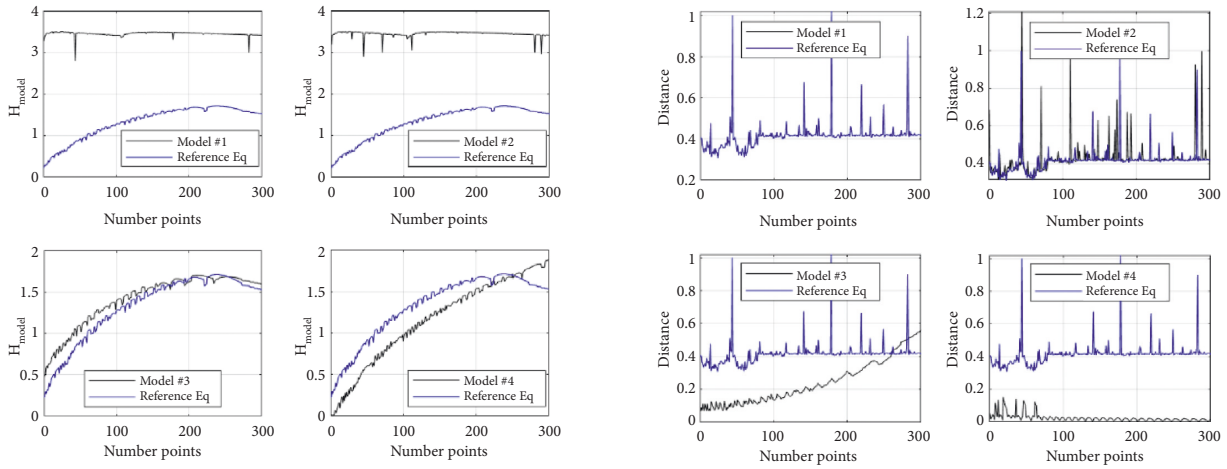

FIGURE 5: Plots of the entropy H and distance D for the models of Table 1 in Section 3.

$$BIC = -2\ln(L) + k\ln(n), \tag{15}$$

where again L is the likelihood of the data given the model, k is the number of estimated parameters in the model, and n is the number of entries in the database. BIC has the same structural form as the AIC and is affected by the same difficulties in practical applications, in particular the challenges posed by the calculation of the likelihood and the quantification of the model complexity.

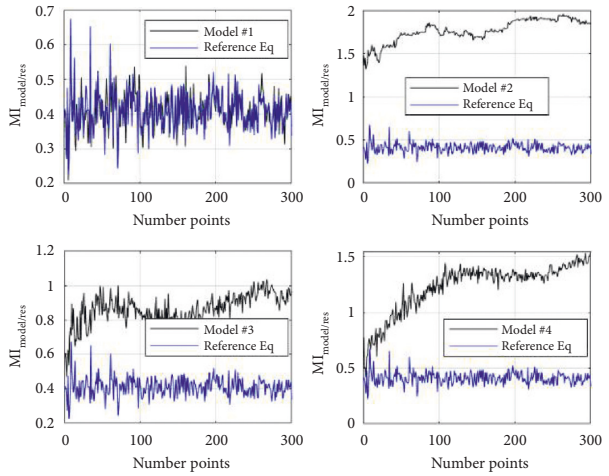

Figure 6: Plots of the mutual information between the models and the residuals for the models of Table 1 in Section 3.

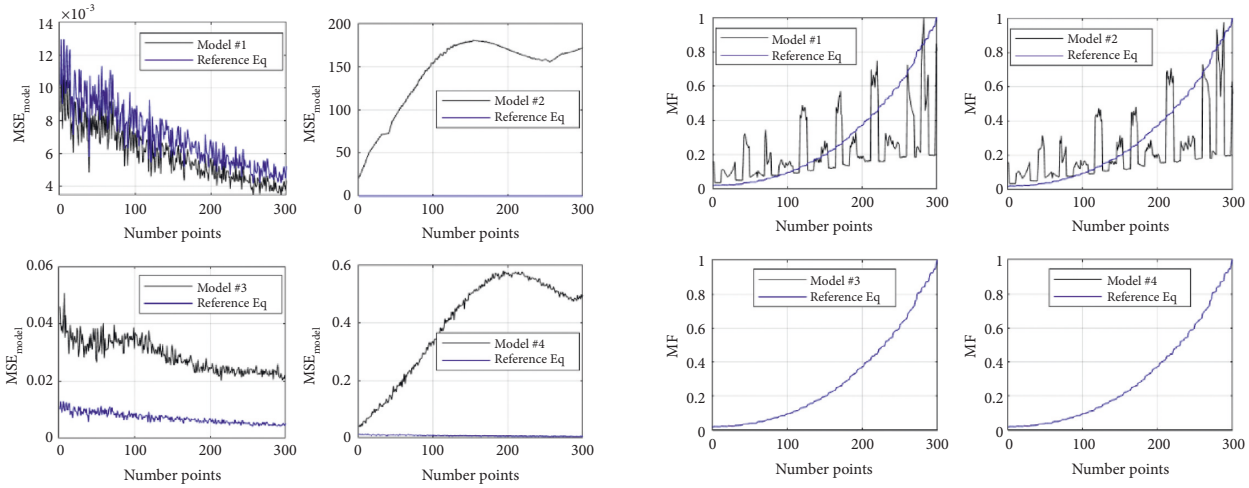

Figure 7: Plots of MSE and MF for the models of Table 1 in Section 3.

Assumptions, similar to the ones leading to (4), allow expressing the BIC criterion as follows:

$$BIC = n \cdot \ln\left(\sigma^2_{(\epsilon)}\right) + k \cdot \ln(n). \tag{16}$$

Even if the conceptual origins of BIC are different, the proposed changes have the same effects, namely, they improve BIC's discriminatory power by including more statistical information about the residuals and by better quantifying the models' complexity. In full analogy to (13), the final upgraded version of the BIC criterion is

$$BIC_{MICx} = n \ln[MSE(1 + MI)] + C_x \ln(n). \tag{17}$$

The tests of the AIC have been performed also for the BIC and they produce basically the same results. The discriminatory capability of BIC_{MICx} is clearly superior to the original version of the indicator, as can be seen in the plots of Appendix B. Of course, given the fact that BIC is based on Bayesian statistics, the argument that the implemented upgrades improve the coherence, with information-theoretic definitions and assumptions, cannot be made. On the other hand, the fact that the proposed modifications improve the quality also of a Bayesian type of selection criterion increases the confidence in the validity of the ideas, which have led to them.

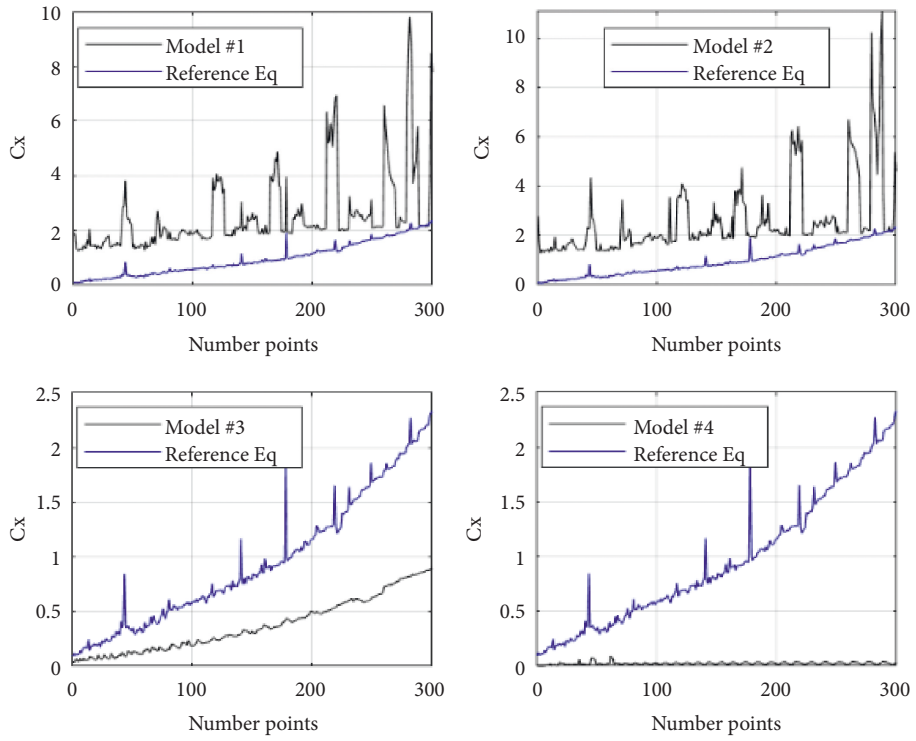

Figure 8: Plots of complexity Cx for the models of Table 1 in Section 3.

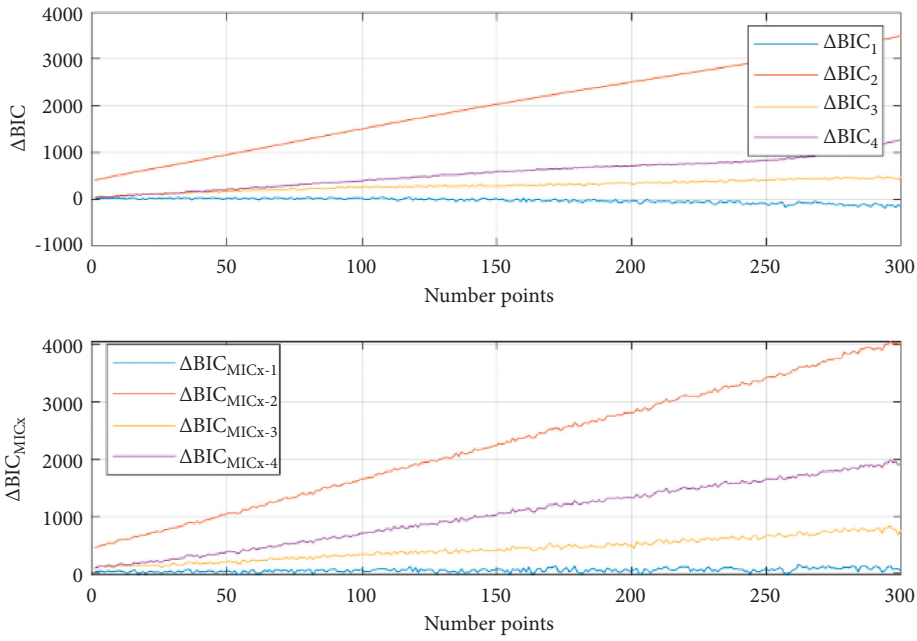

Figure 9: Comparison of the traditional BIC with the new BIC_{MICx} vs. the number of points for the models of Table 1 in Section 3. The plots show the indicator difference between the candidate models and the reference one; therefore negative values indicate that the corresponding indicator would have reached the wrong conclusion about the model to select.

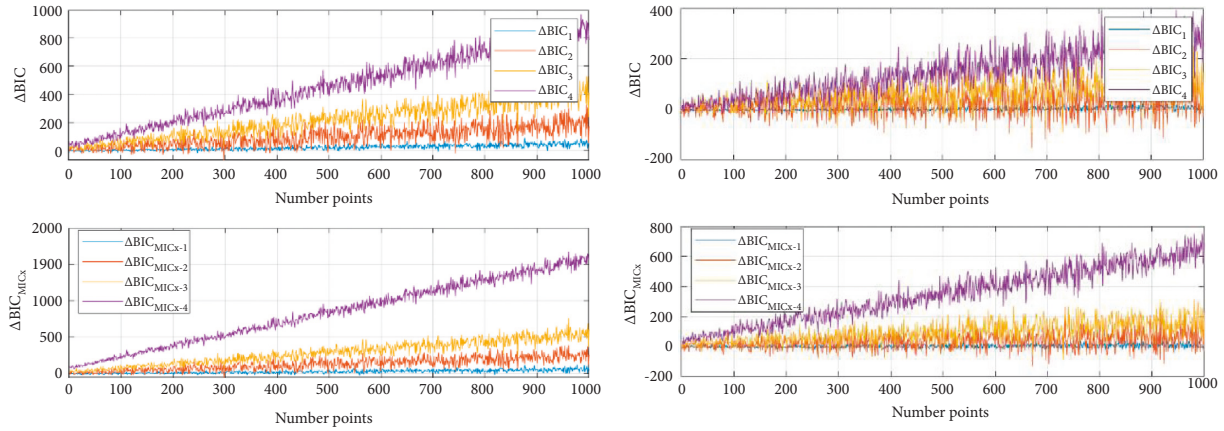

Figure 10: Top two plots: comparison of BIC and BIC_{MICx} for the case of a power law monomial multiplied by a squashing for 15% of Gaussian noise. Bottom: comparison of BIC and BIC_{MICx} for the case of a power law monomial multiplied by a squashing for 30% of Gaussian noise.

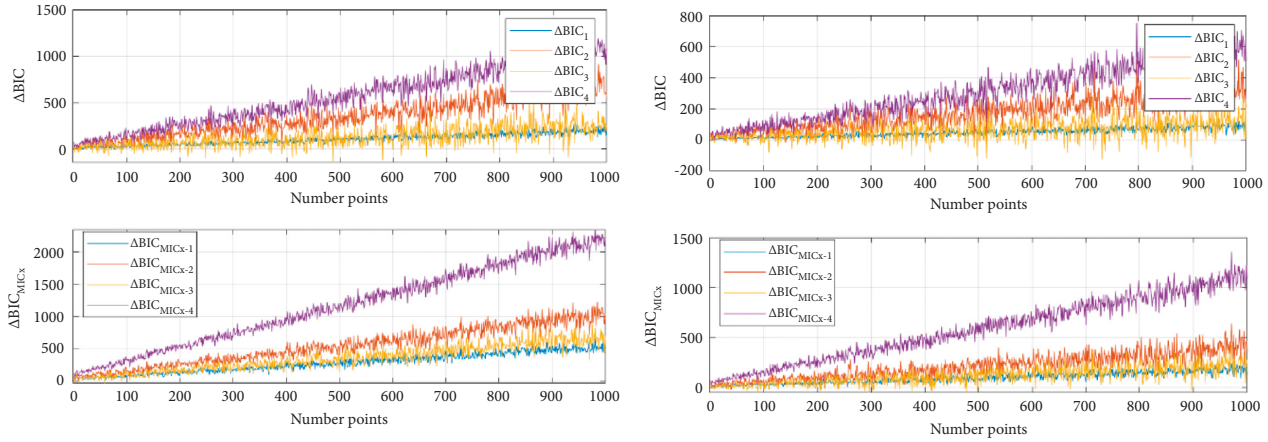

Figure 11: Top two plots: comparison of BIC and BIC_{MICx} for the case of an exponential function for 30% of Gaussian noise plus outliers. Bottom: comparison of BIC and BIC_{MICx} for the case of an exponential function for 60% of Gaussian noise plus outliers.

7. Conclusions

The Akaike Information Criterion was conceived to minimise the out-of-sample error and it is based on information theory. Statistical models are indeed developed to represent the process that generated the data, and the AIC estimates the relative amount of information lost by a given model. On this basis, it is assumed that the better a model, the less information it loses. Unfortunately, the deployment of AIC is problematic because its practical versions are affected by significant limitations. Indeed the most widely used version of AIC is valid under the assumptions that the data are affected by Gaussian, zero-sum additive noise. These hypotheses have to be accepted because, in most practical applications, it is often very difficult, if not impossible, to compute the likelihood of the data given the model. If the processes generating the data do not verify these assumptions, the traditional versions of the AIC can become poorly effective or even misleading.

On the other hand, other information theoretic quantities can be implemented to improve the discrimination potential of the criterion. In particular, the mutual information between the model estimates and the residuals can help reward the goodness of fit. The entropy in its turn can be used to quantify the model complexity. With these upgrades, the proposed version of the AIC has always proved to have much better convergence properties than the traditional version in all respects, including robustness against noise and zero-sum outliers. This has occurred in all the numerical tests performed, some of which consist of very challenging selection tasks, given the fact that some candidate models assume values very similar to the right one in the range covered by the data. The proposed improvements have an equally positive impact on the other criteria of the AIC family, such as TIC and AICc [4]. The extension of the same concepts to the Bayesian information criterion proves the soundness of the basic rationale behind the proposed modifications. The good performance in presence of non-normal noise distributions is particularly encouraging because model assessment in such situations has not yet received a lot of attention in the literature. Indeed, only a few publications have addressed the fact that many existing model selection criteria such as the BIC and Cp may not be suitable for generalized linear model regression, in which the

conditional mean and variance of the response are dependent [14]. Synergies with other formulations of the complexity term would also be very interesting from the methodological point of view [15].

Given the quite positive results obtained with synthetic data, proving their better discriminatory capability, the proposed new versions of the selection criteria are expected to become useful in various fields. They are already being deployed for the investigation of complex systems, ranging from high-temperature plasmas [16–23] to remote sensing of the atmosphere and radar [24–26]. Another promising application seems to be in support of the regularization of recent tomographic inversion methods [27–29]. In these fields, Dimensional Analysis (DA) is a methodology widely used to identify key variables based on physical dimensions. Even if it has been granted some attention recently, in most literature DA is treated as merely a preprocessing tool, creating various statistical problems [30]. The upgrades of the criteria proposed in this work could hopefully help in devising an appropriate statistical methodology that integrates DA and model selection.

Appendix

A. Calculation of the AICMICx and BICMICx Quantities of Section 3

The Figures 5–Figure 9 in this Appendix document all the quantities required to calculate AIC_{MICx} and BIC_{MICx} for the didactic case of Section 3, involving polynomial and sinusoidal models.

B. Performance Details of the BIC and BICMICx Quantities of Section 5

This Appendix documents the performance of BIC_{MICx} for the numerical cases described in Section 5: power laws multiplied by a squashing term and exponentials. Figures 10 and 11 show the comparison of BIC and $BIC_{MICx.}$

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

AM conceived this research; ML participated in the design of the code and interpretation of the results; ML and RR performed the validation of the analysis; AM and MG wrote the paper and participated in the revisions of it. MG provided the funding and supervised the project. All authors read and approved the final manuscript.

References

[1] F. Bailly and G. Longo, *Mathematics and the*, Natural Sciences" Imperial College Press, London, 2011.

[2] B. D'Espagnat, *On Physics and Philosophy*, Princeton University Press, Oxford, 2002.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[4] P. B. Kenneth and D. R. Anderson, *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*, Springer, Berlin, 2nd ed edition, 2002.

[5] G. Claeskens, "Statistical model choice" (PDF)," *Annual Review of Statistics and Its Application*, vol. 3, no. 1, pp. 233–256, 2016.

[6] G. W. Corder and D. I. Foreman, *Nonparametric Statistics for Non-statisticians: A Step-By-Step Approach*, Wiley, Hoboken, 2009.

[7] A. Murari, E. Peluso, F. Cianfrani, P. Gaudio, and M. Lungaroni, "On the use of entropy to improve model selection criteria," *Entropy*, vol. 21, no. 4, p. 394, 2019.

[8] B. K. Natarajan, "On Learning sets and functions," *Machine Learning*, vol. 4, no. 1, pp. 67–97, 1989.

[9] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[10] M. Efatmaneshnik and M. J. Ryan, "A general framework for measuring system complexity," *Complexity*, vol. 21, no. 1, pp. 533–546, 2016.

[11] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, Berlin, 2000.

[12] S. Bonamente, *Statistics and Analysis of Scientific Data"*, Graduate Texts in Physics) Springer Scince+Bussiness Media LLC, Berlin, 2017.

[13] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[14] X. Shen, H. C. Huang, and J. Ye, "Adaptive model selection and assessment for exponential family distributions," *Technometrics*, vol. 46, no. 3, pp. 306–317, 2004.

[15] A. Murari, R. Riccardo, and C. Teddy, "Alternative Definitions of Complexity for Practical Applications of Model Selection Criteria," *Defining and quantifying complexity*, vol. 2021, Article ID 8887171, 8 pages, 2021.

[16] A. Murari, I. Lupelli, P. Gaudio, M. Gelfusa, and J. Vega, "A statistical methodology to derive the scaling law for the H-mode power threshold using a large multi-machine database," *Nuclear Fusion*, vol. 52, no. 6, Article ID 063016, 2012.

[17] A. Murari, I. Lupelli, M. Gelfusa, and P. Gaudio, "Non-power law scaling for access to the H-mode in tokamaks via symbolic regression," *Nuclear Fusion*, vol. 53, no. 4, Article ID 043001, 2013.

[18] 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.

[19] A. Murari, E. Peluso, M. Gelfusa, I. Lupelli, M. Lungaroni, and P. Gaudio, "Symbolic regression via genetic programming for data driven derivation of confinement scaling laws without any assumption on their mathematical form," *Plasma Physics and Controlled Fusion*, vol. 57, no. 1, Article ID 014008, 2015.

[20] A. Murari, E. Peluso, M. Lungaroni, M. Gelfusa, and P. Gaudio, "Application of symbolic regression to the derivation of scaling laws for tokamak energy confinement time in terms of dimensionless quantities," *Nuclear Fusion*, vol. 56, no. 2, Article ID 026005, 2015.

[21] A. Murari, M. Lungaroni, E. Peluso et al., "Adaptive predictors based on probabilistic SVM for real time disruption mitigation on JET," *Nuclear Fusion*, vol. 58, no. 5, Article ID 056002, 2018.

[22] 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.

[23] F. Romanelli and R. Kamendje, "Overview of JET results," *Nuclear Fusion*, vol. 49, no. 10, Article ID 104006, 2009.

[24] P. Gaudio, "New frontiers of forest fire protection: a portable laser system," *FfED WSEAS Transactions on Environment and Development*, vol. 9, no. 3, pp. 195–205, 2013.

[25] P. Gaudio, M. Gelfusa, A. Malizia, and M. Richetta, "Design and development of a compact Lidar/Dial system for aerial surveillance of urban areas," in *Proceedings of the SPIE - The International Society for Optical Engineering*, vol. 8894, Rome, Italy, September 2013.

[26] F. Xin, B. Wang, L. Shumin, X. Song, and W. Chi Hsu, "Adaptive radar waveform design based on weighted MI and the difference of two mutual information metrics," *Complexity*, vol. 2021, Article ID 8947450, 18 pages, 2021.

[27] T. Craciunescu, G. Bonheure, V. Kiptily, A. Murari, I. Tiseanu, and V. Zoita, "A comparison of four reconstruction methods for JET neutron and gamma tomography," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 605, no. 3, pp. 374–383, 2009.

[28] T. Craciunescu and A. Murari, "Geodesic distance on Gaussian manifolds for the robust identification of chaotic systems," *Nonlinear Dynamics*, vol. 86, no. 1, pp. 677–693, 2016.

[29] T. Craciunescu, E. Peluso, A. Murari, and M. Gelfusa, "Maximum likelihood bolometric tomography for the determination of the uncertainties in the radiation emission on JET TOKAMAK," *Review of Scientific Instruments*, vol. 89, no. 5, Article ID 053504, 2018.

[30] W. Shen and D. K. J. Lin, "A conjugate model for dimensional analysis," *Technometrics*, vol. 60, no. 1, pp. 79–89, 2017.